On certain Types of Compound Frequency Distributions in which the Components can be individually described by Binomial Series.

BY KARL PEARSON, F.R.S.

Certain difficulties with regard to the interpretation of negative binomials, which are of constant occurrence in observational frequency series, have suggested the following investigation.

Consider a number \( u \) of binomial series of which the \( sth \) is \( v_s (p_s + q_s)^n \) and let us suppose a frequency series compounded by adding together the \( rth \) terms of all these series, such will be the compound frequency series it is proposed to discuss. We can realise its nature a little more concretely by supposing \( n \) balls drawn out of a bag containing \( N_p \) white and \( N_q \) black balls, \( N \) being very large as compared with \( n \), or else each ball returned before a fresh draw, while the values of \( p \) and \( q \) change discontinuously at the \( v_1 + 1, v_1 + v_2 + 1, v_1 + v_2 + v_3 + 1 \), etc., draws. We shall take as origin the point at which the sum of the first terms of all the binomials may be supposed to be plotted. \( S \) will denote summation to \( u \) terms.

Let \( N = S(v_s) \), and \( N \mu_1', N \mu_2' \) be the moments of the compound system about this origin, \( N \mu_1, N \mu_2 \) its moments about its mean. Thus

\[
N \mu_1' = S(v_s q_s),
\]

\[
N \mu_2' = S(v_s (n p_s q_s + n^2 q_s^2)).
\]

Hence

\[
N \mu_2 = n S(v_s q_s (1 - q_s)) + n^2 \left( \frac{S(v_s q_s^2) - \left( S(v_s q_s) \right)^2}{N} \right)
\]

\[
= N \mu_1' + \frac{n}{N} [n S(v_s q_s (q_s - q_s^2)) - N S(v_s q_s^2)],
\]

since \( N = S(v_s) \). Accordingly, if \( \sigma \) be the standard deviation and \( m \) the mean of the compound series, i.e. \( \mu_2 = \sigma^2, \mu_1 = m \), then

\[
\frac{\sigma^2}{m} = 1 - \frac{1}{N} \left( \frac{n S(v_s q_s (q_s - q_s^2)) - N S(v_s q_s^2)}{S(v_s q_s)} \right).
\]

Now suppose we had endeavoured to fit a binomial \( N(P + Q)^k \) to the compound series, we should have had

\[
m = \kappa Q, \quad \sigma^2 = \kappa PQ,
\]

and accordingly have found

\[
Q = -\frac{1}{N} \left( \frac{n S(v_s q_s (q_s - q_s^2)) - N S(v_s q_s^2)}{S(v_s q_s)} \right),
\]

\[
\kappa = -\frac{n}{N} \left( \frac{S(v_s q_s (q_s - q_s^2)) - N S(v_s q_s^2)}{S(v_s q_s)} \right).
\]

Thus, had we attempted to fit a binomial to the heterogeneous series, we should have found \( Q \) negative and \( P \) greater than unity provided

\[
n S(v_s q_s (q_s - q_s^2)) > N S(v_s q_s^2),
\]

a condition which will frequently be found to be satisfied, especially if \( q_s \) be small and \( n \) large.

In the limit let us take \( n q_s = m_s \) and \( q_s \) vanishingly small, i.e. suppose the \( sth \) binomial to be replaced by the Poisson series

\[
e^{-m_s} \left( 1 + m_s + \frac{m_s^2}{1!} + \frac{m_s^3}{2!} + \ldots \right),
\]

then we have at once

\[
Q = -\frac{S(v_s q_s (m_s - m_s^2)^2)}{NS(v_s m_s)},
\]

\[
P = 1 + \frac{S(v_s q_s (m_s - m_s^2)^2)}{NS(v_s m_s)},
\]

\[
\kappa = -\frac{(S(v_s m_s)^2)}{S(v_s q_s (m_s - m_s^2)^2)}.
\]

Thus, if two or more Poisson's series be combined term by term from the first, then the compound will always be a negative binomial. This theorem was first pointed out to me by "Student" and suggested by him as a possible explanation of negative binomials occurring in.
material which theoretically should obey the Law of Small Numbers, e.g. "Student's" own Haemacytometer counts*. Of course the negative binomial may quite conceivably arise from other sources† than heterogeneity, but if this be the source of its origin in the material of Bortker-

botsch, Mortara and McKendrick‡, it is certainly most unfortunate that such material should have been selected to illustrate Poisson's limit to the binomial.

