A NEW METHOD FOR TESTING TWO BINOMIAL SAMPLES WITH ITS APPLICATION IN STATISTICAL QUALITY CONTROL

by P. V. Krishna Iyer and M. N. Bhattacharyya

Defence Science Laboratory, Ministry of Defence, New Delhi

ABSTRACT

A new method for comparing the proportions observed in two binomial samples involving the characters, say A and B with probabilities p & q has been given in this paper. This method provides the basis for a more reliable comparison than the usual $2 \times 2 \chi^2$—contingency test. Taking the samples in order of occurrence to be $x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n$, the tests depend on the statistics $X_s$ and $Y_s$ which are based on a system of scoring $\Sigma (x_k - y_1)$ or $\Sigma |x_k - y_1|$ where $|k-l| \leq s$, a constant. The score for $(x_k - y_1)$ is zero for combination AA or BB and is +1 or −1 if the combination is AB or BA. Similarly $|x_k - y_1|$ is zero in the first case and +1 in the latter two cases. The tests have been illustrated for examining the percentage of defectives produced in two machines. The applications of the tests in psychological investigations have also been explained with reference to the data discussed by McNemar.

Introduction

The $\chi^2$ test with one degree of freedom is usually used for testing the significance of the difference between the observed proportion or percentage of defectives in two independent samples. The samples are arranged in a $2 \times 2$ contingency table and the $\chi^2$ is calculated as follows:

**Table I**

$2 \times 2$ contingency table

<table>
<thead>
<tr>
<th></th>
<th>Defective</th>
<th>Non-Defective</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample I</td>
<td>$f_1$</td>
<td>$f_2$</td>
<td>$N_1$</td>
</tr>
<tr>
<td></td>
<td>($e_1 = N_2 N_1'/N$)</td>
<td>($e_2 = N_1 N_2'/N$)</td>
<td></td>
</tr>
<tr>
<td>Sample II</td>
<td>$f_3$</td>
<td>$f_4$</td>
<td>$N_2$</td>
</tr>
<tr>
<td></td>
<td>($e_3 = N_2 N_1'/N$)</td>
<td>($e_4 = N_2 N_2'/N$)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$N_1'$</td>
<td>$N_2'$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

In the above table $f$ represents the observed frequency while $e$ stands for the expected frequency calculated on the assumption that the two samples
are homogeneous and the proportions of defectives and non-defectives are \( N_1'/N \) and \( N_2'/N \) respectively.

Then
\[
\chi^2 = \sum_{i=1}^{4} \frac{(f_i - e_i)^2}{e_i} = \frac{(f_1 f_4 - f_2 f_3)^2 N}{N_1 N_2 N_1' N_2'}; \text{ d.f.} = 1
\]

The difference between the two proportions is considered to be significant or not according as \( \chi^2 \) is greater than 3.841, the 5% value (6.635 if 1% level is desired), or less than 3.841.

The two samples can also be compared by calculating the standardised deviate of the difference between the proportion of defectives,
\[
Z = \frac{f_1/N_1 - f_3/N_2}{\sqrt{N_1' N_2' / N^2 \left( \frac{1}{N_1} + \frac{1}{N_2} \right)}}
\]
If \( |Z| > 1.96 \) the sample proportions are significantly different at 5% level.

The above two tests, namely \( \chi^2 \) and \( Z \) are, however, equivalent for comparing two binomial samples.

It may be noted that these tests make only an overall comparison of the two samples ignoring completely the order of occurrence of the individual observations. Theoretically, these tests are valid only for random samples. Hence the correct application of the tests calls for prior testing of randomness of the samples which obviously necessitates the maintenance of the order of occurrence of the individuals at the time of collection of the data. From practical point of view also the order of occurrence of the observations and examination of its randomness seem to be of importance, especially in Statistical Quality Control work (S.Q.C.) while comparing two processes or work in two shifts or two machines etc. It is reasonable that a test which takes the order of occurrence also into consideration will provide a more reliable comparison between two samples. A class of such tests developed by us are briefly discussed below.

