MAXIMALLY SELECTED CHI-SQUARES

BY

RUPERT MILLER and DAVID SIEGMUND

TECHNICAL REPORT NO. 64
DECEMBER 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215-06

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
MAXIMALLY SELECTED CHI-SQUARES

BY

RUPERT MILLER and DAVID SIEGMUND

TECHNICAL REPORT NO. 64
December 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-06

Also prepared under Office of Naval Research Contract N00014-77-C-0306 and issued as Technical Report No. 12, Stanford University, Department of Statistics.

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
MAXIMALLY SELECTED CHI-SQUARES

Rupert Miller and David Siegmund
Stanford University

Two samples can be compared by selecting a cut point and forming a $2 \times 2$ table of numbers of observations above and below the cut point in each sample. When the cut point is selected to maximize the standard $\chi^2$ statistic, the $\chi^2$ percentile points are inappropriate. Actual significance levels are computed for large samples, and correct percentile points are tabulated.

†This research was supported in part by NIH Grant RO1 GM21215 and ONR Contract N00014-77-C-0306.
1. Introduction

The following scenario is not uncommon.

A medical investigator feels that a quantitative variable $X$ may be predictive of an event $E$ (e.g., death from a particular illness). He has values of this variable on two groups of patients for whom the event did and did not occur (e.g., non-survivors and survivors). Let $X_{11}, \ldots, X_{1n_1}$ be the values for the first group, and $X_{21}, \ldots, X_{2n_2}$ the values for the second. He could compare these two data sets by a standard two sample test (e.g., $t$ test or Wilcoxon rank test). However, values of this variable for future patients will be used in a predictive fashion. That is, a threshold or cut point $x^*$ will be decided upon such that in patients having values not exceeding $x^*$ the event $E$ will be considered likely to occur (e.g., death) and in patients exceeding $x^*$ the complementary event $E^C$ will be considered more likely (e.g., survival).

In accordance with the way the variable will be used the investigator searches the data sets for the point which best separates the two groups. Specifically, for each $x$ he forms the $2 \times 2$ table

\[
\begin{array}{cc|cc}
| & & E & E^C \\ \hline
X_{ij} < x & a & b \\ X_{ij} > x & c & d \\
\end{array}
\]  

(1)

He then selects as the cut point $x^*$ the $x$ value which maximizes the standard chi-square statistic.
\[ \chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \]  

where \( N = a+b+c+d \). If the \( \chi^2 \) value is statistically significant, the investigator feels he has found a predictor variable.

How should the significance of the \( \chi^2 \) statistic be evaluated? If the investigator simply compares it with the readily available percentile points of a \( \chi^2 \) variable with one degree of freedom (d.f.), he is clearly underestimating the \( P \) value (i.e., overstating the significance) because he has permitted himself the luxury of searching for the most favorable \( \chi^2 \) giving the largest value to the \( \chi^2 \) statistic. By how much is the \( P \) value underestimated? To what critical value should the \( \chi^2 \) statistic be compared?

Before answering these questions several comments are appropriate.

In practice one would probably insert the standard continuity correction into (2). However, since we shall be examining the behavior of (2) in the large sample setting, this refinement is not important for our study.

In the previous discussion we have referred to \( x^* \) as though it was unique. This, of course, is not the case. Any value in the interval between the two ordered observations which maintain the 2 \( \times \) 2 table with the maximum \( \chi^2 \) can be chosen as \( x^* \).

The square root of the maximally selected \( \chi^2 \) value is the two sample version of a test statistic proposed by Anderson and Darling (1952) for the one sample goodness-of-fit problem.
The potential usefulness of the variable $X$ in discriminating between $E$ and $E^C$ is not determined solely by the statistical significance of (2). Large sample sizes can make a small difference significant. Thus, in assessing the discriminatory ability of $X$, one must also be cognizant of the size of the approximate relative risk, i.e., the cross-product (odds) ratio $ad/bc$.

A criterion other than the $\chi^2$ statistic (2) could be chosen for determining the cut point. One alternative would be the standardized log cross-product ratio

$$\frac{|\log(ad/bc)|}{(a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2}}$$  \hspace{1cm} (3)

The theory developed in this paper is directly applicable to finding the limiting distribution of the maximum value of (3). Hopefully, in practice the maximizing value would not be extremely sensitive to the choice of statistic, although this question merits further study.

