Technical Notes

Probability Binning and Testing Agreement between Multivariate Immunofluorescence Histograms: Extending the Chi-Squared Test
Keith A. Baggerly*
Department of Biostatistics, M. D. Anderson Cancer Center, Houston, Texas

Received 27 June 2001; Accepted 30 June 2001

Background: A key problem in immunohistochemistry is assessing when two sample histograms are significantly different. One test that is commonly used for this purpose in the univariate case is the chi-squared test. Comparing multivariate distributions is qualitatively harder, as the “curse of dimensionality” means that the number of bins can grow exponentially. For the chi-squared test to be useful, data-dependent binning methods must be employed. An example of how this can be done is provided by the “probability binning” method of Roederer et al. (1,2,3).

Methods: We derive the theoretical distribution of the probability binning statistic, giving it a more rigorous foundation. We show that the null distribution is a scaled chi-square, and show how it can be related to the standard chi-squared statistic.

Results: A small simulation shows how the theoretical results can be used to (a) modify the probability binning statistic to make it more sensitive and (b) suggest variant statistics which, while still exploiting the data-dependent strengths of the probability binning procedure, may be easier to work with.

Conclusions: The probability binning procedure effectively uses adaptive binning to locate structure in high-dimensional data. The derivation of a theoretical basis provides a more detailed interpretation of its behavior and renders the probability binning method more flexible. Cytometry 45:141–150, 2001. © 2001 Wiley-Liss, Inc.

Key terms: chi-square tests; goodness-of-fit tests; probability binning

The problem of comparing two samples to see if they could plausibly have come from a common underlying population is an old one, and the inherent importance of the problem has resulted in much research being devoted to it for over a hundred years. The results of this are that we now possess a variety of methods for distinguishing between two univariate samples, and a few ideas as to how to distinguish between two multivariate samples. The imbalance has a simple explanation: Multivariate distributions are qualitatively harder to examine. For reasons outlined below, we shall focus primarily on the chi-squared test and show how it can be extended to multiple dimensions following the “probability binning” procedure outlined by Roederer et al. (1,2,3). In the process, we will derive some theory for the probability binning statistic and show how it is related to the standard chi-squared statistic. This relationship suggests some useful modifications that make the statistic more sensitive and more flexible.

The first efficient statistical test devised for testing the goodness of fit between a sample and a postulated theoretical model was the chi-squared test introduced by Pearson in 1900 (4), which relied on binning the data. For a table with B bins where the model is completely specified, Pearson showed that the distribution of his statistic when there was no difference was $x^2_B$,$\nu$, and using this distribution he was able to construct significance tests. The chi-squared test was soon extended from the case of comparing a sample with a model to the case of comparing two samples.

The extension is achieved by binning both samples (using the same bin boundaries) as follows:

*Correspondence to: Keith A. Baggerly, Department of Biostatistics, M. D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030-4009.
E-mail: kabagg@mdanderson.org
The form of the chi-squared statistic is simply

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \quad \text{where } e_{ij} = \frac{row_i \cdot col_j}{n},$$

where the row and column totals are assumed known. As a historical aside, we note that while the form of this extension was obvious almost from the beginning, the limiting distribution was not. In one of the great statistical battles of the century, Ronald Fisher (5) was able to show that the distribution of the chi-squared statistic in this case was again $\chi^2_{n-1}$; Pearson himself had argued that it should be $\chi^2_{2n-1}$.

It should be noted that in the two-sample context, there is no explicit underlying form for the model that is assumed; a common model must merely exist. Consequently, the chi-squared statistic can also be viewed as a nonparametric test for comparing two samples. For the asymptotic distribution (and cutoffs) to be accurate, however, the expected bin counts in each row should not be too small (five seems to suffice).

In analyzing univariate data, variants of two tests are the most commonly used. These tests are the Kolmogorov-Smirnov (K-S) test and the chi-squared test. The first application of the K-S statistic to immunological data was given by Young (6). Cox et al. (7) note that the chi-squared test may be a more natural one for flow cytometry data, in that the data are already binned, and binning decreases the sensitivity of the K-S test. They also investigate ways of partitioning the overall chi-squared test into a set of bin-by-bin tests, noting that in each bin the observed proportions of counts should be similar to the proportion of total counts. The difference in cell proportions can be plotted to provide a graphical display of the locations of large deviations. Assessing which individual deviations are large is more difficult than it might appear at first, as simply using the threshold that is appropriate for a single test each time will result in a large false positive rate. To check all of them at once, the squared values that are obtained are plotted against the quantiles of the appropriate chi-squared distribution. Generalizations of these tests that still rely on either binning or on the empirical cumulative distribution function (cdf) are given by Bagwell and colleagues (8,9) and Overton (10), among others.