**Definition and Distribution of Statistics \( X_s \) and \( Y_s \)**

Let two binomial samples of size \( n \), arranged in order of occurrence be as follows

<table>
<thead>
<tr>
<th>Sample I</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( \quad \cdots \quad )</th>
<th>( x_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample II</td>
<td>( y_1 )</td>
<td>( y_2 )</td>
<td>( \quad \cdots \quad )</td>
<td>( y_n )</td>
</tr>
</tbody>
</table>

where \( x_r \) and \( y_r \) the \( r^{th} \) observations in samples I and II are either A or B with probabilities \( p \) & \( q \) respectively for both the samples. Then a class of statistics \( X_s \) and \( Y_s \) which will take into account the order of occurrence of the observations in the two samples, is defined by

\[
X_s = \sum_{r=1}^{n} (x_r - y_r) + \sum_{i=1}^{s} \sum_{r=1}^{n-i} (x_r - y_{r+i}) + \sum_{i=1}^{s} \sum_{r=1}^{n-i} (x_{r+i} - y_r)
\]
\[
Y_s = \sum_{r=1}^{n} |x_r - y_r| + \sum_{l=1}^{s} \sum_{r=1}^{n-1} |x_r - y_{r+1}| + \sum_{l=1}^{s} \sum_{r=1}^{n-1} |x_{r+1} - y_r| \]

where \((x_r - y_{r+1})\) is zero, \(+1\) and \(-1\) according as \((i)\) both \(x_r\) & \(y_{r+1}\) are either \(A\) or \(B\), \((ii)\) \(x_r\) is \(A\) & \(y_{r+1}\) is \(B\) and \((iii)\) vice versa, while \(|x_r - y_{r+1}|\) is zero in case \((i)\) and \(+1\) in both the cases \((ii)\) & \((iii)\). It may be noted that \(X_0, Y_0, X_1\) and \(Y_1\) reduce to

\[
X_0 = \sum_{r=1}^{n} (x_r - y_r), \quad Y_0 = \sum_{r=1}^{n} |x_r - y_r|
\]

\[
X_1 = \sum_{r=1}^{n} (x_r - y_r) + \sum_{r=1}^{n-1} (x_r - y_{r+1}) + \sum_{r=1}^{n-1} (x_{r+1} - y_r),
\]

\[
Y_1 = \sum_{r=1}^{n} |x_r - y_r| + \sum_{r=1}^{n-1} |x_r - y_{r+1}| + \sum_{r=1}^{n-1} |x_{r+1} - y_r|
\]

The first two cumulants of \(X_0, X_1, Y_0\) and \(Y_1\) are as follows

**Statistic** \(X_0\)

\[\kappa_1 = 0\]

\[\kappa_2 = 2npq\]

**Statistic** \(X_1\)

\[\kappa_1 = 0\]

\[\kappa_2 = 2(9n-10)pq\]

**Statistic** \(Y_0\)

\[\kappa_1 = 2npq\]

\[\kappa_2 = 2npq (1-2pq)\]

**Statistic** \(Y_1\)

\[\kappa_1 = 2(3n-2) pq\]

\[\kappa_2 = 6npq (3-10 pq) - 4pq (5-18pq)\]

The expectation and variance for the general case when the probabilities for \(A\) and \(B\) are \(p_1, q_1\) and \(p_2, q_2\) for the first and second samples respectively are given below

\[E(X_s) = W_1(p_1 - p_2)\]

\[E(Y_s) = W_1(p_1q_2 + q_1p_2)\]

For \(s \leq n/2\)
\[ V(X_s) = W_1Z_2 + W_2Z_2, \quad V(Y_s) = W_1Z_3 + W_2Z_4 \]

