Model assessment: Bayes assisted tests and tests for discrete data

by

Zheng Sun
M.Sc., University of Victoria, 2010
B.Sc., Simon Fraser University, 2006

A Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Sciences

© Zheng Sun 2014

SIMON FRASER UNIVERSITY
Spring 2014

All rights reserved.
However, in accordance with the Copyright Act of Canada, this work may be reproduced without authorization under the conditions for “Fair Dealing.”
Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.
APPROVAL

Name: Zheng Sun
Degree: Doctor of Philosophy
Title of Thesis: Model assessment: Bayes assisted tests and tests for discrete data

Examining Committee: Dr. Tim Swartz
Chair

__________________________
Dr. Richard Lockhart,
Professor, Senior Supervisor

__________________________
Dr. Michael Stephens,
Professor Emeritus, Senior Supervisor

__________________________
Dr. Joan Hu,
Professor, SFU Examiner

__________________________
Dr. Federico O’Reilly,
Professor, IIMAS, UNAM, Mexico.
External Examiner

Date Approved: January 23rd, 2014
Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the non-exclusive, royalty-free right to include a digital copy of this thesis, project or extended essay[s] and associated supplemental files (“Work”) (title[s] below) in Summit, the Institutional Research Repository at SFU. SFU may also make copies of the Work for purposes of a scholarly or research nature; for users of the SFU Library; or in response to a request from another library, or educational institution, on SFU's own behalf or for one of its users. Distribution may be in any form.

The author has further agreed that SFU may keep more than one copy of the Work for purposes of back-up and security; and that SFU may, without changing the content, translate, if technically possible, the Work to any medium or format for the purpose of preserving the Work and facilitating the exercise of SFU’s rights under this licence.

It is understood that copying, publication, or public performance of the Work for commercial purposes shall not be allowed without the author’s written permission.

While granting the above uses to SFU, the author retains copyright ownership and moral rights in the Work, and may deal with the copyright in the Work in any way consistent with the terms of this licence, including the right to change the Work for subsequent purposes, including editing and publishing the Work in whole or in part, and licensing the content to other parties as the author may desire.

The author represents and warrants that he/she has the right to grant the rights contained in this licence and that the Work does not, to the best of the author’s knowledge, infringe upon anyone's copyright. The author has obtained written copyright permission, where required, for the use of any third-party copyrighted material contained in the Work. The author represents and warrants that the Work is his/her own original work and that he/she has not previously assigned or relinquished the rights conferred in this licence.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2013
Abstract

In this thesis, two areas of goodness-of-fit are discussed and new methodology proposed. In the first, Bayesian methods are introduced to provide a narrow band of alternative continuous distributions when the distribution tested is uniform or normal. A particular use of Bayesian methods allows consideration of the problem of testing the distribution of latent (unobserved) variables when these are connected by a known relationship to a set of observed variables. The technique is used to advance an interesting procedure introduced in Geology by Krumbein and for a modern example, to test the distribution of the frailty term (random effects) in a Cox Proportional Hazards (PH) model.

The second part of the thesis deals with discrete data with particular emphasis on applying Cramér-von Mises statistics. Tests are proposed for $K$ samples in an ordered contingency table. Finally, the $K$ sample procedure is applied to testing the fit of the binary regression model to longitudinal (correlated) data using Generalized estimating equations. A common thread throughout the thesis is the use of the Cramér-von Mises statistics or closely related statistics for testing.
To Kelly, Helena and Michael
“Don’t worry, Gromit. Everything’s under control!”

— *The Wrong Trousers, Aardman Animations, 1993*
Acknowledgments

I would like to thank my supervisors Dr. Richard Lockhart and Dr. Michael Stephens for their great help and patience throughout this thesis. They provided me with valuable advice and I am extremely grateful to have been their student. I also thank Dr. Alberto Contreras for his collaboration. Particularly, I thank Dr. Stephens for introducing me to Statistics when I was an undergraduate. I would like to thank Dr. John Spinelli and Dr. Carl Schwartz for providing me opportunities to work on real world problems. I am also very appreciative of all the assistance I have received from professors and staff and the graduate students in the department, particularly Dr. Tim Swartz and Dr. Joan Hu. I feel very fortunate to have studied in the department of Statistics and Actuarial Science at Simon Fraser University.

To Dad and Mum, thank you for your love and generosity which enabled me to study in Canada. To my wife Kelly Wen Zhang, thank you for your support and for always being by my side.
# Contents

Approval .................................................. ii
Partial Copyright License ................................ iii
Abstract ...................................................... iv
Dedication ..................................................... v
Quotation ...................................................... vi
Acknowledgments ............................................. vii
Contents ...................................................... viii
List of Tables ............................................... xii
List of Figures ............................................... xv

1 Bayes assisted GOF: testing uniformity ..................... 1
   1.1 Gaussian prior ........................................ 2
   1.2 Asymptotic thoery ...................................... 6
      1.2.1 Asymptotic power ................................ 8
   1.3 Example ................................................ 8
   1.4 References ............................................ 9

2 Bayes assisted tests for the normal distribution ............ 10
   2.1 Bayes optimal tests for normality .................... 10
   2.2 Example ............................................... 13
List of Tables

1.1 Significance points (under $H_0$) of the Bayes assisted goodness-of-fit statistic $S$ for finite $n$ sample when the Gaussian process $Z(x)$ is chosen to give $W^2$. We take $a = 1$. Number of Monte Carlo samples is $M=10000$.

1.2 Asymptotic and finite $n = 50$ or 100 powers (under Bayesian alternative) of the Bayes assisted goodness-of-fit statistic $S$ for finite $n$ sample when the Gaussian process $Z(x)$ is chosen to give $W^2$ for various $a$. Number of Monte Carlo samples is $M=1000$.

3.1 Table: Number of sandstone measurements in given intervals.

3.2 Maximum likelihood estimates of Cox PH model with Weibull baseline hazard and lognormal frailty fitted to the Kidney data.

3.3 Comparisons of coefficients used to generate samples with the average coefficients from bootstrap samples and with posterior means.

4.1 Upper-tail percentage points for standardized $\hat{W}_K^2$.

4.2 Upper-tail percentage points for standardized $\hat{A}_K^2$.

4.3 Upper-tail percentage points for standardized $\hat{U}_K^2$.

4.4 Comparison of percentage points for different sets $p$ using the $p$ or $w$ ($w_i = (p_i + p_{i+1})/2$) weights.

4.5 Letter grades of 3 offerings of a statistics course.

4.6 $p$ values of different tests on the grades data.

4.7 Comparisons of asymptotic powers of statistics under different location shift models. Test level $\alpha = 0.10$. Parameters are given in the text.

4.8 Comparisons of asymptotic powers of statistics under different scale shift models. Test level $\alpha = 0.10$. Parameters are given in the text.
4.9 Power comparisons of components of the statistics under extreme value distributions with scale shift models with $p_j = c(0.2, 0.2, 0.2, 0.2, 0.2)$ when $K = 4$. Level $\alpha = 0.10$, $n_i = 50$, $\beta_1 = 3.162$, $\beta_2 = -6.640$, $\beta_3 = 6.640$, $\beta_4 = -3.162$. The rows with an asterisk are values given by Monte Carlo studies using 10000 samples. The rows without an asterisk are values given by asymptotic theory. 49

5.1 Comparison of finite $n$ significance points with asymptotic significance points for $W^2$ and $G^2_W$, when $K = 4$ and $p_j = 0.2$ for $j = 1, \ldots, 5$ and parameters are assumed unknown. M.C. = 10,000 62

5.2 Power study 2: power improvements of likelihood ratio type statistics on CCS type statistics. 64

5.3 Power study 3: power improvements of likelihood ratio type statistics on CCS type statistics. 65

5.4 Letter grades of the last 3 offerings of a STAT course at SFU. 65

5.5 $p$ values of different tests on the grades data. 66

7.1 The Frobenius distances from the “true” covariance matrix $\text{COV}(Y)$ to it different estimates $V_e$, $V_w$, $V_u$ and $V_c$. 96

7.2 Levels of the goodness-of-fit tests at the nominal level 5%(1%) from 1000 replications. The within-subject correlation $\rho = 0.4$ is constant. $W^2_K$ and $A^2_K$ are respectively Cramér-von Mises and Anderson Darling statistics; $X^2_W$ and $X^2_A$ are respectively the Cumulative Chi-square statistics; $G^2_W$ and $G^2_A$ are respectively the likelihood ratio type statistics; $X^2$ and $X^2_u$ are the Pearson chi-square and the unweighted sum of squares; $LQ$ is the Lee and Qaqish statistic. (using the Empirical covariance estimate $V_e$) 99

7.3 Levels of the goodness-of-fit tests at the nominal level 5%(1%) from 1000 replications. The within-subject correlation $\rho = 0.4$ is constant. $W^2_K$ and $A^2_K$ are respectively Cramér-von Mises and Anderson Darling statistics; $X^2_W$ and $X^2_A$ are respectively the Cumulative Chi-square statistics; $G^2_W$ and $G^2_A$ are respectively the likelihood ratio type statistics; $X^2$ and $X^2_u$ are the Pearson chi-square and the unweighted sum of squares; $LQ$ is the Lee and Qaqish statistic. (using the working covariance estimate $V_w$) 100
7.4 Size and power of the goodness-of-fit tests in detecting an omitted quadratic term (dose level) at the nominal level 5%(1%) from 1000 replications. The within-subject correlation \( \rho = 0.4 \) is constant. \( W^2_K \) and \( A^2_K \) are respectively Cramér-von Mises and Anderson Darling statistics; \( X^2_W \) and \( X^2_A \) are respectively the Cumulative Chi-square statistics; \( G^2_W \) and \( G^2_A \) are respectively the likelihood ratio type statistics; \( X^2 \) and \( X^2_u \) are the Pearson chi-square and the unweighted sum of squares; \( LQ \) is the Lee and Qaqish statistic (using the Empirical covariance estimate \( V_e \)).

7.5 Size and power of the goodness-of-fit tests in detecting an omitted interaction (sex \( \times \) dose level) term at the nominal level 5%(1%) from 1000 replications. The within-subject correlation \( \rho = 0.4 \) is constant. \( W^2_K \) and \( A^2_K \) are respectively Cramér-von Mises and Anderson Darling statistics; \( X^2_W \) and \( X^2_A \) are respectively the Cumulative Chi-square statistics; \( G^2_W \) and \( G^2_A \) are respectively the likelihood ratio type statistics; \( X^2 \) and \( X^2_u \) are the Pearson chi-square and the unweighted sum of squares; \( LQ \) is the Lee and Qaqish statistic (using the empirical covariance estimate \( V_e \)).

7.6 The number of smoking relapses (number of participants) at three follow-up points by 4 intervention groups.
List of Figures

1.1 Comparisons of null uniform and several random alternative densities. . . . . 4
1.2 Comparisons of null gamma and several random alternative densities. . . . . 5

3.1 Geometric relationship between $r$ and $x$. . . . . . . . . . . . . . . . . . . . . . 16
3.2 The estimated distribution of $S$ under the null hypothesis that $r$ follows a
gamma distribution. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
3.3 Q-Q plots of the $p$ values from Wald’s statistics for testing the hypothesis
that the indicated coefficient is 0 against quantiles from the standard uniform
distribution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

4.1 Directions chosen by butterflies by different age groups. From left to right:
young, medium, old . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44
4.2 Directions chosen by butterflies at different distances. . . . . . . . . . . . . . 45
4.3 Power study 1. Power comparison of Cramér-von Mises statistics: samples
with different means. Power against total sample size . . . . . . . . . . . . . . 52
4.4 Power study 2. Power comparison of Cramér-von Mises statistics: samples
with different variances. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53
4.5 Power study, power comparison of Cramér-von Mises statistics: increasing
mean differences. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
4.6 Power study, power comparison of Cramér-von Mises statistics: increasing
variance differences. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

5.1 Power study 1: power comparison of likelihood ratio type statistics, CCS type
statistics and other well known statistics for testing samples with different
means. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>Power study 2: power comparison of likelihood ratio type statistics, CCS type statistics and other well known statistics for testing samples with different variances.</td>
</tr>
<tr>
<td>5.3</td>
<td>Power study 3: power comparison of likelihood ratio type statistics, CCS type statistics and other well known statistics for testing samples with different skewness.</td>
</tr>
<tr>
<td>5.4</td>
<td>Power study 4: power comparison of LR type with Nair’s statistics for testing samples with different location and scale.</td>
</tr>
<tr>
<td>5.5</td>
<td>Power study: power comparison of LR statistics with Nair’s statistics for testing samples with increasing mean differences.</td>
</tr>
<tr>
<td>5.6</td>
<td>Power study: power comparison of LR statistics with Nair’s statistics for testing samples with increasing variance differences.</td>
</tr>
</tbody>
</table>
Chapter 1

Bayes Assisted Goodness of fit statistics: tests for uniformity

In a typical goodness-of-fit test, a statistician wishes to test that a random sample \( \mathbf{x} = (x_1, \ldots, x_n) \) comes from a distribution \( F(x : \theta) \), where \( \theta \) is a vector of parameters. Some of these parameters could be unknown. Different statistics are compared for power by taking samples from distinct alternative distributions and counting the proportions which are significant. The powers can be very close, or for one alternative, statistic \( S_1 \) might be better than \( S_2 \) and vice versa for a different alternative. It is difficult to define an overall criterion for optimality. In the first part of this thesis we use average power with respect to a prior on the alternative to give such a criterion.

In Contreras, Lockhart, Stephens and Sun (2013, CLSS), Bayesian procedures were used to define a prior distribution on the alternative hypothesis which is concentrated on alternative distributions whose distance from the null is large enough to be detectable but small enough not to be obvious. We now outline the procedure.

Consider the test for uniformity, \( F_0(x) = x, 0 \leq x \leq 1 \), with density \( f_0(x) = 1, 0 \leq x \leq 1 \). A typical alternative distribution may be written in the form

\[
   f_1(x) = \frac{\exp \{ \epsilon Z(x) \}}{\int_0^1 \exp \{ \epsilon Z(t) \} \, dt}, \tag{1.1}
\]

where \( Z(x) \) is some function and \( \epsilon > 0 \) is a scaling constant to be chosen later to permit large sample theory. Adding a constant to \( Z(x) \) leaves \( f_1(x) \) unchanged so in everything
which follows, it is assumed that \( \int Z(x)dx = 0 \). By taking \( Z(x) \) to be random, that is, by treating \( Z(x) \) as a stochastic process, this description of a random alternative is regarded as a prior distribution on the set of alternative densities. The Neyman-Pearson lemma can then be used to decide which test statistics are optimal. The tests based on such statistics are called Bayes optimal tests.

In CLSS, it is shown that for certain priors of this type, many of the standard quadratic empirical distribution function (EDF) statistics are approximately Bayes optimal.

Brief details of the CLSS ideas are given in this chapter. The statistics are applied to testing for the uniform distribution. Further, the statistics are applied for the normal distribution in Chapter 2, and for testing to the distribution of a latent variable in Chapter 3.

1.1 Gaussian prior

Now suppose \( Z(x) \) is a Gaussian process on \((0,1)\) with mean 0, and covariance function \( \rho_z(s,t) = \text{E}\{Z(s)Z(t)\} \), where \( \text{E} \) denotes expectation. The assumption that \( \int_0^1 Z(t)dt = 0 \) means \( \int_0^1 \rho_z(s,t)dt = 0 \). Provided \( \int \rho_z^2(s,t)dsdt < \infty \), and that \( \int \rho_z(t,t)dt < \infty \), the process \( Z(x) \) may be expressed as

\[
Z(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} g_j(x) \omega_j
\]

(1.2)

where the \( \omega_j \) are independently distributed \( N(0,1) \), and the \( \lambda_j \) are the eigenvalues and the \( g_j(\cdot) \) are corresponding normalized eigenfunctions of the covariance \( \rho_z \). They are the solutions of the integral equation

\[
\int_0^1 \rho_z(s,t)g_j(t)dt = \lambda_j g_j(s).
\]

(1.3)

The conditional density of a sample \( x \) given \( Z(\cdot) \) is

\[
f(x|Z(\cdot)) = \frac{\exp\left\{ \epsilon \sum_{i=1}^{n} Z(x_i) \right\}}{\left[ \int_0^1 \exp\{\epsilon Z(t)\}dt \right]^n}.
\]

(1.4)

To compute the marginal density \( f(x) \) of \( x \), expectations are taken with respect to \( Z(x) \). In general, it is not easy to take expectations to compute the marginal density \( f(x) \); here an approximation will be used. Suppose that, as \( n \to \infty \), \( n\epsilon^2 \) is bounded; the denominator
$D$ of (1.4) can be expanded to give

$$
D = \left\{ \int_0^1 1 + \epsilon Z(t) + \epsilon^2 Z^2(t)/2 \cdot \cdots \, dt \right\}^n
$$

$$
\approx \left\{ 1 + \int_0^1 \epsilon^2 Z^2(t) \, dt/2 \right\}^n
$$

$$
\approx \exp\{n\epsilon^2/2 \int_0^1 Z^2(t) \, dt\}.
$$

The approximate conditional density of $x$ is

$$
f(x|Z(\cdot)) \approx \exp\{\epsilon \sum_{i=1}^n Z(x_i) - n\epsilon^2 \int_0^1 Z^2(t) \, dt/2\}. \quad (1.5)
$$

(Notice the use of the assumption $\int_0^1 Z(t) \, dt = 0$.)

It follows that

$$
f(x) \approx \prod_{j=1}^{\infty} \mathbb{E}\left[ \exp\left\{ \epsilon \sqrt{\lambda_j} \omega_j \sum_{i=1}^n g_j(x_i) - n\epsilon^2 \lambda_j \omega_j^2/2 \right\} \right]; \quad (1.6)
$$

since $\omega_j \sim N(0,1)$, the expectation maybe evaluated directly using the following elementary result:

**Lemma 1** Suppose $z \sim N(0,1)$. Then for any constants $c > 0$ and $b$, it follows that

$$
\mathbb{E}\{\exp(-cz^2/2 + bz)\} = \sqrt{1/(c+1)} \exp[b^2/(2(c+1))].
$$

Following Lemma 1, if $c = n\epsilon^2 \lambda_j$ and $b = \epsilon \sum_{i=1}^n g_j(x_i) \sqrt{\lambda_j}$ then the expected value in equation (1.6) becomes

$$
\frac{1}{\sqrt{1 + n\epsilon^2 \lambda_j}} \exp\left[ \frac{\epsilon^2 \sum_{i=1}^n g_j(x_i)^2 \lambda_j}{2(1 + n\epsilon^2 \lambda_j)} \right]
$$
Figure 1.1: Comparisons of null uniform and several random alternative densities.

Therefore, equation (1.6) becomes

$$f(x) \approx \prod_{j=1}^{\infty} \exp \left[ \frac{n \epsilon^2 \lambda_j \left\{ \sum_{i=1}^{n} g_j(x_i)/\sqrt{n} \right\}^2}{2(1 + n \epsilon^2 \lambda_j)} \right] / \sqrt{1 + n \lambda_j \epsilon^2}$$

$$= \exp\{S(x)\}/\sqrt{\prod_{j=1}^{\infty} (1 + n \epsilon^2 \lambda_j)}, \quad (1.7)$$

where

$$S(x) = \sum_{j=1}^{\infty} \frac{n \epsilon^2 \lambda_j \left\{ \sum_{i=1}^{n} g_j(x_i)/\sqrt{n} \right\}^2}{2(1 + n \lambda_j \epsilon^2)}. \quad (1.8)$$

This is the marginal density $f_1(x)$ on the alternative. The null density is $f(x) = 1$.

Figure 1.1 compares the uniform null density with the realizations of its random alternative density at $\epsilon = 0.5$ and 1 for $n = 100$, when $Z(x)$ is a Gaussian process with covariance $\rho_z(s,t)$ chosen so that $\lambda_j = 1/(\pi^2 j^2)$ and $g_j(s) = \sqrt{2} \cos(\pi j s)$ are the solutions of the integral equation in (1.3).
When the null density is a more general density \( f_0(x) \) with c.d.f. \( F_0(x) \), the alternative density is 
\[
\exp\{S(u)\}/\sqrt{\prod_{j=1}^{\infty} (1 + ne^2 \lambda_j)},
\]
where \( u = (u_1, \ldots, u_n) \) is defined by 
\[
u_i = F_0(x_i).
\]
Similar to Figure 1.1, Figure 1.2 compares the gamma null density with the realizations of its random alternative density with the same Gaussian process as in the uniform case. The figures show that the alternative densities form a band around the null density; the size of the band is determined by \( \epsilon \) and \( n \).

The null hypothesis considered in this Chapter is simple; it contains only the density \( f_0(x) \). The alternative hypothesis is also simple; it contains only the density \( f(x) \). Thus, from the Neyman-Pearson lemma, the approximately Bayes optimal test rejects for large values of the statistic \( S(x) \). The critical region \( S(x) > c^* \) is equivalent to a critical region \( S(x)/(ne^2/2) > c \). So the special case \( ne^2 \to 0 \) leads to rejecting for large values of 
\[
S_0 = \sum_j \lambda_j \left( \sum_i g_j(x_i)/\sqrt{n} \right)^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} \rho_z(x_i, x_k)/n.
\]

by Mercer’s theorem. Note that \( \sum_{i=1}^{n} \sum_{k=1}^{n} \rho_z(x_i, x_k) \) is a V-statistic, as discussed by
1.2 Asymptotic theory

The asymptotic distributions of $S(x)$ can be obtained through the following theorems.

**Theorem 1** Suppose $Z$ is a Gaussian process on $(0, 1)$ with mean 0 and covariance $\rho_z(s, t)$. Assume that $\int_0^1 \rho(s, t)dt = 0$ for all $s$, so $\int_0^1 Z(u)du = 0$. Assume that $\rho_z$ is trace class, that is, that $\int_0^1 \rho_z(t, t)dt < \infty$. Finally assume that $\rho_z(s, t)$ is square integrable, that is $\int_0^1 \int_0^1 \rho_z^2(s, t)dsdt < \infty$.

Let $\lambda_1 > \lambda_2 > \cdots$ be the eigenvalues of $\rho_z$. Under $H_0$, if $n\epsilon^2 \to a \in (0, \infty)$, as $n \to \infty$ then $S \overset{d}{\to} \sum_{j=1}^{\infty} \gamma_j(a)\omega_j^2/2$, where $\gamma_j(a) = \frac{a\lambda_j}{1 + a\lambda_j}$; if $n\epsilon^2 \to 0$, then $S/\sqrt{n\epsilon^2} \overset{d}{\to} \sum_{j=1}^{\infty} \lambda_j \omega_j^2/2$.

**Theorem 2** Suppose $Z$ is a Gaussian process on $(0, 1)$ with mean 0 and covariance $\rho_z(s, t)$. Assume that $\int_0^1 \rho(s, t)dt = 0$ for all $s$. Assume that $\rho_z$ is trace class and square integrable. Let $\lambda_1 > \lambda_2 > \cdots$ be the eigenvalues of $\rho_z$. If $n\epsilon^2 \to a \in (0, \infty)$, as $n \to \infty$, then under the Bayesian alternative hypothesis $S \overset{d}{\to} a \sum_{j=1}^{\infty} \lambda_j \omega_j^2/2$.

**Proof of Theorem 1**: Fix an integer $K$ and put

$$S_K(x) = \sum_{j \leq K} n\epsilon^2 \lambda_j \left\{ \frac{\sum_{i=1}^{n} g_j(X_i)/\sqrt{n}}{2(1 + n\epsilon^2 \lambda_j)} \right\}^2.$$

It is an elementary consequence of the central limit theorem and the facts that $\int g_j(u)du = 0$ and $\int g_j^2(u)du = 1$ that

$$S_K(x) \implies T_{a,K} \equiv \sum_{j \leq K} \gamma_j(a)\omega_j^2/2.$$

If $\nu$ metrizes weak convergence of distributions on the real line and $\mathcal{L}(V)$ denotes the distribution of some statistic $V$ then

$$\nu(\mathcal{L}(S_K), \mathcal{L}(T_{a,K})) \to 0.$$

It follows that there is a sequence $K_n$ tending to infinity so slowly that

$$\nu(\mathcal{L}(S_{K_n}), \mathcal{L}(T_{a,K_n})) \to 0.$$
CHAPTER 1. BAYES ASSISTED GOF: TESTING UNIFORMITY

Now for any sequence $K_n$ tending to infinity we have

$$\mathbb{E} \left[ \gamma_j(n\epsilon^2) \left\{ \sum_i g_j(X_i)/\sqrt{n} \right\}^2 \right] = \sum_{j>K_n} \frac{n\epsilon^2\lambda_j}{(1+n\epsilon^2\lambda_j)} \to 0$$

since $\rho$ is trace class (the condition is equivalent to $\sum_i \lambda_i < \infty$). Similarly

$$\mathbb{E} \left[ \sum_{j>K_n} \gamma_j(a)\omega_j^2 \right] = \sum_{j>K_n} \frac{a\lambda_j}{(1+a\lambda_j)} \to 0.$$

It follows that

$$\nu \left( \mathcal{L}(S), \mathcal{L}(T_a) \right) \leq \nu \left( \mathcal{L}(S), \mathcal{L}(S_{K_n}) \right) + \nu \left( \mathcal{L}(S_{K_n}), \mathcal{L}(T_{a,K_n}) \right) + \nu \left( \mathcal{L}(T_{a,K_n}), \mathcal{L}(T_a) \right) \to 0.$$

This proves Theorem 1.

Rigorous proof of Theorem 2 is work in process. Results from Monte Carlo studies agree with the theorem; see Table 1.2.

In CLSS (2013), it is shown that by choosing $Z(x)$ so that $\lambda_j = 1/(\pi^2j^2)$ and $g_j(s) = \sqrt{2}\cos(\pi js)$, the statistic $S(x)$ becomes equivalent to the Cramér-von Mises goodness-of-fit statistic $W^2$. Similarly, $Z(x)$ can be chosen to make $S(x)$ equivalent to the Anderson Darling statistic $A^2$.

Table 1.1: Significance points (under $H_0$) of the Bayes assisted goodness-of-fit statistic $S$ for finite $n$ sample when the Gaussian process $Z(x)$ is chosen to give $W^2$. We take $a = 1$. Number of Monte Carlo samples is $M=10000$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>25%</th>
<th>10%</th>
<th>5%</th>
<th>2.5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.097</td>
<td>0.159</td>
<td>0.212</td>
<td>0.268</td>
<td>0.352</td>
</tr>
<tr>
<td>50</td>
<td>0.098</td>
<td>0.163</td>
<td>0.218</td>
<td>0.274</td>
<td>0.354</td>
</tr>
<tr>
<td>100</td>
<td>0.097</td>
<td>0.158</td>
<td>0.210</td>
<td>0.263</td>
<td>0.350</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.097</td>
<td>0.162</td>
<td>0.214</td>
<td>0.268</td>
<td>0.340</td>
</tr>
</tbody>
</table>
Table 1.1 shows the asymptotic and Monte Carlo (finite-\(n\)) significance points of a Bayes assisted goodness-of-fit under the null hypothesis that \(x\) is uniformly distributed. The finite-\(n\) points converge to asymptotic points for \(n\) as small as 20. The asymptotic points were obtained following Theorem 1.