As an illustration of selecting the maximum $\chi^2$, consider data on the C3 complement levels of premature babies who did and did not develop necrotizing enterocolitis (NEC). For the babies who developed NEC the mean and standard deviation of the C3 complement levels are $\bar{x}_1 = 35.2$, $s_1 = 10.1$, and for the babies who did not $\bar{x}_2 = 49.7$, $s_2 = 19.8$. The maximum $\chi^2$ statistic is 13.55, which occurs for a cut point between 34 and 35 giving the $2 \times 2$ table.
\[ \begin{array}{cc} \text{NEC} & \text{NO NEC} \\ \hline \leq 34 & 9 & 13 \\ > 35 & 4 & 54 \end{array} \] (4)

The cross-product ratio is \( 9 \times 54 / 4 \times 13 = 9.3 \). Improved discrimination is obtained when gestational age is also included (see Stevenson et al. (1980)).

2. Large Sample Theory

For a given \( x \) the square root of the chi-square statistic (2) can be written as

\[
\left( \chi^2 \right)^{\frac{1}{2}} = \frac{|\hat{F}_1(x) - \hat{F}_2(x)|}{\left[ \hat{F}(x)(1-\hat{F}(x)) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{\frac{1}{2}}},
\]

where

\[
\hat{F}_1(x) = \hat{P}\{X_1 \leq x\} = \frac{a}{a+c},
\]

\[
\hat{F}_2(x) = \hat{P}\{X_2 \leq x\} = \frac{b}{b+d},
\]

\[
\hat{F}(x) = \hat{P}\{X \leq x\} = \frac{a+b}{N},
\]

\[ n_1 = a+c, \quad n_2 = b+d. \] (6)

The empirical distribution functions \( \hat{F}_1, \hat{F}_2, \) and \( \hat{F} \) are estimates of the distribution of \( X \) for population 1, population 2, and the common population under \( H_0: F_1 = F_2 = F \), respectively.
As \( n_1 \to \infty \), \( n_1^{\frac{1}{2}}(\hat{F}_1(x) - F_1(x)) \) converges weakly to a tied-down Wiener process (Brownian bridge) \( W_0(t) \) on \([0,1]\) where \( t = F_1(x) \). The same holds true for \( \hat{F}_2 \). The process \( W_0(t) \) has a multivariate normal distribution for any finite collection of \( t \)'s with

\[
E(W_0(t)) = 0,
\]

\[
\text{Cov}(W_0(s), W_0(t)) = s(1-t), \quad 0 \leq s \leq t \leq 1.
\]

For details see Billingsley (1968, pp. 64–65, 103–108).

It follows from the convergence for \( \hat{F}_1, \hat{F}_2 \) and their independence that under \( H_0: F_1 = F_2 = F \)

\[
\left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} (\hat{F}_1(x) - \hat{F}_2(x))
\]

converges weakly to \( W_0(t) \) as \( n_1, n_2 \to \infty \) where \( t = F(x) \). From weak convergence it also follows that in the limit as \( n_1, n_2 \to \infty \) the distribution of the supremum of (5) over \( x \in [L, U] \) is the same as the distribution of the supremum of

\[
\frac{|W_0(t)|}{(t(1-t))^{\frac{1}{2}}}
\]

over \( t \in [F(L), F(U)] \).

The supremum of (9) needs to be restricted to an interval where \( 0 < F(L), F(U) < 1 \) because by the law of the iterated logarithm for Wiener processes the supremum will be infinite almost surely if either
endpoint 0 or 1 is included in the interval. Thus, in evaluating
the asymptotic probabilities, the search for the largest value of (5)
must be restricted to an interval strictly interior to the support of
F. For simplicity we select the interval \( L = F^{-1}(\varepsilon), U = F^{-1}(1-\varepsilon), \)
\( 0 < \varepsilon < 1/2 \), which is symmetric in the tail areas.

A standard Wiener process \( W(t) \) with \( E(W(t)) = 0, \text{Var}(W(t)) = t \)
can be obtained from a tied-down Wiener process \( W_\delta(t) \) through the
relation

\[
W(t) = (1+t)W_\delta \left( \frac{t}{1+t} \right), \quad 0 \leq t < \infty. \tag{10}
\]

By virtue of (10)

\[
\sup_{\varepsilon < t < 1-\varepsilon} \frac{|W_\delta(t)|}{(t(1-t))^{3/2}} = \sup_{\delta < t < \delta^{-1}} \frac{|W(t)|}{t^{3/2}}, \tag{11}
\]

where \( \delta = \varepsilon/(1-\varepsilon) \), so

\[
P\left\{ \sup_{\varepsilon < t < 1-\varepsilon} \frac{|W_\delta(t)|}{(t(1-t))^{3/2}} > w \right\} = P\left\{ \sup_{\delta < t < \delta^{-1}} \frac{|W(t)|}{t^{3/2}} > w \right\}. \tag{12}
\]

An asymptotic expression for the latter probability in (12) as \( w \to \infty \)
has been given by Dirkse (1975, (3)) and Siegmund (1977, (50)). The
expansion is

\[
P\left\{ \sup_{\delta < t < \delta^{-1}} \frac{|W(t)|}{t^{3/2}} > w \right\} \tag{13}
\]

\[
= 2[1-\Phi(w)] + \left[ \frac{2}{\pi} \frac{1}{w} e^{-w^2/2} \right] \ln(\delta^{-1}) + o(w^{-1} e^{-w^2/2}),
\]

\( 1 

\]
where $\Phi(w)$ is the cumulative standard normal distribution function.