For \( s \geq n/1 \)

\[ V(X_s) = W_1Z_1 + W_2Z_2, \quad V(Y_s) = W_1Z_3 + W_3Z_4 \]

where

\[
\begin{align*}
W_1 &= (n+2ns) - s(s+1), \\
W_2 &= 2\left\{ (2ns^2 + ns) - \frac{1}{3}(5s^3 + 6s^2 + s) \right\}, \\
W_3 &= \frac{n-1}{6} (6ns - n^2 + 2n - 2s (s^2 - 1)), \\
Z_1 &= (p_1q_2 + q_1p_2) - (p_1 - p_2)^2, \\
Z_2 &= \left\{ p_1q_2(p_1 + q_2) + q_1p_2(q_1 + p_2) - 2(p_1 - p_2)^2 \right\}, \\
Z_3 &= (p_1q_2 + q_1p_2) - 2(p_1q_2 + q_1p_2)^2, \\
Z_4 &= \left\{ p_1q_2(p_1 + q_2) + q_1p_2(q_1 + p_2) - 2(p_1q_2 + q_1p_2)^2 \right\}.
\end{align*}
\]

It has been shown on the basis of the first four cumulants that the distributions of \( X_s \) and \( Y_s \) for finite values of \( p \) tend to the normal form as \( n \) tends to infinity. However, when \( p \) is small the distribution of \( Y_s \) tends to the Poisson form.

The above statistics \( X_s \) and \( Y_s \) can be used for testing the difference between two binomial samples. For large values of \( n \), this is done by calculating the standardized deviate of these statistics by assuming \( p \) to be the same for both the samples. This value may be taken to be any of the observed proportions for the two samples separately or the proportion obtained by pooling the two samples together. Among these, that value of \( p \) which gives the maximum power for the alternatives arising from samples I and II is to be preferred. For small samples the probabilities that \( X_s \geq a \) or \( Y_s \geq b \) can be determined from the exact distributions.

**POWER OF THE DIFFERENT TESTS**

The power of the statistics, *i.e.* the chance that the test shows to be significant, when the samples are not homogeneous, was examined for the hypothesis \( p_0 \) and the alternative \( p_1 \) varying from \( 0.1 \) to \( 0.9 \) for \( n = 30 \) and it was found that the power of the statistics \( X_s \) was maximum for \( X_0 \). As regards \( Y_s \), the power increased rapidly as \( s \) increased up to \( s = n/2 \); beyond which the increase in power was much less and the information on the order included in the test became less and less.

The power curves for the tests based on \( X_0 \) and \( Y_{3/2} \) for the hypotheses \( p_0 = 0.1 \) (0.1) 0.5 and alternative \( p_1 = 0.1 \) (0.1) 0.9 are given in graphs I and II for \( n = 30 \) when the probability for the rejection of the hypothesis \( H_0 = p_0 \), when true, is \( 0.05 \). As the power curves for the hypotheses \( p_0 = 0.9, 0.8, 0.7 \) and \( 0.6 \) are the reverse of those for \( 0.1, 0.2, 0.3 \) and \( 0.4 \) respectively they have not been given.

In fact the standardized deviate of \( X_0 \) and \( X^2 \) are algebraically identical. Hence their power curves should be the same. In graph I there is some difference between the power curves for \( p = 0.1 \) & 0.2. This discrepancy is due to the fact that the value of \( n = 30 \) is not big enough for assuming the distribution to be normal.
In the above graphs the continuous curves I to V represent the powers of the statistics $X_0$ and $Y_{n/2}$ for the hypotheses $p = 1, 2, 3, 4$ and $5$ respectively while the dotted curves 1 to 5 show the corresponding powers for
the standardized deviate of \((\sin^{-1}\sqrt{f_1} - \sin^{-1}\sqrt{f_2})\), where \(f_1\) and \(f_2\) are the observed proportion of defectives in the two samples.