1.2.1 Asymptotic power

Asymptotic power under the Bayesian alternatives can be obtained using Theorem 2. For a given \(Z\), the asymptotic power increases as \(a\) increases. For finite \(n\), a Monte Carlo power study with different values of \(a = n\epsilon^2\) was conducted. The powers approach the asymptotic powers as \(n\) increases. The results are in Table 1.2. We are using statistic \(S\) as in (1.8) with \(g_j(s) = \sqrt{2}\cos(2\pi js)\) and \(\lambda_j = 1/(\pi j)^2\). Critical points are computed following Theorem 1 and then powers using Theorem 2.

The steps to generate samples from the Bayes alternative densities are as follows:

1. suppose we want to generate a sample with size \(n\), say \(n = 50\). First generate a sample of size \(n_0\) from \(U(0, 1)\) where \(n_0 \approx 3n = 150\). Call the sample \(x_1, \ldots, x_{150}\)

2. generate a sample of size \(m\), say 100, from \(N(0, 1)\). Call it \(\omega_1, \ldots, \omega_{100}\)

3. for each \(x_i\), compute \(Z(x_i)\) using the expansion in Equation (1.2), and let \(e_i = \exp\{\sqrt{a/nZ(x_i)}\}\). Notice \(e_i, i = 1, \ldots, n_0\) are values of the random density before it is normalized. To sample from this density, rejection sampling was used;

4. sample \(u\) from \(U(0, c)\), where \(c = \max(e_i)\) and check whether or not \(u < e_i\), if this holds, accept \(x_i\) as a sample from \(f_1(x)\), if not, reject \(x_i\) and generate another \(u\) to see if \(x_{i+1}\) is accepted or not.

5. repeat step 4 until the number of \(x_i\) accepted is \(n\).

1.3 Example

Stephens (1986) gave a set of data from Bliss (1967) which consists of weights (in grams) of twenty 21-days-old leghorn chicks. The weights are 156, 162, 168, 182, 186, 190, 190, 196, 202, 210, 214, 220, 226, 230, 230, 236, 236, 242, 246, 270. These were tested for normality assuming known mean \(\mu = 200\) and standard deviation \(\sigma = 35\). By using the probability
Table 1.2: Asymptotic and finite $n = 50$ or 100 powers (under Bayesian alternative) of the Bayes assisted goodness-of-fit statistic $S$ for finite $n$ sample when the Gaussian process $Z(x)$ is chosen to give $W^2$ for various $a$. Number of Monte Carlo samples is $M=1000$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$a$</th>
<th>$1$</th>
<th>$10$</th>
<th>$30$</th>
<th>$50$</th>
<th>$100$</th>
<th>$200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.058</td>
<td>0.183</td>
<td>0.367</td>
<td>0.459</td>
<td>0.667</td>
<td>0.819</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.061</td>
<td>0.175</td>
<td>0.356</td>
<td>0.472</td>
<td>0.653</td>
<td>0.832</td>
<td></td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.062</td>
<td>0.170</td>
<td>0.345</td>
<td>0.463</td>
<td>0.638</td>
<td>0.863</td>
<td></td>
</tr>
</tbody>
</table>

Integral transformation (PIT), the data can be transformed to a set $x$ which should be $U(0,1)$ under $H_0$. The $Z(x)$ was chosen to give $S(x)$ equivalent to $W^2$ as described above, then $S = 0.085$; with $p$ value 0.299 (from Theorem 1). These results agree with Stephens (1986) using $W^2$.

1.4 References


Chapter 2

Bayes assisted tests for the normal distribution

In this Chapter, the test considered is for the normal distribution with unknown mean. Importantly, the normal case differs from the uniform by having unknown parameters under the null hypothesis. For simplicity, the variance is assumed to be 1. Therefore, the null hypothesis is

\[ H_0 : \text{a random sample } x = (x_1, \ldots, x_n) \text{ comes from a } N(\mu, 1), \text{ where } \mu \text{ is unknown.} \]

Since there is now an unknown parameter \( \mu \), the test cannot be reduced to a test for the uniform distribution via the probability integral transformation \( u = \Phi(s) \), where \( \Phi(s) \) is the standard normal cumulative distribution function:

\[
\Phi(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp\left[-\frac{z^2}{2}\right] dz.
\]

2.1 Bayes optimal tests for normality

We proceed by applying a prior on the null hypothesis as well as a prior on the alternative. On the null, define \( f(x|\mu) \) to be the conditional density of \( x \) given \( \mu \)

\[
f(x|\mu) = (2\pi)^{-n/2} \exp \left\{ -\sum_{1}^{n} \frac{(x_i - \mu)^2}{2} \right\}
\]
and suppose \( \mu \) has a prior distribution \( N(0, \tau^2) \), that is, \( \pi(\mu) = 1/(\sqrt{2\pi}\tau) \exp\{-\mu^2/(2\tau^2)\} \). Then the joint density of \( \mu \) and \( x \) is

\[
 f_0(x, \mu) = (2\pi)^{-(n+1)/2}\tau^{-1} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\tau^2}\right\} \exp\{-\mu^2/(2\tau^2)\}.
\]

The marginal joint density of \( x \) on the null hypothesis (averaging over \( \mu \)) is then

\[
 f_0(x) = \int_{-\infty}^{\infty} f_0(x, \mu) d\mu = (2\pi)^{-(n+1)/2}\tau^{-1} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\tau^2}\right\} \exp\{-\mu^2/(2\tau^2)\} d\mu.
\]

After defining null densities, the alternative density is specified by conditioning on \( \mu \) and on a Gaussian process \( Z(x) \) given in Chapter 1. Let \( Z(t); t \in [0,1] \) be a mean 0 Gaussian Process with \( \int_{0}^{1} Z(u) du = 0 \), and square integrable trace class covariance \( \rho(s, t) \), that is \( \int_{0}^{1} \int_{0}^{1} \rho^2(s, t) ds dt < \infty \) and \( \int_{0}^{1} \rho(t, t) dt < \infty \). A conditional alternative density of \( x \) given \( \mu \) and \( Z \) will be

\[
 f(x|\mu, Z) = C_1^{-n}(2\pi)^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\tau^2}\right\} \exp\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z(t_i)\right\}
\]

In this formula, \( t_i = \Phi(x_i - \mu) \) and

\[
 C_1 = 1/\sqrt{2\pi} \int_{-\infty}^{\infty} \exp[-(x - \mu)^2/2] \exp[Z(\Phi(x - \mu))/\sqrt{n}] dx
\]

\[
 = 1/\sqrt{2\pi} \int_{0}^{1} \exp[Z(t)/\sqrt{n}] dt,
\]

using the change of variables \( t = \Phi(x - \mu) \). This density must be approximated, using

\[
 \exp\left(Z(t)/\sqrt{n}\right) \approx 1 + Z(t)/\sqrt{n} + Z^2(t)/(2n).
\]
Then, since $\int_0^1 Z(t)dt = 0$,

$$C_1 = (2\pi)^{-1/2} \int_0^1 \exp (Z(t)/\sqrt{n}) dt$$

$$\approx (2\pi)^{-1/2} \{1 + \int_0^1 Z^2(t)dt/(2n)\}$$

$$\approx (2\pi)^{-1/2} \exp \left( \int_0^1 Z^2(t)dt/2 \right)$$

and

$$f(x|\mu, Z) \approx \exp \left\{ - \sum_{i=1}^n (x_i - \mu)^2/2 \right\} \exp \left\{ \sum_{i=1}^n Z(t_i)/\sqrt{n} - \int Z^2(t)dt/(2n) \right\}.$$  \hspace{1cm} (2.1)

To apply the Neyman-Pearson Lemma, take the expected value of $f(x|\mu, Z)$ first with respect to $Z$ using Lemma 1 in Chapter 1. The expectation gives $f(x|\mu)$

$$f(x|\mu) = C \exp \left\{ - \sum_{i=1}^n (x_i - \mu)^2/2 \right\} \exp \{S(t_1,\ldots, t_n)\},$$  \hspace{1cm} (2.2)

where $C = (2\pi)^{(-n/2)}/\left\{ \prod_{j=1}^\infty (1 + \lambda_j) \right\}^{1/2}$, and $\exp \{S(t_1,\ldots, t_n)\}$ is the $S(x)$ from equation (1.8) with $x_i$ in (1.8) replaced by $t_i = \Phi(x_i - \mu)$.

The next expectation, with respect to $\mu$, becomes

$$f_a(x) = C \int_{-\infty}^\infty \exp \left\{ - \sum_{i=1}^n (x_i - \mu)^2/2 - \mu^2/(2\tau^2) \right\} \exp \{S(t_1,\ldots, t_n)\} d\mu.$$  \hspace{1cm} (2.3)

The Neyman-Pearson test statistic is the ratio $f_a(x)/f_0(x)$. Omitting the term which does not involve data, an equivalent test statistic is

$$T = \frac{\int \exp \left\{ - \sum (x_i - \mu)^2/2 - \mu^2/(2\tau^2) \right\} \exp \{S(t_1,\ldots, t_n)\} d\mu}{\int \exp \left\{ - \sum (x_i - \mu)^2/2 - \mu^2/(2\tau^2) \right\} d\mu}. $$  \hspace{1cm} (2.4)

Let

$$T_1 = \exp \left[ - \frac{1}{2} \left\{ \sum x_i^2 - \frac{(\sum x_i)^2}{n + 1/\tau^2} \right\} \right].$$
The denominator is

\[
T_1 \int \exp \left\{ -\frac{(n + 1/\tau^2)}{2} \left( \mu - \frac{\sum x_i}{n + 1/\tau^2} \right)^2 / 2 \right\} d\mu = T_1 \sqrt{2\pi/(n + 1/\tau^2)},
\]

and similarly, the numerator is

\[
T_1 \int \exp \left\{ -(1/2)(n + 1/\tau^2) \left( \mu - \frac{\sum X_i}{n + 1/\tau^2} \right)^2 \right\} \exp\{S(t_1, \ldots, t_n)\} d\mu
\]

The posterior distribution of \( \mu \) given the data, and conditioning on \( H_0 \), is normal with mean

\[
E(\mu|x) = \frac{n\bar{x}}{n + 1/\tau^2}
\]

and variance

\[
\frac{1}{n + 1/\tau^2}.
\]

Therefore, the numerator can be written as

\[
T_1 \sqrt{2\pi/(n + 1/\tau^2)} \int \exp\{S(t_1, \ldots, t_n)\} f_0(\mu|x) d\mu
\]

and the likelihood ratio test statistic is

\[
T = \int \exp\{S(t_1, \ldots, t_n)\} f(\mu|x) d\mu = E[\exp\{S(t_1, \ldots, t_n)\}|x]. \tag{2.5}
\]

It is important to note that this conditional expectation is computed with respect to a posterior density \( f_0(\mu|x) \) for \( \mu \) which assumes the null hypothesis is true. The resulting integral will need to be calculated numerically to find the value of the test statistic.

### 2.2 Example

We illustrate the test for normality using the chicks data given in Chapter 1, and now suppose that the test is for

\[ H_0 : \text{the sample is from a normal distribution but with mean unknown.} \]
The Gaussian process $Z$ will be the same as in Chapter 1. Since $\sigma$ is assumed to be known, the chick data is standardized to give variance 1. A non-informative prior is assigned to $\mu$ by letting $\tau^2 \to \infty$; therefore the posterior density of $f_{0}(\mu|x)$ is $N(\bar{x}, 1/n)$. To calculate an estimate of $T$ in (2.5), 1000 posterior samples of $[\mu|x]$ are taken and the average calculated. This gives the test statistic $T^* = 1.060$. To approximate the distribution of $T$ under $H_0$, we generated 1000 sets of chick data from $N(\bar{x}, 1)$ and the estimates of $T$ were calculated to give the estimated distribution. Using this distribution, the approximate $p$ value of $T^*$ was 0.842.
Chapter 3

Bayes assisted statistic for testing latent variables

3.1 Introduction

Krumbein (1935) proposed an ingenious method for estimating the moments of the radii \( r \) of spherical rocks embedded in sediments, based on slices through the rocks and the sediments. The data observed are the radii \( x \) of circular cross-sections of the rocks from random slicing, and the radii \( r \) themselves are not easy to measure. However, Krumbein did have one sample of \( r \) and the corresponding \( x \), which will be used in an example later in the Chapter. The method rests on the geometric connection between \( x \) (the observed variable) and \( r \) (a latent variable). In Figure 3.1, let \( y \) be the location of a random slice which is therefore uniformly distributed on \((0, r)\), so that

\[
f(x|r) = \frac{x}{\sqrt{r^2 - x^2}}, x < r. \tag{3.1}
\]

In this Chapter, a procedure is derived to test the fit of a latent (unobserved) variable to a given distribution, based on related observed variables. Using the Bayes assisted methodology given in Chapter 1, MCMC sampling with the EM algorithm is used to calculate the test statistic and its distribution. The procedure is applied to two examples: (1) testing the distribution of \( r \) in Krumbein’s problem; (2) testing the distribution of the frailty term in a survival analysis model.
3.1.1 Test statistic for the distribution of a latent variable

Suppose \( x \) is the observed variable and \( r \) is the unobserved variable. Suppose the null hypothesis is

\[ H_0 : r \text{ follows density } g(r|\theta), \text{ with cdf } G(r|\theta), \]

against

\[ H_1 : r \text{ follows } g(r|\theta)\Lambda(r|\theta), \]

where \( \Lambda(r|\theta) \) is a stochastic process which models the departure of the alternative density from the null distribution. Then \( \Lambda(r|\theta) \) is a general version of \( \exp\{\epsilon Z(x)\} \) from Chapter 1.

Let \( x = (x_1, \ldots, x_n)' \) and \( r = (r_1, \ldots, r_n)' \). Let \( \pi(\theta) \) be a prior distribution for \( \theta \) and let \( h(x|r) \) be the conditional density of \( x \) given \( r \). The conditional null density of \( x \) and \( r \) given \( \theta \) is

\[ h(x|r)g(r|\theta) = \prod_{i=1}^{n} h(x_i|r_i)g(r_i|\theta). \]

The unconditional density of \( x \) on the null hypothesis is

\[ f_0(x) = \int_{\theta} \int_{r} h(x|r)g(r|\theta)\pi(\theta)d\theta dr. \]
Under the alternative, the conditional joint density of the observed values $x$ given $\theta$ and $\Lambda$ is

$$f_1(x|\theta, \Lambda) = \int \cdots \int \prod_{i=1}^{n} h(x_i|r_i) g(r_i|\theta) \Lambda(r_i|\theta) dr_1 \cdots dr_n.$$  

The unconditional joint density of $x$ is obtained by averaging over $\theta$ and $\Lambda$. Averaging first over $\Lambda$ gives

$$f_1(x|\theta) = \int \cdots \int \prod_{i=1}^{n} \left\{ h(x_i|r_i) g(r_i|\theta) \right\} E \left\{ \prod_{i=1}^{n} \Lambda(r_i|\theta) \right\} dr_1 \cdots dr_n.$$  

If $\Lambda(r_i|\theta)$ is chosen to be of the form in Chapters 1 and 2, then

$$\Lambda(r_i|\theta) = \exp \{ \epsilon Z(u_i) \} / \int_0^1 \exp \{ \epsilon Z(t) \} dt,$$

where $u_i = G(r_i; \theta)$ is the probability integral transformation of $r_i$. The expectation of $\prod_{i=1}^{n} \Lambda(r_i|\theta)$ is then approximated as before to give

$$E_\Lambda \left\{ \prod_{i=1}^{n} \Lambda(r_i|\theta) \right\} = S^*(r, \theta),$$

where

$$S^*(r, \theta) = \frac{\exp \{ S(G(r_1, \theta), \ldots, G(r_n, \theta)) \}}{\sqrt{\prod_{j=1}^{\infty} 1 + n\lambda_j \epsilon^2}}$$  \hspace{1cm} (3.2)$$

and $S$ is given in equation (1.8). This expectation $S^*(r, \theta)$ corresponds to $f(x)$ in equation (1.7).

Now average over $\theta$ to get an approximate joint density for $x$ under the alternative. This is given by

$$f_1(x) = \int_0 f_1(x|\theta) \pi(\theta) d\theta = \int \int \cdots \int h(x_i|r_i) g(r_i|\theta) \pi(\theta) S^*(r, \theta) dr_1 \cdots dr_n d\theta.$$  \hspace{1cm} (3.3)
CHAPTER 3. TESTING LATENT VARIABLES

Since the joint density of \([r, x, \theta]\) under the null hypothesis is

\[
f_0(x, r, \theta) = \prod_{i=1}^{n} \{h(x_i| r_i)g(r_i|\theta)\} \pi(\theta),
\]

equation (3.3) becomes

\[
f_1(x) = \int_{\theta} \int_{r} f_0(x, r, \theta) S^*(r, \theta) dr d\theta
\]

\[
= \int_{\theta} \int_{r} f_0(x) \frac{f_0(x, r, \theta)}{f_0(x)} S^*(r, \theta) dr d\theta
\]

\[
= f_0(x) \int_{\theta} \int_{r} f_0(r, \theta|x) S^*(r, \theta) dr d\theta
\]

From the Neyman Pearson Lemma, the approximate Bayes optimal test statistic is the likelihood ratio \(f_1(x)/f_0(x)\):

\[
S = \int_{\theta} \int_{r} f_0(r, \theta|x) S^*(r, \theta) dr d\theta
\]

\[
= \text{E}\{S^*(r, \theta)|x\} \quad (3.4)
\]

The term \(\text{E}\{S^*(r, \theta)|x\}\) is the posterior mean of \(S^*(r, \theta)\) given \(x\). The quantity \(S^*(r, \theta)\) is a goodness-of-fit statistic which would be appropriate for testing the null hypothesis if \(r\) and \(\theta\) had been observed. When they are not observed, the test statistic in equation 3.4 averages the values of \(S^*(r, \theta)\) over all possible \([r, \theta]\) values with weights from posterior distribution of \([r, \theta|x]\). This posterior is computed for the prior on the null hypothesis. The statistic rejects \(H_0\) for large value of \(S\). The problem now is to find the conditional expectation in (3.4).

### 3.1.2 Testing procedure

To compute an approximation of statistic \(S\), the posterior joint density of \(r\) and \(\theta\) is found by Gibbs sampling. The general steps are as follows:

1. assign a prior distribution \(\pi(\theta)\) for \(\theta\). Note that \(\theta\) may be a vector; for example, if the tested null distribution is gamma \((\alpha, \beta)\), then \(\theta\) has two components;

2. generate \(M\) sets of \([r, \theta]\) from the posterior density \(f(r, \theta|x)\), after a burn-in period, by the Gibbs sampling algorithm.
(a) sample from $f(r_i | r^*, \theta, x)$, $i = 1, \cdots, n$ where $r^*$ denotes the set of $r$ omitting $r_i$; this gives $n$ new values of $r$ for a given $\theta$
(b) sample from $f(\theta | r, x)$ similarly
(c) sample iteratively from (a) and (b);

3. compute $S^*$ in (3.2) for the $i$th sample, call it $S^*_i$, where $i = 1, \cdots, M$;
4. the test statistic $S$, based on the given dataset $x$, is then approximately $\sum_{i=1}^{M} S^*_i / M$.

To approximate the null distribution of $S$, the steps are as follows:
1. bootstrap a set of $r$ from the tested null distribution, using an estimate of $\theta$: this may be the posterior mode in step 2 or the estimate from the EM algorithm. An example of the EM algorithm is given in Section 3.2.1 below;
2. generate a set of $x$ from $h(x | r)$;
3. calculate the statistic $S$ from the newly generated $x$ using the Gibbs sampling algorithm above, call it $S_d$;
4. repeat Steps 1–3 $M_1$ times, to get $M_1$ values of $S_d$; these give an estimated distribution of $S$ under $H_0$;
5. to make the test of $H_0$, calculate the $p$ value of $S$.

### 3.2 Krumbein’s problem revisited

To illustrate the test procedure on Krumbein’s problem, suppose the null hypothesis is

$$H_0 : r \sim \text{Gamma}(\alpha, \beta).$$

From equation (3.1),

$$f(x | r) = \frac{x}{r^{\alpha+2} \exp(-r/\beta)},$$

therefore, the joint density of $x, r$ is

$$f(x, r) = \frac{x r^{\alpha-2} \exp(-r/\beta)}{\Gamma(\alpha) \beta^\alpha \sqrt{r^2 - x^2}}, r > x > 0$$

The steps of the testing procedures in Section 3.1.2 become
1. Choose a conjugate prior density for $\alpha$ and $\beta$:

$$\pi(\alpha, \beta | v, q, s, t) \propto v^{\alpha-1} e^{-q/\beta} \frac{1}{\Gamma(\alpha)^{\frac{s}{2}}} \frac{t}{\beta^{\alpha s}},$$  \hspace{1cm} \text{(3.6)}$$

where $v, q, s, t$ are hyper-parameters chosen to make the density $\pi(\alpha, \beta | v, q, s, t)$ close to uniform. The use of a conjugate prior makes the full conditional needed for Gibbs sampling easier.

2. A Gibbs sampler scheme is used which alternates among $r$, $\alpha$, and $\beta$. The full conditionals for the posterior joint density $p(r_i, \alpha, \beta | x_i, v, q, s, t)$ are

$$p(r_i | x_i, \alpha, \beta) \propto r_i^{\alpha-2} e^{-r_i/\beta} \frac{1}{\sqrt{r_i^2 - x_i^2}},$$  \hspace{1cm} \text{(3.7)}$$

and

$$\pi(\alpha | r, \beta) \propto \left(\frac{v \prod_{i=1}^{n} r_i / \beta^{(n+s)}}{\Gamma(\alpha)^{n+t}}\right)^{\alpha};$$  \hspace{1cm} \text{(3.8)}$$

The full conditionals in equations (3.7) and (3.8) are sampled using the Metropolis algorithm within Gibbs. For the prior used here, one can show $\pi(\beta | r, \alpha)$ is

$$\pi(\beta | r, \alpha) \propto e^{-\frac{q + \sum_{i=1}^{n} r_i}{\beta}} \beta^{-\alpha(n+s)},$$

which is the kernel of an Inverse Gamma density with shape parameter $\alpha(n+s) - 1$ and scale parameter $q + \sum_{i=1}^{n} r_i$. Therefore, $\beta$ can be generated from the inverse gamma distribution and a Metropolis algorithm within Gibbs sampling is not necessary.

3.2.1 Finding the posterior mode by the EM algorithm

To approximate the distribution of $S$, estimates of $\alpha$ and $\beta$ in the $r$ distribution are needed. As stated in the general procedure of Section 3.1.2, these may be found from the posterior mode (or mean) of the values obtained above, or by using the EM algorithm, to maximize the posterior. This is an iterative procedure with parameters $\alpha$ and $\beta$ changing at each iteration. Let $\alpha^{(k)}$ and $\beta^{(k)}$ be the values at the $k$th iteration.
As defined in the previous section, under $H_0$, $f(r_i; \alpha, \beta)$ is a gamma distribution and

$$f(x_i; \alpha, \beta) = \int_0^\infty f(r_i; \alpha, \beta) f(x_i | r_i) dr_i$$

$$= \frac{x_i}{\Gamma(\alpha) \beta^\alpha} \int_r^\infty \frac{r_i^{\alpha-2} \exp(-r_i/\beta)}{\sqrt{r_i^2 - x_i^2}} dr_i.$$  

(3.9)

For convenience, define $M(x_i, \alpha, \beta) = \int_0^\infty \frac{r_i^{\alpha-2} \exp(-r_i/\beta)}{\sqrt{r_i^2 - x_i^2}} dr_i$. Then the likelihood of $x$ is

$$\prod_{i=1}^n f(x_i; \alpha, \beta) = \left\{ \frac{1}{\beta^\alpha \Gamma(\alpha)} \right\} \prod_{i=1}^n \{x_i M(x_i, \alpha, \beta)\}.$$  

(3.10)

The joint density of $x$ and $r$ is

$$\prod_{i=1}^n f(x_i, r_i; \alpha, \beta) = \left\{ \frac{1}{\beta^\alpha \Gamma(\alpha)} \right\} \prod_{i=1}^n \left[ \exp\left( -\sum_{i=1}^n r_i/\beta \right) \right] \prod_{i=1}^n \left( \frac{r_i^{\alpha-2} x_i}{\sqrt{r_i^2 - x_i^2}} \right).$$  

(3.11)

This is so-called the augmented likelihood.

Then

$$f(r|x; \alpha, \beta) = \left\{ \frac{1}{\beta^\alpha \Gamma(\alpha)} \right\} \prod_{i=1}^n \frac{r_i^{\alpha-2} \exp(-r_i/\beta)}{\sqrt{r_i^2 - x_i^2}} M(x_i, \alpha, \beta)/\sqrt{r_i^2 - x_i^2}.$$  

(3.12)

The conjugate prior of $\alpha$ and $\beta$ in equation (3.6) is used. For convenience let $s = t$. The log of the augmented posterior is

$$\log\{p(\alpha, \beta | x, r, s, v, q)\} \propto (n + s)[-\alpha \log(\beta) - \log\{\Gamma(\alpha)\}]$$

$$+ (\alpha - 1) \sum_{i=1}^n \log(vr_i) - \left( \sum_{i=1}^n r_i + q \right) / \beta + \sum_{i=1}^n \log \left( \frac{x_i}{r_i \sqrt{r_i^2 - x_i^2}} \right).$$  

(3.13)

The EM algorithm alternates 2 steps: the expectation $E$ step and the maximization $M$ step. For this iterative procedure, on the $E$ step take the expectation of (3.13) with respect
to \( r \) to get

\[
Q\{(\alpha, \beta)|{(\alpha^{(k)}, \beta^{(k)})}\} = (n + s)[-\alpha \log(\beta) - \log\{\Gamma(\alpha)\}]
\]

\[
+ (\alpha - 1) \left\{ \log v + \sum_{i=1}^{n} \frac{1}{M(x_i, \alpha^{(k)}, \beta^{(k)})} \int_{x_i}^{\infty} \log(r_i) r_i^{\alpha^{(k)}-2} \exp(-r_i/\beta^{(k)})/\sqrt{r_i^2 - x_i^2} \, dr_i \right\}
\]

\[
- \frac{1}{\beta} \left\{ q + \sum_{i=1}^{n} \frac{1}{M(x_i, \alpha^{(k)}, \beta^{(k)})} \int_{x_i}^{\infty} r_i r_i^{\alpha^{(k)}-2} \exp(-r_i/\beta^{(k)})/\sqrt{r_i^2 - x_i^2} \, dr_i \right\}
\]

(3.14)

On the \( M \) step, the quantity (3.14) is maximized by setting the following equations to 0 and solving for \( \alpha \) and \( \beta \):

\[
\frac{\partial Q}{\partial \alpha} = (n + s)\{-\alpha \log(\beta) - \Phi(\alpha)\}
\]

\[
+ \log v + \sum_{i=1}^{n} \frac{1}{M(x_i, \alpha^{(k)}, \beta^{(k)})} \int_{x_i}^{\infty} \log(r_i) r_i^{\alpha^{(k)}-1} \exp(-r_i/\beta^{(k)})/\left( r_i \sqrt{r_i^2 - x_i^2} \right) \, dr_i
\]

(3.15)

and

\[
\frac{\partial Q}{\partial \beta} = \alpha \beta (n + s)
\]

\[
- q - \sum_{i=1}^{n} \frac{1}{M(x_i, \alpha^{(k)}, \beta^{(k)})} \int_{x_i}^{\infty} r_i r_i^{\alpha^{(k)}-1} \exp(-r_i/\beta^{(k)})/\left( r_i \sqrt{r_i^2 - x_i^2} \right) \, dr_i
\]

(3.16)

The integrals are evaluated numerically. Finally, \( E \) steps and \( M \) steps are alternated until convergence is achieved.