Now we know that the values about the mean of the successive moment coefficients of the binomial are

\[
\begin{align*}
\mu_1 &= npq, \\
\mu_2 &= npq (p - q), \\
\mu_3 &= npq (p - q)^2, \\
\mu_4 &= npq (1 + 3(n - 2) pq).
\end{align*}
\]

Further the mean is at a distance \( nq \) from the first term \( p^n \). We shall call this distance \( m \).

Let \( \mu_1', \mu_2', \mu_3' \) and \( \mu_4' \) be the moment coefficients round the start of each binomial. Then

\[
\begin{align*}
\mu_1' &= nq, \\
\mu_2' &= npq + n^2q^2, \\
\mu_3' &= npq (p - q) + 3npq \times nq + n^3q^3, \\
\mu_4' &= npq (1 + 3(n - 2) pq) + 4npq (p - q) nq + 6 (npq)n^2q^2 + n^4q^4.
\end{align*}
\]

From these equations we deduce

\[
\begin{align*}
\mu_1' &= nq, \\
\mu_2' - \mu_1' &= n(n - 1)q^2, \\
\mu_3' - 3\mu_2' + 2\mu_1' &= n(n - 1)(n - 2)q^3, \\
\mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1' &= n(n - 1)(n - 2)(n - 3)q^4.
\end{align*}
\]

Now let

\[
\begin{align*}
a_1 &= \mu_1' \\
a_2 &= \mu_2' - \mu_1' \\
a_3 &= \mu_3' - 3\mu_2' + 2\mu_1' \\
a_4 &= \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'.
\end{align*}
\]

Then we have, if \( \lambda_1 = \nu_1/N, \lambda_2 = \nu_2/N \):

\[
\begin{align*}
\frac{1}{n} &= \lambda_1 + \lambda_2, \\
\frac{a_1}{n} &= \lambda_1 q_1 + \lambda_2 q_2, \\
\frac{a_2}{n(n - 1)} &= \lambda_1 q_1^2 + \lambda_2 q_2^2, \\
\frac{a_3}{n(n - 1)(n - 2)} &= \lambda_1 q_1^3 + \lambda_2 q_2^3, \\
\frac{a_4}{n(n - 1)(n - 2)(n - 3)} &= \lambda_1 q_1^4 + \lambda_2 q_2^4.
\end{align*}
\]

Multiply each equation by \( q_1 \) and subtract from that below it and we find:

\[
\begin{align*}
\frac{a_1}{n} - q_1 &= \lambda_1 (q_2 - q_1), \\
\frac{a_2}{n(n - 1)} - \frac{a_1 q_1}{n} &= \lambda_2 (q_2 - q_1), \\
\frac{a_3}{n(n - 1)(n - 2)} - \frac{a_2 q_1}{n(n - 1)} &= \lambda_2 q_2 (q_2 - q_1), \\
\frac{a_4}{n(n - 1)(n - 2)(n - 3)} - \frac{a_3 q_1}{n(n - 1)(n - 2)} &= \lambda_2 q_2^2 (q_2 - q_1).
\end{align*}
\]

\* For an examination of Bortkewitsch and Mortara's instances see L. Whitaker, loc. cit. pp. 49-66. McKendrick has recently reached Poisson's series (Proceedings of the London Mathematical Society, Vol. xiii. (1914), pp. 405 et seq.) without apparently recognising that he was on familiar ground, and has suggested its application to the frequencies obtained by counts of the bacilli ingested by leucoocytes. He has fitted his series not by moments, but from the first two terms, and has failed to recognise that a large proportion of such leucoocyte counts give also negative binomials.
Hence dividing each equation by the preceding one:

\[
\begin{align*}
\frac{a_2}{n(n-1)} - \frac{a_2 q_1}{n} & = \frac{a_2}{n(n-1)(n-2)} - \frac{a_2 q_1}{n(n-1)} \\
\frac{a_1}{n - q_1} & = \frac{a_2}{n(n-1)} - \frac{a_2 q_1}{n} \\
\frac{a_3}{n(n-1)(n-2)} - \frac{a_2 q_1}{n(n-1)} & = \frac{a_3}{n(n-1)(n-2)} - \frac{a_2 q_1}{n(n-1)} \\
\frac{a_4}{n(n-1)(n-2)} & = q_2.
\end{align*}
\]

Writing \(q_1 q_2 = P_2\) and \(q_1 + q_2 = P_1\), we obtain

\[
\begin{align*}
& a_2 - (n - 1) a_1 P_1 + n (n - 1) P_2 = 0, \\
& a_3 - (n - 2) a_2 P_1 + (n - 1) (n - 2) a_1 P_2 = 0, \\
& a_4 - (n - 3) a_3 P_1 + (n - 2) (n - 3) a_2 P_2 = 0,
\end{align*}
\]

three equations to determine \(n, P_1\) and \(P_2\).