The expansion (13) can be used to approximate the probabilities under $H_0$ that the $\chi^2$ statistic (2) actually exceeds the nominally listed critical values for a $\chi^2$ variable with one d.f. when the maximum of the $\chi^2$ is chosen over the central $1-2\varepsilon$ proportion of the underlying distribution. In Table 1 this has been done for $\alpha = .10, .05, .01, .001$ and $\varepsilon = 1/3, 1/4, 1/5, 1/20, \text{ and } 1/100$. The entries in the upper right hand corner of the table have not been reported because it is felt that the accuracy of the expansion (13) for these combinations of $\alpha$ and $\varepsilon$ may be questionable.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\varepsilon$</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
<th>1/10</th>
<th>1/20</th>
<th>1/100</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.25</td>
<td>.33</td>
<td>.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.17</td>
<td>.24</td>
<td>.28</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.01</td>
<td>.05</td>
<td>.08</td>
<td>.10</td>
<td>.15</td>
<td>.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.001</td>
<td>.008</td>
<td>.013</td>
<td>.016</td>
<td>.024</td>
<td>.032</td>
<td>.050</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Asymptotic probabilities that $\max \chi^2$ exceeds $\chi^2_{2\alpha}$ where the max is over $F^{-1}(\varepsilon)$ to $F^{-1}(1-\varepsilon)$.

One would expect the probabilities in Table 1 to substantially exceed their respective $\alpha$'s, but the magnitude of the exceedance is surprising. Selection of the cut point greatly enhances the value of the $\chi^2$ statistic, at least in large samples.

Clearly it is inappropriate to use the percentile point $\chi^2_{2\alpha}$ of a $\chi^2$ variable with one d.f. in conjunction with a maximally selected
\( \chi^2 \) statistic. What are the appropriate critical points? Table 2a gives a small set for \( \alpha = .10, .05, .01 \) and \( \varepsilon = 1/3, 1/4, 1/5, 1/10, 1/20. \)

The square root of a \( \chi^2 \) variable with one d.f. is a normal deviate. Since it is easier to remember normal critical constants than chi-square constants, it is often simpler in practice to work with \( (\chi^2)^{1/2} \). Table 2b gives the critical constants in terms of this variable. For comparing how much increase in the critical constant is required to maintain the significance level under maximal selection, the rows of Table 2b should be compared with 1.64, 1.96, and 2.58, respectively.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \varepsilon )</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
<th>1/10</th>
<th>1/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td></td>
<td>5.24</td>
<td>6.12</td>
<td>6.61</td>
<td>7.61</td>
<td>8.26</td>
</tr>
<tr>
<td>.05</td>
<td></td>
<td>6.86</td>
<td>7.75</td>
<td>8.23</td>
<td>9.21</td>
<td>9.85</td>
</tr>
<tr>
<td>.01</td>
<td></td>
<td>10.46</td>
<td>11.36</td>
<td>11.84</td>
<td>12.80</td>
<td>13.41</td>
</tr>
</tbody>
</table>

Table 2a. Upper \( \alpha \) percentile points of \( \max \chi^2 \) where the max is over \( F^{-1}(\varepsilon) \) to \( F^{-1}(1-\varepsilon) \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \varepsilon )</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
<th>1/10</th>
<th>1/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td></td>
<td>2.29</td>
<td>2.47</td>
<td>2.57</td>
<td>2.76</td>
<td>2.87</td>
</tr>
<tr>
<td>.05</td>
<td></td>
<td>2.62</td>
<td>2.78</td>
<td>2.87</td>
<td>3.04</td>
<td>3.14</td>
</tr>
<tr>
<td>.01</td>
<td></td>
<td>3.23</td>
<td>3.37</td>
<td>3.44</td>
<td>3.58</td>
<td>3.66</td>
</tr>
</tbody>
</table>

Table 2b. Upper \( \alpha \) percentile points of \( \max(\chi^2)^{1/2} \) where the max is over \( F^{-1}(\varepsilon) \) to \( F^{-1}(1-\varepsilon) \).
Table 2a was obtained from Table 2b by squaring the entries. In computing the entries for Table 2b the standard asymptotic expansion

\[ 1 - \Phi(w) = \frac{1}{(2\pi)^{\frac{3}{2}}w} e^{-\frac{w^2}{2}} + O(\frac{1}{w^3} e^{-\frac{w^2}{2}}) \quad (14) \]

as \( w \to \infty \) was utilized. When \( (2\pi)^{-\frac{3}{2}} w^{-1} \exp\{-w^2/2\} \) is substituted for \( 1 - \Phi(w) \) in (13) the resulting expression is easily computed on a programmable calculator or minicomputer. Searching for the correct critical value is a simple matter.