It happens that Krumbein had one sample of both \( x \) and corresponding \( r \). It is not recorded which \( x \) goes with which \( r \). The values are given in Table 3.1, grouped into cells.

Using these \( x \) data, the value of the statistic is \( S = 0.199 \).

<table>
<thead>
<tr>
<th>Table 3.1: Table: Number of sandstone measurements in given intervals.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell midpoint (in mm)</td>
</tr>
<tr>
<td>06</td>
</tr>
<tr>
<td>( x )</td>
</tr>
<tr>
<td>( r )</td>
</tr>
</tbody>
</table>

The EM algorithm was used to estimate \( \alpha \) and \( \beta \) and the values obtained were close to
the posterior means. The estimated distribution of $S$ is given in Figure 3.2 and the $p$-value of $S$ is 0.043. A 2-parameter gamma distribution for the variable $r$ is rejected at level 0.05.

Since the $r$ values are available for this set, they can be tested directly using Cramér-von Mises statistics for discrete data described in Spinelli, Lockhart and Stephens (2007). Again, a 2-parameter gamma distribution does not fit, however, when a 3-parameter gamma is fitted, the estimates are $\hat{\alpha} = 1.708, \hat{\beta} = 0.050$ and the location parameter $\hat{\eta} = 0.107$. The goodness-of-fit statistic $W^2$ has $p$ value 0.442 and $A^2$ has $p$ value 0.321 indicating a good fit.

### 3.3 Testing a frailty distribution.

Frailty models are random effects models for time-to-event data. They are commonly seen in the literature on survival analysis. It is often convenient to assume $r_i$ follows a lognormal or gamma distribution. However, Shih and Louis (1995) showed that different frailty distributions induce quite different dependence structures. Therefore, it is necessary to examine the adequacy of the frailty distribution for intra-cluster dependence. A frailty term can be
viewed as a latent variable. In this section, the procedure to test latent variables is applied to test the fit of a frailty term to a given distribution.

Example: Kidney dialysis data. McGilchrist and Aisbett (1991) gave the recurrence times to kidney infection for 38 patients using portable dialysis equipment. A recurrence time of infection is the time from the moment a catheter is inserted into a patient to when it is removed due to infection. If the catheter was removed for reasons other than kidney infection, the time of infection is regarded as censored. When patients experienced any form of discomfort, they were told to stop using the equipment and to go to hospital. For each patient, first and second recurrence of kidney infection were recorded. After the occurrence of the first infection, patients were treated in the hospitals and sufficient time (ten weeks) was allowed for the infection to be cured before the catheter was inserted to the patient for the second time. In this case, each patient is a cluster of size 2 and there is a frailty term $r_i$ for patient $i$. Conditioning on the common frailty term, it is assumed that the first and the second recurrence times are independent.

The data set contains three risk variables (covariates): age, sex and disease. The diseases are Glomerulo Neptiritis (GN), Acute Neptiritis (AN) and Polycryatic Kidney Disease (PKD). Suppose the Cox’s Proportion Hazards model is fitted to the time-to-event data. Let $y_{ij}$ be the observed recurrence time to infection and $\delta_{ij}$ be the indicator ($\delta_{ij} = 1$, event occurred; $\delta_{ij} = 0$, right censored) for $i = 1, \ldots, 38$ and $j = 1, 2$. Suppose $r_i$ is a random effect (frailty term) for patient $i$. With the Cox model, the likelihood contribution of the $i$th patient in the $j$th recurrence is (conditional on $r_i$)

$$L_{ij} = \{h_0(y_{ij})r_i \exp(\mu_{ij})\}^{\delta_{ij}} \exp\{-H_0(y_{ij})r_i \exp(\mu_{ij})\}.$$

In the analysis, a Weibull baseline hazard is used, $h_0(y_{ij}) = \lambda \rho y_{ij}^{\rho-1}$, with cumulative hazard $H_0(y_{ij}) = \lambda y_{ij}^\rho$. The parameter $\mu_{ij}$ is a linear function of 5 covariates:

$$\mu_{ij} = \beta_1 AGE_{ij} + \beta_2 SEX_i + \beta_3 DISEASE_{i1} + \beta_4 DISEASE_{i2} + \beta_5 DISEASE_{i3},$$

where $AGE_{ij}$ is a continuous covariate, $SEX_i$ has values (0 = male, 1 = female), $DISEASE_{ik}$, for $k = 1, 2, 3$ are dummy variables representing present or absence of 3 different diseases. It is often assumed the frailty term $r_i$ follows a lognormal$(0, \tau)$ distribution. This model was fitted to the kidney dialysis data using the R package parfm, which maximizes the marginal likelihood by integrating out the frailty parameter. The results are in Table 3.2. The table
shows that $SEX$ is the only significant risk variable in the model and the risk of infection was higher for male patients.

Table 3.2: Maximum likelihood estimates of Cox PH model with Weibull baseline hazard and lognormal frailty fitted to the Kidney data.

<table>
<thead>
<tr>
<th></th>
<th>ESTIMATE</th>
<th>SE</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>0.315</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>1.162</td>
<td>0.110</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.013</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>factor(sex)F</td>
<td>-1.877</td>
<td>0.427</td>
<td>0.000</td>
</tr>
<tr>
<td>age</td>
<td>0.003</td>
<td>0.011</td>
<td>0.792</td>
</tr>
<tr>
<td>factor(disea)GN</td>
<td>0.121</td>
<td>0.486</td>
<td>0.803</td>
</tr>
<tr>
<td>factor(disea)AN</td>
<td>0.606</td>
<td>0.467</td>
<td>0.194</td>
</tr>
<tr>
<td>factor(disea)PKD</td>
<td>-1.142</td>
<td>0.732</td>
<td>0.119</td>
</tr>
</tbody>
</table>

3.3.1 Test for the frailty distribution

The model includes a random effect (frailty term) which is assumed to have a lognormal distribution, with $\tau$ unknown. For testing the hypothesized frailty distribution,

$$Ho : r_i \sim \text{lognormal}(0, \tau),$$

the general steps for testing the latent variable in Section 3.1.2 are followed. The model parameters are given “non-informative” priors: $\lambda, \beta_l \sim N(0, 10000)$ for $l = 1, \ldots, 5$ and $\tau \sim \text{Gamma}(\text{shape}= 0.0001, \text{rate}= 0.0001)$. The shape parameter $\rho$ of the Weibull distribution is assigned a Gamma$(1, 0.0001)$ prior. To generate samples from the posterior densities, the Metropolis algorithm within the Gibbs sampler is required. To improve efficiency, the block updating strategy suggested by Gamerman (1996) was used.

After the first 1000 samples (burn in period), 9000 posterior samples were used to calculate $S^*$ in equation (3.2) and hence find $S$, the average $S^*$. The value of $S$ is 0.165.

To approximate the $p$-value of $S$, $M$ bootstrap samples are generated from the fitted model in Table 3.2. The steps to generate censored survival data are in Section 3.3.2. For each new sample, a Bayes assisted goodness of fit statistic is calculated. The distribution of $S$ is approximated by $M = 100$ values. When test statistic $S$ is compared with the
approximated distribution of $S$, the p-value is 0.86. Thus $H_0$ is not rejected and a lognormal distribution fits the frailty random effect.

### 3.3.2 Generating data:

To generate samples from survival models with censoring algorithm 7.2 of Davison and Hinkley (1997) was used. The steps are as follows:

1. Generate $r_i$ from lognormal $(0, \hat{\tau})$, for $i = 1, \ldots, 38$.

2. Conditional on $r_i$, generate $Y_{ij}^{0*}$ from the estimated failure time survivor function \[ \{1 - \hat{F}_0(y)\} r_i \exp(\hat{\mu}_{ij}), \] where $\hat{F}_0(y) = 1 - \exp(-H_0(y))$.

3. if $\delta_{ij} = 1$ (not censored), set $C_{ij}^* = y_{ij}$, and if $\delta_{ij} = 0$, generate $C_{ij}^*$ from conditional censoring distribution \[ \frac{\hat{G}(y) - \hat{G}(y_{ij})}{1 - \hat{G}(y_{ij})}, \] where $\hat{G}(y)$ is a Kaplan Meier estimate of the censoring function.

4. Set $Y_{ij}^* = \min(Y_{ij}^{0*}, C_{ij}^*)$, and let $\delta_{ij} = 0$ if $Y_{ij}^* = Y_{ij}^{0*}$ and zero otherwise.

5. the combination $(Y_{ij}^*, \delta_{ij}^*)$ constitutes the bootstrap value for patient $i$ and occurrence $j$.

6. Generate samples for patient $i, i = 1, \ldots, 38$.

A check was made to compare the estimates used in generating the bootstrap samples with the average estimates given by the fitted model when the frailty distribution was lognormal. The comparisons are in Table 3.3. The comparisons suggest the samples were generated correctly.

| Table 3.3: Comparisons of coefficients used to generate samples with the average coefficients from bootstrap samples and with posterior means |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | $\lambda$      | $\rho$         | $\sigma$       | $\beta_1$      | $\beta_2$      | $\beta_3$      | $\beta_4$      | $\beta_5$      |
| Original       | -4.34          | 1.16           | 0.56           | -1.88          | 0.003          | 0.12           | 0.61           | -1.14          |
| Bootstrap      | -4.14          | 1.22           | 0.49           | -1.97          | 0.003          | 0.13           | 0.61           | -1.14          |
| Pos. Means     | -4.70          | 1.20           | 0.70           | -2.00          | 0.003          | 0.10           | 0.60           | -0.90          |
3.3.3 Sensitivity analysis

When the frailty model is fitted, it is common to test if $\beta_i = 0, i = 1, \ldots, 5$. The test level will be affected if a wrong model is fitted to the distribution of the frailty. To show this, a small sensitivity study was conducted under the framework of the Kidney dialysis example. There are two important questions: (1) when the frailty distribution is correct, but the sample size is relative small, is the level of the test $\beta_i = 0$ (Wald’s statistic) correct? (2) if a wrong frailty distribution is used, is the level strongly affected?

The steps are as follows:

1. generate data from the fitted PH model in Table 3.2 with lognormal frailty (the correct model) using the method described in Section 3.3.2;

2. fit three models, one with a lognormal frailty, one with a gamma frailty and the other with a stable frailty;

3. repeat steps 1 and 2 $M$ times and calculate the $p$—values for each of the $\beta$ tests;

4. test the $p$—values for uniformity by a Q-Q plot.

Figure 3.3.3 shows comparisons of the levels of the Wald’s statistic (left: $\beta_{age} = 0$, right: $\beta_{age} = 0$) when different frailty distributions are assumed. The plots, which are plots of the quantiles of the simulated $p$-values against Uniform$[0,1]$ quantiles, show only the lower test levels, as these are of major interest.

Comments. With reference to (a) when the correct frailty was lognormal and fitted frailty model is also lognormal, the empirical level of the Wald’s test (blue) is higher than the nominal level. This could be due to the small sample size. With reference to (b) when the frailty distribution is lognormal but a gamma is fitted, the levels (green) are close to those of a lognormal fit; when a stable distribution is fitted, the levels (red) are worse and the discrepancy get larger as the nominal level increases.

3.4 Conclusion

Frailty models and other random effects models are widely used in the literature, but the assumed distribution of the random effect cannot usually be tested. In this work, Bayesian methods are used to assist Neyman-Pearson theory to test these distributions. The test
Figure 3.3: Q-Q plots of the $p$ values from Wald’s statistics for testing the hypothesis that the indicated coefficient is 0 against quantiles from the standard uniform distribution

Test levels of Ho: $\beta_1(AGE)=0$

Test levels of Ho: $\beta_2(SEX)=0$

statistics depend on the Bayesian prior and this prior is chosen to lead to the Cramér-von Mises $W^2$, which is known to give tests with good power. The general procedure could be applied to testing other latent variable problems occurring in biostatistics, for example random effect logistic regression.
3.5 References


Chapter 4

K sample Cramer von Mises Statistics for ordered contingency tables

In this Chapter $K$-sample tests are given for discrete or grouped data presented as counts in the cells of a $K$ by $m$ contingency table. The test statistics are discrete versions of the classical Cramér-von Mises statistics $W^2$, $U^2$, and $A^2$.

The counts may occur in several ways, the best-known when data are counts in categories (voters for political parties, or students with different grades, etc.). Counted data may also occur when continuous data are grouped, perhaps because the values are measured only within a limited range (angles at which migratory birds disappear from sight) or when times are grouped for display purposes (incidences of disease each month, or accidents each hour or day of the week).

Suppose there are $K$ samples, with counts recorded in the rows of the contingency table; the null hypothesis to be tested is

$H_0$ : the $K$-samples have the same parent population, discrete or continuous.

Equivalently, on $H_0$, the probability $p_j$ of an observation falling into cell $j$ is the same for all samples.

The well-known statistic for testing $H_0$ is Pearson’s $X^2$ statistic, often called $\chi^2$ from its asymptotic distribution. This statistic gives the same value for all orderings of the cells.

However, the cells may be ordered naturally by some criterion; for example, grades in
an examination from \( F \) to \( A \), tones in photographs from light to dark, or opinions from a survey (strongly opposed, opposed, neutral, in favour, strongly in favour) on a new policy in different regions. The table is then called an ordered contingency table. Here, statistics are examined which depend on the cell order, with the aim of obtaining better power than the \( X^2 \) statistic.

For a single sample of grouped data, Choulakian, Lockhart and Stephens (1994), here called CLS, gave tests of Cramér-von Mises type when cell probabilities are known, and Spinelli, Lockhart and Stephens (2007, called SLS) extended these to the case where parameters must be estimated from the sample. The Cramér-von Mises statistics depend on the order of the cells.

### 4.1 \( K \)-Sample Cramér-von Mises Statistics

In this Section the Cramér-von Mises test statistics are defined as follows: suppose the data for the \( i \)th sample are recorded in row \( i \); there are \( m \) cells, and the data are counts \( o_{ij}, \ i = 1, \ldots, K, \ j = 1, \ldots, m \). Let \( n_i = \sum_{j=1}^{m} o_{ij} \) be the number of observations in the \( i \)th sample; \( N = \sum_{i=1}^{K} n_i \) is the total sample size.

Let \( p_j \) be the probability that an observation falls into cell \( j \). On \( H_0 \), \( p_j \) is the same for all samples. Define \( Z_{ij} = S_{ij} - T_{ij} \) where \( S_{ij} = \sum_{l=1}^{j} o_{il} \) and \( T_{ij} = n_i \sum_{l=1}^{j} p_l \). Let \( \bar{Z}_i = \sum_{j=1}^{m} Z_{ij}p_j \), and let \( H_j = \sum_{i=1}^{K} p_i \), for \( j = 1, \ldots, m \).

The discrete \( K \) sample statistics are then defined by

\[
W_K^2 = \sum_{i=1}^{K} n_i^{-1} \sum_{j=1}^{m} Z_{ij}^2 p_j, \tag{4.1}
\]

\[
A_K^2 = \sum_{i=1}^{K} n_i^{-1} \sum_{j=1}^{m} Z_{ij}^2 p_j / \{H_j(1 - H_j)\}, \tag{4.2}
\]

and

\[
U_K^2 = \sum_{i=1}^{K} n_i^{-1} \sum_{j=1}^{m} (Z_{ij} - \bar{Z}_i)^2 p_j, \tag{4.3}
\]

Note that since \( Z_{im} = 0 \) and \( H_m = 1 \) the last term of \( A_K^2 \) is \( 0/0 \) and is set to 0.

Let \( \mathbf{p}' = (p_1, p_2, \ldots, p_m) \). In practice, the \( p_j \) are usually unknown, and are estimated
from the samples by

\[ \hat{p}_j = \frac{\sum_{i=1}^{K} o_{ij}}{N}. \]  

(4.4)

Let \( \mathbf{\hat{p}}' \) be the estimated \((\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)\); \( \mathbf{\hat{p}} \) will be used in the above definitions when \( p \) is unknown. In the rest of this chapter, suppose \( T \) is one of the Cramér-von Mises statistics calculated from data with known \( p \), and suppose the statistic is labelled \( \hat{T} \) when it is calculated using \( \mathbf{\hat{p}} \).

Scholz and Stephens (1987) studied the \( K \) sample Anderson-Darling statistic \( A^2 \) for continuous data, and standardised the statistic to give a test statistic with percentage points which do not depend significantly on the cell probabilities. A similar standardization will be used here for the grouped Cramér-von Mises statistics. For continuity the test procedure will be given in this Section, and theory deferred to Section 4.2.

### 4.1.1 \( K \)-sample Test Procedure

The procedure will be given only for the practical case when \( p \) is estimated.

1. Calculate the desired Cramér-von Mises statistic \( \hat{T} \) from Section 4.1, using \( \mathbf{\hat{p}} \);
2. Find the estimated mean \( \hat{\mu} \) and standard deviation \( \hat{\sigma} \) of the asymptotic distribution of \( \hat{T} \), as given equations (4.12) and (4.13) in Section 4.2;
3. Standardize the test statistic using

\[ \hat{T}_K^* = \frac{\hat{T} - \hat{\mu}}{\hat{\sigma}}; \]  

(4.5)

4. Refer \( \hat{T}_K^* \) to the percentage points of the appropriate Table 4.1, 4.2, or 4.3; \( H_0 \) is rejected at level \( \alpha \) if \( \hat{T}_K^* \) is greater than the \( \alpha \) point in the table.
Table 4.1: Upper-tail percentage points for standardized $\hat{W}_K^2$

<table>
<thead>
<tr>
<th>$K \setminus \alpha$</th>
<th>0.25</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.295</td>
<td>1.252</td>
<td>2.012</td>
<td>2.791</td>
<td>3.838</td>
</tr>
<tr>
<td>3</td>
<td>0.431</td>
<td>1.320</td>
<td>1.977</td>
<td>2.626</td>
<td>3.476</td>
</tr>
<tr>
<td>4</td>
<td>0.486</td>
<td>1.336</td>
<td>1.943</td>
<td>2.532</td>
<td>3.292</td>
</tr>
<tr>
<td>5</td>
<td>0.518</td>
<td>1.340</td>
<td>1.916</td>
<td>2.469</td>
<td>3.175</td>
</tr>
<tr>
<td>6</td>
<td>0.538</td>
<td>1.341</td>
<td>1.896</td>
<td>2.424</td>
<td>3.093</td>
</tr>
<tr>
<td>7</td>
<td>0.553</td>
<td>1.341</td>
<td>1.879</td>
<td>2.389</td>
<td>3.032</td>
</tr>
<tr>
<td>8</td>
<td>0.564</td>
<td>1.340</td>
<td>1.866</td>
<td>2.361</td>
<td>2.983</td>
</tr>
<tr>
<td>10</td>
<td>0.579</td>
<td>1.337</td>
<td>1.845</td>
<td>2.319</td>
<td>2.910</td>
</tr>
<tr>
<td>20</td>
<td>0.613</td>
<td>1.327</td>
<td>1.790</td>
<td>2.214</td>
<td>2.734</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.674</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
</tr>
</tbody>
</table>

Table 4.2: Upper-tail percentage points for standardized $\hat{A}_K^2$

<table>
<thead>
<tr>
<th>$K \setminus \alpha$</th>
<th>0.25</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.312</td>
<td>1.263</td>
<td>2.011</td>
<td>2.774</td>
<td>3.798</td>
</tr>
<tr>
<td>3</td>
<td>0.442</td>
<td>1.324</td>
<td>1.972</td>
<td>2.609</td>
<td>3.441</td>
</tr>
<tr>
<td>4</td>
<td>0.494</td>
<td>1.337</td>
<td>1.936</td>
<td>2.515</td>
<td>3.261</td>
</tr>
<tr>
<td>5</td>
<td>0.525</td>
<td>1.340</td>
<td>1.910</td>
<td>2.454</td>
<td>3.148</td>
</tr>
<tr>
<td>6</td>
<td>0.544</td>
<td>1.341</td>
<td>1.889</td>
<td>2.409</td>
<td>3.068</td>
</tr>
<tr>
<td>7</td>
<td>0.558</td>
<td>1.340</td>
<td>1.873</td>
<td>2.375</td>
<td>3.008</td>
</tr>
<tr>
<td>8</td>
<td>0.568</td>
<td>1.339</td>
<td>1.860</td>
<td>2.348</td>
<td>2.961</td>
</tr>
<tr>
<td>10</td>
<td>0.583</td>
<td>1.336</td>
<td>1.839</td>
<td>2.307</td>
<td>2.890</td>
</tr>
<tr>
<td>20</td>
<td>0.616</td>
<td>1.326</td>
<td>1.786</td>
<td>2.206</td>
<td>2.720</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.674</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
</tr>
</tbody>
</table>

**Comment 1.** The asymptotic distribution of $\hat{T}$ for $K$ samples ($p$ estimated) is the same as that of $T$ for $K-1$ samples ($p$ known).

**Comment 2.** Ideally the exact mean and variance of $\hat{T}$ would be used in the standardization above. When $p$ is known, these parameters may be calculated and formulas are given in Section 4.3. However, when $p$ is not known, calculation of the exact variance is very difficult: see Scholz and Stephens (1987) for such a calculation for $A^2$ for continuous data. Therefore the asymptotic mean and variance have been used above.
Table 4.3: Upper-tail percentage points for standardized $\hat{U}_K^2$

<table>
<thead>
<tr>
<th>$K \setminus \alpha$</th>
<th>0.25</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.435</td>
<td>1.322</td>
<td>1.975</td>
<td>2.620</td>
<td>3.463</td>
</tr>
<tr>
<td>3</td>
<td>0.520</td>
<td>1.340</td>
<td>1.914</td>
<td>2.464</td>
<td>3.166</td>
</tr>
<tr>
<td>4</td>
<td>0.555</td>
<td>1.340</td>
<td>1.877</td>
<td>2.384</td>
<td>3.023</td>
</tr>
<tr>
<td>5</td>
<td>0.574</td>
<td>1.338</td>
<td>1.852</td>
<td>2.334</td>
<td>2.936</td>
</tr>
<tr>
<td>6</td>
<td>0.586</td>
<td>1.336</td>
<td>1.834</td>
<td>2.298</td>
<td>2.875</td>
</tr>
<tr>
<td>7</td>
<td>0.595</td>
<td>1.333</td>
<td>1.820</td>
<td>2.271</td>
<td>2.829</td>
</tr>
<tr>
<td>8</td>
<td>0.603</td>
<td>1.331</td>
<td>1.808</td>
<td>2.248</td>
<td>2.790</td>
</tr>
<tr>
<td>10</td>
<td>0.613</td>
<td>1.327</td>
<td>1.791</td>
<td>2.216</td>
<td>2.737</td>
</tr>
<tr>
<td>20</td>
<td>0.634</td>
<td>1.316</td>
<td>1.749</td>
<td>2.140</td>
<td>2.612</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.674</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
</tr>
</tbody>
</table>

Comment 3. The tables were calculated from the asymptotic distribution of $\hat{T}$, given in Section 4.2, assuming $K$ samples and $m = 5$ cells. The true $p_j$ were all equal to $1/m$. Various combinations of $m$ and true $p$ show that the distribution of standardized $\hat{T}_K^*$ is almost independent of these parameters, except for extreme cases of $p$. A small demonstration is in Table 4.4, where points are given for $\hat{W}_4^2$, using the four sets $p = (0.2, 0.2, 0.2, 0.2, 0.2)'$, $p = (0.05, 0.2, 0.5, 0.2, 0.05)'$, $p = (0.5, 0.25, 0.125, 0.0625, 0.03125)'$, and $p = (0.333, 0.167, 0, 0.167, 0.333)'$. It may be seen that there is little variation in the percentage points for the last three sets compared with the first, which is that used to create Table 4.1.

Comment 4. Finally, Monte Carlo studies with 10,000 samples of different sizes showed that points for finite samples converge quickly to the asymptotic, as has been observed in the continuous versions of $W^2$, $A^2$ and $U^2$.

Thus, although the test level will not be strictly the nominal $\alpha$, and the calculated $p$-level of the test will not be strictly accurate, the simplicity of the standardization and the tables should usually outweigh any small errors in the true level of the test.

Comment 5. As $K \to \infty$, the distribution of $T_K^*$ tends to standard normal by the Central Limit Theorem.
4.1.2 Reverse Ordering

For ordered contingency tables the data (for example, grades) might be given in the order A to F or in the order F to A, and tones of photographs could be light to dark or equally dark to light. Lockhart, Spinelli and Stephens (2007) observed that, as defined, a one-sample grouped-data Cramér-von Mises statistic will take a different value if the cell orderings and counts are completely reversed. This can be avoided by replacing the cell weights \( p_i \) in the definitions by \( w_i = \frac{(p_i + p_{i+1})}{2} \), with \( p_{m+1} = p_1 \). In making the test, Tables 4.1, 4.2, and 4.3 should still be used, since the percentage points obtained using weights \( p \) or \( w \) will be almost the same and any error in \( \alpha \) from the tabulated value will be negligible. This can be seen from Table 4.4, where percentage points obtained using weights \( p \) or \( w \) are compared for \( K = 4 \) and for four sets of \( p \). For \( p_j \) all equal, the statistics will take the same value.