Eliminating \(- P_1\) and \(P_2\) we find:

\[
\begin{vmatrix}
 a_2 & (n - 1) a_1 & n (n - 1) \\
 a_3 & (n - 2) a_2 & (n - 1) (n - 2) a_1 \\
 a_4 & (n - 3) a_3 & (n - 2) (n - 3) a_2
\end{vmatrix} = 0, \quad \text{..................................(xii)}
\]

which expanded gives us the cubic for \(n\):

\[
\begin{align*}
n^3 (2a_1 a_2 a_3 - a_3^2 - a_1^2 a_4 + a_2 a_4 - a_2^3) + n^2 (-12 a_1 a_3 a_2 + 7 a_3^2 + 4 a_1 a_2^2 a_4 - 3 a_3 a_2^4 + 4 a_2^4) \\
+ n (22 a_1 a_2 a_3 - 16 a_3^2 - 5 a_1^2 a_4 + 2 a_2 a_4 - 3 a_3^2) + (-12 a_1 a_2 a_3 + 12 a_2^2 + 2 a_1 a_2^4) = 0. \quad \text{..................................(xii)bis}
\end{align*}
\]

A root of this cubic substituted in the first two equations of (xi) will give \(P_1\) and \(P_2\) and then the quadratic

\[
p^2 - P_1 p + P_2 = 0 \quad \text{..................................(xiii)}
\]

will determine the two values \(q_1\) and \(q_2\) corresponding to the value of \(n\). The first two equations of (viii) then complete the solution by providing \(\lambda_1\) and \(\lambda_2\).

Until the roots of the cubic (xii)bis have been discussed we can only assume that three solutions are possible. As a matter of fact in the examples so far dealt with some of these solutions have usually to be discarded.

For the special case of Poisson's limit to the binomial, we make \(n\) indefinitely large, \(q\) indefinitely small, and \(nq = m\) finite. Hence equations (viii) become

\[
\begin{align*}
1 & = \lambda_1 + \lambda_2, \\
 a_1 & = \lambda_1 m_1 + \lambda_2 m_2, \\
 a_2 & = \lambda_1 m_1^2 + \lambda_2 m_2^2, \\
 a_3 & = \lambda_1 m_1^3 + \lambda_2 m_2^3, \\
 a_4 & = \lambda_1 m_1^4 + \lambda_2 m_2^4,
\end{align*}
\]

leading to

\[
\begin{align*}
& a_2 - a_1 Q_1 + Q_2 = 0, \\
& a_3 - a_2 Q_1 + a_1 Q_2 = 0, \\
& a_4 - a_3 Q_1 + a_2 Q_2 = 0,
\end{align*}
\]

if

\[
Q_1 = m_1 + m_2, \quad Q_2 = m_1 m_2.
\]

Thus we find:

\[
\begin{align*}
Q_1 & = (a_3 - a_2 a_2)/(a_2 - a_1^2), \\
Q_2 & = (a_4 a_3 - a_2^2)/(a_2 - a_1^2),
\end{align*}
\]

subject to the condition*

\[
a_4 (a_3 - a_2^2) + 2 a_1 a_2 a_3 - a_2^3 - a_3^2 = 0. \quad \text{..................................(xv)}
\]

Hence \(m_1\) and \(m_2\) are roots of

\[
m^2 (a_2 - a_1^2) - m (a_3 - a_1 a_2) + a_3 a_1 - a_2^2 = 0. \quad \text{..................................(xvi)}
\]

* Of course equations of condition hold for the 5th and higher moments in the case of the two binomial components. But they are of small service as the probable errors of these high moments are usually very considerable.
Further
\[ \lambda_1 = \frac{\mu_1' - m_2}{m_1 - m_2}, \quad \lambda_2 = \frac{\mu_1' - m_1}{m_2 - m_1}, \]
which determine \( \lambda_1 \) and \( \lambda_2 \).

