ON THE PROBABLE ERROR OF A COEFFICIENT OF CONTINGENCY WITHOUT APPROXIMATION.

BY ANDREW W. YOUNG, M.A. AND KARL PEARSON, F.R.S.

(1) Introductory.

There have been two memoirs dealing with the probable error of a coefficient of contingency, namely that by Blakeman and Pearson in 1906* and that by Pearson in 1914†. In the former paper the authors started from the expression for the mean square contingency

\[ \phi^2 = S \left( \frac{n_{ss'}}{n_{s}, n_{s'}} \right) - 1, \]

and varied \( n_{ss'}, n_s, \) and \( n_{s'} \) but neglected the squares and products of these variations. The result was lengthy, and the arithmetical work laborious. In 1914 Pearson gave reasons for considering \( n_s \) and \( n_{s'} \) as constant during the sampling and got a much simpler value for \( \sigma_\phi^2 \). The result in actual numerical cases did not differ widely from the much more elaborate formula of the earlier memoir. Recent work in other directions has, however, shown that caution must be used in neglecting the square and product terms of the variations due to random sampling, and the object of the present paper is to consider the variation of \( \phi^2 \) on the hypothesis of Pearson's 1914 note but without approximation.

Let a population of size \( M \) be grouped into \( c \) divisions—for example, the cells of a contingency table—and let the contents of the \( sth \) division be \( m_s \). Let a sample of size \( N \) be taken at random from the population and let \( n_s \) be the contents of the \( sth \) division according to the same grouping.

We shall here consider the variation of the quantity \( \phi^2 \) defined by

\[ 1 + \phi^2 = S \left( \frac{n_s^2}{N \lambda_s} \right) \]

(i)

where \( \lambda_s \) is a number connected with the \( sth \) division and is for the present restricted only by the condition

\[ S (\lambda_s) = N \]

(ii)

—a condition which enables us to write

\[ \phi^2 = S \left\{ \frac{(n_s - \lambda_s)^2}{N \lambda_s} \right\} \]

(iii)

as equivalent to (i).

† Biometrika, Vol. x. p. 570.
These undetermined numbers \( \lambda \) are thus in general of the nature of weights and may be chosen in a variety of ways. The most important particular case is that of the population being grouped in a contingency table with, say, two variates. The \( st \)th division will be, say, the cell \((u, v)\) and if \( \lambda \) be taken to be \( N \frac{m_u m_v}{M^2} \), \( m_u \) and \( m_v \) being as usual the marginal totals of the \( u \)th row and \( v \)th column of the population \( M \), \( \phi^2 \) will be the mean square contingency*. Other cases will be discussed later.

The object of the present paper is to investigate the variation of the quantity \( \phi^2 \) as determined from the samples of the population. We take the numbers \( \lambda \) to be a property of the whole population and accordingly to have no variation as long as the size \( N \) of the samples is constant. It is true that in most cases in practice there will be only one sample and that the values of the numbers \( \lambda \) will have to be deduced from that sample and will therefore deviate from the values which would be used if the sampled population were known. But what we are seeking is the variability of the samples on the understanding that the distribution of the whole population is definite although in practice we know only the approximation to that distribution which is given by our sample. If we had wanted the variability of the calculated values of \( \phi^2 \) deduced from a large number of random samples, then we should have taken into account the variation of the \( \lambda \)'s as well as of the \( n \)'s. In this lies the difference between the discussions in the two earlier papers of 1906 and 1914.

This investigation follows that of the second paper, but we shall here give the full expressions without approximation, i.e. without neglecting the square of \( \delta n \) as was done in 1914. It will appear from the numerical examples worked out later that this squared term makes a fairly great difference and, even if this were not so, it is always preferable to have such formulae in full in order to decide the legitimacy of neglecting any terms. This is especially the case in statistical theory where neglect of the later terms of a Taylor expansion often leads to false results.

(2) Mean Value of \( \phi^2 \).

Let \( \bar{\phi}^2 \) be the mean value of \( \phi^2 \) and let \( \bar{n} \) be the mean value of \( n \), i.e. the value which would be given by taking a very large number of samples. Then we can write

\[
\frac{\bar{n}}{N} = m,
\]

Also if we define \( \delta \phi^2 \) and \( \delta n \) by the equations

\[
\phi^2 = \phi^2 + \delta \phi^2,
\]

\[
n = \bar{n} + \delta n,
\]

we have

\[
1 + \bar{\phi}^2 + \delta \phi^2 = S \left( \frac{\bar{n}^2}{N \lambda} \right) + 2S \left( \frac{\bar{n} \delta n}{N \lambda} \right) + S \left( \frac{(\delta n)^2}{N \lambda} \right) \quad \ldots \ldots \ldots \ldots \ldots \ldots (iv).
\]

Sum all such equations for a large number of samples and divide by the number of samples. Then, since \( \text{Mean } \delta n_s = 0 \) and \( \text{Mean } \delta \sigma^2 = 0 \), we have

\[
1 + \bar{\sigma}^2 = S \left( \frac{\bar{\sigma}^2}{N \lambda_s} \right) + \text{Mean} \left( \frac{(\delta n_s)^2}{N \lambda_s} \right) \quad \text{...............(v).}
\]

The expression \( \text{Mean } (\delta n_s)^2 \) is typical of several which we have to use in what follows and it will be useful to state or prove all the needful formulae before proceeding further.

(3) **Formulae regarding Products of Deviations.**

The deviations \( \delta n_s \) arrange themselves according to a hypergeometrical series and the moment coefficients of this series are known to be*

\[
\begin{align*}
\mu_2 &= x_1 N pq \\
\mu_3 &= x_1 x_2 N pq (p - q) \\
\mu_4 &= x_1 N pq (3 x_2 N pq + x_4)
\end{align*}
\]

where

\[
\begin{align*}
x_1 &= 1 - \frac{N - 1}{M - 1} \\
x_2 &= 1 - \frac{2 (N - 1)}{M - 2} \\
x_3 &= \left( 1 - \frac{2}{N} \right) \left[ 1 - \frac{N - 1}{M - 2} \left( \frac{N - 10}{N - 2} + \frac{9}{M - 3} \right) \right] \\
x_4 &= 1 - 6 \frac{N - 1}{M - 2} \left( 1 - \frac{N - 2}{M - 3} \right)
\end{align*}
\]

and in the present case

\[
\begin{align*}
p &= \frac{m_s}{M} = \frac{\bar{n}_s}{N} \\
q &= 1 - \frac{m_s}{M} = 1 - \frac{\bar{n}_s}{N}
\end{align*}
\]

\[
\text{...............(vii).}
\]

When \( M \) is very large as compared with \( N \), as in the majority of cases in practice, we may write \( x_1 = x_2 = x_3 = x_4 = 1 \).