Table 4.4: Comparison of percentage points for different sets \( p \) using the \( p \) or \( w \) \((w_i = (p_i + p_{i+1})/2)\) weights

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>.750</th>
<th>.90</th>
<th>.95</th>
<th>.975</th>
<th>.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell probabilities</td>
<td>2</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>( p )</td>
<td>0.486</td>
<td>1.336</td>
<td>1.943</td>
<td>2.532</td>
<td>3.292</td>
</tr>
<tr>
<td>( w )</td>
<td>0.486</td>
<td>1.336</td>
<td>1.943</td>
<td>2.532</td>
<td>3.292</td>
</tr>
<tr>
<td>cell probabilities</td>
<td>.05</td>
<td>.2</td>
<td>.5</td>
<td>.2</td>
<td>.05</td>
</tr>
<tr>
<td>( p )</td>
<td>0.484</td>
<td>1.326</td>
<td>1.932</td>
<td>2.525</td>
<td>3.299</td>
</tr>
<tr>
<td>( w )</td>
<td>0.507</td>
<td>1.330</td>
<td>1.913</td>
<td>2.480</td>
<td>3.213</td>
</tr>
<tr>
<td>cell probabilities</td>
<td>.5</td>
<td>.25</td>
<td>.125</td>
<td>.0625</td>
<td>.03125</td>
</tr>
<tr>
<td>( p )</td>
<td>0.470</td>
<td>1.326</td>
<td>1.946</td>
<td>2.555</td>
<td>3.349</td>
</tr>
<tr>
<td>( w )</td>
<td>0.472</td>
<td>1.325</td>
<td>1.943</td>
<td>2.549</td>
<td>3.341</td>
</tr>
<tr>
<td>cell probabilities</td>
<td>.333</td>
<td>.167</td>
<td>0</td>
<td>.167</td>
<td>.333</td>
</tr>
<tr>
<td>( p )</td>
<td>0.472</td>
<td>1.325</td>
<td>1.944</td>
<td>2.550</td>
<td>3.342</td>
</tr>
<tr>
<td>( w )</td>
<td>0.475</td>
<td>1.325</td>
<td>1.940</td>
<td>2.543</td>
<td>3.330</td>
</tr>
</tbody>
</table>

4.2 Theory of the Tests

For the distribution theory of the tests, the statistics will be put into matrix form and the following well-known theorem — see for example Rao (1965, p.149) — will be used.
Theorem 3  If $X = (X_1, \cdots, X_m)$ has the multivariate normal distribution $N(0, \Sigma)$ and $Y = XCX'$ for some symmetric matrix $C$, then the distribution of $Y$ is that of $Y_\infty$ and where

$$Y_\infty \sim \sum_{j=1}^{m} \lambda_j Z_j^2,$$  \hspace{1cm} (4.6)

where the $Z_j^2$ are independent $\chi^2$ random variables and the $\lambda_1, \cdots, \lambda_m$ are the eigenvalues of $C \Sigma$ or equivalently of $C^{1/2} \Sigma C^{1/2}$.

Let $I$ be an $m \times m$ identity matrix and $1'$ be a $1 \times m$ vector of ones. Suppose for the present that $p$ is known. Let $D$ be the $m \times m$ diagonal matrix with entries $D_{j,j} = p_j, j = 1, \cdots, m$, and let $E$ be the $m \times m$ diagonal matrix with entries $E_{j,j} = (p_j + p_{j+1})/2, j = 1, \cdots, m$. Also let $H$ be a diagonal matrix with $H_{j,j} = 1/\{H_j(1 - H_j)\}$ for $j = 1, \ldots, m - 1$ and $H_{m,m} = 0$. Let $o_i' = \{o_{i1}, \ldots, o_{im}\}$ and suppose $P$ is the $m \times m$ partial sum matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

For sample $i$, define $Z_i' = \{Z_{i1}, \cdots, Z_{im}\} = P(o_i - n_i p)$. Define $\Sigma = D - pp'$. Then $\Sigma_0 = n_i \Sigma$ is the covariance of $o_i$ and the covariance of $Z_i$ is $\Sigma_i = P \Sigma_0 P' = n_i P(D - pp') P'$.

Let $Y_i = Z_i/\sqrt{n_i}$. As $n_i \to \infty$, following the Central Limit Theorem, $Y_i$ has the multivariate normal distribution $N(0, \Sigma_i/n_i)$.

Suppose $t_i$ is one of the Cramér-von Mises statistics for the $i$th sample; then the statistic $t_i$ can be written as $t_i = Y_i'MY_i$ for an appropriate symmetric matrix $M$. When $M = D$, $t_i$ is discrete $W^2$; when $M = DH$, $t_i$ is discrete $A^2$; when $M = (I - D11')D(I - 11'D)$, $t_i$ is discrete $U^2$. The $K$ sample statistic becomes $T_K = \sum_{i=1}^{K} t_i$. When the weights $w_i$ of Section 4.1.2 are used, the $M$ above become $M = E$ for discrete $W^2$, $M = EH$ for discrete $A^2$, and $M = (I - D11')E(I - 11'D)$ for discrete $U^2$. For the asymptotic distribution of $t_i$, the following one sample theory was given by Choulakian, Lockhart and Stephens (1994).
Let $\lambda_j, j = 1, \cdots, m - 1$ be the $m - 1$ nonzero eigenvalues of $Q^* = M^{1/2} \Sigma_i/n_i M^{1/2}$ and $w_j$ be the corresponding normalized (length 1) eigenvectors. Then

$$t_i = Y_i' MY_i = \sum_{j=1}^{m-1} \lambda_j (w_j Y_i)^2,$$

(4.7)

and following Theorem 3, the asymptotic distribution of $t_i$ is that of $t_\infty$ where

$$t_\infty \sim \sum_{j=1}^{m-1} \lambda_j c_j^2,$$

(4.8)

where the $c_j^2$ are independent $\chi^2_1$ random variables.

The term $(w_j Y_i)^2$ in (4.7) is called the $j$th component of $t_i$ and the components of the Cramér-von Mises statistics are discussed later in the chapter.

For $K$ samples, with known $p_j$, the statistics $t_i, i = 1, 2, \cdots, K$ are independent, and the $K$ sample overall statistic is $T = \sum_{i=1}^{K} t_i$. The limiting distribution of $T$ is thus that of

$$T_\infty = \sum_{i=1}^{K} \sum_{j=1}^{m} \lambda_j c_j^2 = \sum_{j=1}^{m-1} \lambda_j s_j^2,$$

(4.9)

where $s_j^2$ are independent $\chi^2_K$ random variables.

In this case ($p$ known) a $K$ sample test seems unnecessary - a practical statistician would surely test each sample separately.

### 4.2.1 Matrix form for $K$ sample theory.

When $p$ must be estimated, it is useful to put the $K$-sample theory in matrix form.

Define $Y$ be a $mK \times 1$ vector: $Y' = \{Y_1', \cdots, Y_K'\} = \{Z_1'/\sqrt{n_1}, \cdots, Z_K'/\sqrt{n_K}\}$; and define a block diagonal matrix

$$M' = \begin{pmatrix} M & 0 & 0 & \cdots & 0 \\ 0 & M & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & M \end{pmatrix}.$$

Since under the null hypothesis, vector $p$ is the same for every sample, the statistic $T$ can
be written as
\[ T = Y' M^* Y; \]
the covariances of \( Y_i \) are the same for \( i = 1, \ldots, K \). Define \( Q = P(D - pp')P' \), and \( Q^* = M^{1/2} Q M^{1/2} \). When \( p \) known, for \( i \neq j \), \( Y_i \) and \( Y_j \) are independent and therefore \( Y \) is asymptotically normal with mean 0 and covariance \( \Sigma_Y \), where
\[
\Sigma_Y = \begin{pmatrix}
Q & 0 & 0 & \ldots & 0 \\
0 & Q & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & Q
\end{pmatrix}.
\]
The matrix \( \Sigma_Y \) can be written as \( Q \otimes I \), where \( \otimes \) denotes the Kronecker product and \( I \) is a \( K \times K \) identity matrix. Again, following Theorem 3, the asymptotical distribution of \( T \) is that of \( T_\infty \) where
\[
T_\infty \sim \sum_{j=1}^{K \times m} \lambda_j c_j^2,
\]
where the \( c_j^2 \) are independent \( \chi_1^2 \) random variables and the \( \lambda_1, \ldots, \lambda_{Km} \) are the eigenvalues of \( M^{1/2} \Sigma_Y M^{1/2} \), which can be written as \( Q^* \otimes I \). These are clearly \( K \) copies of the \( \lambda_j \) given in (4.8).

When \( \hat{p} \) is used, the estimates will be used in all the above formulas in the obvious way. Let the typical statistic now be \( \hat{T}_K \), and let \( \hat{Y}, \hat{Z}_i, \hat{Q}^* \) and \( \hat{Q} \) be corresponding estimates. The \( \hat{Y}_i \) are correlated and the covariance \( \Sigma_{\hat{Y}} \) of \( Y \) becomes
\[
\Sigma_{\hat{Y}} = Q \otimes B,
\]
where
\[
B = \begin{pmatrix}
1 - 1/K & -1/K & \ldots & -1/K \\
-1/K & 1 - 1/K & \ldots & -1/K \\
\vdots & \vdots & \ddots & \vdots \\
-1/K & -1/K & \ldots & 1 - 1/K
\end{pmatrix}.
\]
To obtain the eigenvalues, the following results are used.

Results:
Suppose \( \delta_i, i = 1, \ldots, m \) are the eigenvalues of an arbitrary \( m \times m \) matrix \( U \) and similarly
suppose $\mu_j, j = 1, \cdots, k$ are eigenvalues of a $k$ by $k$ matrix $V$. Then $\delta_i\mu_j$ are the $mk$ eigenvalues of $mk$ by $mk$ matrix $U \otimes V$.

With reference to $\Sigma = \Sigma \otimes B$, matrix $B$ is idempotent with eigenvalues 1 (with multiplicity $K - 1$) and one 0. Therefore, the eigenvalues of $M^{1/2} \Sigma \chi M^{1/2}$ are the $\lambda_j$ in (4.8) above, occurring with multiplicity $K - 1$.

To summarize: when $p$ is known, eigenvalues of $Q^* \otimes I$ are the $\lambda_j$ in (4.8), with multiplicity $K$; when $p$ is unknown, eigenvalues of $Q^* \otimes B$ are the same $\lambda_j$, but with multiplicity $K - 1$. Thus, the asymptotic distribution of $\hat{T}_K$ is

$$\hat{T}_\infty = \sum_{j=1}^{m-1} \lambda_j s_j^2$$

(4.11)

where $s_j^2$ are independent $\chi^2_{K-1}$ random variables. The mean of $\hat{T}_\infty$ is

$$\mu = (K - 1) \sum_{j=1}^{m-1} \lambda_j$$

$$= \text{trace}(Q)$$

$$= \sum_i Q_{ii}$$

(4.12)

and the variance is

$$\sigma^2 = 2(K - 1) \sum_{j=1}^{m-1} \lambda_j^2$$

$$= \text{trace}(Q^2)$$

$$= \sum_i \sum_j Q_{ij}^2$$

(4.13)

These results are expressed in the following theorem.

**Theorem 4** Let $T_K$ be one of the Cramér-von Mises statistics $W_2^K$, $U_2^K$, or $A_2^K$, calculated from $p$ known, and let $\hat{T}_K$ be the value using $\hat{p}$. The limiting distribution of $\hat{T}_K$ is the same as that of $T_{K-1}$.

This theorem is the grouped data version of a theorem by Kiefer (1959), who showed the same result for continuous data.
4.2.2 Construction of tables

The first three cumulants of $\hat{T}_\infty$ are $\kappa_1 = (K - 1) \sum_{j=1}^{m-1} \lambda_j$, $\kappa_2 = 2(K - 1) \sum_{j=1}^{m-1} \lambda_j^2$, and $\kappa_3 = 8(K - 1) \sum_{j=1}^{m-1} \lambda_j^3$. The cumulants can be used to approximate the distribution by a statistic $S$ with distribution $a + b\chi^2_p$, where the $a$, $b$ and $p$ are chosen to match the first three cumulants of $S$ with those of $\hat{T}_\infty$. This gives $a = \kappa_1 - bp$, $b = \kappa_3 / (4\kappa_2)$, and $p = 8\kappa_3^2 / \kappa_2^3$. The points in tables 4.1-4.3 were obtained by approximating the standardized distribution of $\hat{T}_*\infty = (\hat{T}_\infty - \kappa_1) / \sqrt{\kappa_2}$, which has mean 0 and variance 1, and third cumulant $\kappa_3^* = \kappa_3 / (\kappa_2)^{1.5}$. The approximation $a + b\chi^2_p$ for $\hat{T}_*\infty$ has $b = \kappa_3^* / 4$, $p = 8 / (\kappa_3^*)^2$, and $a = -bp$.

4.3 Exact means and variances.

When $p$ is known, the means and variances of the Cramér-von Mises statistics can be calculated exactly. The formulas differ only in the matrix $M$.

For one sample, let $n$ be the sample size and define $d_j = o_{ij} - np_j$, $j = 1, \cdots, m$ and let $d' = \{d_1, \cdots, d_m\}$. Define matrix $Q = P'MA$; then the mean of the one sample Cramér-von Mises statistic is

$$E(t_i) = n_i^{-1} \sum_{s=1}^{m} \sum_{t=1}^{m} Q_{st} E(d_s d_t)$$

where $E$ denotes expectation. This formula is equivalent to one sample mean. The variance of the one sample statistic is

$$V(t_i) = n_i^{-2} \sum_{s=1}^{m} \sum_{t=1}^{m} \sum_{l=1}^{m} \sum_{g=1}^{m} Q_{st} Q_{lg} \{E(d_s d_t d_l d_g) - E(d_s d_t) E(d_l d_g)\}.$$  \hspace{1cm} (4.14)

There are various possibilities for the terms in $E(d_s d_t d_l d_g) - E(d_s d_t) E(d_l d_g)$, as follows:

$$E(d_s d_t) = \begin{cases} 
-np_s p_t & \text{if } s \neq t \\
np_s (1 - p_t) & \text{if } s = t 
\end{cases}.$$
and

\[ E(d_s d_t d_l d_g) = \begin{cases} 
3n(n - 2)p_sp_t p_g, & \text{if } s \neq t \neq l \neq g, \\
n(n - 2)p_sp_t (3p_s - 1), & \text{if } s = g \neq l \neq t, \\
p_sp_t \{3(n - 2)p_sp_t - (n - 2)(p_s + p_t) + n - 1\}, & \text{if } s = g \neq t = l, \\
p_sp_t \{3(n - 2)(p_s - 1) - 1\}, & \text{if } s = g = l \neq t, \\
p_sp \{3p_s(p_s - 1) - 6p_s(p_s - 1) - 1\}, & \text{if } s = t = l = g 
\end{cases} \]

and other combinations similarly.

For \( K \) samples, since \( \mathbf{p} \) is known, the samples are independent. Then the mean is

\[ E(T_K) = \sum_{i=1}^{K} E(t_i) \]

and the variance is

\[ V(T_K) = \sum_{i=1}^{K} V(t_i). \]

For the case when \( \mathbf{p} \) must be estimated from the samples as in (4.4), the samples are no longer independent. The mean of \( T_K \) is calculated as before but using \( \hat{\mathbf{p}} \), but as stated above, the variance becomes very difficult to calculate.

### 4.4 Related work

Taguchi (1974) modified the standard \( X^2 \) by grouping the first \( J \) columns and the last \( m - J \) columns to make \( m - 1 \) tables, for \( J = 1, \cdots, m - 1 \), each of dimension \( K \times 2 \). The \( X^2 \) statistics from these tables are weighted to give an overall test statistic, say \( T_E \). Subsequently, Nair (1986, 1987) modified Taguchi’s \( T_E \) to give a statistic \( T_C \). Nair also suggested components of the test statistics to give better power for location shifts between the samples and similarly for scale shifts.

The connections between the Cramér-von Mises statistics and Taguchi and Nair statistics can be noted. The Taguchi statistic is

\[ T_E = \sum_{i=1}^{K} n_i^{-1} \sum_{j=1}^{m} \frac{Z_{ij}^2}{H_j(1 - H_j)}; \tag{4.15} \]

this is clearly related to \( A_K^2 \). Nair suggested the following adaptation

\[ T_C = \sum_{i=1}^{K} n_i^{-1} \sum_{j=1}^{m} \frac{Z_{ij}^2}{m}, \tag{4.16} \]

which is \( W_K^2 \) with all \( p_j = 1/m \). Also, when \( \mathbf{M} = \mathbf{H} \), \( T_K \) is \( T_E \), and when all \( p_j \) equal.
$M = I/m$, $T_K$ is $T_C$.

### 4.5 Components of $K$ sample the Cramér-von Mises statistics.

In equation (4.7), the quantity $(w_j'Y_i)^2$ is called the $j$th component of the statistic. Similarly, for $K$ samples. The term $s_j^2$ in Section 4.2 is called the $j$th component of the statistic. On the null, when $p_j$ are known, the components of $K$ sample statistics are asymptotically $\chi^2_{(K)}$ distributed and when $p_j$ are unknown, they are asymptotically $\chi^2_{(K-1)}$ distributed.

Nair (1986) suggested to use the first components of $T_E$ and $T_C$ as tests for location and the second components of $T_E$ and $T_C$ as tests for dispersion. He also gave closed form approximations to these components. As CLS pointed out, the individual components can be used to describe certain features of the data, and the importance of each component is decided by the weights assigned to them. Components of the Cramér-von Mises statistics are examined in this Section.

In Section 4.8, various power studies will show that the first three components of the statistics were able to detect, respectively location (mean), scale (variance) and skewness among the $K$ samples. However, how to use components as goodness-of-fit statistics is not clear. CLS pointed out several problems with using components: it is difficult to decide how many components one should use, because this depends on the tested distribution and also the type of alternative it is desired to detect. Therefore it is hard to control the correct $\alpha$ level of the overall procedure. In the case of the Cramér-von Mises statistics, CLS recommended to use the entire statistic as a gate keeper, only using the components when the estimate statistic is significant. Several examples in Section 4.6 are analyzed following the recommendation.

### 4.6 Examples

**Example 1: Letter grades**

The counts of letter grades of 3 offerings of a Statistics course are given in Table 4.5. It is interesting to see if the grades were given in a consistent way.

The $p$ values associated with the different statistics are in Table 4.6. The two omnibus statistics $W^2_3$ and $A^2_3$ rejected, at the 0.10 level, the null hypothesis that the grades follow the
Table 4.5: Letter grades of 3 offerings of a statistics course.

<table>
<thead>
<tr>
<th></th>
<th>A+</th>
<th>A</th>
<th>A-</th>
<th>B+</th>
<th>B</th>
<th>B-</th>
<th>C+</th>
<th>C</th>
<th>C-</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offering 1</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Offering 2</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Offering 3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.6: $p$ values of different tests on the grades data.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>value</th>
<th>$p$ values (asym)</th>
<th>$p$ values (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_3^2$</td>
<td>0.736</td>
<td>0.060</td>
<td>0.046</td>
</tr>
<tr>
<td>$A_3^2$</td>
<td>3.639</td>
<td>0.072</td>
<td>0.079</td>
</tr>
<tr>
<td>$W_1$</td>
<td>6.010</td>
<td>0.050</td>
<td>0.052</td>
</tr>
<tr>
<td>$W_2$</td>
<td>2.006</td>
<td>0.367</td>
<td>0.381</td>
</tr>
<tr>
<td>$W_3$</td>
<td>0.129</td>
<td>0.937</td>
<td>0.937</td>
</tr>
<tr>
<td>$W_4$</td>
<td>2.097</td>
<td>0.350</td>
<td>0.354</td>
</tr>
<tr>
<td>$A_1$</td>
<td>5.832</td>
<td>0.054</td>
<td>0.054</td>
</tr>
<tr>
<td>$A_2$</td>
<td>2.501</td>
<td>0.286</td>
<td>0.292</td>
</tr>
<tr>
<td>$A_3$</td>
<td>0.243</td>
<td>0.885</td>
<td>0.890</td>
</tr>
<tr>
<td>$A_4$</td>
<td>0.120</td>
<td>0.942</td>
<td>0.944</td>
</tr>
</tbody>
</table>

same distribution. The components of $W_3^2$ and $A_3^2$ were calculated and only first components were significant. This suggested the distributions of grades were different in location.

Example 2: Butterfly directions

The following dataset concerns directions taken by butterflies, and was given by Dr. Marcio Zikan Cardoso of Universidade Federal do Rio Grande do Norte in Brazil, in a private communication. There are two species (Heliconius erato and Heliconius melpomene). The butterflies were taken a distance $D$ meters from their home environment (a forest) into one of two unfamiliar habitats: an open field or a coconut plantation. The direction flown by each butterfly towards its home was then recorded. For a typical butterfly, the angle where it disappears from sight can be shown as a point $P$ on a circle with center $O$ and radius 1; the direction taken by the butterfly is the unit vector $OP$. Because of the difficulty of accurate measurements, the angles are recorded in cells of 10 degree width.

The primary goal of the study is to determine whether the age of a butterfly affects its ability to find its homeward direction. Therefore, the butterflies were split into three age
groups, labelled young, intermediate and old. There were 58 young, 41 intermediate and 31 old. Figure 4.1 gives plots of the points for the three age groups. For the purposes of illustration, the angles for the two species are recorded together.

The statistic $U^2_k$ is used for directional data because its value does not depend on the origin for the angles. The value of $U^2_3$ is 0.082 with $p$-value = 0.909 based on the asymptotic theory in Section 4.2; a Monte Carlo study gave $p$-value = 0.912. The results indicate that age does not seriously affect the distributions of the angles in each group.

There were five distances $D = 0, 15, 30, 60, 100$ used in the experiment. The sets of directions for each distance are regarded as $K = 5$ samples. It is also of interest to see if the directions chosen by butterflies are different when they are released at different distances.

The value of $U^2_5$ equals 0.707 with $p$ value = 0.005. Thus it is clear that the five samples have different distributions. Although not a formal test procedure, the values of $U^2_k$ for each sample assuming the estimated overall distribution is correct, are respectively: 0.188, 0.050, 0.083, 0.066, 0.320 and the $p$ values are 0.019, 0.575, 0.265, 0.401, 0.001. There is a clear suggestion that the samples for $D = 0$ and 100 are different from the others. Figure 4.2 shows the five directional plots.
CHAPTER 4. K SAMPLE CRAMER VON MISES STATISTICS

Figure 4.2: Directions chosen by butterflies at different distances.
4.7 Power Comparisons: Contiguous Alternatives

Suppose samples are generated from grouped contiguous alternatives. Let \( x_0, x_1, \ldots, x_m \) be cell boundaries from small to large and suppose \( F(x) \) and \( f(x) \) are the distribution and density functions of a random variable \( x \) from which the null cell probabilities are calculated by \( p_j = F(x_j) - F(x_{j-1}), j = 1, \ldots, m \). Define \( p_{ij} \) to be close to \( p_i \) for large \( N \)

\[
 p_{ij} = p_j + N^{-1/2} h_{ij} + o(1/N), \tag{4.17}
\]

\( i = 1, \ldots, K, j = 1, \ldots, m. \)

Following Nair (1986), a location shift alternative is defined by

\[
 p_{ij} = F(x_j + \mu_i/\sqrt{N}) - F(x_{j-1} + \mu_i/\sqrt{N}),
\]

where \( \mu_i \) is a location shift parameter for sample \( i \).

It then follows

\[
 h_{ij} = \sqrt{N}(p_{ij} - p_j) = \sqrt{N}\{F(x_j + \mu_i/\sqrt{N}) - F(x_{j-1} + \mu_i/\sqrt{N}) - (F(x_j) - F(x_{j-1}))\} = \mu_i\frac{F(x_j + \mu_i/\sqrt{N}) - F(x_j) - \{F(x_{j-1} + \mu_i/\sqrt{N}) - F(x_{j-1})\}}{\mu_i/\sqrt{N}}
\]

(4.18)

and as \( N \to \infty \),

\[
 h_{ij} \approx \mu_i(f(x_j) - f(x_{j-1})). \tag{4.19}
\]

Then equation (4.17) becomes

\[
 p_{ij} \approx p_j + \mu_i(f(x_j) - f(x_{j-1}))/\sqrt{N}. \tag{4.20}
\]

Similarly, a scale alternatives is defined by

\[
 p_{ij} = F\left(\frac{x_j}{e^{\beta_i/\sqrt{N}}}\right) - F\left(\frac{x_{j-1}}{e^{\beta_i/\sqrt{N}}}\right),
\]

where \( \beta_i \) is a location shift parameter for sample \( i \). Following equation (4.17), the term \( h_{ij} \)
is

\[
\begin{align*}
    h_{ij} &= \sqrt{N}(p_{ij} - p_j) \\
    &= \sqrt{N} \left\{ \frac{x_j}{e^{\beta_i/\sqrt{N}}} - F\left( \frac{x_{j-1}}{e^{\beta_i/\sqrt{N}}} \right) - (F(x_j) - F(x_{j-1})) \right\} \\
    &= \sqrt{N} \left\{ F\left( \frac{x_j}{e^{\beta_i/\sqrt{N}}} \right) - F(x_j) - \left[ F\left( \frac{x_{j-1}}{e^{\beta_i/\sqrt{N}}} \right) - F(x_{j-1}) \right] \right\}, \tag{4.21}
\end{align*}
\]

Let \( \epsilon = \beta_i/\sqrt{N} \). As \( N \to \infty \),

\[
\begin{align*}
    F\left( \frac{x_j}{e^{\epsilon}} \right) &= F(x_j(1 - \epsilon)) + o(1), \\
    &\approx F(x_j) + f(x_j)x_j\epsilon
\end{align*}
\]

Now equation (4.21) becomes

\[
\begin{align*}
    h_{ij} &\approx \sqrt{N} \left[ F(x_j) + f(x_j)x_j\beta_i/\sqrt{N} - F(x_j) - \left\{ F(x_{j-1}) + f(x_{j-1})x_{j-1}\beta_i/\sqrt{N} - F(x_{j-1}) \right\} \right] \\
    &\approx \beta_i \{ f(x_j)x_j - f(x_{j-1})x_{j-1} \}. \tag{4.23}
\end{align*}
\]

Then the equation (4.17) can be written as

\[
    p_{ij} = p_j + \beta_i(x_jf(x_j) - x_{j-1}f(x_{j-1}))/\sqrt{N} \tag{4.24}
\]

where \( \beta_i \) is a scale shift for sample \( i \). Let \( \mathbf{h}_i = (h_{i1}, \ldots, h_{im})' \).