Both the K-S and the chi-squared tests were initially designed to look at univariate distributions. Both tests can be extended to deal with multivariate data, though with difficulty.

The difficulty in extending the K-S test lies in the fact that there is no uniquely defined "direction of increase" in which to accumulate the data and thus form a cdf. Therefore, extending the test requires a redefinition of the form of the statistic, and the costs of implementation and interpretation suffer dramatically.

By contrast, the form of the chi-square tests lends itself to multidimensional extension fairly nicely. The main problem here is how to bin the data. Naive gridding along each axis produces either lots of bins, most of which will be empty, or a reasonable number of bins with fairly coarse resolution along each axis. One solution is to use bins of unequal sizes, so as to capture a reasonable amount of data in each bin while keeping the total number of bins fairly small.

Choosing the bins to adapt to the location of the data is known as adaptive or "data-dependent" binning. Such binning of the data has been advocated before, but the procedure has not attained widespread use. This is possibly because of the lack of widespread computing power now. A brief discussion (primarily focused on the univariate case) is given in Moore’s (11) overview of chi-squared tests in the monograph on goodness-of-fit techniques by D’Agostino and Stephens (12). As general heuristics, they recommend choosing bins with roughly equal amounts of data and letting the number of bins increase slowly with the overall sample size. The optimal number of bins can be shown to increase at the rate of $n^{2/5}$, but the advantage of "optimality" in this case is small. As a rule, they suggest using a number of bins between $1.88n^{2/5}$ and $3.76n^{2/5}$, $2n^{2/5}$ is a workable intermediate. This should be regarded as a very rough guide. Moore (13) and Moore and Spruill (14) examined the case where the multivariate bins are rectangular in shape with sides parallel to the coordinate axes; some of the theoretical issues associated with dealing with differently shaped regions in higher dimensions are addressed by David Pollard (15). The effect of choosing the data in a dependent fashion on the the asymptotic distribution is actually fairly small. Effectively, as Moore (11) puts it, "Random cells are used in chi-squared tests by 'forgetting' that the cells are data dependent and proceeding as if fixed cells had been chosen.” This is not a license to wreak havoc, as “the random cell boundaries must converge in probability to a set of fixed boundaries.” Thus, probability binning is okay, but choosing gerrymandered cells specifically to include counts from one sample and omit those from the other is not.
control sample. The median and variance of the data along each of the coordinate axes are computed, and the data are split in half along the axis of greatest variance. These halves are then considered as individual data sets, and each is split again as above. The process continues until the number of cells in each bin is sufficiently small. This results in rectangular bins with sides parallel to the coordinate axes, but the sizes of the rectangles vary substantially, being smaller in regions where the data are concentrated and larger where the data are sparse.

The probability binning metric suggested by Roederer et al. (1,2,3) is similar in form to Pearson’s chi-squared test, with a few differences. These differences are most clearly illustrated by considering a 2 × B contingency table, where the rows correspond to “control” and “sample” treatments and the columns correspond to bins (partitions of the possible sample space). For concreteness, we focus on the case B = 2, and we have changed our notation slightly. In the Pearsonian case discussed above there is no mathematical distinction between the sample row and the control row. Probability binning seeks to emphasize the distinct nature of the control sample by binning the table so that the counts in the control row, not the column totals, are treated as fixed. Further, the control counts are partitioned into equal sized bins:

<table>
<thead>
<tr>
<th>Sample</th>
<th>s1</th>
<th>s2</th>
<th>n_c = row1 = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>50</td>
<td>50</td>
<td>n_s = row2 = 100</td>
</tr>
<tr>
<td>Total</td>
<td>?</td>
<td>?</td>
<td>total = 200</td>
</tr>
</tbody>
</table>

Finally, there is a different statistic used for computing agreement. The “PB metric” is

\[ \chi^{PB}_2 = \sum_{i=1}^{B} \left( \frac{(s_i - c_i)^2}{s_i + c_i} \right), \]

where B is the number of bins/columns (here 2), c_i denotes the proportion of the control treatment in bin i, c_i/n_c, (which is fixed at 1/B), and s_i denotes the proportion of the sample treatment in bin i, s_i/n_s. So, summation is here over two cells, not four, and we are using proportions, not counts. This statistic has a bounded range; it cannot fall below 0 or exceed 2.