It is thus quite easy to resolve a series into the sum of two Poisson's binomial limits provided the roots of the above quadratic are real. As illustration I take "Student’s" first count of yeast cells on the 400 squares of a haemacytometer. He found:

<table>
<thead>
<tr>
<th>No. of yeast cells</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>213</td>
<td>128</td>
<td>37</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td>400</td>
</tr>
</tbody>
</table>

giving: \( \text{mean} = \mu_1' = 6825, \quad \mu_2 = 8117, \quad \mu_3 = 1.0876. \)

From the above values of "Student" I determined
\[ \mu_1' = 1.2775, \quad \mu_2 = 3.0675. \]

Whence \( a_1 = -6825, \quad a_2 = -5950, \quad a_3 = -6000, \)
and the resulting quadratic is
\[ -129,194m^2 - 193,913m + 655,475 = 0. \]

The roots are \( m_1 = 3847 \) and \( m_2 = 1.1163 \), leading to \( \lambda_1 = 59.295 \) and \( \lambda_2 = 40.705 \), or, the series has for its two components
\[ \nu_1 = 237.18, \quad m_1 = 3847, \]
\[ \nu_2 = 162.82, \quad m_2 = 1.1163. \]

Calculating out the Poisson’s series for these components we have:

<table>
<thead>
<tr>
<th>No. of yeast cells</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Compt.</td>
<td>161.44</td>
<td>62.11</td>
<td>11.95</td>
<td>1.53</td>
<td>0.15</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2nd Compt.</td>
<td>53.32</td>
<td>32.22</td>
<td>12.36</td>
<td>3.45</td>
<td>0.77</td>
<td>0.14</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Round totals</td>
<td>215</td>
<td>122</td>
<td>44</td>
<td>14</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>213</td>
<td>128</td>
<td>37</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The test for "goodness of fit" for these six groups gives \( \chi^2 = 2.82 \) and \( P = 0.73 \), or the fit is very good. The negative binomial gave \( P = 0.52 \) and a single Poisson’s series only \( P = 0.04 \). But the double Poisson’s series places of course one constant more at our disposal than the binomial, and we can do still better with a double binomial, as we have four constants and only six frequencies, while the double Poisson has three constants to six frequencies. It is clear that neither of the above components forming 41 % and 59 % of the total number of cells, and having their means at 3847 and 1.1163 instead of 6825, gives any idea of a dominant constitution in the solution sampled. If in this case heterogeneity accounts for the negative binomial, then the difference of the components is not slight, and the heterogeneity being gross would indicate some considerable failure in technique.

If we assume that the counts with a haemacytometer ought to follow the Poisson distribution, —and this seems to be theoretically probable,—then the criterion of the binomial might well be adopted to ascertain the possibility of some failure in technique. The actual binomial in "Student's" first case should be \( \left( \frac{14.75}{11.8} \right)^{273} \) and any binomial with \( p \) very small and \( np = 6825 \) would effectively represent the series; we could not anticipate getting \( n = 273 \) and \( p = \frac{1}{1.05} \) closely from the data, but we might certainly anticipate a positive binomial, if the theory of a Poisson distribution be correct. If on the other hand we say in this and many similar cases that the negative binomial arises from heterogeneity, then it appears to me that we have saved our theory at the expense of our technique. I propose now to test this point further by considering the component binomials. If the theory of heterogeneity be correct, unless it be very
manifold, we might anticipate two binomial components, $v_1 (p_1 + q_1)^n + v_2 (p_2 + q_2)^n$, with $n$ positive and large, both $q_1$ and $q_2$ being small, while $n_1 q_1$ and $n_2 q_2$ would be approximately $\cdot 3847$ and $\cdot 1163$, the frequencies $v_1$ and $v_2$ being roughly in the ratio of 3 to 2.

Returning to "Student's" data we find $\mu_4' = 8.9275$, whence

$$a_1 = -6825, \quad a_2 = -5950, \quad a_3 = -6000, \quad a_4 = -4800.$$

Substituting in (iii) we obtain for the cubic:

$$-021,3269n^3 + -028,2321n^2 + 363,3020n + -051,0825 = 0.$$

This has three real roots, approximately.

$$n' = 4.89997, \quad n'' = -1.14234, \quad \text{and} \quad n''' = -3.43390.$$

We will consider in succession these cases: (i) $n' = 4.89997$. The first two equations of (xi) provide

$$-5950 - 2.66173P_1 + 19.10974P_2 = 0,$$

$$-6000 - 1.72548P_1 + 7.71894P_2 = 0,$$

leading to $P_1 = -55304, \quad P_2 = -04590$, and the quadratic

$$q^2 - 55304q + -04590 = 0,$$

whence we deduce the binomial factors

$$q_1 = -5414, \quad P_1 = -5486, \quad \text{and} \quad q_2 = -1017, \quad P_2 = -8983.$$