We can now make use of these formulae to derive the following:

(a) **Mean \( (\delta n_s)^2 \).** This is \( \mu_2 \) in the notation of (vi) and

\[
= x_1 \bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) \quad \text{.........................(a)}.
\]

(b) **Mean \( \delta n_s \delta n_{s'} \) where \( s \) and \( s' \) differ.** Suppose first that \( \delta n_s \) remains constant and investigate the mean of \( \delta n_{s'} \) for this constant value of \( \delta n_s \). Now the distribution

of \( n_s \) for \( n_s \) constant is clearly given by the sub-hypergeometrical series with 
\( N' = N - \bar{n}_s - \delta n_s \) as total population and 
\[
\begin{align*}
p' &= \frac{\bar{n}_s'}{N - \bar{n}_s}, \\
q' &= 1 - \frac{\bar{n}_s'}{N - \bar{n}_s},
\end{align*}
\]
so that the Mean \( \delta n_s' \) for \( \delta n_s \) constant is 
\[
(N - \bar{n}_s - \delta n_s) \left( \frac{\bar{n}_s'}{N - \bar{n}_s} - \bar{n}_s' \right) = -\delta \frac{N}{N}.
\]
Hence 
\[
\text{Mean } \delta n_s \delta n_s' = -\text{Mean } \delta n_s \cdot \frac{\delta n_s \bar{n}_s'}{N - \bar{n}_s}.
\]
\[
= -\text{Mean } \delta n_s^2 \cdot \frac{\bar{n}_s'}{N - \bar{n}_s},
\]
\[
= -\chi_1 \bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) \frac{\bar{n}_s'}{N - \bar{n}_s} \text{ from (a)}
\]
\[
= -\chi_1 \frac{\bar{n}_s \bar{n}_s'}{N} \text{..................................(b).}
\]

(c) Mean \( (\delta n_s)^3 \) Directly from (vi) and (vii) 
\[
\text{Mean } (\delta n_s)^3 = \chi_1 \chi_2 \bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) \left( 1 - \frac{2\bar{n}_s}{N} \right) \text{..................(c).}
\]

(d) Mean \( (\delta n_s)^2 \delta n_s' \). Using the process of double summation as in (b) we have 
\[
\text{Mean } (\delta n_s)^2 \delta n_s' = -\text{Mean } (\delta n_s)^2 \cdot \frac{\delta n_s \bar{n}_s'}{N - \bar{n}_s}.
\]
\[
= -\frac{\bar{n}_s'}{N - \bar{n}_s} \cdot \chi_1 \chi_2 \bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) \left( 1 - \frac{2\bar{n}_s}{N} \right) \text{ from (c)}
\]
\[
= -\chi_1 \chi_2 \frac{\bar{n}_s \bar{n}_s'}{N} \left( 1 - \frac{2\bar{n}_s}{N} \right) \text{.................................(d).}
\]

(e) Mean \( (\delta n_s)^4 \). As in (c) we have immediately 
\[
\text{Mean } (\delta n_s)^4 = \chi_1 \bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) \left( 3\bar{n}_s \left( 1 - \frac{\bar{n}_s}{N} \right) + \chi_4 \right) \text{.............(e).}
\]

(f) Mean \( (\delta n_s)^2 (\delta n_s')^2 \). We again use the double summation as in (b), but in this case the algebra is much more troublesome. From the constants of the sub-hypergeometrical as given in (b) we have 
\[
\chi_1' N' p' q' = \left( 1 - \frac{N - \bar{n}_s - \delta n_s - 1}{M - m_s - 1} \right) \left( N - \bar{n}_s - \delta n_s \right) \frac{\bar{n}_s'}{N - \bar{n}_s} \left( 1 - \frac{\bar{n}_s'}{N - \bar{n}_s} \right).
\]

* Since \( S(\delta n_s) \) must be zero, we can regard the Mean \( \delta n_s' \) for a given \( \delta n_s \) as being the result of a distribution of a deviate \(-\delta n_s\) distributed over all the divisions except the \( s' \)th. The portion due to the \( s' \)th is then \( \frac{n_s'}{N - \bar{n}_s} \times (\delta n_s) \), as obtained above.
But this is the Mean \((\delta'n_s)^2\), where \(\delta'n_s\) is measured from the mean of \(n_s\) in the case where \(\delta n_s\) is fixed and we must reduce to the general mean, i.e. where \(\delta n_s\) is not given, to obtain the mean value of \((\delta n_s)^2\) for constant \(\delta n_s\). This is done by adding the square of the difference between these means of \(n_s\), namely \(\frac{\overline{n_s} \delta n_s}{N - \overline{n_s}}\).

We thus obtain that the Mean \((\delta n_s)^2\) for constant \(\delta n_s\)

\[
(1 - \frac{N - \overline{n_s} - \delta n_s - 1}{M - m_s - 1}) (N - \overline{n_s} - \delta n_s) \frac{\overline{n'_s}}{N - \overline{n_s}} (1 - \frac{\overline{n'_s}}{N - \overline{n_s}}) + \frac{\overline{n'_s}^2 (\delta n_s)^2}{(N - \overline{n_s})^2} = \frac{N (1 - \frac{N}{M}) - (1 - \frac{2N}{M}) M N \delta n_s}{M (N - \overline{n_s}) - \frac{N^2 (\delta n_s)^2}{M^2 (N - \overline{n_s})^2}} + \frac{\overline{n'_s}^2 (\delta n_s)^2}{(N - \overline{n_s})^2},
\]

when we substitute \(M \overline{n'_s}/N\) for \(m_s\).