Let’s take a closer look at the PB metric. First off, in the notation of the Pearsonian table,

\[ \chi^{PB}_2 = \sum_{i=1}^{B} \left( \frac{(s_i/n_i - (c_i/n_c))^2}{s_i + c_i} \right) = \frac{1}{n_i} \sum_{i=1}^{B} \left( \frac{(s_i - n_i * (c_i/n_c))^2}{s_i + c_i} \right) = \frac{1}{n_i} \sum_{i=1}^{B} \frac{(o_i - e_i)^2}{o_i + e_i}. \]

The last formulation emphasizes the similarity to a standard chi-squared test; it uses the fact that the expected sample count in bin i is simply (n_i/n_c) * c_i if the control and sample treatments have the same distribution. The scaling by n_s is new and the denominator has an added term in it. What is the role of the additional o_i in the denominator? For large cells this will be close to e_i under the null hypothesis, so we get an additional scaling of the standard chi-squared test. If we choose to bin the data according to the counts in the control sample, and then work only with the top row, then it can be shown that the adjusted statistic

\[ \sum_{i=1}^{B} \frac{(o_i - e_i)^2}{e_i \left( 1 + \frac{n_i}{n_c} \right)} \]

has a χ_{B-1}^2 distribution. A derivation is given in the appendix. In general, if the two distributions are the same then e_i + o_i = 2e_i + 8, where δ is small relative to e_i. Consequently, when n_s = n_c, the above reduces to

\[ \sum_{i=1}^{B} \frac{(o_i - e_i)^2}{e_i + o_i} = \frac{2n_s n_c}{n} * \text{PB metric.} \]

Consequently, the PB metric has a scaled χ^2 distribution.

To assess significance using the PB metric, Roederer et al. suggest the use of

\[ T(\chi') = \max \left( 0, \frac{\chi'^2 - \chi^2}{\sigma_{\chi'^2}} \right), \]

where

\[ \chi'^2 = \frac{B}{n_{min}} \quad \text{and} \quad \sigma_{\chi'^2} = \frac{\sqrt{B}}{n_{min}}, \]

so

\[ T(\chi') = \max \left( 0, \frac{n_{min} \chi'^2 - B}{B} \right) \]

where n_{min} = min(n_s, n_c). This suggestion is based on the observation that T(\chi’) (before the max operation) has a roughly normal distribution. The fact that T(\chi’) has an approximately normal distribution follows from the fact that a χ_{B-1}^2 distribution can be written as the sum of B − 1 independent χ^2 terms; the central limit theorem then asserts that this will converge to a normal distribution if the number of bins is large. If the number of bins is small, the agreement with normality will be poor.
The theoretical results suggest the need for some slight modifications. Specifically, as the mean of a $\chi^2_{B-1}$ is $B - 1$, the mean of $\chi^2$ is

$$\bar{x}^2 = \frac{(B - 1)n}{2n_1n_e};$$

as the variance of a $\chi^2_{B-1}$ is $2(B - 1)$, the variance and standard deviation of $\chi^2$ are

$$\sigma_{\chi^2}^2 = \frac{2(B - 1)n^2}{4n_1n_e^2} \quad \text{and} \quad \sigma_{\chi^2} = \frac{\sqrt{2(B - 1)n}}{2n_1n_e}. $$

When $n_e = n_s$, the mean reduces to the term given by Roederer et al. $(1,2,3)$ modulo the shift from $B$ to $B - 1$; the standard deviation has an additional factor of $\sqrt{2}$. Taken together, these facts suggest that in its current form, the probability binning statistic is not as precise as it could be. The numerator in $T(\chi')$ has a consistent negative bias, which will tend to suggest that differences are not as large as they actually are. The denominator is consistently too small, which will tend to suggest that differences are significant even when they are not. Which of these two effects dominates depends on the relative sizes of $n_e$ and $n_s$. Incorporating the theoretical results means that more accurate bounds can be set.

The theoretical matchup between the PB metric and Pearson’s chi-squared test suggests the following modifications: either adjust the parameters in the normal approximation, getting

$$T_i(\chi') = \frac{2n_1n_e}{n} \bar{x}^2 - (B - 1) \sqrt{2(B - 1)}, $$

or redefine $\chi^2$ by including a factor of $\frac{2n_1n_e}{n}$ and use the exact chi-squared distribution rather than the approximate normal distribution. We prefer the latter option. While this omits the max operation, excessively small $\chi^2$ values can also be informative. (Based on such a result, Fisher (16) was able to show that Mendel’s original data on peas had likely been “cooked”!)

Alternatively, we can work with the standard Pearson chi-squared statistic for a $2 \times B$ table, namely

$$\chi^2_{Pearson} = \sum_{i=1}^2 \sum_{j=1}^B \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \quad \text{where} \quad e_{ij} = \text{row}_i \ast \text{col}_j / n.$$ 

The results outlined in Moore $(11)$ suggest that the non-random nature of the counts in the control row leave the distribution essentially unaffected, so this statistic has a $\chi^2_{B-1}$ distribution whether we choose the bins to have equal numbers in the control row or equal numbers in each bin, as long as the counts per bin are large enough.

To investigate these modifications, we examined the behavior of six statistics in a small simulation of bivariate data.