In order to compare power of the statistics, the distribution \( F(x) \) will be selected first, and then parameters \( \mu_i \) or \( \beta_i \) will be chosen to give the parent population for each sample.

As an example, suppose \( K = 4 \) samples are generated from a standard logistic distribution where \( F(x) = 1/(1 + \exp(-x)) \). Let \( \mu_1 = -4.472, \mu_2 = -2.236, \mu_3 = 2.236, \) and \( \mu_4 = 4.472 \). Assume first that a Cramér-von Mises statistic \( T \) is calculated from the \( K \) samples assuming the null probabilities \( p_j \) are known. Using classical theory, the distribution of \( T \) as \( N \to \infty \) is of the form

\[
    T_\infty = \sum_{i=1}^{K} \sum_{j=1}^{m-1} \lambda_j z_j^2 = \sum_{j=1}^{m-1} \lambda_j \delta_j^2,
\]
where \( \lambda_j \) are those in the null distribution and \( s_j^2 \) are independent non-central \( \chi^2 \) random variables with non-centrality parameter \( \delta_j^2 \).

Let \( \gamma_i = -\sqrt{n_i} \sqrt{N} p_i \), where \( P \) is the partial sum matrix; then \( \delta_i = w_i' \gamma_i \), where \( w_i \) is the eigenvector corresponding to \( \lambda_i \), as defined in Section 4.2. The asymptotic power is the probability of \( T_\infty \) exceeding the \( \alpha \)-level percentage point of the null distribution of \( T \).

To approximate the distribution of \( T_\infty \) under contiguous alternatives, the approximation \( a + b \chi_p^2 \) can be employed. In this case, \( \kappa_j = 2^{j-1}(j-1)! \sum_{i=1}^{m-1} (j \delta_i^2 + K^*) \lambda_i^j \). When \( p_j \) are known and statistic \( T \) is calculated, \( K^* = K \). When \( p_j \) are unknown and statistic \( \hat{T} \) is calculated, \( K^* = K - 1 \).

For calculating the power for finite samples, suppose \( p_j \) are known and the statistic used is \( T \); suppose its value for a given set of samples is \( T_0 \). To decide if \( T_0 \) is significant, the distribution of \( T \) must be found. This may be done by a Monte Carlo study. Suppose \( M \) such sample sets are drawn from the parent logistic populations, and suppose \( n_T \) is the number of significant samples; the power of \( T \) is estimated by \( n_T/M \). Similarly, the power of \( \hat{T} \) may be calculated.

Table 4.7: Comparisons of asymptotic powers of statistics under different location shift models. Test level \( \alpha = 0.10 \). Parameters are given in the text.

<table>
<thead>
<tr>
<th></th>
<th>Logistic</th>
<th>Logistic*</th>
<th>Normal</th>
<th>Normal*</th>
<th>Extreme value</th>
<th>Extreme value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_k^2 )</td>
<td>0.470</td>
<td>0.454</td>
<td>0.867</td>
<td>0.853</td>
<td>0.788</td>
<td>0.742</td>
</tr>
<tr>
<td>( A_k^2 )</td>
<td>0.467</td>
<td>0.443</td>
<td>0.869</td>
<td>0.855</td>
<td>0.797</td>
<td>0.756</td>
</tr>
<tr>
<td>( U_k^2 )</td>
<td>0.334</td>
<td>0.332</td>
<td>0.652</td>
<td>0.664</td>
<td>0.629</td>
<td>0.598</td>
</tr>
<tr>
<td>( T_C )</td>
<td>0.470</td>
<td>0.457</td>
<td>0.867</td>
<td>0.855</td>
<td>0.788</td>
<td>0.747</td>
</tr>
<tr>
<td>( T_E )</td>
<td>0.467</td>
<td>0.453</td>
<td>0.869</td>
<td>0.856</td>
<td>0.797</td>
<td>0.756</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>0.305</td>
<td>0.291</td>
<td>0.679</td>
<td>0.659</td>
<td>0.626</td>
<td>0.569</td>
</tr>
</tbody>
</table>

Table 4.7 gives powers for \( K = 4 \) samples, and for the standard normal distribution, standard extreme values \( (F(x) = 1 - e^{-e^x}) \) distribution and standard logistic distribution with location shifts \( \mu_1 = -4.472, \mu_2 = -2.236, \mu_3 = 2.236, \text{and} \mu_4 = 4.472 \). The null probabilities for 5 cells are all \( p_j = 0.2 \). The table gives asymptotic powers and also Monte Carlo results, shown with an asterisk, based on 10000 sets of samples, each of size 20. They show powers close to those given by using the asymptotic distribution, so that this distribution is adequate for determining power, thus avoiding a two-level (Monte Carlo inside
Table 4.8: Comparisons of asymptotic powers of statistics under different scale shift models. Test level $\alpha = 0.10$. Parameters are given in the text.

<table>
<thead>
<tr>
<th>Test level</th>
<th>$\alpha = 0.10$. Parameters are given in the text.</th>
<th>Logistic</th>
<th>Logistic*</th>
<th>Normal</th>
<th>Normal*</th>
<th>Extreme value</th>
<th>Extreme value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_k^2$</td>
<td></td>
<td>0.283</td>
<td>0.283</td>
<td>0.309</td>
<td>0.299</td>
<td>0.328</td>
<td>0.315</td>
</tr>
<tr>
<td>$A_k^2$</td>
<td></td>
<td>0.341</td>
<td>0.322</td>
<td>0.378</td>
<td>0.362</td>
<td>0.375</td>
<td>0.360</td>
</tr>
<tr>
<td>$U_k^2$</td>
<td></td>
<td>0.606</td>
<td>0.579</td>
<td>0.652</td>
<td>0.636</td>
<td>0.595</td>
<td>0.602</td>
</tr>
<tr>
<td>$T_C$</td>
<td></td>
<td>0.283</td>
<td>0.287</td>
<td>0.309</td>
<td>0.299</td>
<td>0.328</td>
<td>0.330</td>
</tr>
<tr>
<td>$T_E$</td>
<td></td>
<td>0.341</td>
<td>0.316</td>
<td>0.378</td>
<td>0.359</td>
<td>0.375</td>
<td>0.384</td>
</tr>
<tr>
<td>$X^2$</td>
<td></td>
<td>0.535</td>
<td>0.500</td>
<td>0.585</td>
<td>0.546</td>
<td>0.525</td>
<td>0.521</td>
</tr>
</tbody>
</table>

Table 4.8 gives asymptotic powers for $K = 4$ samples, and Monte Carlo powers for sample sizes 20, 20, 20, 20 and for the 3 different distributions. Null probabilities and test size are as for Table 4.7. The scale shifts are equal to $\beta_1 = -3.162$, $\beta_2 = -3.037$, $\beta_3 = 3.037$, $\beta_4 = 3.162$.

Table 4.9 gives the asymptotic powers of the first four components of the statistics under scale shift alternatives. The second components of all the other statistics have good power. The asymptotic powers are checked with Monte Carlo studies.

Table 4.9: Power comparisons of components of the statistics under extreme value distributions with scale shift models with $p_j = c(0.2, 0.2, 0.2, 0.2, 0.2)$ when $K = 4$. Level $\alpha = 0.10$, $n_i = 50$, $\beta_1 = 3.162$, $\beta_2 = -6.640$, $\beta_3 = 6.640$, $\beta_4 = -3.162$. The rows with an asterisk are values given by Monte Carlo studies using 10000 samples. The rows without an asterisk are values given by asymptotic theory.

<table>
<thead>
<tr>
<th>Component</th>
<th>$A_k^2$</th>
<th>$W_k^2$</th>
<th>$U_k^2$</th>
<th>$T_E$</th>
<th>$T_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$First$</td>
<td>0.263</td>
<td>0.273</td>
<td>0.917</td>
<td>0.263</td>
<td>0.273</td>
</tr>
<tr>
<td>$First^*$</td>
<td>0.266</td>
<td>0.291</td>
<td>0.713</td>
<td>0.263</td>
<td>0.293</td>
</tr>
<tr>
<td>$Second$</td>
<td>0.988</td>
<td>0.984</td>
<td>0.729</td>
<td>0.988</td>
<td>0.984</td>
</tr>
<tr>
<td>$Second^*$</td>
<td>0.982</td>
<td>0.969</td>
<td>0.657</td>
<td>0.987</td>
<td>0.979</td>
</tr>
<tr>
<td>$Third$</td>
<td>0.122</td>
<td>0.112</td>
<td>0.225</td>
<td>0.122</td>
<td>0.112</td>
</tr>
<tr>
<td>$Third^*$</td>
<td>0.150</td>
<td>0.166</td>
<td>0.198</td>
<td>0.135</td>
<td>0.143</td>
</tr>
<tr>
<td>$Fourth$</td>
<td>0.124</td>
<td>0.227</td>
<td>0.134</td>
<td>0.124</td>
<td>0.227</td>
</tr>
<tr>
<td>$Fourth^*$</td>
<td>0.139</td>
<td>0.253</td>
<td>0.168</td>
<td>0.118</td>
<td>0.206</td>
</tr>
</tbody>
</table>
4.8 Power Comparisons: Discrete Alternatives

In this Section, the powers of the statistics are compared when $H_0$ is tested for various separate distributions. For a typical power study for a statistic $\hat{T}$ the steps are as follows:

1. Generate $K$ samples with sizes $n_i, i = 1, \cdots, K$ from the alternative distributions for the $K$ samples and group each sample into $m$ cells using given boundaries;

2. Estimate the cell probabilities $\hat{p}$ from the samples and compute statistic $\hat{T}$; suppose its value is $\hat{T}_0$ this is the basis for an inner MC run to find if $\hat{T}_0$ is significant, as follows.

3. Generate $K$ samples from $\hat{p}$ in (2), recalculate $\hat{p}$ and compute $\hat{T}$;

4. Repeat (3) $M_n = 10000$ times and hence estimate the distribution of $\hat{T}$ and find the 0.90 percentile $\hat{T}_{0.90}$

5. Reject $H_0$ if the value of $\hat{T}_0$ obtained in (2) is greater than $\hat{T}_{0.90}$ obtained in (4).

6. Repeat (1) to (5) $M = 1000$ times and let $N_T$ be the number of times $\hat{T}$ rejects $H_0$; then $P_T = N_T/M$ is the estimated power of $\hat{T}$.

Power study 1 Differences in mean. For this study, $k = 4$ samples were generated from four normal distributions with variance 1 but with different means, $\mu_1 = 0.375$, $\mu_2 = 0.125$, $\mu_3 = -0.125$ and $\mu_4 = -0.375$. The cell boundaries were $(-\infty, -2, -1, 0, 1, 2, \infty)$. A wide range of sample sizes was used. The power results are plotted against $N$, the total of the 4 sample sizes, in Figure 5.1. Power increases with $N$ as expected, and all the tests gave powers greater than 0.9 for $N > 160$.

Note that in Figure 4.4 Pearson’s $X^2$ statistic outperformed all the other statistics. This is not surprising because the $X^2$ statistic gives heavier weight at the second component, which is known to be powerful against scale shift alternative, than any other Cramér-von Mises statistics.

Power study 2 Differences in scale. Samples were generated from normal distributions with mean 0 and variances: $\sigma_1^2 = 0.35$, $\sigma_2^2 = 0.7$, $\sigma_3^2 = 1.05$ and $\sigma_4^2 = 1.4$. The cell boundaries were $(-\infty, -2, -1, 0, 1, 2, \infty)$. Results are plotted in Figure 4.4.

Finally, it is worth showing how the power increase with increasing location or scale differences. These are shown in Figures 4.5 and 4.6 for fixed $n_i = 20$ for all samples. The powers increase as expected, and there is little change in the relative ranking of the different statistics.
4.9 Summary and conclusion

Multisample Cramér-von Mises statistics are defined for testing that grouped or discrete data come from the same parent distribution. Asymptotic distribution theory is given, and an easy standardization for a typical statistic enables a simple table to be used for testing, when cell probabilities are known or in the more practical case when they are unknown. The statistics are compared with closely related statistics suggested by Taguchi and Nair. Some tables are given to show powers of the statistics against various patterns of cell probabilities. These show better powers than the conventional chi-squared test.
Figure 4.3: Power study 1. Power comparison of Cramér-von Mises statistics: samples with different means. Power against total sample size.
Figure 4.4: Power study 2. Power comparison of Cramér-von Mises statistics: samples with different variances.
Figure 4.5: Power study, power comparison of Cramér-von Mises statistics: increasing mean differences.
Figure 4.6: Power study, power comparison of Cramér-von Mises statistics: increasing variance differences.
4.10 References


Chapter 5

Alternative tests based on likelihood ratio and other multisample tests

In the previous chapter, $K$ samples are to be tested to have the same distribution. Sample counts are in the rows of a $K \times m$ contingency table, called an ordered contingency table because there is a natural ordering of the cells.

Statistics called Cumulative Chi-Squared (CCS) statistics were introduced by Taguchi (1974) and studied by Takeuchi and Hirotsu (1987) and further developed by Nair (1986a, 1986b, 1987) among others. It will be shown below these are related to the Cramér-von Mises statistics discussed in Chapter 4. The statistics are calculated as follows: (1) collapse columns 1 to $j$ into one column and columns $j + 1$ to $m$ into another column, for $j = 1, \ldots, m - 1$, to give $m - 1$ tables with the dimension $K \times 2$; (2) for each table, compute the Pearson $X^2$ statistic; (3) compute a weighted sum of the individual $X^2$ statistics. CCS statistics were proposed for contingency tables where the columns have a natural ordering because, in contrast to the classic Pearson $X^2$ statistic, the ordering is taken into account by sequential collapsing of the columns. In this Chapter, the individual $X^2$ statistics in the CCS are replaced by likelihood ratio statistics to give a family of omnibus tests and the proposed tests are shown to give generally improved power.

The test statistics are defined in Section 5.1.2. Asymptotic distributions and percentage points of the new statistics are given in Section 5.2. Power studies with different types of
alternatives are in Section 7.2.1. All the above statistics are omnibus statistics. Comparisons are also made with directional statistics specially suggested for location or scale shifts between the samples. Some examples are analyzed in Section 5.4 to demonstrate the use of these statistics.

5.1 Test Statistics

The null hypothesis to be tested is:

\[ H_0 : \text{the } K \text{ samples have the same distribution.} \]

Let \( p_{ij} \) is the probability of an observation in sample \( i \) falling into cell \( j \) and let \( H_{ij} = \sum_{l=1}^{j} p_{il} \) be the cumulative probability to cell \( j \). Let \( n_i \) be the number of observations in the \( i \)th sample and \( N = \sum_{i=1}^{K} n_i \) be the total sample size. For the \( i \)th sample, suppose \( o_{ij} \) are the counts in the \( j \)th cell, for \( i = 1, \ldots, K; j = 1, \ldots, m \). The cumulative observed counts are \( S_{ij} = \sum_{l=1}^{j} o_{il} \) and the cumulative observed relative frequency is \( S_{ij}/n_i = \hat{H}_{ij} \), the estimate of \( H_{ij} \).

On \( H_0 \), all \( p_{ij} = p_j \). Then the common cumulative probability is \( H_j = \sum_{l=1}^{j} p_l \) and \( \hat{H}_j = \sum_{i=1}^{K} S_{ij}/N \) is an estimate of \( H_j \) and \( \hat{p}_j = \sum_{i=1}^{K} o_{ij}/N \) estimates \( p_j \).

5.1.1 CCS tests

The Taguchi statistic \( T_E \) is calculated as follows:

\[
T_E = \sum_{j=1}^{m-1} \sum_{i=1}^{K} n_i (\hat{H}_{ij} - \hat{H}_j)^2 / [\hat{H}_j(1 - \hat{H}_j)],
\]

and a related version \( T_C \), proposed by Nair (1986b), is

\[
T_C = \sum_{j=1}^{m-1} \sum_{i=1}^{K} n_i (\hat{H}_{ij} - \hat{H}_j)^2 / m.
\]

These statistics are CCS statistics. Suppose, for \( j = 1, \ldots, m - 1 \), columns 1 to \( j \) are collapsed into one column and similarly for columns \( j + 1 \) to \( m \), to give a \( K \times 2 \) table, called the \( j \)th collapsed table. The usual Pearson \( X^2 \) statistic computed from the \( j \)th collapsed table...
The CCS statistics are
\[ T^2 = \sum_{j=1}^{m-1} w_j X_j^2, \] (5.4)
where \( w_j \) is a weight on the \( j \)th \( \chi^2 \) statistic. When \( w_j = 1 \), \( T^2 \) is \( T_E \), and when \( w_j = \hat{H}_j(1 - \hat{H}_j)/m \), \( T^2 \) is \( T_C \). Nair (1987) pointed out that the CCS statistics are closely related to the two sample versions of Cramér-von Mises statistics. When \( p_j \) and \( H_j \) are known, the Cramér-von Mises \( W^2_K \) statistic
\[ W^2_K = \sum_{j=1}^{m-1} \sum_{i=1}^{K} n_i (\hat{H}_{ij} - H_j)^2 p_j \] (5.5)
is \( T^2 \) with \( w_j = H_j(1 - H_j)p_j \), and the Anderson Darling statistic \( A^2_K \),
\[ A^2_K = \sum_{j=1}^{m-1} \sum_{i=1}^{K} n_i (\hat{H}_{ij} - H_j)^2 p_j / [H_j(1 - H_j)] \] (5.6)
is \( T^2 \) with \( w_j = p_j \). Statistic \( A^2_K \) was first studied by Scholz and Stephens (1987). When \( p_j, H_j, \) and \( H_{ij} \) are unknown, they should be replaced by the MLEs \( \hat{p}_j, \hat{H}_j, \) and \( \hat{H}_{ij} \).

For one-sample continuous data, Zhang (2002) proposed using the likelihood ratio statistics on the collapsed samples instead of \( X^2 \) and Zhang and Wu (2007) extended this idea to \( K \) sample goodness-of-fit statistics.

### 5.1.2 Likelihood ratio tests

In this Section, the use of the likelihood ratio instead of \( X^2 \) will be discussed.

The log likelihood function of the \( j \)th \( K \times 2 \) collapsed table is
\[ l_j = \sum_{i=1}^{K} n_i \{ \hat{H}_{ij} \log(H_{ij}) + (1 - \hat{H}_{ij}) \log(1 - H_{ij}) \}, \] (5.7)
and under $H_0$, $H_{ij} = H_j$ for all samples. Let $l_{j0}$ be the log likelihood function under $H_0$:

$$l_{j0} = \sum_{i=1}^{K} n_i \{ \hat{H}_{ij} \log(H_j) + (1 - \hat{H}_{ij}) \log(1 - H_j) \}. \quad (5.8)$$

When $p_j, H_j$ and $H_{ij}$ are unknown, they should be replaced by the MLEs $\hat{p}_j, \hat{H}_j$ and $\hat{H}_{ij}$ from Section 5.1. Then the likelihood ratio computed from the $j$th collapsed table is:

$$G^2_j(\hat{H}_{ij}, \hat{H}_j) = 2 \sum_{i=1}^{K} n_i \{ \hat{H}_{ij} \log \left( \frac{\hat{H}_{ij}}{\hat{H}_j} \right) + (1 - \hat{H}_{ij}) \log \left( \frac{1 - \hat{H}_{ij}}{1 - \hat{H}_j} \right) \}, \quad (5.9)$$

and the proposed likelihood ratio statistics are

$$G^2 = \sum_{j=1}^{m-1} w_j G^2_j, \quad (5.10)$$

where $w_j$ is the weight on the $G^2_j$ statistic. With $w_j$ equal to $p_j, H_j(1 - H_j)p_j, H_j(1 - H_j)$ or $H_j(1 - H_j)/m$, respectively, new statistics analogous to $A^2_K, W^2_K, T_E$ and $T_C$ are obtained. They will be called $G^2_A, G^2_W, G^2_{TE}$, and $G^2_{TC}$.

In many situations, two statisticians might reverse the order of the categories; for example, grades may be in columns from $F$ to $A^+$, or from $A^+$ to $F$. It would be desirable to make the same inference from the data. The omnibus statistics discussed above will give the same values in both cases. In contrast, the Pearson $\chi^2$ statistic has the same value for all orderings of the columns.

### 5.2 Asymptotic Theory

In this section, it is shown that on $H_0$, $G^2_j$ is asymptotically equivalent to $X^2_j$ and therefore $G^2$ is asymptotically equivalent to $T^2$. This conclusion requires the assumption that $n_m = \min(n_i)$ tends to infinity.

It is a standard consequence of the central limit theorem that

$$\hat{H}_{ij}^{(n_i)} = H_j + \epsilon_{ij}, \text{ where } \epsilon_{ij} = O_p(n_i^{-\frac{1}{2}}). \quad (5.11)$$
So averaging over \(i\) results in

\[
\hat{H}_j^{(n_i)} = H_j + \bar{\epsilon}_j, \quad \text{where} \quad \bar{\epsilon}_j = \sum_{i=1}^{K} n_i \epsilon_{ij} / N = O_p(n_m^{-\frac{1}{2}}).
\] (5.12)

Expanding \(G^2_j(\hat{H}_{ij}, \hat{H}_j)\) in a Taylor series about \((H_j, H_j)\) to the second order gives, after some algebra,

\[
G^2_j(\hat{H}_{ij}, \hat{H}_j) = K \sum_{i=1}^{K} n_i \frac{(\hat{H}_{ij} - \hat{H}_j)^2}{H_j(1 - H_j)} \{1 + o_p(1)\} + o_p(\sum_{i=1}^{K} ||\hat{H}_{ij} - H_j||^2 + \sum_{i=1}^{K} ||\hat{H}_j - H_j||^2) \tag{5.13}
\]

Following equations (5.11) and (5.12) above

\[
\frac{\hat{H}_j^{(n_i)}}{\hat{H}_j} = \frac{H_j + o_p(1)}{\hat{H}_j} = 1 + o_p(1), \tag{5.14}
\]

and

\[
\frac{1 - \hat{H}_j^{(n_i)}}{1 - \hat{H}_j} = \frac{1 - H_j + o_p(1)}{1 - H_j} = 1 + o_p(1). \tag{5.15}
\]

Substituting \(\hat{H}_{ij}^{(n_i)}\) for \(\hat{H}_{ij}\) and \(\hat{H}_j^{(n_i)}\) for \(\hat{H}_j\) in (5.13)

\[
G^2_j(\hat{H}_{ij}^{(n_i)}, \hat{H}_j^{(n_i)}) = \sum_{i=1}^{K} n_i \frac{(\hat{H}_{ij}^{(n_i)} - \hat{H}_j^{(n_i)})^2}{\hat{H}_j^{(n_i)}(1 - \hat{H}_j^{(n_i)})} \{1 + o_p(1)\} \\
+ o_p(\sum_{i=1}^{K} ||\hat{H}_{ij} - H_j||^2 + \sum_{i=1}^{K} ||\hat{H}_j - H_j||^2)
\]

\[
= X_j^2 + o_p(1).
\]

Therefore, \(G^2_j\) is asymptotically equivalent to \(X_j^2\) and for the same weights \(w_j\), the corresponding likelihood ratio statistic \(G^2\) is asymptotically equivalent to \(T^2\).

Nair (1987) has given the asymptotic theory for \(T_E\) and \(T_C\) and in Chapter 4, the theory was given for \(W_K^2\) and \(A_K^2\). Therefore, the asymptotic significance points of \(W_K^2\) and \(G_W^2\) can be found.
Table 5.1: Comparison of finite \( n \) significance points with asymptotic significance points for \( W^2 \) and \( G_{W}^2 \), when \( K = 4 \) and \( p_{j} = 0.2 \) for \( j = 1, \ldots, 5 \) and parameters are assumed unknown. M.C.=10,000

<table>
<thead>
<tr>
<th>( n_i )</th>
<th>( \text{STAT/} \alpha )</th>
<th>0.25</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>( W_{4}^2 )</td>
<td>0.621</td>
<td>0.8443</td>
<td>1.000</td>
<td>1.341</td>
</tr>
<tr>
<td></td>
<td>( G_{W}^2 )</td>
<td>0.677</td>
<td>0.932</td>
<td>1.108</td>
<td>1.549</td>
</tr>
<tr>
<td>20</td>
<td>( W_{4}^2 )</td>
<td>0.616</td>
<td>0.851</td>
<td>1.026</td>
<td>1.377</td>
</tr>
<tr>
<td></td>
<td>( G_{W}^2 )</td>
<td>0.641</td>
<td>0.890</td>
<td>1.076</td>
<td>1.481</td>
</tr>
<tr>
<td>30</td>
<td>( W_{4}^2 )</td>
<td>0.608</td>
<td>0.834</td>
<td>1.001</td>
<td>1.360</td>
</tr>
<tr>
<td></td>
<td>( G_{W}^2 )</td>
<td>0.622</td>
<td>0.858</td>
<td>1.025</td>
<td>1.407</td>
</tr>
<tr>
<td>50</td>
<td>( W_{4}^2 )</td>
<td>0.617</td>
<td>0.841</td>
<td>0.996</td>
<td>1.310</td>
</tr>
<tr>
<td></td>
<td>( G_{W}^2 )</td>
<td>0.624</td>
<td>0.853</td>
<td>1.011</td>
<td>1.342</td>
</tr>
<tr>
<td>100</td>
<td>( W_{4}^2 )</td>
<td>0.603</td>
<td>0.827</td>
<td>0.981</td>
<td>1.348</td>
</tr>
<tr>
<td></td>
<td>( G_{W}^2 )</td>
<td>0.608</td>
<td>0.831</td>
<td>0.988</td>
<td>1.362</td>
</tr>
<tr>
<td>( \infty )</td>
<td></td>
<td>0.611</td>
<td>0.841</td>
<td>1.005</td>
<td>1.369</td>
</tr>
</tbody>
</table>

In Table 5.1, Monte Carlo points for various \( n \), and asymptotic points are given for \( W^2_K \) and \( G_{W}^2 \). The points are for 5 cells with equal known probabilities. It may be seen that for \( n_i > 10 \), the asymptotic points for \( W^2_K \) maybe used with good accuracy. Similarly, the asymptotic points maybe used for \( G_{W}^2 \) for \( n_i > 30 \).

Similar results were observed for \( A_{K}^2 \) and \( G_{A}^2 \), for \( T_E \), and \( G_{TE}^2 \), and for \( T_C \), and \( G_{TC}^2 \). In general, the CCS statistics converge to the limit faster than the likelihood ratio statistics.

### 5.3 Power comparisons

In this section, the power of different tests is examined when testing \( H_0 \) against several fixed alternatives at 5% level.