Thus we have to evaluate

\[
\text{Mean} (\delta n_s)^2 (\delta n_s)^2 = \frac{N}{M} (1 - \frac{N}{M}) \text{Mean} (\delta n_s)^2 - \frac{1 - 2N}{M} \frac{N}{N - \overline{n_s}} \text{Mean} (\delta n_s)^3
\]

\[
- \frac{N^2 \text{Mean} (\delta n_s)^4}{M^2 (N - \overline{n_s})^2} \frac{M^2 \overline{n'_s} (N - \overline{n_s} - \overline{n'_s})}{N^2 (M(N - \overline{n_s}) - 1)} + \frac{\overline{n'_s}^2 (\delta n_s)^2}{(N - \overline{n_s})^2} \text{Mean} (\delta n_s)^4.
\]

Substituting from \((a), (c)\) and \((e)\), we find

\[
\text{Mean} (\delta n_s)^2 (\delta n_s)^2 = \chi_1 \frac{\overline{n_s} \overline{n'_s}}{N} \frac{N}{M} (1 - \frac{N}{M}) \frac{M^2 (N - \overline{n_s}) (N - \overline{n_s} - \overline{n'_s})}{N^2 (M(N - \overline{n_s}) - 1)}
\]

\[
- \chi_2 \left(1 - \frac{2N}{M}\right) \frac{M (N - 2\overline{n_s}) (N - \overline{n_s} - \overline{n'_s})}{N^2 (M(N - \overline{n_s}) - 1)} - 3 \chi_3 \frac{\overline{n_s} (N - \overline{n_s} - \overline{n'_s})}{N(M(N - \overline{n_s}) - 1)}
\]

\[
- \chi_4 \frac{N - \overline{n_s} - \overline{n'_s}}{(N - \overline{n_s})(M(N - \overline{n_s}) - 1)} + 3 \chi_3 \frac{\overline{n_s} \overline{n'_s}}{N} + \chi_4 \frac{\overline{n'_s}}{N - \overline{n_s}}\right\}. \tag{f}
\]

This expression must be symmetrical in \(s\) and \(s'\) and this will be the case only if the quantities \(M/N(N - \overline{n_s}) - 1\) and \(N - \overline{n_s}\) in the denominator cancel with factors in the numerator. By taking the two terms in \(\chi_4\) together we get rid of the \(N - \overline{n_s}\) factor and after a laborious expansion in substituting the values of the \(\chi_s\)'s we reduce the whole expression to the comparatively simple form

\[
\chi_1 \frac{\overline{n_s} \overline{n'_s}}{N} \chi_3 N \left(1 + \frac{\overline{n_s} + \overline{n'_s}}{N} + \frac{3 \overline{n_s} \overline{n'_s}}{N^2}\right) + \chi_4 \right\} \tag{f}.
\]

This agrees with the value obtained by Isserlis from the differential equation to the hypergeometric series and thus confirms his result obtained by a totally different procedure.*

On the Probable Error of a Coefficient of Contingency

With these formulae we can now proceed to discuss the mean value of $\phi^2$ and its variability.

(4) Mean Value of $\phi^2$ (continued).

At the end of (2) we had arrived at the equation

$$1 + \bar{\phi}^2 = S \left( \frac{\bar{n}_s^2}{N \lambda_s} \right) + \text{Mean} \left( \frac{(\delta n_s)^2}{N \lambda_s} \right),$$

and substituting from (a) we now obtain

$$1 + \bar{\phi}^2 = S \left( \frac{\bar{n}_s^2}{N \lambda_s} \right) + x_1 \frac{1}{N} S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right) \quad (\text{viii}).$$

We can usually put $x_1 = 1$ and write

$$1 + \bar{\phi}^2 = \left( 1 - \frac{1}{N} \right) S \left( \frac{\bar{n}_s^2}{N \lambda_s} \right) + \frac{1}{N} S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right) \quad (\text{ix}).$$

In the particular case of the contingency table

$$1 + \bar{\phi}^2 = \left( 1 - \frac{1}{N} \right) S \left( \frac{\bar{n}_u^2 \bar{n}_v^2}{N \bar{n}_u \bar{n}_v} \right) + \frac{1}{N} S \left( \frac{N \bar{n}_u \bar{n}_v}{\bar{n}_u \bar{n}_v} \right) \quad (\text{x}).$$

Now $S \left( \frac{\bar{n}_u^2 \bar{n}_v^2}{\bar{n}_u \bar{n}_v} \right) = 1 + \phi_s^2$, where $\phi_s^2$ is the mean square contingency for the whole population, so that the mean value of $\phi^2$ as determined from a large number of samples is in excess of the true mean square contingency by

$$\frac{1}{N} \left( S \left( \frac{N \bar{n}_u \bar{n}_v}{\bar{n}_u \bar{n}_v} \right) - (1 + \phi_s^2) \right).$$


From the equations

$$1 + \bar{\phi}^2 + \delta \phi^2 = S \left( \frac{\bar{n}_s^2}{N \lambda_s} \right) + \frac{2}{N} S \left( \frac{\bar{n}_s \delta n_s}{\lambda_s} \right) + \frac{1}{N} S \left( \frac{(\delta n_s)^2}{\lambda_s} \right),$$

and

$$1 + \bar{\phi}^2 = S \left( \frac{\bar{n}_s^2}{N \lambda_s} \right) + x_1 \frac{1}{N} S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right),$$

we have

$$N \delta \phi^2 = S \left( \frac{(\delta n_s)^2}{\lambda_s} \right) + 2 S \left( \frac{\bar{n}_s \delta n_s}{\lambda_s} \right) - x_1 S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right) \quad (\text{xi}).$$

Squaring, summing for a large number of samples and dividing by the number of samples, we have

$$N^2 \sigma^2_{\phi^2} = \text{Mean} \left[ S \left( \frac{(\delta n_s)^2}{\lambda_s} \right)^2 + 4 \left( S \left( \frac{\bar{n}_s \delta n_s}{\lambda_s} \right)^2 \right) + x_1^2 \left( S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right)^2 \right) \right.$$}

$$+ 4 S \left( \frac{(\delta n_s)^2}{\lambda_s} \right) S \left( \frac{\bar{n}_s \delta n_s}{\lambda_s} \right) - 2 x_1 S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right) \cdot S \left( \frac{\delta n_s^2}{\lambda_s} \right) \right]$$
\[= \text{Mean} \left[ S \left( \frac{(\bar{n}_s)^2}{\lambda_s^2} \right) + 4 S \left( \frac{(\bar{n}_s)(\bar{n}_s')}{\lambda_s \lambda_s'} \right) + 4 S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) \right. \]
\[+ 4 S S \left( \frac{\bar{n}_s^2(\bar{n}_s' - \bar{n}_s)}{\lambda_s \lambda_s'} \right) + \left. 4 S \left( \frac{\bar{n}_s^3}{\lambda_s^2} \right) + 4 S S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) \right] \]
\[- \chi^2 \left( \frac{\bar{n}_s^2(\bar{n}_s - \bar{n}_s')}{\lambda_s^2} \right) - \chi^2 S S \left( \frac{(N - \bar{n}_s)(N - \bar{n}_s')}{\lambda_s \lambda_s'} \right) \]

where \( S \) denotes the same as \( S \), i.e. summation for all values of \( s \), and \( S S \) denotes \( s' \) for all values of \( s = s' \) followed by summation for all values of \( s^* \).