The first five statistics are $\chi^2$, $T(\chi')$, $T_i(\chi')$, $\frac{2n_1n_e}{n} \chi^2$, and Pearson’s $\chi^2$. For the first five statistics, we binned the data according to the probability binning procedure. The sixth statistic is again Pearson’s $\chi^2$, but in this case we combined the control data with the sample data and used the combined data set to partition the space. This corresponds to a symmetric treatment of the control and the sample data, and may be more appropriate if we do not have an a priori reason for preferring the control. We considered three different combinations of sample sizes: $n_s = 256$ and $n_e = 512$; $n_s = 512$ and $n_e = 512$; and $n_s = 512$ and $n_e = 256$. These correspond to control to sample ratios of 2 to 1, 1 to 1, and 1 to 2, respectively. In each case we generated $n_e + n_s$ observations from a bivariate normal distribution, binned the data into 16 bins, and computed the various statistics.

The numerical results are given in Table 1. For each of the statistics we recorded the mean, $\mu$, the standard error of the mean, $\sigma/\sqrt{5000}$, and the variance, $\sigma^2$. With respect to $\chi^2$, these values are not particularly informative, as the values to be expected change as we change $n_e$ and $n_s$. Some transformation is needed. For both of the $T$ statistics we would like these values to be near 0 and 1, respectively, as we want these statistics to have a distribution that is close to a standard normal. In the case of $T(\chi')$, there is an obvious large negative bias; as the value of the mean is several SEM below zero: $-1.2939/0.0138 = -93.76$, which has a $p$ value of zero. This problem is worse when the sample sizes are unequal. Further, the variance is not quite what we want; it is too large when $n_e = n_s$ but appears to be about right in the other two cases considered. These phenomena correspond to what we expected. In the numerator of $T(\chi')$, we are (a) subtracting off a number that is too large ($B$ instead of $B - 1$), and (b) subtracting it from a number that is too small $\left(\frac{1}{n_{min}} \chi^2\right)$ instead of $\left(\frac{2n_1n_e}{n} \chi^2\right)$. The change in the size of the effect as $n_e$ and $n_s$ go out of balance is caused by (b); the two fractions are equal when $n_e = n_s$, and the first fraction becomes relatively smaller (more biased) as we move further from the balanced scenario. In

RESULTS

To investigate these modifications, we examined the behavior of six statistics in a small simulation of bivariate data.

The first five statistics are $\chi^2$, $T(\chi')$, $T_i(\chi')$, $\frac{2n_1n_e}{n} \chi^2$, and Pearson’s $\chi^2$. For the first five statistics, we binned the data according to the probability binning procedure. The sixth statistic is again Pearson’s $\chi^2$, but in this case we combined the control data with the sample data and used the combined data set to partition the space. This corresponds to a symmetric treatment of the control and the sample data, and may be more appropriate if we do not have an a priori reason for preferring the control. We considered three different combinations of sample sizes: $n_s = 256$ and $n_e = 512$; $n_s = 512$ and $n_e = 512$; and $n_s = 512$ and $n_e = 256$. These correspond to control to sample ratios of 2 to 1, 1 to 1, and 1 to 2, respectively. In each case we generated $n_e + n_s$ observations from a bivariate normal distribution, binned the data into 16 bins, and computed the various statistics. This procedure was repeated 5,000 times for each set of sample sizes. The numerical results are given in Table 1.

For each of the statistics we recorded the mean, $\mu$, the standard error of the mean, $\sigma/\sqrt{5000}$, and the variance, $\sigma^2$. With respect to $\chi^2$, these values are not particularly informative, as the values to be expected change as we change $n_e$ and $n_s$. Some transformation is needed. For both of the $T$ statistics we would like these values to be near 0 and 1, respectively, as we want these statistics to have a distribution that is close to a standard normal. In the case of $T(\chi')$, there is an obvious large negative bias; as the value of the mean is several SEM below zero: $-1.2939/0.0138 = -93.76$, which has a $p$ value of zero. This problem is worse when the sample sizes are unequal. Further, the variance is not quite what we want; it is too large when $n_e = n_s$ but appears to be about right in the other two cases considered. These phenomena correspond to what we expected. In the numerator of $T(\chi')$, we are (a) subtracting off a number that is too large ($B$ instead of $B - 1$), and (b) subtracting it from a number that is too small $\left(\frac{1}{n_{min}} \chi^2\right)$ instead of $\left(\frac{2n_1n_e}{n} \chi^2\right)$. The change in the size of the effect as $n_e$ and $n_s$ go out of balance is caused by (b); the two fractions are equal when $n_e = n_s$, and the first fraction becomes relatively smaller (more biased) as we move further from the balanced scenario. In
Instead of an bias that appears to decrease as the ratio of reduced, and the variance is closer to 1 throughout. There problems are largely corrected. The bias below zero is of denominator. To clarify the point, consider the behavior of

This accounts for the variance being too large when ns is too small (\( \sigma / \sqrt{5000} \)). When we shift from \( n_c = 256 \), \( n_s = 512 \) and \( n_c = 512 \), \( n_s = 512 \) to \( n_c = 512 \), \( n_s = 256 \), the variance of an approximately correct variance is again just about enough that the biased variance drop significantly below 1.