As stated above, the statistics so far discussed are omnibus. There are some tests that are powerful against certain alternatives only. Nair (1986a) proposed to use the first components of \( T_E \) and \( T_C \), (defined in Chapter 4) which he called \( U_{E,1}^2 \) and \( U_{C,1}^2 \), as tests for location difference. For scale alternatives, Nair proposed to use the second components, called \( U_{E,2}^2 \) and \( U_{C,2}^2 \). He showed that when \( p_1 = \cdots = p_m \), \( U_{E,1}^2 \) is the discrete version of the Kruskal-Wallis test (1952), a multi-sample version of the Wilcoxon test and \( U_{E,2}^2 \) is the discrete version of Mood’s test (1954). Box and Jones (1990) proposed a method of scoring
the categories. The steps are as follows:

1. equally spaced scores are assigned to the \( m \) columns. For example, the score in column \( j \) is \( s_j = j \).

2. the frequency in cell \( j \) for the \( i \)th sample is \( o_{ij} \).

3. create an "extended" \( i \)th sample consisting of values \( s_j \) with multiplicity \( o_{ij} \). For example, with four columns, if the \( i \)th sample has cell frequencies 1, 2, 3, 2, the extended sample is 1, 2, 2, 3, 3, 4, 4;

4. Test \( H_0 \) by the \( F \) statistic from the extended samples using standard analysis of variance.

The power of all the statistics has been examined in several studies. In the first study, samples were generated from four normal distributions with variance 1 but with different means, \( \mu_1 = 0.125, \mu_2 = 0.375, \mu_3 = -0.125 \) and \( \mu_4 = -0.375 \). The observations were grouped into cells whose boundaries were \( -\infty, -1.5, -0.5, 0.5, 1.5, \infty \). A very large range of sample sizes was used; the power plots are against only the overall sample size. In the second study, samples were generated from normal distributions with mean 0 and different variances: \( \sigma_1^2 = 0.35, \sigma_2^2 = 0.7, \sigma_3^2 = 1.05 \) and \( \sigma_4^2 = 1.4 \). The cell boundaries were \( -\infty, -2, -1, 0, 1, 2, \infty \). In the third study, samples were generated from Pearson curves with common mean=0, variance =1, kurtosis = 2, different skewness = \( -0.5, -0.25, 0.25 \) and 0.5. The cell boundaries were \( -\infty, -1.5, -0.5, 0.5, 1.5, \infty \).

In the fourth study, samples were generated from normal distributions with different means: \( \mu_1 = -0.375, \mu_2 = -0.125, \mu_3 = 0.125 \) and \( \mu_4 = 0.375 \), and different variances: \( \sigma_1^2 = 0.35, \sigma_2^2 = 0.7, \sigma_3^2 = 1.05 \) and \( \sigma_4^2 = 1.4 \). The cells boundaries were the same as in study 1.

The steps of the power studies are as follows:

1. Generate 4 samples with sizes \( n_i, i = 1, \cdots, 4 \) from the alternative distributions given and group each sample into \( m \) cells;

2. Calculate the estimate \( \hat{p} \) from the samples and compute the four CCS type statistics, the 4 likelihood ratio type statistics, and the other statistics described in Section 7.2.1

3. Generate 4 samples from \( \hat{p} \) in (2), recalculate \( \hat{p} \) and compute the statistics;

4. Repeat (3) 10,000 times and find the 95 percentiles for each statistics.

5. Reject \( H_0 \) if the values of the statistics obtained in (2) is greater than the 95 percentiles obtained in (4).

6. Repeat (1) to (5) 1000 times, and the number of times a statistic rejects \( H_0 \) divided by 1000 is the estimated power of this statistic.
Table 5.2: power study 2: power improvements of likelihood ratio type statistics on CCS type statistics.

<table>
<thead>
<tr>
<th>N</th>
<th>40</th>
<th>80</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>180</th>
<th>200</th>
<th>250</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta(G_{W_2}^{TE} - \beta(W_{K}^2))$</td>
<td>-0.01</td>
<td>0.04</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
<td>0.11</td>
<td>0.08</td>
<td>0.08</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>$\beta(G_{A_2}^{TE} - \beta(A_{K_2}^2))$</td>
<td>0.01</td>
<td>0.07</td>
<td>0.15</td>
<td>0.09</td>
<td>0.13</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\beta(G_{T_E}^{TE} - \beta(T_E))$</td>
<td>0.02</td>
<td>0.09</td>
<td>0.19</td>
<td>0.12</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\beta(G_{T_C}^{TE} - \beta(T_C))$</td>
<td>0.01</td>
<td>0.08</td>
<td>0.14</td>
<td>0.13</td>
<td>0.18</td>
<td>0.07</td>
<td>0.09</td>
<td>0.05</td>
<td>0.04</td>
<td>0.02</td>
</tr>
</tbody>
</table>

In Figure 5.1, all the statistics gave similar powers except Mood’s test, which is designed for different scales.

In Figure 5.2, $G_{TE}$ and Mood’s test have the highest powers, particularly when $N$ is small. For the weights given in Section 5.1.1, likelihood ratio tests are more powerful than the CCS statistics. Using the likelihood ratio type statistics as omnibus tests appears to improve the power over standard CCS tests when the samples have scale differences; the power differences are less obvious for location shifts.

The proposed omnibus tests, except for $G_{TE}^2$, have less power than Mood’s statistic, Nair’s statistics $U_{E,2}^2$ and $U_{C,2}^2$, which are not shown in the figures. However, note that Mood’s statistic, and $U_{E,2}^2$ and $U_{C,2}^2$ are tests specifically for scale difference. Box and Jones’s $F$ test and $KW$ are tests for location differences and fail to detect scale differences.

In Figure 5.3, with samples generated from distributions with same mean and variances, but with different skewness, $G_{TE}$ and $T_E$ have the highest power. All tests for location differences or scale differences fail to detect skewness differences.

In Figure 5.4, with samples generated from distributions with different location and different scale, the omnibus statistics perform better than Nair’s statistics.

Tables 5.2 and 5.3 show the power differences between the likelihood ratio type statistics and the CCS type statistics for different $N$ and for scale different and skewness different alternatives. The likelihood ratio tests generally have more power than the CCS type tests. When testing against the scale shift alternative, the powers of LRT tests are around 10% to 20% higher than those of the CCS tests for $100 < N < 120$.

Finally, it is worth showing how the power increases with increasing location or scale differences. These are shown in Figures 5.5 and 5.6 for fixed $n_i = 20$ for all $i$. The powers
Table 5.3: Power study 3: power improvements of likelihood ratio type statistics on CCS type statistics.

<table>
<thead>
<tr>
<th>N</th>
<th>40</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>180</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
</tr>
</thead>
<tbody>
<tr>
<td>β((G_{W}^{2}) - (W_{K}^{2}))</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.07</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>β((G_{A}^{2}) - (A_{K}^{2}))</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.06</td>
<td>0.07</td>
<td>0.03</td>
<td>0.10</td>
<td>0.11</td>
<td>0.25</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>β((G_{T}^{2}) - (T_{E}))</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.17</td>
<td>0.11</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>β((G_{T}^{2}) - (T_{C}))</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.08</td>
<td>0.01</td>
<td>0.09</td>
<td>0.14</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 5.4: Letter grades of the last 3 offerings of a STAT course at SFU.

<table>
<thead>
<tr>
<th></th>
<th>A+</th>
<th>A-</th>
<th>B+</th>
<th>B-</th>
<th>C+</th>
<th>C-</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offering 1</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Offering 2</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Offering 3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

increase as expected. Also the figures show that there is little change in the relative ranking of the different statistics.

More tables and figures related to the power studies are in a research report available from the author.

5.4 Examples

The letter grades of three recent offerings of a Statistics course are in Table 5.4. The course was taught by different instructors each time. Administrators are interested to see if the grades were given in a consistent way.

The \(p\) values associated with the different statistics are in table 5.5. At the 5% level, none of the statistics reject \(H_0\) that the grades follow the same distribution.

5.5 Summary and conclusions

In this Chapter, many omnibus tests for testing the hypothesis that \(K\) samples come from the same parent distributions have been investigated when the samples are recorded as counts in cells. The emphasis has been on comparing statistics where the cells have a
natural ordering. Standard multisample statistics based on Pearson’s Chi-square test can often be improved using likelihood ratio. Asymptotic percentage points of the new statistics can be used with good accuracy and programs are available from the author. Many power studies demonstrate the performance of the different statistics.

Table 5.5: $p$ values of different tests on the grades data.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>$p$ values (asym)</th>
<th>$p$ values (m.c.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^2_{W^2}$</td>
<td>0.058</td>
<td>0.057</td>
</tr>
<tr>
<td>$G^2_{A^2}$</td>
<td>0.069</td>
<td>0.070</td>
</tr>
<tr>
<td>$G^2_{TE}$</td>
<td>0.068</td>
<td>0.070</td>
</tr>
</tbody>
</table>
Figure 5.1: Power study 1: power comparison of likelihood ratio type statistics, CCS type statistics and other well known statistics for testing samples with different means.
Figure 5.2: Power study 2: power comparison of likelihood ratio type statistics, CCS type statistics and other well known statistics for testing samples with different variances.
Figure 5.3: Power study 3: power comparison of likelihood ratio type statistics, CCS type statistics and other well known statistics for testing samples with different skewness.
Figure 5.4: Power study 4: power comparison of LR type with Nair’s statistics for testing samples with different location and scale.
Figure 5.5: Power study: power comparison of LR statistics with Nair’s statistics for testing samples with increasing mean differences.
Figure 5.6: Power study: power comparison of LR statistics with Nair’s statistics for testing samples with increasing variance differences.
CHAPTER 5. LIKELIHOOD RATIO TYPE TESTS

5.6 References


Chapter 6

Testing fit for a GEE regression model with correlated binary responses

In this chapter, goodness of fit will be discussed for a regression model with correlated binomial responses. The model parameters will be estimated using Generalized Estimating Equations (GEE). This method was introduced by Liang and Zeger (1986) to deal with correlated measures, which often arise in longitudinal studies. A review of the development of the GEE approach is given by Ziegler, Kastner, and Blettner (1998).

In a longitudinal study, subjects are followed and measured multiple times. The repeated measures therefore have a natural ordering. Within a subject, the correlation between two consecutive measures are usually higher than those between measures far apart. Hence it is reasonable to suppose that if the fit is poor at time $t$, then it is likely to be poor at the next time $t + 1$. As described in Chapter 4, the Cramér-von Mises statistics compare the cumulative difference between observed and expected (fitted) counts for each subject and therefore take this “hangover” effect into account. In this Chapter, the Cramér-von Mises and the Pearson’s $X^2$ statistics will be adapted to test fit for Binomial longitudinal observations. In addition, following the idea in Chapter 5, the likelihood ratio type statistics and the Cumulative Chi-squared type statistics are extended to test fit for this type of data. The levels and powers of these statistics will be studied. This chapter is organized as follows: a brief introduction of GEE is given in Section 6.1; the Cramér-von Mises statistics are
introduced in Section 6.2. In Section 6.2.1, asymptotic theory is given to approximate the finite-n distributions of the Cramér-von Mises statistics. Sections 6.3 and 6.4 introduce the Cumulative Chi-squared type statistics and the likelihood ratio type statistics respectively with the aim of providing better power. Section 7.3 gives a hypothetical example to illustrate the statistics. Finally, simulation studies will be performed to evaluate the levels and powers of these statistics in Section 7.2.1.

In order to introduce GEE, consider the following example. Suppose there are $K$ study units or clusters: in each cluster, there are $m_i$ measurement. Suppose the response variable $Y_{i,t}$ at cluster $i$ and measurement $t_i$ are binomial distributed with parameter $n_{i,t}$, for $i = 1, \ldots, K$ and $t_i = 1, \ldots, m_i$. Further, suppose $Y_{i,t}$ are correlated within the same cluster. Let $Y_i = (Y_{i,1}, \ldots, Y_{i,m_i})'$ denote the vector of random variables for subject $i$.

The problem is to model $\pi_{it}$, the probability of success at cluster $i$ and measurement $t$ by regressing on the covariates in vector $x_i$. The model is

$$\logit(\pi_{it}) = x_i' \beta.$$  

(6.1)

Assuming the measurements over time are independent, the steps are as follows (GLM):

1. $Y_{i,t}$ are Binomial $(n_{i,t}, \pi_{it})$
2. construct the log-likelihood $l(\pi_{it}, Y_{i,t})$
3. replace $\pi_{it}$ by solving equation 6.1
4. maximize $l$ with respect to $\beta$ using the score function and hence get the model estimate of $\pi_{it}$.

### 6.1 Generalized estimating equations

In the problem considered, the repeated measurements are correlated. Therefore, the GEE procedures can be used.

**Notation.** It will useful to define notation for the entries of a matrix which has an index, for example, a matrix, say $A_i$ associated with the $i$th cluster. The row $r$, column $c$ entry of $A_i$ will be labelled $a_{irc}$. Similarly, the components of a vector of length $m_i$, $Y_i$ will be $Y_{it}, t = 1, \ldots, m_i$. For cluster $i$ at time $t$, there is a vector $x_{it}, i = 1, \ldots, K$, and $t = 1, \ldots, m_i$,
with components \( x_{itr}, r = 1, \ldots, p \); these can include both cluster-level covariates and time-dependent covariates. In general, the expectation \( E(Y_{it} | x_{it}) = \mu_{it} = n_{i} \pi_{it} \) is related to \( x_{it} \) through a link function \( g(\cdot) \)

\[
g(\pi_{it}) = x'_{it} \beta.
\]

and the variance-mean relationship is specified by a function

\[
V(Y_{it}) = \phi \mu_{it} (n_{i} - \mu_{it}) / n_{i},
\]

where \( \phi \) is parameter to deal with over- or under-dispersion. In the case when \( Y_{it} \) is a multivariate binomial random variable, the natural link function is the logit: \( \text{logit}(\pi_{it}) = \log(\pi_{it} / (1 - \pi_{it})) \).

Suppose \( D_{i} \) is a matrix of partial derivatives so that \( d_{irc} = \partial \mu_{it} / \partial \beta_{c} \), for \( r = 1, \ldots, m_{i} \) and \( c = 1, \ldots, p \), where \( p \) is the number of regression coefficients. Suppose further that \( R_{i} \) is a specified correlation matrix of \( Y_{i} \), called the working correlation matrix. Popular choices of \( R_{i} \) are (for the propose of illustration, let \( m_{i} = 3 \))

- the independence matrix

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

- the exchangeable matrix

\[
\begin{pmatrix}
1 & \alpha & \alpha \\
\alpha & 1 & \alpha \\
\alpha & \alpha & 1
\end{pmatrix}
\]

- the unspecified matrix

\[
\begin{pmatrix}
1 & \alpha_{1} & \alpha_{2} \\
\alpha_{1} & 1 & \alpha_{3} \\
\alpha_{2} & \alpha_{3} & 1
\end{pmatrix}
\]

- and the AR(1) matrix:

\[
\begin{pmatrix}
1 & \alpha & \alpha^{2} \\
\alpha & 1 & \alpha \\
\alpha^{2} & \alpha & 1
\end{pmatrix}
\]
where the $\alpha$ and $\alpha_i$ are specified parameters. Also let $A_i$ be a $m_i \times m_i$ diagonal matrix with entries $a_{it} = \{n_i \pi_{it}(1 - \pi_{it})\}$.

Then $V_i$ is the working variance-covariance matrix of $Y_i$:

$$V_i = \phi A_i^{1/2}R_i A_i^{1/2}.$$

Finally, let $n = \sum_{i=1}^{K} n_i$ be the total number of subjects in the data. It is assumed $\nu_i = \lim_{n \to \infty} \frac{n_i}{n}$ lies strictly between 0 and 1.

### 6.1.1 Estimation of parameters

Let $y_i = [y_{i1}, \ldots, y_{im_i}]$ be the observed values of $Y_i$. Estimates $\hat{\beta}$ of the regression coefficients $\beta$ can be found by solving the following set of estimating equations:

$$U = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) = 0. \quad (6.2)$$

Using the estimates $\hat{\beta}$, the parameters $\phi$ and $\alpha$ can be estimated using moment methods (Liang and Zeger, 1986). Then the working matrices $V_i$ are recalculated and equations (6.2) are solved again. This results in an iterative scheme that estimates $\beta$ while fixing $(\phi, \alpha)$ and then estimates $(\phi, \alpha)$ while fixing $\beta$. The procedure produces consistent estimates even when the correlation structure is mis-specified. Prentice (1988) suggested estimating $\alpha$ through another set of estimating equations to improve efficiency. Liang and Zeger show that $K^{1/2}(\hat{\beta} - \beta)$ is asymptotically normal with mean 0 and covariance matrix $\text{COV}(\hat{\beta}) = \lim_{K \to \infty} J^{-1}C J^{-1}$ where $J = \sum_{i=1}^{K} D_i' V_i^{-1} D_i$, and $C = \sum_{i=1}^{K} D_i' V_i^{-1} \text{COV}(Y_i)V_i^{-1} D_i$. Replacing $D_i$ and $V_i$ by their consistent estimators $\hat{D}_i$ and $\hat{V}_i$, which are functions of the final $\hat{\beta}$, and $(\hat{\phi}, \hat{\alpha})$, and estimating $\text{COV}(Y_i)$ by $(y_i - \mu_i)(y_i - \mu_i)'$ yields a so-called sandwich estimator for $\text{COV}(\beta)$. Again, this estimator is robust against mis-specification of the working correlation matrix $R_i$. Hence the asymptotic normality of $K^{1/2}(\hat{\beta} - \beta)$ can be used to make inferences about $\beta$. 

---

CHAPTER 6. GEE: BINOMIAL REGRESSION

77
6.2 Cramér-von Mises Goodness of fit tests

For longitudinal data, measures from different clusters (\(Y_i\) and \(Y_j\) for \(i \neq j\)) are independent, but the components in \(Y_i\) are correlated. After fitting a regression model under GEE, the \(K\) sample Cramér-von Mises statistics, studied in Chapter 4, can be extended to test fit of such data. Each cluster is considered as one of the \(K\) samples. The discrepancies between the cumulative observed and cumulative expected values within a cluster are first calculated; these discrepancies are combined for all the clusters using weights, to give the Cramér-von Mises statistics. The null hypothesis is

\[
H_0 : \mu = \mu(\beta).
\]

The alternative hypothesis is a union of many possible model failures:

- the link function is specified incorrectly;
- the variance function is specified incorrectly;
- there is interaction and/or quadratic departure in the linear model;
- the distributional assumption of the response \(Y\) is wrong.

The Cramér-von Mises statistics are defined as follows: for the \(i\)th cluster, define the cumulative observes \(S_{it} = \sum_{t=1}^{\tau} Y_{i,t}\) and the cumulative expecteds \(T_{it} = \sum_{t=1}^{t} \mu_{it}\), for \(t = 1, \ldots, m_i\). Let \(N_i = \sum_{t=1}^{m_i} Y_{i,t}\) be the total of the observed values and \(M_i = \sum_{t=1}^{m_i} \mu_{it} = T_{im_i}\) be the total of the expected values in the \(i\)th cluster. Note that when \(\mu_{it}\) is unknown and is estimated from GEE, \(N_i\) is usually not equal to \(M_i\). Define \(H_{it} = T_{it}/M_i\) and the difference

\[
Z_{it} = S_{it} - T_{it};
\]

then the Cramér-von Mises statistics are

\[
T_0 = n^{-1} \sum_{i=1}^{K} \sum_{t=1}^{m_i} Z_{it}^2 w_{it},
\]

where \(w_{it}\) are the weights given to the \(Z_{it}\). For example, when

\[
w_{it} = \mu_{it}/M_i,
\]
the statistic is the usual Cramér-von Mises \(W^2_K\) statistic for discrete data given in Chapter 4
and when

\[ w_{it} = \mu_{it}/\{M_i H_{it}(1 - H_{it})\}, \]  

(6.6)

the statistic is similar to the \(A^2\) statistic. Since \(M_i = T_{im_i}\), the weight \(w_{im_i}\) in the \(A^2_K\) statistic is set to 0. Note that when \(m_i = 1\), the problem reduces to an ordinary binomial regression with independent responses and \(W^2_K\) is a statistic suggested by Hosmer et al (1997), which is based on an unweighted sum of residual squares. In practice, \(\pi_{it}\) and therefore \(\mu_{it}\) are usually unknown. They are estimated from the regression model by solving

\[ \text{logit}(\hat{\pi}_{it}) = x'_{it} \hat{\beta} \]  

to get \(\hat{\pi}_{it} = 1/(1 + \exp (-x'_{it} \hat{\beta}))\).

The statistic is then called \(T_2\).

### 6.2.1 Asymptotic theory: Cramér-von Mises

For the distribution theory of the tests, the statistic will be put into matrix form. For cluster \(i\), define vectors \(\mu_i' = (\mu_{i1}, \ldots, \mu_{im_i})\) and \(Z_i' = (Z_{i1}, \ldots, Z_{im_i})\).

Define \(Z\) to be \(Z = (Z_1', \ldots, Z_K')\); thus \(Z\) is a long vector obtained by stacking the individual \(Z_{it}\). Also let \(Y\) and \(\mu\) be similar stacked vectors \(Y = (Y_1', \ldots, Y_K')'\) and \(\mu = (\mu_1', \ldots, \mu_K')'\).

Define a block diagonal matrix \(P = \text{diag}(P_1, \ldots, P_K)\). For cluster \(i\), define \(Q_i\) to be a \(m_i \times m_i\) diagonal matrix with entries \(q_{itt} = w_{it}, t = 1, \ldots, m_i; w_{it}\) are weights defined in equation (6.5) for \(W^2_K\) and in equation (6.6) for \(A^2_K\). Let \(Q\) be a block diagonal matrix with entries \(Q_i\) for \(i = 1, \ldots, K\). Further, define \(x_i = (x_{i1}, \ldots, x_{im_i})'\) be the covariates for cluster \(i\) and \(X = (x_1', \ldots, x_K')'\). Let \(M = \sum_{i=1}^{K} m_i\) and define \(A\) and \(V\) to be \(M \times M\) block diagonal matrices \(A = \text{diag}(A_1, \ldots, A_K)\) and \(V = \text{diag}(V_1, \ldots, V_K)\). The matrix \(V\) is called the working covariance matrix of \(Y\). It follows the estimating function \(U\) in (6.2) can be written as \(U = X'A^{-1}(Y - \mu)\).

### 6.2.2 Parameters known

When parameters are known, the Cramér-von Mises statistics are

\[ T_0 = n^{-1}(Y - \mu)'P'QP(Y - \mu) \]
\[ = n^{-1}Z'QZ, \]  

(6.7)

The following theorem gives the asymptotic distribution of \(T_0\).
Theorem 5 Define a block diagonal matrix $V^* = \text{diag}\{\nu_1 P_1 V_1 P_1', \ldots, \nu_K P_K V_K P_K'\}$. Assume that the model $\mu = \mu(\beta) = (\mu_1',\ldots,\mu_K')'$ holds with all parameters known and that $\nu_i = \lim_{n\to\infty} \frac{n_i}{n}$ and $0 < \nu_i < 1$. As $n \to \infty$, the statistic $T_0 \xrightarrow{d} \sum_{k=1}^{M} \lambda_k s_k^2$, where $M = \sum_{i=1}^{K} m_i$, the $\lambda_k$ are the non-zero eigenvalues of $QV^*$, and the $s_k$ are independent standard normal random variables.

Proof:

Since all parameters are known, the matrix $V$ is the covariance matrix of $Y$. For each $Y_i$ in $Y$, the expectation $E(Y_i) = 0$ and $\text{COV}(Y_i) = n_i V_i$. Since $Y_i$ are sums of multivariate Bernoulli random variables, $Y_i \xrightarrow{d} \text{MVN}(0, n_i V_i)$ as $n_i \to \infty$ by the Central Limit Theorem. Therefore $Z_i \xrightarrow{d} \text{MVN}(0, n_i P_i' V_i P_i')$. For any two different clusters, $Z_i$ and $Z_j$ are independent. It follows that $n^{-1/2} Z \xrightarrow{d} \text{MVN}(0, V^*)$, where $V^*$ is a block diagonal matrix $V^* = \text{diag}\{\nu_1 P_1 V_1 P_1', \ldots, \nu_K P_K V_K P_K'\}$.

Recall $T_0 = n^{-1} Z' Q Z$; then following Theorem 3, $T_0 \xrightarrow{d} \sum_{k=1}^{M} \lambda_k s_k^2$, where the $\lambda_k$ are the eigenvalues of $QV^*$ or equivalently $V^{*1/2} Q V^{*1/2}$, and the $s_k$ are independent standard normal random variables.

6.2.3 Unknown parameters

To develop asymptotic theory, it is convenient to define statistics $T_1$ by replacing the unknown parameters in $Z$ but not in $Q$: $T_1 = n^{-1} Z' Q Z$. 

Theorem 6 Assume that the model $\mu = \mu(\beta)$ holds with parameters estimated from the GEE model and that $\nu_i = \lim_{n\to\infty} \frac{n_i}{n}$ and $0 < \nu_i < 1$. Suppose $M = \sum_{i=1}^{K} m_i$ and $I$ is a $M \times M$ identity matrix. Define a block diagonal matrix $\Sigma^* = \text{diag}\{\nu_1 P_1' \Sigma_1 P_1', \ldots, \nu_K P_K' \Sigma_K P_K'\}$ and the matrices $H = -A X \frac{\partial U(\beta)}{\partial \beta} X' A V^{-1}$ and $Q^* = (I - H) Q (I - H)$. Then as $n \to \infty$, the statistic $T_1 \xrightarrow{d} \sum_{k=1}^{M} \lambda_k s_k^2$, where the $s_k$ are independent standard normal random variables and the $\lambda_k$ are the non-zero eigenvalues of $\Sigma^{*1/2} Q^* \Sigma^{*1/2}$. When $R_i$ is the identity matrix, the term $\frac{\partial U(\beta)}{\partial \beta}$ simplifies to $-X' A V^{-1} A X$, but for a general $R_i$, the term $\frac{\partial U(\beta)}{\partial \beta}$ is given in equation (6.14).
CHAPTER 6. GEE: BINOMIAL REGRESSION

Proof:

The stacked vector $\mu$ has length $M$; suppose its components are $(\mu_1, \ldots, \mu_M)$. Recall that $\beta = (\beta_1, \ldots, \beta_p)'$. Then define the $M \times p$ matrix $\frac{\partial \mu}{\partial \beta}$ to have the row $r$, column $c$ entry $\frac{\partial \mu_r}{\partial \beta_c}$. Similar notation is used for other matrices of partial derivatives.