Substituting from the equations \( (a), (b), (c), (d), (e), (f) \), we have
\[ N^2 \sigma^2_{\varphi^i} = S \left\{ \frac{1}{\lambda_s^2} \chi_1 \bar{n}_s \left( \frac{N - \bar{n}_s}{N} \right) \left( 3X_3 \bar{n}_s \left( \frac{N - \bar{n}_s}{N} \right) + \chi_4 \right) \right\} \]
\[+ 4 S \left\{ \frac{\bar{n}_s^2}{\lambda_s^2} \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N} \right) \right\} \]
\[+ 4 S \left\{ \frac{\bar{n}_s^2}{\lambda_s^2} \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]
\[- \chi^2 S \left\{ \frac{\bar{n}_s^2}{\lambda_s^2} \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - \chi^2 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]
\[= S \left\{ \chi_1 \left( \frac{X_3 \bar{n}_s}{N} \right) + \left( \frac{3X_3 + 4X_2 - X_1}{N} \right) \bar{n}_s^2 + \left( \frac{4 - (6X_3 + 12X_2 - 2X_1)}{N^2} \right) \bar{n}_s^4 \right\} \]
\[+ 4 S \left\{ \bar{n}_s^2 \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]
\[+ 4 S \left\{ \frac{\bar{n}_s^2}{\lambda_s^2} \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]
\[= S \left\{ \chi_1 \left( \frac{X_3 \bar{n}_s}{N} \right) + \left( \frac{3X_3 + 4X_2 - X_1}{N} \right) \bar{n}_s^2 + \left( \frac{4 - (6X_3 + 12X_2 - 2X_1)}{N^2} \right) \bar{n}_s^4 \right\} \]
\[+ 4 S \left\{ \bar{n}_s^2 \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]
\[+ 4 S \left\{ \frac{\bar{n}_s^2}{\lambda_s^2} \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} - 4 S S \left\{ \bar{n}_s \bar{n}_s' \chi_1 \chi_2 \left( \frac{N - \bar{n}_s}{N^2} \right) \right\} \]

after expansion and rearrangement.

Now it is evident that in numerical work the double summation would involve much extra labour, but we can get rid of it by using the identities
\[ \left( S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) \right) = S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) + S S \left( \frac{\bar{n}_s \bar{n}_s'}{\lambda_s \lambda_s'} \right), \]
\[ S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) S \left( \frac{\bar{n}_s}{\lambda_s} \right) = S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) + S S \left( \frac{\bar{n}_s \bar{n}_s'}{\lambda_s \lambda_s'} \right) = S \left( \frac{\bar{n}_s^3}{\lambda_s} \right) + \frac{1}{2} S S \left( \frac{\bar{n}_s \bar{n}_s'}{\lambda_s \lambda_s'} \right), \]
\[ \left( S \left( \frac{\bar{n}_s^2}{\lambda_s^2} \right) \right)^2 = S \left( \frac{\bar{n}_s^4}{\lambda_s^4} \right) + S S \left( \frac{\bar{n}_s^2 \bar{n}_s'^2}{\lambda_s \lambda_s'} \right), \]

and so reducing all to single summations.

* As this notation may be somewhat unusual, it may be better to make it clear by taking a case with three variates only, for example:
\[ \left( S n_s \right)^2 = (n_1 + n_2 + n_3)^2 = n_1^2 + n_2^2 + n_3^2 + (n_2 + n_3) n_1 + (n_2 + n_3) n_2 + (n_1 + n_2) n_3 = S n_s^2 + S S n_s n_s'. \]
On the Probable Error of a Coefficient of Contingency

This leads to
\[ N^2 \sigma^2_{\phi^4} = x_1 x_4 S \left( \frac{\bar{n}_e}{\lambda_e^2} \right) + x_1 \left( x_3 - x_1 + \frac{x_4}{N} \right) \left\{ S \left( \frac{\bar{n}_e}{\lambda_e} \right) \right\}^2 \]
\[ + x_1 \left( 2x_3 + 4x_2 - \frac{2x_4}{N} \right) S \left( \frac{\bar{n}_e}{\lambda_e} \right) \]
\[ + x_1 \left( 2x_1 - \frac{2x_3 - 4x_2}{N} \right) S \left( \frac{\bar{n}_e}{\lambda_e} \right) \]
\[ + x_1 \left( \frac{4 - \left( 4x_3 + 8x_2 \right)}{N} \right) S \left( \frac{\bar{n}_e}{\lambda_e} \right) \]
\[ + x_1 \left( -\frac{4}{N} + (3x_3 + 8x_2 - x_1) \right) \left\{ S \left( \frac{\bar{n}_e}{\lambda_e} \right) \right\}^2 \]
...(xii).

(6) **Standard Deviation of \( \phi^4 \). Approximate Formulae.**

The result of the preceding section is an exact one since we have neglected no terms in arriving at it, but as mentioned before we can usually take \( M \) to be very large compared with \( N \) and make \( x_1 = x_2 = x_3 = x_4 = 1 \). With this simplification equation (xii) becomes

\[ \sigma^2_{\phi^4} = \frac{4}{N} \left[ S \left( \frac{\bar{n}_e^2}{N\lambda_e^2} \right) - \left\{ S \left( \frac{\bar{n}_e}{N\lambda_e} \right) \right\}^2 \right] \]
\[ + \frac{1}{N^2} \left[ 6S \left( \frac{\bar{n}_e^2}{\lambda_e^2} \right) - 4S \left( \frac{\bar{n}_e}{\lambda_e} \right) S \left( \frac{\bar{n}_e}{\lambda_e} \right) - 10 \left\{ S \left( \frac{\bar{n}_e^2}{N\lambda_e} \right) \right\}^2 - 12S \left( \frac{\bar{n}_e^3}{N\lambda_e^2} \right) \right] \]
\[ + \frac{1}{N^3} \left[ S \left( \frac{N\bar{n}_e^3}{\lambda_e^2} \right) + \left\{ S \left( \frac{\bar{n}_e}{\lambda_e} \right) \right\}^2 - 2S \left( \frac{\bar{n}_e^3}{\lambda_e^2} \right) \right] \] .................(xiii).