It is perhaps worth noting that one can in principle use the tables of Keilson and Ross (1975) to obtain the probability (12) exactly. However, substantial additional computation is required, and since this probability is already an approximation to the probability of interest, the additional "accuracy" seems not worth the effort.

On the conservative side, one may use the upper bound

\[ P\left\{ \sup_{0 \leq t \leq t_1} \frac{|W(t)|}{t^{\frac{3}{2}}} > w \right\} \leq 2[1 - \Phi(w)] + \frac{w}{(2\pi)^{\frac{3}{2}}} e^{-\frac{w^2}{2}} \left[ \log(t_1/t_0) + 2w^{-2} \right], \quad (15) \]

which may be proved by sharpening slightly the argument of Ito and McKean (1965, p. 34). Some experimentation with the Keilson-Ross tables suggests that the approximation (13) usually underestimates the true probability, and that for \( w \) less than about 3, the upper bound (15) is quite accurate. Use of this bound to construct Table 2b would lead to the entry 2.29 being replaced by 2.44 and the entry 3.66 by 3.68.
The reader may question how to use Tables 2a,b since the points \( F^{-1}(\varepsilon) \) and \( F^{-1}(1-\varepsilon) \) are unknown. The proposal would be to use the sample quantiles from the combined sample to estimate \( F^{-1}(\varepsilon) \) and \( F^{-1}(1-\varepsilon) \). Thus, for \( \varepsilon = 1/4 \) the maximum \( \chi^2 \) would be sought just over the middle half of the combined data. Since the sample quantiles consistently estimate the population quantiles, this modification does not alter the asymptotic theory we have just described.

For the example discussed at the end of the Introduction the maximum \( \chi^2 \) was 13.55. This occurred with 13 observations below the cut point and 67 above it. The value 13.55 would be significant at nearly the \( P = .0001 \) level when ordinary \( \chi^2 \) tables are used, but we know from Table 1 that this significance is exaggerated. If a decision had been made to not search for the maximum \( \chi^2 \) beyond the central 80\% of the data, then 13.55 exceeds the 1\% significance point 12.80 from Table 2a. Making allowance for the non-large NEC sample and the estimation of the population percentiles, one might still feel that the maximal \( \chi^2 \) statistic was significant near the 1\% level. Incidentally, by a two sample \( t \) test the NEC and NO NEC populations are significantly different at the 1\% level.

3. Final Comments

With the aid of a large computer it would be possible to do a Monte Carlo study to determine the correct critical values of the maximal \( \chi^2 \) statistic for use with small to moderate sized samples. All that would be required is generation of uniform random variables and calculation of the \( \chi^2 \) statistic (2) at each point in the combined
sample. With small to moderate sized samples it might not be important to restrict attention to preselected central portions of the data.

Such an extensive undertaking would be somewhat antithetical to the purpose of this note, which is to give some useful guidelines on the interpretation of maximally selected $\chi^2$ values. To test for differences between populations by this procedure would have poor power for alternatives which commonly occur in practice.

Maximally selected $\chi^2$ values can arise in more complicated ways. With more than one predictor variable linear discriminant analysis or logistic regression can determine the angles of the plane which best separates the two populations. Where the plane should be located is then determined by Bayesian considerations, control of one or the other error probabilities, the estimated cross-product ratio, the $\chi^2$ statistic, or some other criterion. Non-linear classification procedures also lead to $2 \times 2$ tables and selected $\chi^2$ values.

Another instance occurs in comparing two treatments when a variable such as age may affect the difference between the treatments. For example, above a certain age there may be no difference between the treatments on the outcome (e.g., death/survival) whereas for younger patients there may be a difference. In this case the investigator may construct two $2 \times 2$ tables for younger and older patients, respectively, and the determination of where to split on age is influenced by how the tables look for various choices of the separation point.

It would be interesting to know how the distribution of the $\chi^2$ statistic (2) under $H_0$ is affected by the selection process in these more complicated settings. The distribution of the maximal $\chi^2$ will
not always be distribution-free of the underlying $F$ in these settings, but some guidelines on how large $\chi^2$ must be to warrant attention after selection would be useful.

Acknowledgement

We thank Fred Huffer and Tom Sellke for some assistance in the preparation of this manuscript.

References