Table 1

<table>
<thead>
<tr>
<th>( n_c = 256 ), ( n_s = 512 )</th>
<th>( n_c = 512 ), ( n_s = 256 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 )</td>
<td>( T(\chi') )</td>
</tr>
<tr>
<td>( \mu )</td>
<td>0.0423</td>
</tr>
<tr>
<td>( \sigma / \sqrt{5000} )</td>
<td>2.152e-04</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>2.3156e-04</td>
</tr>
<tr>
<td>( \alpha = 0.05 )</td>
<td>0.0066</td>
</tr>
<tr>
<td>K-S</td>
<td>36.0161</td>
</tr>
</tbody>
</table>

Results of 5,000 simulations of partitioning a bivariate distribution into \( B = 16 \) bins using probability binning. The nominal distribution of \( T(\chi') \) and \( T_2(\chi') \) is a standard normal; that of \( 2n_{nc}n_{ns} \frac{1}{n} \chi^2 \), \( \chi^2 \) (PB) and \( \chi^2 \) (comb) is \( \chi^2_{512} \). For each of these five statistics, we computed the mean (\( \mu \)), SEM (\( \sigma / \sqrt{5000} \)), variance (\( \sigma^2 \)), fraction deemed significant at a 95% confidence level (\( \alpha = 0.05 \)), and the K-S distance between the cdf of the simulation results and the cdf of the corresponding theoretical distribution. The overall agreement with theory is best for \( \chi^2 \) (PB) and \( \chi^2 \) (comb).

terms of the variance, we know that we are scaling by a number that is too small (\( \sqrt{B} \) instead of \( \sqrt{2(B - 1)} \)). This accounts for the variance being too large when \( n_c = n_s \). In the other two cases considered, the appearance of an approximately correct variance is again caused by (b) above; when we multiply \( \chi' \) by a number that is too small, both the mean and the variance are reduced. At the points considered, this reduction in variance is just about enough that the biased variance estimate is now approximately correct! If the imbalance between \( n_c \) and \( n_s \) is allowed to get even more extreme than that shown here, the variance of \( \chi' \) will actually drop significantly below 1.

When we shift from \( T(\chi') \) to \( T_2(\chi') \), many of the problems are largely corrected. The bias below zero is reduced, and the variance is closer to 1 throughout. There is still some bias that appears to decrease as the ratio of \( n_c \) to \( n_s \) increases. This bias is largely caused by an effect that we glossed over in our derivation of the distribution of \( \chi^2 \); namely the presence of an \( o_i \) instead of an \( e_i \) in the denominator. To clarify the point, consider the behavior of

\[
A_i = \frac{(o_i - e_i)^2}{o_i} \quad \text{versus} \quad B_i = \frac{(o_i - e_i)^2}{e_i}.
\]

If \( o_i > e_i \), than \( A < B \), and conversely. While this might appear to suggest that things should cancel out, we need to consider both the relative magnitudes of \( A \) and \( B \) and the types of values that we expect from \( o_i \). The \( o_i \) can generally be modeled by Poisson counts. What this means is that we will tend to get counts below the mean (\( o_i < e_i \)) more frequently than counts above the mean, but when we do get counts above the mean they tend to be farther from the mean. As an example, consider the case where

\[
e_i = (5, 5, 5)
\]

\[
o_i = (9, 3, 3)
\]

corresponding to one large positive deviation and two small negative deviations. Here,
As means and variances can be distorted by a few outlying values, it is also informative to see how close we come to matching our nominal significance values. Thus, we also computed how frequently we got a significant result when our nominal confidence level was 95%, corresponding to \( \alpha = 0.05 \). Ideally, we would expect to have these values very close to 5%, values much lower than this indicate that the test is overly conservative, so that we are not picking up differences that are there, and values much larger than this indicate that the test is too liberal, so that we are picking up too many differences that are not there. We did not compute this for \( \chi' \) directly, as all significance tests involve other scaling factors. Again, there are problems with \( T(\chi') \), for which the fraction significant is too small when \( n_c \) and \( n_s \) are unbalanced, and too large when they are balanced. When things are unbalanced, this problem is caused by the large bias of \( T(\chi') \), and when things are balanced, this problem is caused by the standard deviation being too small. When we shift to \( T^*(\chi') \), we see the test being negligibly conservative when \( n_c = 256 \) and \( n_s = 512 \), and slightly liberal otherwise. This slight liberality is because of the fact that a chi-squared distribution has a longer right tail than a normal, so when we use a normal to approximate this distribution we will say that too many things are “significantly large.” This effect is
cancelled when $n_e = 256$ and $n_i = 512$ by the slight negative bias of the statistic.