In the great majority of cases it will be impossible to make rigorous use of this formula since we have no other knowledge of the whole population than what is given by the sample. In particular the \( \bar{n} \)'s are usually unknown and we must simply make use of the approximations at our disposal, namely the \( n \)'s of the observed sample.

Again, it will usually happen that while \( \bar{n} \) may be fairly large \( \bar{n} = \lambda \) will be small and it will give formulae which are much more convenient for computation if we write \( \psi = \bar{n} - \lambda \) and substitute \( \psi + \lambda \) for \( \bar{n} \) in equation (xiii), remembering that \( S (\lambda) = N \).

After some reduction the formula becomes

\[ \sigma^2_{\phi^4} = \frac{4}{N} \left[ S \left( \frac{\psi^2}{N\lambda_e^2} \right) + \phi^2 - \phi^4 \right] \]
\[ + \frac{1}{N^2} \left[ 6S \left( \frac{\psi^2}{\lambda_e^2} \right) + (8 - 4\phi^2) S \left( \frac{\psi}{\lambda_e} \right) - 12S \left( \frac{\psi^3}{N\lambda_e^2} \right) + (2 - 4\phi^2)c - 22 - 56\phi^4 - 10\phi^4 \right] \]
\[ + \frac{1}{N^3} \left[ S \left( \frac{N\psi^3}{\lambda_e^2} \right) + S \left( \frac{\psi}{\lambda_e} \right) - 2S \left( \frac{\psi^2}{\lambda_e^2} \right) + 2 (c - 2) S \left( \frac{\psi}{\lambda_e^2} \right) + c (c - 2) \right] \] ...............(xiv),

where \( c \) is the number of classes or categories in the population in question.
(7) First Application. Contingency.

As mentioned in (1), if we regard the sth division as the \((u, v)\) cell of a contingency-table and if we take

\[
\lambda_s = N \frac{m_u m_v}{M^2} = \frac{\bar{n}_u \bar{n}_v}{N},
\]

then

\[
1 + \phi^2 = S \left( \frac{n_{uv}}{\bar{n}_u \bar{n}_v} \right),
\]

and

\[
\phi^2 = S \left( \frac{(n_{uv} - \bar{n}_u \bar{n}_v)^2}{\bar{n}_u \bar{n}_v} \right) = \text{Mean square contingency}
\]

Accordingly, with the notation

\[
\psi_s = n_{uv} - \lambda_s = n_{uv} - \frac{\bar{n}_u \bar{n}_v}{N},
\]

and

\[
\phi^2 = S \left( \frac{\psi_s^2}{N \lambda_s} \right) = S \left( \frac{(n_{uv} - \bar{n}_u \bar{n}_v)^2}{\bar{n}_u \bar{n}_v} \right),
\]

the equation (xiv) gives the standard deviation of the mean square contingency, when \(M\) is very large as compared with \(N\).

The terms enclosed in the first bracket of (xiv) are exactly those of Pearson’s 1914 paper in Biometrika, so that the second and third brackets contain the terms arising from the squares and higher products of \(\delta n_s\).

Of the total correction due to the higher approximation it is of interest to find how much is due to the change of mean and consequently the change of origin of \(\delta \phi^2\), when the square of \(\delta n_s\) is not neglected. The true mean is given by

\[
1 + \bar{\phi}^2 = \frac{1}{N} S \left( \frac{\bar{n}_s^2}{\lambda_s} \right) + \frac{1}{N} S \left( \frac{\bar{n}_s}{\lambda_s} \left( 1 - \frac{\bar{n}_s}{N} \right) \right),
\]

and using the observed values of \(n_s\) as the best approximation available for \(\bar{n}_s\)

\[
1 + \bar{\phi}^2 = (1 + \phi^2) \left( 1 - \frac{1}{N} \right) + \frac{1}{N} S \left( \frac{\psi_s + \lambda_s}{\lambda_s} \right)
\]

\[
= (1 + \phi^2) + \frac{1}{N} \left[ S \left( \frac{\psi_s}{\lambda_s} \right) - \phi^2 + c - 1 \right],
\]

so that the difference between the true mean and the approximate mean obtained by neglecting squares of \(\delta n_s\) is

\[
\bar{\phi}^2 - \phi^2 = \frac{1}{N} \left[ S \left( \frac{\psi_s}{\lambda_s} \right) - \phi^2 + c - 1 \right] .
\]

In accordance, then, with the formula for change of second moment with change of origin we get the effect of the change of mean on \(\sigma^2_s\), by subtracting

\[
\frac{1}{N^2} \left[ S \left( \frac{\psi_s}{\lambda_s} \right) - \phi^2 + c - 1 \right]^2
\]

from the approximate value.
On the Probable Error of a Coefficient of Contingency

In the examples given below it will be seen that this is only a small part of the total correction and thus the main part of the correction is due to the retention of the squares and products in the value of \((\delta \phi^2)^2\) used in (5).

(8) Numerical Illustrations.

I. Contingency between Handwriting and Intelligence in Girls.

The probable errors of the contingency constants in this table have been worked out both in the 1905 and in the 1914 papers and below is given a table showing the effect of the corrective terms of the present discussion.

The new summations required are found to be

\[
S \left( \frac{\psi_s}{\lambda_s} \right) = 12.788, \quad S \left( \frac{N}{\lambda_s} \right) = 15982, \\
S \left( \frac{\psi_s^2}{\lambda_s^2} \right) = 93.144, \quad S \left( \frac{N\psi_s}{\lambda_s^2} \right) = 57270,
\]

and in the following equation the numerical values of the various terms of equation (xiv) are given in the same order as their corresponding algebraic terms:

\[
\sigma^2_\phi = \frac{4}{1801} [1.4865 + 0.0580 - 0.0918] \\
+ \frac{1}{(1801)^2} [558.864 + 97.404 - 1.784 + 58.205 - 22 - 5.365 - 0.092] \\
+ \frac{1}{(1801)^2} [57270 + 15982 + 164 - 186 + 870 + 1224].
\]

The other calculations are summarised in the table below:

**TABLE I.**

\[
\phi^2 = 0.09580, \quad C_2 = \sqrt{\frac{\phi^2}{1 + \phi^2}} = 0.2957.
\]