The first order Taylor expansion of the fitted value $\hat{\mu}$ gives

$$\hat{\mu} \approx \mu + \frac{\partial \mu}{\partial \beta} (\hat{\beta} - \beta). \quad (6.9)$$

Define $\eta_{it} = x_{it}' \beta$ and $\eta'_i = (\eta_{i1}, \ldots, \eta_{im_i})$ and a stacked vector $\eta = (\eta'_1, \ldots, \eta'_K)'$. It follows that $\frac{\partial \mu}{\partial \beta} = \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta} = \frac{\partial \mu}{\partial \eta} X$. Since the “logit” link function is used, $\pi_{it} = \{1 + \exp(-\eta_{it})\}^{-1}$ and

$$\frac{\partial \mu_{it}}{\partial \eta_{it}} = \frac{n_i \pi_{it}}{n_i \pi_{it}} = n_i \pi_{it} e^{-\eta_{it}} / (1 + e^{-\eta_{it}})^2 = n_i \pi_{it} (1 - \pi_{it}). \quad (6.10)$$

Thus

$$\frac{\partial \mu_1}{\partial \eta_1} = \begin{pmatrix}
    n_1 \pi_{i1} (1 - \pi_{i1}) & 0 & \ldots & 0 \\
    0 & n_1 \pi_{i2} (1 - \pi_{i2}) & \ldots & 0 \\
    \vdots & \vdots & \vdots & \vdots \\
    0 & 0 & \ldots & n_i \pi_{im_i} (1 - \pi_{im_i})
\end{pmatrix},$$

which is the matrix $A_i$. It then follows $\frac{\partial \mu}{\partial \eta} = \text{diag}\{A_1, \ldots, A_K\} = A$ and $\frac{\partial \mu}{\partial \beta} = A \beta$. Therefore, from equation (6.9)

$$\hat{\mu} - \mu \approx AX(\hat{\beta} - \beta). \quad (6.11)$$

Another first order Taylor expansion of the Quasi-score function $U(\beta)$ at $\beta$ gives

$$\begin{align*}
U(\hat{\beta}) & \approx U(\beta) + \frac{\partial U(\beta)}{\partial \beta} (\hat{\beta} - \beta) \\
& = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i) + \frac{\partial U(\beta)}{\partial \beta} (\hat{\beta} - \beta) \\
& = X' A V^{-1} (Y - \mu) + \frac{\partial U(\beta)}{\partial \beta} (\hat{\beta} - \beta). \quad (6.12)
\end{align*}$$
For the independence working correlation matrix,

\[
\frac{\partial U(\beta)}{\partial \beta} = \sum_{i=1}^{K} x_i' A_i V_i^{-1} A_i x_i = -X' AV^{-1} AX. \tag{6.13}
\]

For any general working correlation matrix \( R_i \), the term \( \frac{\partial U(\beta)}{\partial \beta} \) is more complicated and is given by Pan (2001) for correlated binary data. With slight modification for the correlated binomial data, it can be shown

\[
\frac{\partial U(\beta)}{\partial \beta} = \sum_{i=1}^{K} \left\{ \frac{m_i}{2} \left( \sum_{t=1}^{m_i} (1 - 2\pi_{it}) x_{it} x_{it}' - m_i \right) - \sum_{t=1}^{m_i} \frac{1}{2} (Y_{it} - \mu_{it}) v_{it}^{-1/2} (1 - 2\pi_{it}) (x_{iA_i^{-1/2} R_i^{-1}})' x_{it}' - x_{iA_i^{-1} x_i} \right\}, \tag{6.14}
\]

where \( b_{it} \) is the \( t \)th element of \( (Y_i - \mu_i)' A_i^{-1/2} R_i^{-1} \), \( v_{it} = n_i \pi_{it} (1 - \pi_{it}) \), and \( (x_{iA_i^{-1/2} R_i^{-1}})' \) is the \( t \)th column of \( x_{iA_i^{-1} R_i(A)^{-1}} \).

Following (6.12), \( U(\hat{\beta}) = 0 \), and \( \hat{\beta} - \beta \approx -\left( \frac{\partial U(\beta)}{\partial \beta} \right)^{-1} X' AV^{-1} (Y - \mu) \); an approximation to the residuals of the GEE regression model can be written as

\[
Y - \hat{\mu} = Y - \mu - (\hat{\mu} - \mu) \\
\approx Y - \mu + AX \left( \frac{\partial U(\beta)}{\partial \beta} \right)^{-1} X' AV^{-1} (Y - \mu) \\
= \{I + AX \left( \frac{\partial U(\beta)}{\partial \beta} \right)^{-1} X' AV^{-1}\} (Y - \mu) \\
= (I - H) (Y - \mu). \tag{6.15}
\]

where \( H = -AX \left( \frac{\partial U(\beta)}{\partial \beta} \right)^{-1} X' AV^{-1} \).

By substituting \( \hat{Z} = (I - H) P(Y - \mu) \) into equation (6.8), the statistic \( T_1 \) can be written as

\[
T_1 = n^{-1} (Y - \mu)' P (I - H) Q (I - H) P (Y - \mu) \\
= n^{-1} Z' Q' Z. \tag{6.16}
\]
As \( n_1 \to \infty \), \( Z_1 \stackrel{d}{\to} \text{MVN}(0, n_i P_i^T \Sigma_i P_i) \), when \( i \neq j \), \( Z_i \) and \( Z_j \) are independent. It follows that \( n^{-1/2} Z \stackrel{d}{\to} \text{MVN}(0, \Sigma^*) \), where \( \Sigma^* = \text{diag}\{\nu_1 P_1^T \Sigma_1 P_1, \ldots, \nu_K P_K^T \Sigma_K P_K\} \). Following Theorem 3, the asymptotic distribution of \( T_1 \) is \( T_1 \stackrel{d}{\to} \sum_{k=1}^{M} \lambda_k s_k^2 \), where the \( \lambda_k \) are the eigenvalues of \( Q^* \Sigma^* \), and the \( s_k \) are independent standard normal random variables. In practice, \( \nu_i \) is estimated by \( \hat{\nu}_i = n_i/n \) and different estimators for \( \Sigma^* \) is discussed in section 6.2.4.

Note that the special case where the working correlation matrix is the identity matrix is useful in practice, since the estimates of \( \beta \) are consistent estimates even when the working correlation matrix is mis-specified. Many papers see for example Albert (1998) and Pan (2001), have suggested using the independence matrix as the working correlation matrix when there is no clear information about the correlation schedule within a subject.

In practice, the parameters are unknown and the corresponding statistics are:

\[
T_2 = n^{-1} \hat{Z}' \hat{Q} \hat{Z},
\] (6.17)

in which all the unknown parameters are replaced by their estimates obtained from the GEE, using “hat” as before. The statistic \( T_2 \) can be shown to have the same asymptotic distribution as \( T_1 \).

**Theorem 7** Under the conditions of Theorem 6, the asymptotic distribution of the statistic \( T_2 \) is the same as that of \( T_1 \).

**Proof:**

From equation (6.17),

\[
T_2 = n^{-1} \hat{Z}' \hat{Q} \hat{Z} = n^{-1} \hat{Z}' Q \hat{Z} + n^{-1} \hat{Z}' (\hat{Q} - Q) \hat{Z} = T_1 + n^{-1} \hat{Z}' (\hat{Q} - Q) \hat{Z}
\] (6.18)

As in the proof of Theorem 6, the quantity \( n^{-1/2} \hat{Z} \) is bounded in probability since it converges in distribution to a normal distribution with mean 0 and a covariance matrix that does not depend on \( n \). To show \( T_2 \) and \( T_1 \) have the same asymptotic distribution, it is sufficient to show \( \hat{Q} - Q \xrightarrow{p} 0 \). The GEE ensures that the estimates of \( \beta \) are consistent, \( \hat{\beta} - \beta \xrightarrow{p} 0 \). There are several continuous functions that link \( \hat{\beta} - \beta \) to \( \hat{Q} - Q \), for example, the linear predictor, the “logit” link function and the weight functions of the statistics.
Therefore, \( \hat{Q} - Q \) is a continuous function of \( \hat{\beta} - \beta \). Since convergence in probability is preserved under continuous transformations, \( \hat{Q} - Q \xrightarrow{P} 0 \) and \( T_2 \) and \( T_1 \) have the same asymptotic distribution.

### 6.2.4 Estimating \( \text{COV}(Y_i) \)

For correlated Binomial responses, the variance-covariance matrix \( \text{COV}(Y_i) \) may be estimated in different ways:

- When the working correlation matrix \( R_i \) is specified correctly, it is natural to consider the estimated working covariance \( V_{w_i} = \hat{A}_i^{-1/2} R_i \hat{A}_i^{1/2} \) as an estimate of \( \text{COV}(Y_i) \). However, very often \( R_i \) is mis-specified.

- An estimate that has appeared in the literature is \( V_{c_i} = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' \). This is the sample covariance estimate based on sample size 1.

- The third estimate of \( \text{COV}(Y_i) \) is similar to the estimator proposed by Pan (2001). The estimator of the covariance of \( Y_i \) is

  \[ V_{u_i} = \hat{A}_i^{-1/2} R_u \hat{A}_i^{-1/2}, \]

where

\[
R_u = \frac{1}{K} \sum_{i=1}^{K} \hat{A}_i^{-1/2} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' \hat{A}_i^{-1/2}.
\]

The latter is the unstructured correlation matrix estimate due to Liang and Zeger (1986). Note that \( R_u \) does not depend on any assumption on the specific structure on the true correlation matrix \( R_i \).

### 6.3 Cumulative Chi-square statistics (CCS) (for balanced data)

As was discussed in Chapter 5, the Cramér-von Mises statistics are Cumulative Chi-square Statistics (CCS) and the power of the Cramér-von Mises statistics can be improved notably by their corresponding likelihood ratio type (LRT) statistics. The CCS and LRT statistics are calculated as follows:
(1) for each of the subjects 1 to \( K \), sum the 1st to the \( j \)th repeated measurements and then sum the \( j + 1 \)th to \( m \)th measurements, for \( j = 1, \ldots, m - 1 \), to give \( m - 1 \) tables with the dimension \( K \times 2 \);

(2a) for the CCS statistic, for each table, compute Pearson \( X^2 \) statistic;

(2b) for the LRT statistic, for each table, compute the likelihood ratio statistic;

(3) compute a weighted sum of the individual \( X^2 \) or likelihood ratio statistics.

In this section, the CCS statistics and the LRT goodness of fit statistics are studied for the correlated Binomial regression models. For simplicity, it is assumed the data are balanced and therefore \( m_i = m \). Let \( N_i = \sum_{t=1}^{m} y_{it} \) be the total number of observations and \( M_i = \sum_{t=1}^{m} \mu_{it} \) be the total number of expects in the \( i \)th subject. As was shown in chapter 5, the Cramér-von Mises statistics are CCS statistics when \( N_i = M_i \). However, note that for finite samples, it is usually the case that the total number of observed counts is different than the total number of the expected counts, \( N_i \neq M_i \), and then the Cramér-von Mises statistics for the GEE models are not the CCS statistics. Since parameter estimates from the GEE are always consistent under mild regularity conditions, it is easily shown that \( M_i/N_i \rightarrow 1 \), as \( n \rightarrow \infty \).

In step (2a) above, the \( X^2 \) statistic for \( (t < m_i) \) is computed as follows: for the \( t \)th collapsed table, compute the \( \chi^2 \) statistic:

\[
\chi^2_t = \sum_{i=1}^{K} \frac{(S_{it} - T_{it})^2}{T_{it}} + \sum_{i=1}^{K} \frac{(N_i - S_{it}) - (M_i - T_{it}))^2}{M_i - T_{it}}
\]

\[
= \sum_{i=1}^{K} \frac{(S_{it} - T_{it})^2}{T_{it}} + \frac{(S_{it} - T_{it})^2}{M_i - T_{it}} + \frac{(M_i - N_i)^2}{M_i - T_{it}} + \frac{2(S_{it} - T_{it})(M_i - N_i)}{M_i - T_{it}}
\]

\[
= \sum_{i=1}^{K} \frac{M_i(S_{it} - T_{it})^2}{T_{it}(M_i - T_{it})} + \sum_{i=1}^{K} \frac{(N_i - M_i)(N_i - M_i - 2(S_{it} - T_{it}))}{M_i - T_{it}}. \quad (6.19)
\]
With weights, \( \chi^2_t = \sum_{i=1}^{K} w^*_it \left\{ \frac{(S_{it}-T_{it})^2}{T_{it}} + \frac{(N_i-M_i)-(M_i-T_{it})}{M_i-T_{it}} \right\} \). The CCS statistics \( S^2 \) are written as

\[
S^2 = m - 1 \sum_{t=1}^{m-1} \chi^2_t = \sum_{t=1}^{m-1} \sum_{i=1}^{K} w^*_it M_i(S_{it}-T_{it})^2 T_{it}(M_i-T_{it}) + \sum_{t=1}^{m-1} \sum_{i=1}^{K} \frac{w^*_it(N_i-M_i)(N_i-M_i-2(S_{it}-T_{it}))}{M_i-T_{it}}
\]

\[
= T^2 + \sum_{t=1}^{m-1} \sum_{i=1}^{K} \frac{w^*_it(N_i-M_i)(N_i-M_i-2(S_{it}-T_{it}))}{M_i-T_{it}}.
\] (6.20)

The weights of the CCS statistics are chosen so the statistics become their corresponding Cramér-von Mises statistics when \( N_i = M_i \). Therefore, the weights of CCS statistics are related to the weights of Cramér-von Mises statistics through \( w^*_it = \frac{T_{it}\mu_{it}(M_i-T_{it})}{(nM_i)} \). When \( w^*_it = T_{it}\mu_{it}(M_i-T_{it})/(nM_i^2) \), the statistic has weights that are similar to those in \( W^2_K \) and when \( w^*_it = \mu_{it}/n \), the statistic's weights are similar to those of \( A^2 \). Furthermore, when \( w^*_it = M_i/n \), the statistic is similar to Taguchi's statistic.

### 6.3.1 Asymptotic theory: CCS statistics

For the distribution theory of the tests, again the statistics will be put into matrix form. The goal is to write the statistics as quadratic forms in \( Y - \mu \):

Define the following \( K \times KM \) matrices \( \Omega_1, \ldots, \Omega_M \) where we use the typographical shorthand \( K^* = K - 1 \):

\[
\Omega_1 = \begin{pmatrix}
1 & 2 & \ldots & m & m+1 & m+2 & \ldots & K^*m & K^*m+1 & K^*m+2 & \ldots & Km \\
1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 
\end{pmatrix},
\]

The matrix \( \Omega_1 \) operates on the vector \( Y - \mu \) to give \( (S_{11} - T_{11}, \ldots, S_{K1} - T_{K1})' \).
The matrix $\Omega_2$, operates on $Y - \mu$ to give $(S_{12} - T_{12}, \ldots, S_{K2} - T_{K2})'$

\[
\Omega_2 = \begin{pmatrix}
1 & 2 & \ldots & m & m + 1 & m + 2 & \ldots & K^*m & K^*m + 1 & K^*m + 2 & \ldots & Km \\
1 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 0 \\
\end{pmatrix},
\]

The matrix $\Omega_M$, operating on a vector $Y - \mu$ gives $(N_1 - M_1, \ldots, N_K - M_K)'$, since by definition $S_{im} = N_i$ and $T_{im} = M_i$ for $i = 1, \ldots, K$.

Recall $Z_{ij} = S_{ij} - T_{ij}$ and define $K$ dimension vectors $\zeta_t$ and $E$: $\zeta_t = (Z_{1t}, \ldots, Z_{Kt})'$, $t = 1, \ldots, m$ and $E = (N_1 - M_1, N_2 - M_2, \ldots, N_K - M_K)'$. Define the vector $\Lambda$ of length $2Km$ by $\Lambda = (\zeta_1', E', \zeta_2', E', \ldots, \zeta_K', E')'$. Finally, put $\Omega = (\Omega_1', \Omega_m', \Omega_2', \Omega_m', \ldots, \Omega_m', \Omega_m')'$. Then $\Omega$ is a $2Km \times Km$ matrix. It can be shown that

\[
\Lambda = \Omega(Y - \mu).
\]  \hspace{1cm} (6.21)

The CCS statistic can be written as

\[
S^2 = n^{-1} \Lambda' J^* \Lambda = n^{-1} (Y - \mu)' \Omega' J^* \Omega (Y - \mu),
\]  \hspace{1cm} (6.22)

where $J^*$ is a block diagonal matrix with blocks $(J^*_1, \ldots, J^*_m)$ and each of the $J^*_t$, $t = 1, \ldots, m$ matrices has four equal size quadrants:

\[
J^*_t = \begin{pmatrix}
J^{*(1)}_t & J^{*(2)}_t \\
J^{*(3)}_t & J^{*(4)}_t \\
\end{pmatrix}
\]
where

\[ J_t^{(1)} = \text{diag}(w_t^* M_1 / \{T_{1t}(M_1 - T_{1t})\}, \ldots, w_t^* M_K / \{T_{Kt}(M_K - T_{Kt})\}) , \]

\[ J_t^{(2)} = J_t^{(3)} = \text{diag}(-w_t^* M_1 / \{T_{1t}(M_1 - T_{1t})\}, \ldots, -w_t^* M_K / \{T_{Kt}(M_K - T_{Kt})\}) , \]

and

\[ J_t^{(4)} = \text{diag}(w_t^* M_1 / \{T_{1t}(M_1 - T_{1t})\}, \ldots, w_t^* M_K / \{T_{Kt}(M_K - T_{Kt})\}) . \]

Define \( S_0^2 \) to be the CCS statistics when parameters are known. The following theorem gives the asymptotic distribution of \( S_0^2 \).

**Theorem 8** Define a block diagonal matrix \( V_S^* = \text{diag}\{\nu_1 V_1, \ldots, \nu_K V_K\} \), where \( V_i, i = 1, \ldots, K \) are defined in Section 6.2.1. Assume that the model \( \mu = \mu(\beta) \) holds with parameters known and that \( \nu_i = \lim_{n \to \infty} \frac{n_i}{n} \) and \( 0 < \nu_i < 1 \). Let \( Q_j^* = \Omega' J^* \Omega \). Then

\[ S_0^2 \overset{D}{\rightarrow} \sum_{k=1}^M \lambda_k s_k^2 , \]

where \( M = \sum_{i=1}^K m_i \), the \( \lambda_k \) are the eigenvalues of \( Q_j^{1/2} V_S^* Q_j^{1/2} \), and the \( s_k \) are independent standard normal random variables.

The proof is similar to that of Theorem 5.

Define \( S_1^2 \) to be the CCS statistics when parameters are estimated. The following theorem gives the asymptotic distribution of \( S_1^2 \).

**Theorem 9** Define a block diagonal matrix \( \Sigma_S^* = \text{diag}\{\nu_1 \Sigma_1, \ldots, \nu_K \Sigma_K\} \), where \( \Sigma_i, i = 1, \ldots, K \) are defined in Section 6.2.1. Assume that the model \( \mu = \mu(\beta) \) holds with parameters estimated from the GEE model and that \( \nu_i = \lim_{n \to \infty} \frac{n_i}{n} \) and \( 0 < \nu_i < 1 \). Let \( Q_j^* = \Omega' J^* \Omega \). Then

\[ S_1^2 \overset{D}{\rightarrow} \sum_{k=1}^M \lambda_k s_k^2 , \]

where \( M = \sum_{i=1}^K m_i \) and the \( \lambda_k \) are the eigenvalues of \( Q_j^{1/2}(I - H) \Sigma_S^*(I - H)Q_j^{1/2} \), where \( H = -AX \frac{\partial U(\beta)}{\partial \beta} A' \) and the \( s_k \) are independent standard normal random variables. When \( R_i \) is the identity matrix, the term \( \frac{\partial U(\beta)}{\partial \beta} \) simplifies to \( -X' AV^{-1} AX \), but for a general \( R_i \), the term \( \frac{\partial U(\beta)}{\partial \beta} \) is given in equation (6.14).
6.4 Likelihood ratio type statistics

As was shown in Chapter 5, when the likelihood ratio (LR) statistics are computed (instead of the Chi-square statistic) from the collapsed $2 \times K$ tables, the weighted sum of these statistics is called a likelihood ratio type (LRT) statistic. The LRT statistic from the $t$th collapsed table is

$$G^2_t = 2 \sum_{i=1}^{K} \left\{ S_{it} \log \left( \frac{S_{it}/N_i}{T_{it}/M_i} \right) + (N_i - S_{it}) \log \left( 1 - \frac{S_{it}/N_i}{1 - T_{it}/M_i} \right) \right\}.$$ 

By a Taylor series expansion as in Section 5.2, it is found that

$$G^2_t \approx \sum_{i=1}^{K} w^*_it N_i M^2_i \frac{(S_{it}/N_i - T_{it}/M_i)^2}{T_{it}(M_i - T_{it})}$$

$$= \sum_{i=1}^{K} w^*_it \left\{ \frac{M^2_i(S_{it} - T_{it})^2}{T_{it}N_i(M_i - T_{it})} + \frac{T_{it}(M_i - N_i)^2}{N_i(M_i - T_{it})} - 2 \frac{M_i(S_{it} - T_{it})(N_i - M_i)}{N_i(M_i - T_{it})} \right\}. \quad (6.23)$$

The likelihood ratio type statistics is

$$G^2 = \sum_{t=1}^{m-1} G^2_t$$

$$\approx \sum_{t=1}^{m} \sum_{i=1}^{K} w^*_it \left\{ \frac{M^2_i(S_{it} - T_{it})^2}{T_{it}N_i(M_i - T_{it})} + \frac{T_{it}(M_i - N_i)^2}{N_i(M_i - T_{it})} - 2 \frac{M_i(S_{it} - T_{it})(N_i - M_i)}{N_i(M_i - T_{it})} \right\}. \quad (6.23)$$

The weights of the LRT statistics are chosen to be the same as those of the CCS statistics since the two families have very similar structure.

6.4.1 Asymptotic theory: LRT statistics

Using the matrices defined in Section 6.3.1, the LRT statistic can be written as

$$G^2 = n^{-1} \Lambda' J \Lambda = n^{-1} (Y - \mu)' \Omega' J \Omega (Y - \mu), \quad (6.24)$$
CHAPTER 6. GEE: BINOMIAL REGRESSION

where \( J \) is a block diagonal matrix with blocks \((J_1, \ldots, J_m)\) and each of the \( J_t, t = 1, \ldots, m \) matrices has four equal size quadrants:

\[
J_t = \begin{pmatrix}
J_t^{(1)} & J_t^{(2)} \\
J_t^{(3)} & J_t^{(4)}
\end{pmatrix},
\]

where

\[
J_t^{(1)} = \text{diag}(w_{1t}^* M_1^2 / \{T_1 N_1 (M_1 - T_1)\}, \ldots, w_{Kt}^* M_K^2 / \{T_K N_K (M_K - T_K)\}),
\]

\[
J_t^{(2)} = J_t^{(3)} = \text{diag}(-w_{1t}^* M_1 / \{N_1 (M_1 - T_1)\}, \ldots, w_{Kt}^* M_K / \{N_K (M_K - T_K)\}),
\]

and

\[
J_t^{(4)} = \text{diag}(w_{1t}^* T_1 / \{N_1 (M_1 - T_1)\}, \ldots, w_{Kt}^* T_K / \{N_K (M_K - T_K)\}).
\]

Define \( G_2^0 \) to be the LRT statistics when parameters are known. The following theorem approximates the asymptotic distribution of \( G_2^0 \).

**Theorem 10** Define a block diagonal matrix \( V_* \mathbf{S} = \text{diag}\{\nu_1 \mathbf{V}_1, \ldots, \nu_K \mathbf{V}_K\} \). where \( \mathbf{V}_i, i = 1, \ldots, K \) are defined in Section 6.2.1. Assume that the model \( \mu = \mu(\beta) \) holds with parameters known and that \( \nu_i = \lim_{n \to \infty} \frac{\nu_i}{n} \) and \( 0 < \nu_i < 1 \). Let \( Q_J = \Omega' J \Omega \). Then

\[
G^2 \xrightarrow{D} \sum_{k=1}^{M} \lambda_k s_k^2,
\]

where \( M = \sum_{i=1}^{K} m_i \) and the \( \lambda_k \) are the eigenvalues of \( Q_J^{1/2} V_* \mathbf{S} Q_J^{1/2} \) and the \( s_k \) are independent standard normal random variables.

The proof is similar to Theorem 5.

Define \( G_2^2 \) to be the LRT statistic when parameters are estimated. The following theorem approximates the asymptotic distribution of \( G_2^2 \).

**Theorem 11** Define a block diagonal matrix \( \Sigma_* \mathbf{S} = \text{diag}\{\nu_1 \Sigma_1, \ldots, \nu_K \Sigma_K\} \). where \( \Sigma_i, i = 1, \ldots, K \) are defined in Section 6.2.1. Assume that the model \( \mu = \mu(\beta) \) holds with parameters estimated from the GEE model and that \( \nu_i = \lim_{n \to \infty} \frac{\nu_i}{n} \) and \( 0 < \nu_i < 1 \). Let \( Q_J = \Omega' J \Omega \). Then

\[
S^2_2 \xrightarrow{D} \sum_{k=1}^{M} \lambda_k s_k^2,
\]
CHAPTER 6. GEE: BINOMIAL REGRESSION

where \( M = \sum_{i=1}^{K} m_i \) and the \( \lambda_k \) are the eigenvalues of \( Q J_{1/2} (I - H) \Sigma^*_1 (I - H) Q J_{1/2} \), where \( H = -AX \frac{\partial U(\beta)}{\partial \beta} X'AV^{-1} \) and the \( s_k \) are independent standard normal random variables. When \( R_i \) is the identity matrix, the term \( \frac{\partial U(\beta)}{\partial \beta} \) simplifies to \(-X'AV^{-1}AX\), but for a general \( R_i \), the term \( \frac{\partial U(\beta)}{\partial \beta} \) is given in equation (6.14).

The proof is similar to the proofs of Theorem 6 and 7.