When we shift to the various chi-squared statistics, the fit gets better; the values indicate a very mild conservative tendency, but this is of a scale that could be attributable to Monte Carlo error. The exception to this is again the case of $\frac{2n_i n_e}{n} \chi^2$, where we get notably fewer significant results than we expect. Again, this is attributable to the observed bias.

Finally, because in each simulation case we have an ideal theoretical distribution to compare it with, we computed the K-S distance from $T(\chi')$ and $T_x(\chi')$ to the standard normal distribution, and from the chi-squared statistics to a $\chi^2_{15}$ distribution. We would like these values to be small, with values less than about 1.5 corresponding to values that we might expect. Obviously, there are problems with both of the normal approximations, and the problem is exacerbated when the sample sizes are unbalanced. The problem is much worse for $T(\chi')$ than for $T_x(\chi')$. This corresponds to the fact that not only are we using a normal distribution to fit a chi-squared distribution, we are not choosing the parameters of the normal so as to achieve the best possible fit. The values for the chi-squared tests are acceptable, with the worst case occurring for the scaled version of $\frac{2n_i n_e}{n} \chi^2$ when $n_e = 256$ and $n_i = 512$.

Based on these results, we favor using Pearson’s $\chi^2$ statistic, using either either type of binning, followed by $\frac{2n_i n_e}{n} \chi^2$. Using a chi-squared distribution as opposed to a normal makes a discernible difference here.

**DISCUSSION**

Binning multivariate data in a data-dependent way allows us to focus our statistical methods on the regions of most interest, rather than focusing too much on clouds of empty space. Probability binning outlines the method, and development of the underlying theory shows how the method can be refined into a more precise and effective tool. In particular, we note that using the regular chi-squared statistic with the bins chosen either by probability binning or by binning the combined data yields good results. These are the procedures that we recommend. They are simple to apply, and because of the nature of their definition they will remain stable even if we are not able to split either sample into “exactly equal sized” bins because of the discreteness of the raw data. This criterion may apply with bivariate data; as the dimension of the data increases this becomes less of a concern.

There are two caveats we wish to express: The first is simply the point made clearly by Cox et al. (7) and Finch (17): that “statistical significance” is not necessarily equivalent to “practically meaningful.” The size of the effect must be looked at. Further, as Cox et al. (7) also note, for these tests to be trusted, we need to remove other sources of “normal variation” that we are not interested in to the greatest extent possible.

The second has to do with the Kolmogorov-Smirnov test. The difficulty in extending the K-S test is that there is no uniquely defined “direction of increase” in which to accumulate the data and thus form a cdf. Consequently, the definition needs to be restated in a different form. This was done by Wolfowitz (18), who noted that the K-S test can be thought of as follows. Starting with a line, we split the line in half and look for the split point such that the proportions of the two groups falling in a given half are maximally discrepant. If we have bivariate data, we start with a plane, and split the plane in half such that the proportions of the two groups falling in a given half are maximally discrepant. This idea of maximally discrepant half-spaces is the multivariate generalization of the K-S test. The difficulty with this extension is that there is no simple way in which to compute the best value. Beran and Millar (19) showed how to approximate the best value by Monte Carlo: choose a point at random on the unit sphere to define a direction, compute the 1-d K-S test for this projection, and record the value. Choose another direction and repeat the process. After a few hundred such choices, the maximum obtained value provides a good approximation to the half-space distance. Walther (20) provides a more efficient way of implementing the Monte Carlo sampling. The problem with this test is that even though it is well-defined, it involves two nested layers of simulation: the first for the computation of the statistic and the second for derivation of the distribution to compare this value with to see if it is “large.” This added computational difficulty causes us to favor binning methods at this time, but as computer power continues to increase this may also become a useful addition to the toolbox.

**APPENDIX**

**Derivation of the Form of the $\chi^2$ Statistic in the PB Context**

We now derive the form of the $\chi^2$ statistic in the probability binning context, which largely corresponds to deriving the scaling factor to be used in the denominator.