<table>
<thead>
<tr>
<th></th>
<th>Blakeman and Pearson (1905)</th>
<th>1st Term of (xiv) or Pearson (1914)</th>
<th>1st and 2nd Terms of (xiv)</th>
<th>All Terms of (xiv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma_{\phi}^2)</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Probable error of (\phi^2)</td>
<td>-0.02023</td>
<td>-0.02286</td>
<td>-0.02709</td>
<td>-0.02729</td>
</tr>
<tr>
<td>(\sigma_C^2)</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Probable error of (C)</td>
<td>-0.0192</td>
<td>-0.02171</td>
<td>-0.02573</td>
<td>-0.02593</td>
</tr>
</tbody>
</table>

In this table the work has been carried out to four significant figures with a view to showing the corrective effects of the various terms. It is apparent that the fineness of approximation given by (xiv) in full is more than is required in practice,

* Incorrectly given as -0.0042 in Blakeman and Pearson's paper, *loc cit.* footnote p. 196.
but there is a considerable difference between the values given according as the second term is used or not and it seems that in some cases it would be advisable to calculate this term or at least the most important terms in it, viz.

\[ \frac{1}{N^2} \left( 6s \left( \frac{\psi_s^2}{\lambda_s^2} \right) + 8s \left( \frac{\psi_s}{\lambda_s} \right) + 2c \right) \] ...

(xv).

Using this approximation in the above case, we obtain \( \sigma^2 \phi = 0.02730 \), a result which—by a mere chance, of course—is almost exactly that given by the full expression in (xiv).

II. Contingency between the Hair-colours of Pairs of Female Cousins.

In the example just given the total number \( N \) in the sample was fairly large, viz. 1801, and it might be expected that in smaller samples the corrective terms would be of increased importance. With a view to testing this a contingency-table given by Miss Elderton in her Memoir on "The Measure of the Resemblance of First Cousins"* was selected. There are 36 cells in this table and the total number in the sample is only 218, there being several cells with zero or very small content.

The Table is given in full on p. 226 along with the quantities required for the calculation of \( \phi^2 \) and \( \sigma^2 \); it is there evident from the figures how large a proportion of the variability depends on the cells of small content. This is of course to be expected but the importance of having large numbers in all the cells is not always appreciated. In this particular case physiological reasons would prevent us from clubbing together the "Fairs" and "Reds" and with the fewness of the observations at our disposal we must use the table simply as it stands.

The scheme followed in each cell of the table is shown in the last column, and in the marginal totals are given the values of all the summations required for (xiv). These are

\[ \phi^2 = S \left( \frac{\psi_s^2}{N\lambda_s} \right) = 0.14895, \]

\[ S \left( \frac{\psi_s}{N\lambda_s} \right) = -1.8481, \]

\[ S \left( \frac{N\psi_s}{\lambda_s^2} \right) = 19.162, \]

\[ S \left( \frac{N}{\lambda_s} \right) = 5170.9, \]

\[ S \left( \frac{\psi_s^3}{N\lambda_s^2} \right) = 0.08277. \]

When these are substituted in equation (xiv), we have, preserving the algebraic order as before,

\[ \sigma^2 \phi = \frac{4}{218} \left[ 0.08277 + 0.14895 - 0.02218 \right] \]

\[ + \frac{1}{(218)^2} \left[ 114.9714 - 13.6837 - 0.9932 + 50.5515 - 22 - 8.3411 - 0.2219 \right] \]

\[ + \frac{1}{(218)^3} \left[ -2336.8 + 5170.9 + 3.4 - 38.3 + 125.7 + 1224 \right], \]

and the whole work is again summarised in Table III (p. 227):