As seen from the above graphs, the standardized deviate of \(X_0\) and the contingency \(\chi^2\) test (almost equivalent to square of the standardized deviate of \((\sin^{-1}\sqrt{f_1} - \sin^{-1}\sqrt{f_2})\) appear to be equally powerful for testing two binomial samples. It may, however, be noted that while the existing test is an overall test ignoring the order of occurrence of the individual observations, the statistic \(X_0\) takes that also into account. For \(Y_{\alpha/2}\) the test is slightly more powerful than the usual \(\chi^2\) test for hypothesis \(p_0=q_0=0.5\) and alternatives near about \(p_1=0.3\) to \(0.7\). For other values of \(p_0\), the more powerful regions are as shown below.

<table>
<thead>
<tr>
<th>Hypothesis, (p_0)</th>
<th>Approximate region (w_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1)</td>
<td>(p_1 &gt; 0.1)</td>
</tr>
<tr>
<td>(0.2)</td>
<td>(p_1 &gt; 0.3)</td>
</tr>
<tr>
<td>(0.3)</td>
<td>(0.2 &lt; p_1 \leq 0.3) and near about (0.6)</td>
</tr>
<tr>
<td>(0.4)</td>
<td>(0.2 &lt; p_1 &lt; 0.4)</td>
</tr>
</tbody>
</table>

**Application of the Tests**

*Example I*—Suppose thirty articles are collected from each of two machines according as the order of production. Let the inspection results be as follows:—

Machine I \(BBABBBABBABBBBABBABBBBABBABBB\),

Machine II \(BBBBBBBBBBBBBBBBBABBABBBBBBBBBBBBB\),

Where \(A\) stands for defective and \(B\) for non-defective article.

The object is to examine whether the two samples are homogeneous *i.e.* the percentage of defectives in the two machines are same and are produced at random.

The observed proportion of defectives in the first sample is \(0.2\) and that in the second sample is \(0.1\). The hypothesis to be tested in this case is that the two samples are homogeneous. If this be true, either of the sample proportions can be taken to be the estimate of the population proportion \(p\). But the object of the test is to find out the real difference between the two samples, if there be any. In the present case this will be done in a more striking manner if the population value of \(p\) is taken to be \(0.1\) instead of \(0.2\) because the power for \(p_0=0.1\) is more than that for \(p_0=0.2\) for all alternatives \((p_1=0.1\) to \(0.9\). Thus assuming \(p=0.1\) and \(q=0.9\) and using the formulae (7) and (8) the standardized deviates of the statistics \(Y_s\) for \(s=0, 1, 5, 10, 15, 20, 25\) and \(29\) and of the
statistic $X_0$ along with the usual large sample binomial test are calculated and
given in the following table.

**Table 3**

*Standardized deviates of the statistics $Y_s$, $X_0$ and the binomial test*

<table>
<thead>
<tr>
<th>Value of $s$</th>
<th>Value of Statistic</th>
<th>Expected value</th>
<th>Standard deviation</th>
<th>Coefficient of variation</th>
<th>Standardized deviate $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_s$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>5.40</td>
<td>2.10</td>
<td>38.88</td>
<td>1.76 not significant</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>15.84</td>
<td>5.82</td>
<td>36.74</td>
<td>1.57 Do.</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>54.00</td>
<td>19.09</td>
<td>35.35</td>
<td>1.72 Do.</td>
</tr>
<tr>
<td>10</td>
<td>139</td>
<td>93.60</td>
<td>33.13</td>
<td>35.40</td>
<td>1.37 Do.</td>
</tr>
<tr>
<td>15</td>
<td>180</td>
<td>124.20</td>
<td>43.76</td>
<td>35.23</td>
<td>1.48 Do.</td>
</tr>
<tr>
<td>20</td>
<td>217</td>
<td>145.80</td>
<td>50.80</td>
<td>34.84</td>
<td>1.40 Do.</td>
</tr>
<tr>
<td>25</td>
<td>232</td>
<td>158.40</td>
<td>54.84</td>
<td>34.62</td>
<td>1.34 Do.</td>
</tr>
<tr>
<td>29</td>
<td>234</td>
<td>162.00</td>
<td>56.03</td>
<td>34.59</td>
<td>1.28 Do.</td>
</tr>
<tr>
<td>$X_0$</td>
<td>3</td>
<td>0</td>
<td>2.32</td>
<td></td>
<td>1.29 Do.</td>
</tr>
<tr>
<td>Usual large sample binomial test</td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
<td>Do.</td>
</tr>
</tbody>
</table>
The replies in this case may be classified in a $2 \times 2$ frequency table as shown in Table 4.