To compute the scaling factor, we consider yet another version of the $\chi^2$ test, namely that which arises when a theoretical model has been specified and we have just a single sample to work from. In this case, if there are $B$ bins, there are likewise $B$ known probabilities $p_1$ through $p_B$ which sum to 1.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>$o_1$</th>
<th>$o_2$</th>
<th>...</th>
<th>$o_B$</th>
<th>$n$</th>
</tr>
</thead>
</table>

The expected counts are simply $np_i$. The distribution of observed counts in a cell under the null hypothesis is binomial($n$, $p_i$), which is approximately normal if the expected counts are large enough. Squaring $(o_i - e_i)$ for each cell gives a scaled $\chi^2$ random variable; when the contributions from each cell are scaled to be of the same order of magnitude and
summed, we get a $\chi^2_{d-1}$ random variable, where one degree of freedom has been lost to account for the fact that there is one linear constraint applied to the cell counts (they must sum to $n$). The form of this was proposed by Pearson in 1900 (4), and the commonly used mathematical derivation was given by Fisher in 1922 (5). What is different here is that the $p_i$s, which correspond to the chance that a new observation will fall in bin $i$, are fixed and known if we start with a model, but they are random if we start with probability binning.

Assume the null hypothesis holds—both the control and the treatment samples are drawn from the same underlying population. In the probability binning model, there are two sources of random variation: variation in $p$ about $\bar{p}$, and the binomial variation associated with allocating units into two bins. What can we say about the variation in $p$? We are estimating a quantile of a distribution. When we take a single draw from a distribution, that corresponds to a draw from a uniform distribution on the distribution’s quantiles; the fraction of the distribution to the left of the value we see is uniformly distributed on [0, 1]. If we take several draws and sort them, then the fraction to the left of the central order statistic is no longer uniformly distributed; it can be shown (see, e.g., Rice [21, p. 102] or Hogg and Craig [22]) that the distribution of the $d$th order statistic (out of $n$) is given by

$$\frac{n!}{(d - 1)!(n - d)!} x^{d-1}(1 - x)^{n-d}, \quad 0 \leq x \leq 1,$$

which is a beta distribution with parameters $d$ and $n - d + 1$. In the case where we are splitting things into $B$ equal bins on the basis of a control sample of size $n_c$, the distribution of the probability associated with each bin is approximately beta($n_c/B$, $(B - 1) * (n_c/B)$). The mean and variance of a beta($\alpha$, $\beta$) distribution are

$$\mu = \frac{\alpha}{\alpha + \beta}, \quad \text{which here becomes } \frac{1}{B},$$

$$\sigma^2 = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}, \quad \text{which here becomes } \frac{B - 1}{B^2} \frac{1}{n_c + 1}.$$

Now, when we look at the distribution of counts in a given cell, instead of working with a binomial($n$, $p_i$) distribution, we are working with a beta-binomial($n$, $\alpha_i$, $\beta_i$) distribution, which has the same mean (so we are subtracting off the same amount) but a larger variance (so the squared deviations will be larger on average).

We will attack this problem following Fisher’s approach—we will treat the cells as initially independent (so that $n$ can vary) and then introduce the restriction to get the scaling factors. In Fisher’s case, if we let $X_i$ denote the number observed in cell 1, then the distribution of $X_i$ given $n$ is binomial, and the variance in $n$ is equal to the variance of the sum of the $X_i$. This means that $n$ has a mean and variance, which we shall denote $N$ and $\sigma_n^2$, respectively. We find the variance of $X_i$ given our knowledge of the conditional distribution as follows:

$$V(X_i) = E(X_i^2) - E(X_i)^2 \quad \text{by definition}$$

$$E(X_i|n) = np_i, \quad \text{binomial distribution}$$

$$E(X_i^2|n) = np_iq_i + n^2p_i^2, \quad \text{binomial, with } q_i = 1 - p_i$$

$$E(X_i) = E[E(X_i|n)] = E[np_i]$$

$$= np_1, \quad \text{conditional expectation}$$

$$E(X_i^2) = E[E(X_i^2|n)] = E[np_iq_i + n^2p_i^2]$$

$$= np_1q_i + (\sigma_i^2 + N^2)p_i^2$$

$$V(X_i) = np_1q_i + \sigma_i^2 p_i.$$
This is the expected value used to scale things in the standard Pearson $\chi^2$ statistic. Now, let’s try it with the beta-binomial in place. Here, $p_i$ has a beta distribution with mean $P_i$ and variance $\sigma^2_{p,i}$ and the distribution of $X_i$ given both $n$ and $p_i$ is binomial.

\[ V(X_i) = E(X_i^2) - E(X_i)^2 \] by definition

\[ E(X_i | n, p_i) = np_i \] binomial distribution

\[ E(X_i^2 | n, p_i) = np_i + n^2p_i^2 \]

binomial, with $q_i = 1 - p_i$

\[ E(X_i) = E[E(E(X_i | n, p_i))] \]

\[ = E[E(np_i | n)] \]

\[ = E(np_i) = NP_i \] conditional expectation

\[ E(X_i^2) = E[E(E(X_i^2 | n, p_i))] \]

\[ = E[E(np_i q_i + n^2p_i^2 | n)] \]

\[ = E(np_i + (\sigma^2_{p,i} + p_i^2)(n^2 - n)) \]

\[ = NP_i + (\sigma^2_{p,i} + p_i^2)(\sigma^2_n + N^2 - N) \]

\[ V(X_i) = N(P_i - P_i^2) - NP_{p,i} + N^2\sigma^2_{p,i} \]