### TABLE II. Contingency between the Hair-colours of Female Cousins

<table>
<thead>
<tr>
<th>Tints</th>
<th>Very Dark</th>
<th>Dark Brown</th>
<th>Brown</th>
<th>Light Brown</th>
<th>Fair</th>
<th>Red</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>11</td>
<td>3:5</td>
<td>11</td>
<td>1:5</td>
<td>0</td>
<td>36</td>
<td>= ( S_{n_p} )</td>
</tr>
<tr>
<td>36:67</td>
<td>23:57</td>
<td>33:42</td>
<td>22:96</td>
<td>55:00</td>
<td>263:93</td>
<td>1-2230= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
<tr>
<td>-5139</td>
<td>-1894</td>
<td>-4634</td>
<td>-1585</td>
<td>-6214</td>
<td>-263:93</td>
<td>-1-0000= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
<tr>
<td>-2041</td>
<td>-9359</td>
<td>-3933</td>
<td>-1043</td>
<td>-3861</td>
<td>-1-0000</td>
<td>1-0000= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
<tr>
<td>18:84</td>
<td>4:46</td>
<td>15:49</td>
<td>3:64</td>
<td>3:41</td>
<td>-263:93</td>
<td>-1-2230= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
<tr>
<td>-0.0720</td>
<td>-0.00152</td>
<td>-0.0642</td>
<td>-0.0109</td>
<td>-0.0072</td>
<td>-0.00379</td>
<td>-0.02705= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
<tr>
<td>-0.0370</td>
<td>-0.00029</td>
<td>-0.00289</td>
<td>-0.0017</td>
<td>-0.00436</td>
<td>-0.00379</td>
<td>-0.00697= ( S \left( \psi_i/\lambda_i \right) )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11</th>
<th>13</th>
<th>12</th>
<th>10</th>
<th>9</th>
<th>1</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5185</td>
<td>-3300</td>
<td>-1043</td>
<td>-5044</td>
<td>-2129</td>
<td>-5855</td>
<td>-534:72</td>
</tr>
<tr>
<td>-5044</td>
<td>-9717</td>
<td>-134</td>
<td>-134</td>
<td>-1267</td>
<td>-556:07</td>
<td>-001702</td>
</tr>
<tr>
<td>-0.00152</td>
<td>-0.00055</td>
<td>-0.00046</td>
<td>-0.00020</td>
<td>-0.00127</td>
<td>-0.02220</td>
<td>-0.00006</td>
</tr>
<tr>
<td>3:5</td>
<td>12</td>
<td>8</td>
<td>9:75</td>
<td>3:25</td>
<td>3</td>
<td>39:5</td>
</tr>
<tr>
<td>6:523</td>
<td>10:147</td>
<td>7:157</td>
<td>10:419</td>
<td>4:349</td>
<td>0:906</td>
<td>397:00</td>
</tr>
<tr>
<td>-4634</td>
<td>-1826</td>
<td>-1178</td>
<td>-1578</td>
<td>-642</td>
<td>-2527</td>
<td>2-3111</td>
</tr>
<tr>
<td>-2147</td>
<td>-3333</td>
<td>-1139</td>
<td>-1548</td>
<td>-1043</td>
<td>-534:72</td>
<td>5-6711</td>
</tr>
<tr>
<td>-0.00642</td>
<td>-0.00155</td>
<td>-0.00046</td>
<td>-0.00020</td>
<td>-0.00127</td>
<td>-0.02220</td>
<td>-0.001702</td>
</tr>
<tr>
<td>-0.00298</td>
<td>-0.00029</td>
<td>-0.00005</td>
<td>-0.00001</td>
<td>-0.0005</td>
<td>-0.00006</td>
<td>-0.00006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9</th>
<th>10</th>
<th>9:75</th>
<th>16:5</th>
<th>3:25</th>
<th>10</th>
<th>57:5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5185</td>
<td>-3300</td>
<td>-1043</td>
<td>-5044</td>
<td>-2129</td>
<td>-5855</td>
<td>-534:72</td>
</tr>
<tr>
<td>-5044</td>
<td>-9717</td>
<td>-134</td>
<td>-134</td>
<td>-1267</td>
<td>-556:07</td>
<td>-00083</td>
</tr>
<tr>
<td>-0.00109</td>
<td>-0.00070</td>
<td>-0.00020</td>
<td>-0.00054</td>
<td>-0.00127</td>
<td>-0.02220</td>
<td>-0.00006</td>
</tr>
<tr>
<td>-0.00017</td>
<td>-0.00228</td>
<td>-0.00001</td>
<td>-0.00005</td>
<td>-0.00025</td>
<td>-0.00009</td>
<td>-0.00009</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1:5</th>
<th>9</th>
<th>3:25</th>
<th>9:25</th>
<th>1</th>
<th>0</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>55:00</td>
<td>35:36</td>
<td>50:12</td>
<td>34:44</td>
<td>82:51</td>
<td>396:38</td>
<td>1-5744</td>
</tr>
<tr>
<td>-6214</td>
<td>-4598</td>
<td>-2527</td>
<td>-4614</td>
<td>-6215</td>
<td>-1-0000</td>
<td>1-0000</td>
</tr>
<tr>
<td>-3861</td>
<td>-2114</td>
<td>-6939</td>
<td>-2129</td>
<td>3863</td>
<td>1-0000</td>
<td>1-0000</td>
</tr>
<tr>
<td>-0.00702</td>
<td>-0.00598</td>
<td>-0.00127</td>
<td>-0.00618</td>
<td>-0.00468</td>
<td>-0.00252</td>
<td>-0.00452</td>
</tr>
<tr>
<td>-0.00436</td>
<td>-0.00027</td>
<td>-0.00032</td>
<td>-0.00028</td>
<td>-0.000291</td>
<td>-0.00252</td>
<td>-0.000431</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>3</th>
<th>0</th>
<th>0</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1-0000</td>
<td>-2212</td>
<td>2:3111</td>
<td>-2419</td>
<td>-1-0000</td>
<td>-1-0000</td>
</tr>
<tr>
<td>1-0000</td>
<td>-4893</td>
<td>-5:3142</td>
<td>-685</td>
<td>1-0000</td>
<td>1-0000</td>
</tr>
<tr>
<td>-0.00379</td>
<td>-0.00029</td>
<td>-0.0220</td>
<td>-0.0035</td>
<td>-0.00252</td>
<td>-0.00053</td>
</tr>
<tr>
<td>-0.00379</td>
<td>-0.00006</td>
<td>-0.05131</td>
<td>-0.0009</td>
<td>-0.00262</td>
<td>-0.00053</td>
</tr>
</tbody>
</table>

Totals

Marginal Totals are the same as for vertical margin

\( N = 218 \)

\( \phi^2 = 14895 \)

\( S(\lambda_i) = 5170.9; \ S(\psi_i/\lambda_i) = -1.8481; \ S(\psi_i/\lambda_i)^2 = 19.162; \ S(\psi_i/\lambda_i)^2 = -2336.8; \ S(\psi_i/\lambda_i)^2 = -0.08277. \)
TABLE III.

\[ \phi^2 = 0.14895, \quad C_2 = \sqrt{\frac{\phi^2}{1 + \phi^2}} = 0.36005. \]

<table>
<thead>
<tr>
<th>Various formulae used</th>
<th>1st Term of equation (xiv)</th>
<th>1st and 2nd Terms of equation (xiv)</th>
<th>All Terms of equation (xiv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_2 )</td>
<td>0.062006</td>
<td>0.079848</td>
<td>0.082317</td>
</tr>
<tr>
<td>Probable error of ( \phi^2 )</td>
<td>0.041822</td>
<td>0.053857</td>
<td>0.055522</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.16902</td>
<td>0.21765</td>
<td>0.22438</td>
</tr>
<tr>
<td>Probable error of ( C )</td>
<td>0.11400</td>
<td>0.14680</td>
<td>0.15134</td>
</tr>
</tbody>
</table>

The relative importance of the three terms of equation (xiv) is not markedly different in this table of small total content from what it was in the case of the Handwriting-Intelligence table and we cannot base different conclusions on the two cases.

Again, using the approximation given by selecting the large terms from the second bracket of (xiv), viz.

\[ 6S (\frac{\psi_2^2}{\lambda_s^2}) + 8S (\frac{\psi_2}{\lambda_s}) + 2c, \]

we obtain

\[ \sigma_2 = 0.0864, \]

which as in the previous example is a reasonable approximation to the full expression result.

(9) Second Application. Test for Zero Contingency.

Suppose that we may expect in the sampled population an absence of contingency or correlation between the variates considered. In other words we will suppose

\[ \lambda_s = \bar{n}_u \bar{n}_v = \bar{n}_s. \]

If, however, we take a sample from that population, the quantity

\[ \phi^2 = S \frac{(n_s - \bar{n}_s)^2}{N \bar{n}_s}, \]

which is the mean square contingency in the case of a population with zero correlation, would certainly not vanish. The problem then arises: How great may the quantity \( \phi^2 \) be without making it highly improbable that the sample in question is really a sample from a population of uncorrelated material?