6.5 Other goodness-of-fit statistics.

6.5.1 Pearson’s \( X^2 \) statistic and unweighted \( X^2 \) statistic.

The well known Pearson’s \( X^2 \) statistic is defined as follows

\[
X^2 = \sum_{i=1}^{K} \sum_{t=1}^{m_i} \frac{(Y_{it} - n_i \pi_{it})^2}{n_i \pi_{it} (1 - \pi_{it})},
\]

or in matrix form \( X^2 = (Y - \mu)' W (Y - \mu) \), where \( W \) is a block diagonal matrix with \( W_i = \text{diag}\{(n_i \pi_{i1} (1 - \pi_{i1}))^{-1}, \ldots, (n_i \pi_{imi} (1 - \pi_{imi}))^{-1}\} \), for \( i = 1, \ldots, K \). Following the theory in Section 6.2.1, it is easy to see that as \( n \to \infty \), the statistic \( X^2 \overset{d}{\to} \sum_{k=1}^{M} \lambda_k s_k^2 \), where the \( s_k \) are independent standard normal random variables. When the parameters are known, the \( \lambda_k \) are the non-zero eigenvalues of \( V^{*1/2} W V^{*1/2} \), where matrix \( V^* \) is defined in Theorem 5; when the parameters are estimated from the GEE model, the \( \lambda_k \) are the non-zero eigenvalues of \( \Sigma^{*1/2} (I - H) W (I - H) \Sigma^{*1/2} \); the matrices \( I, H \) and \( \Sigma^* \) are defined in Theorem 6. Various authors (see Hosmer et al (1997) and Copas (1989)) suggested using an unweighted version of the \( X^2 \) statistic:

\[
X^2_u = (Y - \mu)' (Y - \mu).
\]  

(6.25)

Similarly, one can compute the asymptotic distributions for \( X^2_u \).
6.6 References


Chapter 7

Testing fit for a GEE regression model with correlated binary responses

In this Chapter, goodness of fit tests developed in Chapter 6 will be applied to test fit for a logistic regression model with correlated responses. The situation considered here is when the number of possible covariate combinations is finite and is much smaller than the total sample size.

Suppose there are $N$ subjects. Each subject is considered as a cluster. Suppose there are $K$ combinations of covariate values, the subjects are divided into $K$ groups as follows: within a group, the subjects have the same combination of time-independent covariates (group level covariates). For example, there could be two time-independent covariates: sex (F or M) and dose level (0 ml, 1 ml, 2 ml and 3 ml). The subjects with combination (F, 0ml) form cluster $c_1$, those with (F,1ml) form cluster $c_2$ and so on till the final group $c_8$ consists of subjects (M, 3 ml). Thus there are $K = 8$ groups. Suppose in group $c_i$, there are $n_i$ subjects $s_{ij}, j = 1, \ldots, n_i$. For each subject $s_{ij}$, there are $m_i$ binary response values $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijm_i})'$ and covariate matrix $x_i = (x_{i1}, \ldots, x_{im_i})'$. Suppose there are $p$ regression coefficients in the model, then $x_{it} = (x_{it1}, \ldots, x_{itp})'$. When there are only time-independent covariates in the model, covariates $x_{it}$ would be the same for all $t$, but when there are another covariates have different values depending on when they are measured, covariates $x_{it}$ would be different. For example, a time covariate could be time
CHAPTER 7. GEE: LOGISTIC REGRESSION

94

t itself, indicating when each measurement is taken. Each subject is measured \( m_i \) times which could be slightly different, but for simplicity, it is assumed the measurements are at equally spaced times \( t = 1, \ldots, m_i \). Note that there is no subject-level covariates, so given a covariate combination \( i \) and a measurement time \( t \), subjects \( s_{ij} \) have the same probability of success \( \pi_{it} \) for \( j = 1, \ldots, n_i \). Because all subjects in the cluster are measured at the same times, the number \( Y_{i:t} = \sum_{j=1}^{n_i} Y_{ijt} \) is the number of successes at time \( t \). In this notation, the dot indicates the subscript over which summation is being conducted.

For a given group \( c_i \), define the aggregate vector \( Y_i = (Y_{i1}, \ldots, Y_{im_i})' \). Since \( Y_{ij}^* \) is a correlated multivariate Bernoulli random variable with covariance matrix \( \text{COV}(Y_{ij}^*) = \Sigma_i \) then \( Y_i \) is a correlated multivariate Binomial random variable with covariance matrix \( \text{COV}(Y_i) = n_i \Sigma_i \). The vectors \( Y_i \) for \( i = 1, \ldots, K \) are independent.

In a longitudinal study, observations are correlated when they come from the same subject and are otherwise independent. For such observations, the logistic regression model can be specified by a logistic link function \( \text{logit}(\pi_{it}) = x_{it}' \beta \), where \( \pi_{it} = E(Y_{ijt}|x_{it}) \), for \( j = 1, \ldots, n_i \) and by the following mean-variance relationship \( \text{V}(Y_{ijt}|x_{it}) = \phi \pi_{it}(1 - \pi_{it}) \).

The regression coefficients \( \beta \)'s will be estimated using Generalized Estimating Equations.

7.1 Estimation of parameters

Recall from Chapter 6, the matrix \( D_i \) contains partial derivatives so that \( d_{irc} = \frac{\partial \pi_{it}}{\partial x_{ict}} \), for \( r = 1, \ldots, m_i \) and \( c = 1, \ldots, p \) the matrix \( R_w \) is a specified correlation matrix of \( Y_{ij}^* \), called the working correlation matrix. Also let \( A_i \) be a \( m_i \times m_i \) diagonal matrix with entries \( a_{itt} = \{n_i \pi_{it}(1 - \pi_{it})\} \). Then \( V_i \) is the working variance-covariance matrix of \( Y_{ij}^* \) for \( j = 1, \ldots, m_i \) :

\[
V_i = \phi A_i^{1/2} R_w A_i^{1/2}.
\]

Let \( y_{ij}^* = (y_{ij1}, \ldots, y_{ijm_i})' \) be the observed values of \( Y_{ij}^* \) and let \( \pi_i = (\pi_{i1}, \ldots, \pi_{im_i})' \). The estimation equations are

\[
U = \sum_{i=1}^{K} \sum_{j=1}^{n_i} D_i' V_i^{-1} (y_{ij}^* - \pi_i) = 0. \tag{7.1}
\]

It is easy to see that the double summation essentially sums over all subjects (clusters). Therefore when subjects \( s_{ij} \) have the same \( \pi_i, D_i \) and \( V_i \) for all \( j \), equation (7.1) is the
logistic regression GEEs, see for example, Halekoh et. al. (2006) or equation (1) in Pan (2002). By summing over $j$, equation (7.1) can be written as

$$U = \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \pi_i) = 0. \quad (7.2)$$

Recall that $y_i = (\sum_{j=1}^{n_i} y_{ij1}, \cdots, \sum_{j=1}^{n_i} y_{ijm})'$. This is the estimating equation in equation (6.2); therefore the goodness of tests for GEE Binomial regression can be used for testing the GEE logistic regression with finite covariate combinations (no subject level covariates) based on the $y_i$. Asymptotic theory remains the same as in Chapter 6, but $\text{COV}(Y_i) = n_i \Sigma_i$ can be estimated from the individual subjects $y_{ijt}$: a consistent estimator of $\Sigma_i$ is

$$\hat{\Sigma}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij}^* - \hat{\mu}_i^*)(Y_{ij}^* - \hat{\mu}_i^*)', \quad (7.3)$$

and $\hat{\mu}_i^* = 1/n_i \sum_{j=1}^{n_i} Y_{ij}^*$. The matrix $\hat{\Sigma}_i$ does not depend on any assumption on the working correlation matrix $R_i$.

### 7.1.1 Comparisons of estimators for $\text{COV}(Y)$

In this Section, the performance of different estimators of $\text{COV}(Y)$ were evaluated. The following block diagonal matrices are the different estimates of $\text{COV}(Y)$:

$$V_e = \text{diag}\{V_{e1}, \cdots, V_{eK}\},$$

$$V_w = \text{diag}\{V_{w1}, \cdots, V_{wK}\},$$

$$V_c = \text{diag}\{V_{c1}, \cdots, V_{cK}\},$$

and

$$V_u = \text{diag}\{V_{u1}, \cdots, V_{uK}\}.$$

The estimators $V_u$, $V_w$ and $V_c$ were introduced in Section 6.2.4 and the estimator $V_e$ was in equation (7.3). Notice $V_e$ is computed from individual subject responses $y_{ij}^*$ while the other estimators are computed from the group responses $y_i$. 
Define the Frobenius distance between two equal size square matrices $B$ and $C$ as

$$F_{B,C} = \sqrt{\text{trace}\{(B - C) \ast (B - C)\}'}$$  \hspace{1cm} (7.4)$$

A Monte Carlo study that compared the closeness of the above estimates to the “true” $\text{COV}(Y)$ was conducted. To compute the closeness of the above estimates to the “true” $\text{COV}(Y)$ the steps are as follows:

1. Samples with $n_i = 100$ were generated from one of the models above.
2. Fit the “true” model to samples.
3. The “true” $\text{COV}(Y) = \text{diag}\{V_1, \cdots, V_K\}$ where $V_i = \phi A_i^{1/2}R_iA_i^{1/2}$ is computed from the true parameters.
4. For each of the estimates of $\text{COV}(Y)$, its Frobenius distance to the “true” $\text{COV}(Y)$ is calculated.

The averages of each distance are summarized in Table 7.1.

Table 7.1: The Frobenius distances from the “true” covariance matrix $\text{COV}(Y)$ to its different estimates $V_e, V_w, V_u$ and $V_c$.

<table>
<thead>
<tr>
<th></th>
<th>$V_e$</th>
<th>$V_w$</th>
<th>$V_u$</th>
<th>$V_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>30.04</td>
<td>53.82</td>
<td>108.30</td>
<td>249.42</td>
</tr>
</tbody>
</table>

7.2 The goodness of marginal fit statistic

The goodness of marginal fit (GOMF) test is suggested by Lee and Qaqish (2004). The theory rests on the fact that in the $i$th cluster, the distribution of correlated Bernoulli variable $Y_{ij}^* = (Y_{ij1}, \cdots, Y_{ijm_i})$ can be viewed as a multinomial with $d_i = 2^{m_i}$ cells. Let $\pi_i^* = (\pi_{i1}^*, \cdots, \pi_{id_i}^*)$ denote the probabilities of the multinomial distribution. For example, $\pi_{i1}^*$ is the probability that all the Bernoulli variables equal 0 and $\pi_{id_i}^*$ is the probability that all the Bernoulli variables equal 0, except $Y_{ij1} = 1$ etc.. Suppose $n_i$ independent and identical observations are put into these cells, and let $O_i$ be a $1 \times d$ vector denotes
the cell frequencies. Suppose $C_i$ is a $d \times m_i$ matrix with entries 0 or 1 and each row contains one of the possible values of $Y_{ij}^*$. The Binomial variables $Y_i = O_i C_i$. Since $O_i \sim \text{Mult}(n_i, \pi_i)$, then $\text{COV}(O_i) = n_i D_{\pi_i} = \hat{\pi}_i C_i$. Hence $\text{COV}(Y_i) = n_i C_i (D_{\hat{\pi}_i} - \hat{\pi}_i \hat{\pi}_i') C_i'$. Note that this is equivalent to the empirical covariance estimator $\hat{\Sigma}_e$ in (7.3).

To define the GOMF statistic, the following vectors are needed: define

$$P = (Y_1'/n_1, \ldots, Y_K'/n_k)$$

and

$$\pi = (\mu_1'/n_1, \ldots, \mu_K'/n_k).$$

Further, define $\Omega = V - Q$, where

$$Q = D(D' \mathbf{V}^{-1} D)^{-1} D'$$

and

$$V = \text{COV}(P) = \text{diag}(\Gamma_1/n_1, \ldots, \Gamma_K/n_K),$$

where $\Gamma_i = C_i (D_{\hat{\pi}_i} - \hat{\pi}_i \hat{\pi}_i') C_i'$. Lee and Qaqish suggested using $\hat{\pi}_i = (O_i + 2^{-m_i} \mathbf{1})/(n_i + 1)$, which stabilizes $\Gamma_i$ for small samples and is asymptotically equivalent to $O_i/n_i$. The GOMF statistics is

$$G^2 = (P - \hat{\pi})' \hat{\Omega}^{-1} (P - \hat{\pi}).$$

Under the null hypothesis, the statistic $G^2$ follows asymptotically a $X^2$ distribution with degrees of freedom $df = \sum_{i=1}^K m_i - s$, where $s$ is the number of regression coefficients in the model. Note that this $G^2$ cannot be calculated from the aggregated correlated Binomial data; it requires that one observe the multivariate Bernoulli data.

### 7.2.1 Level and power.

This section examines the levels of the tests as well as the powers of the tests for detecting quadratic and interaction departures from the null model.

Suppose $x_1$ and $x_2$ are cluster level covariates and $x_3$ is a time dependent covariate. For example, $x_1$ is gender of the patients, say $x_1 = 0$ for a male and $x_1 = 1$ for a female; $x_2$ is the dose level (say 0, 1ml, 2ml and 3ml) so that there are $K = 8$ combinations and $x_3$ is
the time of measures. The model under $H_0$ is

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_t,$$

The regression coefficients are chosen to be $\beta_0 = -0.680, \beta_1 = 0.283, \beta_2 = -0.305$ and $\beta_3 = 0.150$. The correlation structure considered is exchangeable with $\rho = 0.4$. For simplicity, each cluster has the same number of subjects ($n_i = 400$), $i = 1, \cdots K$. Each unit is measured the same number of times ($m_i = 5$). Therefore, within a cluster $i$, there are $n_i$ multivariate ($m_i$-variate) binary samples with success probabilities $\pi_{it}$. These samples can be generated by thresholding a normal distribution as suggested by Leisch, Weingessel and Hornik (1998).

A GEE logistic regression model with main effects (sex, dose level etc) and the exchangeable working correlation matrix is fitted and the goodness-of-fit statistics above for GEE models are computed.

The steps are as follows

- (1a) for the levels of the tests, generate $M$ sets of response $Y$ from the null model in equation (7.5)

- (1b) for the powers against an interaction alternative, generate $M$ sets of response $Y$ from model:

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_t + \beta_4 x_{i1} x_{i2}$$

- (1c) for the powers against a quadratic alternative, generate $M$ sets of response $Y$ from model:

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_t + \beta_4 x_{i2}^2$$

- for each set of $Y$, fit the model $\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_t$

- compute the goodness-of-fit statistics.

- compute their $p$ values based on the asymptotic theory

- compute the proportion of the $M$ $p$ values that are less than the nominal level (5% or 1%).

Comments:
1. In Table 7.2, where the empirical covariance estimate \( V_e \) is used, the levels of all the statistics, using asymptotic points, are close to the nominal levels even for small sample sizes, except for the statistic \( LQ \).

2. In Table 7.3, where the working covariance estimate \( V_w \) is used, the levels of all the statistics are smaller than the nominal level. Tests that are based on these levels will be too liberal. This is not surprising because in general \( V_w \) performs worse than \( V_e \) when estimating the COV (\( Y \)), see Table 7.1.

3. In Table 7.4 where the quadratic term is omitted in the analysis, the LRT statistics (\( G^2_W \) and \( G^2_A \)) are much better than the others including the Cramér-von Mises statistics.

4. In Table 7.5 where the interaction term is omitted in the analysis, the LRT statistics are much worse. The other statistics (except \( LQ \)) have similar powers, rising to 70% to 80% with \( \beta = 0.1 \), whereas LRT is 7%.

Table 7.2: Levels of the goodness-of-fit tests at the nominal level 5%(1%) from 1000 replications. The within-subject correlation \( \rho = 0.4 \) is constant. \( W^2_K \) and \( A^2_K \) are respectively Cramér-von Mises and Anderson Darling statistics; \( X^2_W \) and \( X^2_A \) are respectively the Cumulative Chi-square statistics; \( G^2_W \) and \( G^2_A \) are respectively the likelihood ratio type statistics; \( X^2 \) and \( X^2_u \) are the Pearson chi-square and the unweighted sum of squares; \( LQ \) is the Lee and Qaqish statistic. (using the Empirical covariance estimate \( V_e \))

<table>
<thead>
<tr>
<th>Statistics</th>
<th>( n_i = 50 )</th>
<th>( n_i = 100 )</th>
<th>( n_i = 400(V_e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W^2_K )</td>
<td>0.049 (0.013)</td>
<td>0.054(0.008)</td>
<td>0.047 (0.015)</td>
</tr>
<tr>
<td>( A^2_K )</td>
<td>0.046 (0.014)</td>
<td>0.059(0.009)</td>
<td>0.047 (0.013)</td>
</tr>
<tr>
<td>( X^2_W )</td>
<td>0.053 (0.01)</td>
<td>0.049(0.006)</td>
<td>0.047 (0.016)</td>
</tr>
<tr>
<td>( X^2_A )</td>
<td>0.053 (0.01)</td>
<td>0.048(0.007)</td>
<td>0.047 (0.016)</td>
</tr>
<tr>
<td>( G^2_W )</td>
<td>0.043 (0.007)</td>
<td>0.054(0.008)</td>
<td>0.041 (0.006)</td>
</tr>
<tr>
<td>( G^2_A )</td>
<td>0.044(0.007)</td>
<td>0.054(0.005)</td>
<td>0.036 (0.006)</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>0.046(0.006)</td>
<td>0.049(0.005)</td>
<td>0.045 (0.012)</td>
</tr>
<tr>
<td>( X^2_u )</td>
<td>0.046(0.005)</td>
<td>0.05(0.004)</td>
<td>0.04 (0.011)</td>
</tr>
<tr>
<td>( LQ )</td>
<td>0.002(0.000)</td>
<td>0.01(0.002)</td>
<td>0.03(0.008)</td>
</tr>
</tbody>
</table>
Table 7.3: Levels of the goodness-of-fit tests at the nominal level 5%(1%) from 1000 replications. The within-subject correlation \( \rho = 0.4 \) is constant. \( W^2_K \) and \( A^2_K \) are respectively Cramér-von Mises and Anderson Darling statistics; \( X^2_W \) and \( X^2_A \) are respectively the Cumulative Chi-square statistics; \( G^2_W \) and \( G^2_A \) are respectively the likelihood ratio type statistics; \( X^2 \) and \( X^2_u \) are the Pearson chi-square and the unweighted sum of squares; \( LQ \) is the Lee and Qaqish statistic. (using the working covariance estimate \( V_w \))

<table>
<thead>
<tr>
<th>Statistics</th>
<th>( n_i = 50 )</th>
<th>( n_i = 100 )</th>
<th>( n_i = 400(V_e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W^2_k )</td>
<td>0.023(0.001)</td>
<td>0.026(0.001)</td>
<td>0.024(0.001)</td>
</tr>
<tr>
<td>( A^2_k )</td>
<td>0.031(0.003)</td>
<td>0.040(0.004)</td>
<td>0.031(0.003)</td>
</tr>
<tr>
<td>( X^2_W )</td>
<td>0.029(0.001)</td>
<td>0.031(0.002)</td>
<td>0.028(0.001)</td>
</tr>
<tr>
<td>( X^2_A )</td>
<td>0.028(0.001)</td>
<td>0.030(0.002)</td>
<td>0.027(0.001)</td>
</tr>
<tr>
<td>( G^2_W )</td>
<td>0.021(0.002)</td>
<td>0.030(0.004)</td>
<td>0.017(0.005)</td>
</tr>
<tr>
<td>( G^2_A )</td>
<td>0.020(0.002)</td>
<td>0.030(0.005)</td>
<td>0.019(0.006)</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>0.046 (0.001)</td>
<td>0.05(0.004)</td>
<td>0.043(0.009)</td>
</tr>
<tr>
<td>( X^2_u )</td>
<td>0.048(0.003)</td>
<td>0.044(0.003)</td>
<td>0.037(0.006)</td>
</tr>
</tbody>
</table>

7.3 Example

The Cramér-von Mises, CCS and LRT statistics are used now to test fit to a GEE model fitted to a dataset given by Brandon et al (2004). Lee and Qaqish (2004) fitted a standard GEE model and a modified GEE model to the data.

The dataset is part of a much larger study on the effectiveness of different interventions to prevent smoking relapse. A relapse means a participant had smoked during the immediately preceding followed-up periods. The 431 patients were randomly assigned to four groups: (1) Participants received the series of eight Forever Free booklet through the mail immediately after enrollment and 1, 2, 3, 5, 7, 9 and 12 months after. This was called “high contact and high content” condition; (2) Participants received the single simplified Forever Free booklet followed by seven brief letters sent at the same intervals. This was the “high contact and low content” condition; (3) Participants received all eight Forever Free booklets at once upon enrollment and no further contact until the first check up at 12 month. This was “low contact and high content” condition; (4) Participants received only the single simplified booklet at the time of enrollment and no further contact. This was the “low contact and low content” condition. These four conditions formed the \( K = 4 \) clusters. The patients’ smoking relapse status were checked longitudinally at 12, 18 and 24 months. Therefore, \( m_i = 3 \). The number of patients and the number of smoking relapse are summarized in
Table 7.4: Size and power of the goodness-of-fit tests in detecting an omitted quadratic term (dose level) at the nominal level 5%(1%) from 1000 replications. The within-subject correlation $\rho = 0.4$ is constant. $W^K_2$ and $A^K_2$ are respectively Cramér-von Mises and Anderson Darling statistics; $X^2_W$ and $X^2_A$ are respectively the Cumulative Chi-square statistics; $G^2_W$ and $G^2_A$ are respectively the likelihood ratio type statistics; $X^2$ and $X^2_u$ are the Pearson chi-square and the unweighted sum of squares; $LQ$ is the Lee and Qaqish statistic (using the Empirical covariance estimate $V_e$).

<table>
<thead>
<tr>
<th>Statistics</th>
<th>$\beta_4 = 0$</th>
<th>$\beta_4 = 0.03$</th>
<th>$\beta_4 = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W^K_2$</td>
<td>0.047 (0.015)</td>
<td>0.074 (0.018)</td>
<td>0.108 (0.033)</td>
</tr>
<tr>
<td>$A^K_2$</td>
<td>0.047 (0.013)</td>
<td>0.09 (0.028)</td>
<td>0.181 (0.056)</td>
</tr>
<tr>
<td>$X^2_W$</td>
<td>0.047 (0.016)</td>
<td>0.096 (0.022)</td>
<td>0.243 (0.061)</td>
</tr>
<tr>
<td>$X^2_A$</td>
<td>0.047 (0.016)</td>
<td>0.096 (0.022)</td>
<td>0.237 (0.061)</td>
</tr>
<tr>
<td>$G^2_W$</td>
<td>0.041 (0.006)</td>
<td>0.482 (0.264)</td>
<td>0.965 (0.877)</td>
</tr>
<tr>
<td>$G^2_A$</td>
<td>0.036 (0.006)</td>
<td>0.486 (0.263)</td>
<td>0.97 (0.878)</td>
</tr>
<tr>
<td>$X^2$</td>
<td>0.045 (0.012)</td>
<td>0.124 (0.029)</td>
<td>0.403 (0.158)</td>
</tr>
<tr>
<td>$X^2_u$</td>
<td>0.04 (0.011)</td>
<td>0.125 (0.023)</td>
<td>0.368 (0.116)</td>
</tr>
<tr>
<td>$LQ$</td>
<td>0.03 (0.008)</td>
<td>0.239 (0.115)</td>
<td>0.744 (0.497)</td>
</tr>
</tbody>
</table>

Table 7.6. In particular, the question of interest is whether the content of the booklets and the repeated contact affect smoking relapse.

A logistic regression model for correlated responses (smoking relapses) was fitted with covariates intervention style (e.g. Low contact, high content) and time in months. The working correlation matrix is assumed to be “ar(1)”.

Comments: (1) All statistics suggest the model gives adequate fit the $p$ values are 0.72 for $A^K_2$, 0.75 for $W^K_2$, 0.69 for $X^2_W$, 0.63 for $X^2_A$, 0.53 for $G^2_W$ and 0.54 for $G^2_A$. (2) The model implies that probability of smoking relapse is lower for participants who have high content intervention and the the probability is higher as time passes; (3) The model also implies that probability of smoking relapse is not significant with high or low contact intervention; (4) The robust S.E. and the naive S.E. are quite close and this implies the ”ar1” working correlation matrix is probably adequate.
Table 7.5: Size and power of the goodness-of-fit tests in detecting an omitted interaction (sex × dose level) term at the nominal level 5%(1%) from 1000 replications. The within-subject correlation ρ = 0.4 is constant. \( W^2_K \) and \( A^2_K \) are respectively Cramér-von Mises and Anderson Darling statistics; \( X^2_W \) and \( X^2_A \) are respectively the Cumulative Chi-square statistics; \( G^2_W \) and \( G^2_A \) are respectively the likelihood ratio type statistics; \( X^2 \) and \( X^2_u \) are the Pearson chi-square and the unweighted sum of squares; \( LQ \) is the Lee and Qaqish statistic (using the empirical covariance estimate \( V_e \)).

<table>
<thead>
<tr>
<th>Statistics</th>
<th>( \beta_4 = 0.05 )</th>
<th>( \beta_4 = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W^2_K )</td>
<td>0.226 (0.09)</td>
<td>0.829 (0.627)</td>
</tr>
<tr>
<td>( A^2_K )</td>
<td>0.215 (0.089)</td>
<td>0.797 (0.580)</td>
</tr>
<tr>
<td>( X^2_W )</td>
<td>0.223 (0.089)</td>
<td>0.834 (0.639)</td>
</tr>
<tr>
<td>( X^2_A )</td>
<td>0.223 (0.089)</td>
<td>0.838 (0.641)</td>
</tr>
<tr>
<td>( G^2_W )</td>
<td>0.056 (0.012)</td>
<td>0.07 (0.011)</td>
</tr>
<tr>
<td>( G^2_A )</td>
<td>0.055 (0.012)</td>
<td>0.072 (0.011)</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>0.214 (0.085)</td>
<td>0.816 (0.602)</td>
</tr>
<tr>
<td>( X^2_u )</td>
<td>0.215 (0.085)</td>
<td>0.808 (0.600)</td>
</tr>
<tr>
<td>( LQ )</td>
<td>0.108 (0.032)</td>
<td>0.462 (0.233)</td>
</tr>
</tbody>
</table>

7.4 References


Table 7.6: The number of smoking relapses (number of participants) at three follow-up points by 4 intervention groups.

<table>
<thead>
<tr>
<th>Contact</th>
<th>Content</th>
<th>12</th>
<th>18</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>16(82)</td>
<td>19(91)</td>
<td>19(90)</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>17(70)</td>
<td>23(81)</td>
<td>26(84)</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>11(81)</td>
<td>19(93)</td>
<td>21(96)</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>23(87)</td>
<td>31(91)</td>
<td>32(96)</td>
</tr>
</tbody>
</table>
Chapter 8

Future work

This thesis has focused on two aspects of model assessment. In the first, Bayes procedures are used with classical theory to develop Bayes optimal tests for uniformity and normality. Ongoing and future work includes:

1. Developing tests for other distributions
2. Tests when parameters are unknown
3. Tests for discrete distributions
4. Develop tests for other latent variable models, for example time series, regression models with random effects
5. Distribution theory, especially asymptotic for all the above

In the second part of the thesis, various aspects of testing discrete data were examined. Further work includes:

1. Cramér-von Mises statistics for the poisson regression model using GEE
2. Cramér-von Mises similarly for the logistic regression model