The first two equations of (viii) give

$$1 = \lambda_1 + \lambda_2, \quad -6825/4.89997 = -051,355\lambda_1 + -101,685\lambda_2,$$

leading to

$$\lambda_1 = -101,535, \quad \lambda_2 = -892,465,$$

or, in a population of 400,

$$v_1 = 43.014, \quad v_2 = 356.986.$$

Accordingly the compound series is given by

$$43.014 (0.5486 + 0.4514)^{-1.14234} + 356.986 (0.8983 + 0.1017)^{-3.43390},$$

with means of the components at

$$m_1 = 2.2118 \quad \text{and} \quad m_2 = -4983.$$

We see that neither the sizes of the component populations nor their means have any relation to the previously discussed component Poisson series; further the present series* diverge widely from Poisson series, $n$ is not large nor $q_1$ or $q_2$ very small. Calculated to the nearest whole numbers we obtain:

<table>
<thead>
<tr>
<th>No. of yeast cells</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Compt.</td>
<td>2</td>
<td>9</td>
<td>15</td>
<td>12</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2nd Compt.</td>
<td>211</td>
<td>117</td>
<td>26</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Combination</td>
<td>213</td>
<td>126</td>
<td>41</td>
<td>15</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Observed</td>
<td>213</td>
<td>128</td>
<td>37</td>
<td>18</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

which leads to $\chi^2 = 1.27$ and $P = 0.93$.

Thus the fit is excellent, but it does not correspond to the heterogeneity of a double Poisson series.

(ii) $n'' = -1.14234$. Here the first two equations of (xi) provide

$$-5950 - 77965P_1 + -16280P_2 = 0,$$

$$-6000 - 1.27469P_1 + 1.75727P_2 = 0,$$

and give

$$P_1 = -0.81529, \quad P_2 = 0.24996,$$

with

$$q^2 - 0.81529q + 0.24996 = 0.$$

* It should be noted that such fractional binomial series tend ultimately to become negative, although with negligibly small frequencies.
This gives imaginary values of \( q_1 \) and \( q_2 \) and thus the solution can for the present purpose be discarded.

(iii) \( n''' = -3.43390 \). We deduce
\[
-3.950 + 3.02614P_1 + 15.22557P_2 = 0, \\
-6.000 + 3.23317P_1 + 16.44372P_2 = 0,
\]
giving
\[
P_1 = -1.21441, \quad P_2 = -0.20229
\]
and
\[
q^2 + 1.21441q + 0.20229 = 0.
\]
Hence \( q_1 = -1.0151, \quad P_1 = 2.0151, \quad q_2 = -1.993, \quad P_2 = 1.993. \)

The \( \lambda \) equations are
\[
1 = \lambda_1 + \lambda_2, \quad -0.6845/3.4339 = -1.0151\lambda_1 - 1.993\lambda_2,
\]
leading to
\[
\lambda_1 = -0.00043, \quad \lambda_2 = -9.99957,
\]
or,
\[
v_1 = 0.0172, \quad v_2 = 399.9828.
\]

Thus the component series is
\[
0.0172 (2.0151 - 1.0151)^{-3.4339} + 399.9828 (1.993 - 1.993)^{3.4339},
\]
with means at
\[
m_1 = 3.4858 \quad \text{and} \quad m_2 = 0.6847.
\]

The first of these components is negligible, it contains roughly only \( 0.02 \) individuals in 400, and the second is sensibly identical with the negative binomial obtained by “Student,” i.e.
\[
400 (1.1893 - 1.1893)^{-3.6054},
\]
with slightly modified constants. It provides:

<table>
<thead>
<tr>
<th>No. of yeast cells</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated</td>
<td>214</td>
<td>122</td>
<td>45</td>
<td>14</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Observed</td>
<td>213</td>
<td>128</td>
<td>37</td>
<td>18</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

leading to \( \chi^2 = 3.12 \) and \( P = 0.68 \), which for all practical purposes is as good as the double Poisson.

Conclusions. It having been suggested that the appearance of negative binomials as better “fits” than Poisson’s series for material that is supposed to follow the law of small numbers is due to heterogeneity, formulae have been provided for testing whether this heterogeneity is due to a second component. If so this component should be small and the first component should substantially agree with the primary Poisson’s series. The smallness of the second component would measure the goodness of the technique in haemacytometer or opsonic index counts. Applied to “Student’s” first series of counts of yeast cells we obtain (a) two Poisson’s series neither of which dominates the data or approximates to the primary Poisson’s series; (b) two positive binomials, neither of which has any approach to a Poisson series or any agreement with the components of (a); and lastly (c) two negative binomials, one of which dominates the series, and agrees with the primary negative binomial. This investigation as far as it goes suggests either that “Student’s” first count is really described homogeneously by a negative binomial, or, if it be heterogeneous, then the heterogeneity is manifold, and no weight can be given to the results of fitting by the primary Poisson’s series.

The general numerical discussion by the formulae of this paper of a variety of data assumed to follow the “law of small numbers” is in hand and will shortly be published.