At this point, we still need to solve for $\sigma^2_n$, making use of the fact that $\sigma^2_n = \sum V(X_i)$. At this point we also introduce $Q_i = 1 - P_i$ in the notation.

\[ \sigma^2_n = N \sum_i P_i Q_i + \sigma^2_n \sum_i P_i^2 + (N^2 - N) \sum_i \sigma^2_{p,i} \]

\[ + \sigma^2_n \sum_i \sigma^2_{p,i} \]

Rearranging, this gives

\[ \sigma^2_n \left( 1 - \sum_i P_i^2 - \sum_i \sigma^2_{p,i} \right) = N \sum_i P_i Q_i + (N^2 - N) \sum_i \sigma^2_{p,i} \]

which we can use to isolate $\sigma^2_n$ again

\[ \sigma^2_n = \frac{N \sum_i P_i Q_i + (N^2 - N) \sum_i \sigma^2_{p,i}}{1 - \sum_i P_i^2 - \sum_i \sigma^2_{p,i}} \]

\[ = \frac{N \left( \sum_i P_i Q_i - \sum_i \sigma^2_{p,i} \right) + N^2 \sum_i \sigma^2_{p,i}}{\sum_i P_i Q_i - \sum_i \sigma^2_{p,i}} \]

as $1 - \sum_i P_i^2 = \sum_i (1 - P_i^2) = \sum_i q_i$.

\[ = \frac{N^2 \sum_i \sigma^2_{p,i}}{\sum_i P_i Q_i - \sum_i \sigma^2_{p,i}} \]

Now, to finish things off we make use of the fact that the distribution of $p_i$ is beta, so

\[ \sigma^2_{p,i} = \frac{\alpha}{\alpha + \beta} \cdot \frac{\beta}{\alpha + \beta} \cdot \frac{1}{\alpha + \beta + 1} = P_i Q_i \frac{1}{n_i + 1} \cdot \]

Plugging this in to the expression for $\sigma^2_n$ above gives

\[ \sigma^2_n = N \left( \sum_i P_i Q_i \right) \left( \frac{1}{n_i + 1} \right)^{-1} = N + \frac{N^2}{n_i} \]

Taking this and plugging it in to the expression for $V(X_i)$ gives

\[ V(X_i) = NP_i Q_i - N \sigma^2_{p,i} + N^2\sigma^2_{p,i} + \sigma^2_n \sigma^2_{p,i} + P_i^2 \sigma^2_n \]

\[ = NP_i Q_i + \frac{1}{n_i} \frac{N^2}{n_i} \]

\[ + P_i Q_i \left( \frac{1}{n_i + 1} \right) \left( N^2 - N + \frac{N^2}{n_i} \right) \]

\[ = NP_i + \frac{N^2}{n_i} P_i + P_i Q_i \left( \frac{N^2}{n_i} - N \right) \]

\[ = NP_i + \frac{N^2}{n_i} P_i + P_i Q_i \left( 1 - \frac{N^2}{n_i} \right) \]

\[ = NP_i + \frac{N^2}{n_i} P_i (1 + Q_i) \]

\[ = NP_i \left( 1 + \frac{N}{n_i} \right) \]
Using the fact that \( N \) refers to the size of the sample being allocated, which here is \( n_s \), we get the scaling factor that should be used in the denominator of the \( \chi^2 \) statistic in the fixed control binning schema:

\[
n_s P_i \left( 1 + \frac{n_i}{n_s} \right) = e_i \left( 1 + \frac{n_i}{n_s} \right).
\]

Thus, the statistic

\[
\sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i \left( 1 + \frac{n_i}{n_s} \right)}
\]

has a \( \chi^2_{n_s-1} \) distribution. Does this make qualitative sense? If we increase \( n_s \) a lot, making it effectively infinite, then the control sample is large enough to give us a perfect estimate of the underlying distribution and we get the classic \( \chi^2 \) test back out again. Conversely, if \( n_s = n_c \), then we get as much variation from the random location of the bin splits as we do from the allocation of the treatment observations to those bins. When \( n_s = n_c \), the above reduces to

\[
\sum_{i=1}^{n} \frac{(o_i - e_i)^2}{2e_i} = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i + o_i} = n_s \cdot \text{PB metric}.
\]

**LITERATURE CITED**

4. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil Mag 1900;5:157–175.