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison of proportions in two samples</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Before Less</td>
</tr>
<tr>
<td>Lecture More</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The problem is to test whether the lecture was effective in convincing the soldiers of the difficulties of war against Japan thereby making a significant change in their opinion.

McNemar’s solution is quoted below.

"Before the lecture, 70 men out of 205 thought that the war would last less than a year, whereas after the lecture this number dropped to 36. The comparison which we wish to make is that between the two frequencies 70/205 and 36/205. There are several ways in which the test may be derived. Perhaps the easiest is to note that both numerators, 70 and 36 contain the $36(a)$ men who persisted in thinking that the war would last less than a year. Hence, equality of the numerators would imply that the same number of men changed from ‘Less’ to ‘More’ as changed from ‘More’ to ‘Less’. In other words, if the lecture is without effect half the persons who changed their minds to change in one direction and half in the other. Thus the test can be made by testing whether the numbers (b) and (c) are binomial successes and failures out of $n=(b+c)$ trials, with probability $\frac{1}{2}$. For this

$$
\chi^2 = \frac{(b-\frac{1}{2}n)^2}{\frac{1}{2}n} + \frac{(c-\frac{1}{2}n)^2}{\frac{1}{2}n} = \frac{(b-c)^2}{(b+c)} = 34
$$

with 1 degree of freedom. A correction for continuity can be applied by subtracting 1 from the absolute value of the numerator before squaring".

Let us examine the application of the statistics $X_a$ and $Y_a$ in the present example. As there is definite matching in the present case, the matched differences $(x_r - y_r)$ or $|x_r - y_r|$ where $x_r$ and $y_r$ are the replies “more” or “less” of the $r$th soldier before and after lecture, can only be considered. Thus either of the statistics $X_a$ and $Y_a$ only is to be used for comparing two samples. Designating the reply “more” by 1 and “less” by 0 the data of Table 4 can be rearranged as in Table 5.
TABLE 5

Replies of Soldiers

<table>
<thead>
<tr>
<th>Before lecture</th>
<th>After lecture</th>
<th>Number of soldier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>135</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The above table shows that the number of soldiers replying (1,1), (0,1) (0,0) and (1,0) are 135, 34, 36 & 0 respectively. Thus the values of $X_o$ and $Y_o$ are $-34$ and $34$ in this case. The proportions of 1 i.e. "more" before and after the lecture are 135/205 & 169/205 respectively. Assuming that the proportion of soldiers replying "more" before and after the lecture remained the same and taking that to be 135/205 the standardized deviates of statistics $X_o$ and $Y_o$ are calculated and given below

\[
Z = \frac{X_o}{\sqrt{2npq}} = -3.54
\]

and

\[
Z = \frac{Y_o - 2npq}{\sqrt{n[2pq - 4p^2q^2]}} = -8.17
\]

These values are significant at 1 per cent level indicating thereby that the proportions of soldiers replying "more" after the lecture changed significantly. It is interesting to note the interpretations of statistics $X_o$ and $Y_o$ in the present case. The statistics $X_o$ represents the difference between the number of soldiers who changed their opinion in one direction, say from "more" to "less" and that of those changing in the other direction i.e. from "less" to "more", while $Y_o$ represents the total number of soldiers who changed their opinion after the lecture.

References


M/s53Amry—8