First of all, the mean value of \( \phi^2 \) as determined from a large number of samples would be

\[ \bar{\phi}^2 = \chi_1 \frac{1}{N} (c - 1) \]
On the Probable Error of a Coefficient of Contingency

as is obtained by substitution of \( \lambda = \tilde{n}_z \) in (viii), or, if \( \chi_1 = 1 \),

\[
\tilde{\phi}^2 = \frac{1}{N} (c - 1).
\]

In the same way we derive from equation (xiii)

\[
\sigma_{\phi^2}^2 = \frac{X_1}{N^2} \left( \frac{c}{H} + \frac{c(c - 2)}{N - 1} \right) + \frac{X_3 (c^2 - 1) - X_1 (c - 1)^2}{N^2} \quad \text{.........(xvii)},
\]

where \( H \) is the harmonic mean of the mean cell contents, or for the usual particular case when \( M \) is very large as compared with \( N \)

\[
\sigma_{\phi^2}^2 = \frac{1}{N^2} \left\{ \frac{c}{H} + \frac{c(c - 2)}{N} + 2 (c - 1) \right\} \quad \text{.................(xviii)},
\]

an expression which is very easily calculated especially as \( \frac{1}{H} \) will usually be small compared with \( c \) and hence a good rough approximation for a fairly large table will be got from

\[
\sigma_{\phi^2}^2 = \frac{2c}{N^2} \quad \text{.........................(xix)}.
\]

Thus if we take twice the standard deviation as a limit to the probability of a deviation being that of a random sample, we have as a rough upper limit to the value, which \( \phi^2 \) may be expected to take in any sample,

\[
\frac{1}{N} (c - 1) + 2 \frac{\sqrt{2c}}{N}
\]

(10) Numerical Illustration.

In the example of the Contingency-table for Handwriting and Intelligence in Girls

\[
\tilde{\phi}^2 = \frac{1}{N} (c - 1) = 0.01943,
\]

and when calculated from the more exact formula (xviii)

\[
\sigma_{\phi^2} = 0.004879,
\]

the approximation given by (xix) being 0.0046.

Hence in accordance with our assertion above, we should regard any observed value of \( \phi^2 \) which exceeds 0.01943 + 2 \times 0.00488, i.e. 0.02919 or, say, 0.03, as being incompatible with zero contingency. The observed value of \( \phi^2 = 0.0958 \).

The corresponding mean value of \( C \)—the coefficient of contingency—is

\[
\sqrt{\frac{0.01943}{0.01943}} = 1.3806,
\]

and the upper limit for \( C \) according to our assertion is 1.684 or, say, 1.7. The observed value of \( C \) is 1.2957. Clearly there is definite association between intelligence and handwriting.
Summary of Formulae.

It will be convenient for purposes of reference to have all the formulae collected into one section.

**General Formulae.**

For \( \phi^2 \) defined by

\[
1 + \phi^2 = S \left( \frac{n_s}{N \lambda_s} \right)
\]

or

\[
\phi^2 = S \left( \frac{(n_s - \lambda_s)^2}{N \lambda_s} \right)
\]

where \( N \) is the number in a sample, \( n_s \) is the number in the \( s \)th division of that sample and \( \lambda_s \) is a number connected with the \( s \)th division satisfying the condition \( S (\lambda_s) = S (n_s) = N \), we have proved that for an "infinite" sampled population

\[
1 + \phi^2 = \left( 1 - \frac{1}{N} \right) S \left( \frac{n_s^2}{N \lambda_s} \right) + \frac{1}{N} S \left( \frac{n_s}{\lambda_s} \right) \]

.................................\( (A) \),

and

\[
\sigma^2_{\phi^2} = \frac{4}{N} \left[ S \left( \frac{\psi_s^2}{N \lambda_s^2} \right) + \phi^2 - \phi^4 \right]
\]

\[
+ \frac{1}{N^2} \left[ 6S \left( \frac{\psi_s^2}{\lambda_s^2} \right) + (8 - 4\phi^2) S \left( \frac{\psi_s}{\lambda_s} \right) - 12S \left( \frac{\psi_s^3}{N \lambda_s^2} \right) + (2 - 4\phi^2) c - 22 - 56\phi^2 - 10\phi^4 \right]
\]

\[
+ \frac{1}{N^3} \left[ S \left( \frac{N \psi_s}{\lambda_s^2} \right) + S \left( \frac{\lambda_s}{\lambda_s} \right) \right] - 2S \left( \frac{\psi_s^2}{\lambda_s^2} \right) + 2 (c - 2) S \left( \frac{\psi_s}{\lambda_s} \right) + c (c - 2)
\]

.................................\( (C) \),

where \( c \) is the number of divisions or cells and \( \psi_s = n_s - \lambda_s \), or, with a fair amount of approximation,

\[
\sigma^2_{\phi^2} = \frac{4}{N} \left[ S \left( \frac{\psi_s^3}{N \lambda_s^2} \right) + \phi^2 - \phi^4 \right] + \frac{1}{N^2} \left[ 6S \left( \frac{\psi_s^2}{\lambda_s^2} \right) + 8S \left( \frac{\psi_s}{\lambda_s} \right) + 2c \right] \]

.................................\( (D) \).

**Contingency.**

In the case of a contingency-table the \( s \)th division may be taken to be the \((u, v)\) cell and \( \lambda_s = \frac{n_u n_v}{N} \lambda_s \), where \( n_u \) and \( n_v \) are the marginal totals of the \( u \)th row and the \( v \)th column. Formulae (A), (B), (C), (D) are then directly applicable.

**Test for zero contingency.**

When there is zero contingency in the total population

\[
\lambda_s = \frac{n_u n_v}{N} = \bar{n}_s,
\]

and (B) reduces to

\[
\bar{\phi}^2 = \frac{1}{N} (c - 1) \]

.................................\( (B') \),

* As usual the bar over a letter denotes "the mean value of." It is to be noted that usually there is only one sample and the value of \( n_s \) in that sample has to be taken as \( \bar{n}_s \).
On the Probable Error of a Coefficient of Contingency

and (C) to

\[ \sigma^2_{\phi^2} = \frac{1}{N^2} \left\{ \frac{c}{H} + \frac{c(c - 2)}{N} + 2(c - 1) \right\} \]

where \( c \) is the number of cells in the table and \( H \) is the harmonic mean of the cell contents.

Rough approximations in the case of zero contingency are given by

\[ \tilde{\phi}^2 = \frac{c}{N}, \]

and

\[ \sigma^2_{\phi^2} = \frac{2c}{N^2}, \]

and from these we can derive as a rough upper limit to the value of \( \phi^2 \) given by a random sample from a population of zero contingency

\[ \frac{c}{N} + 2 \frac{\sqrt{2c}}{N} = \frac{c}{N} \left( 1 + \frac{2.8284}{\sqrt{c}} \right). \]