ON INDICES OF DISPERSION

By Paul G. Hoel

University of California, Los Angeles

1. Introduction. In biological sciences the index of dispersion for the binomial and Poisson distributions is very useful for testing homogeneity of certain types of data. For example, the dilution technique in making blood counts finds it useful. Recently there have been attempts to use it to determine allergies by observing the change in the blood count after allergic foods have been taken. Here the sample may consist of only a few readings; consequently it is important to know how accurate this index is when applied to small samples. After inspecting the application of the Poisson index to such counts, I was surprised to see the lack of agreement with theory. At first it appeared that the fault lay with the chi-square approximation which is used on this index, but later it was clear that the assumption of a basic Poisson distribution was at fault. It now appears that statisticians will need to be careful about citing blood counts as examples of data following a Poisson distribution.

This paper is the result of investigating the accuracy of the chi-square approximation for the distribution of these indices. Previous work on this problem seems to have consisted in some sampling experiments [1] for small values of the parameters involved, and in some theoretical work [2] in which the sampling distribution is considered only for a fixed sample mean. Although sampling distributions ordinarily differ very little from the distributions obtained by assuming the mean of the sample fixed, for small degrees of freedom the difference may be appreciable and therefore requires investigation. In this paper the accuracy of the chi-square approximation is investigated by finding expressions for the descriptive moments of the distribution which are correct to terms of order N^{-3}. These expressions are obtained by means of Fisher's semi-invariant technique.

2. Moments of the distribution. Employing Fisher's notation [3], let the binomial index of dispersion be denoted by z, then z may be written as:

$$z = \frac{\Sigma(x-\bar{x})^2}{\bar{x}\left(1-\frac{\bar{x}}{n}\right)} = \frac{(N-1)k_2}{k_1\left(1-\frac{k_1}{n}\right)} = \frac{N-1}{\kappa_1\left(1-\frac{\kappa_1}{n}\right)}\,\frac{k_2}{\left(1+\frac{k_1-\kappa_1}{\kappa_1}\right)\left(1-\frac{k_1-\kappa_1}{n-\kappa_1}\right)}.$$

Letting $w = k_1 - \kappa_1$, $y = k_2$, $a = n - \kappa_1$, $b = \dfrac{N-1}{\kappa_1\left(1-\frac{\kappa_1}{n}\right)}$, z may be expanded as follows:

$$\begin{aligned} z &= \frac{by}{\left(1 + \frac{w}{\kappa_1}\right)\left(1 - \frac{w}{a}\right)} \\ &= by\left\{1 - \frac{w}{\kappa_1} + \frac{w^2}{\kappa_1^2} - \cdots\right\}\left\{1 + \frac{w}{a} + \frac{w^2}{a^2} + \cdots\right\} \\ &= by\left\{1 + w\left(\frac{1}{a} - \frac{1}{\kappa_1}\right) + w^2\left(\frac{1}{a^2} - \frac{1}{a\kappa_1} + \frac{1}{\kappa_1^2}\right) + \cdots\right\} \\ &= b\{y + c_1 wy + c_2 w^2 y + c_3 w^3 y + \cdots\}, \end{aligned}$$

where the definition of c_i is obvious. As will be seen later, these expansions are valid for obtaining the expected values of powers of z; hence

$$\begin{aligned} E(z) &= b\ \{\mu_{01} + c_1\mu_{11} + c_2\mu_{21} + \cdots\} \\ E(z^2) &= b^2\ \{\mu_{02} + 2c_1\mu_{12} + (2c_2 + c_1^2)\mu_{22} + (2c_3 + 2c_2c_1)\mu_{32} + \cdots\} \\ E(z^3) &= b^3\ \{\mu_{03} + 3c_1\mu_{13} + (3c_2 + 3c_1^2)\mu_{23} + (3c_3 + 6c_2c_1 + c_1^3)\mu_{33} + \cdots\} \\ E(z^4) &= b^4\ \{\mu_{04} + 4c_1\mu_{14} + (4c_2 + 6c_1^2)\mu_{24} + (4c_3 + 12c_2c_1 + 4c_1^3)\mu_{34} + \cdots\}. \end{aligned} \tag{1}$$

Since only the first four moments of z are to be found, it will be necessary to evaluate the μ_{ij} for $j = 1, 2, 3, 4$ and for $i = 0, 1, 2, \cdots$ as far as necessary to give the desired degree of accuracy.

First consider the relation between the moments μ_{ij} and the semi-invariants κ_{ij} which are defined in terms of the μ_{ij} by the following formal identity in t and τ.

$$e^{\frac{\kappa_{10}t + \kappa_{01}\tau}{1!} + \frac{\kappa_{20}t^2 + 2\kappa_{11}t\tau + \kappa_{02}\tau^2}{2!} + \cdots} = 1 + \frac{\mu_{10}t + \mu_{01}\tau}{1!} + \frac{\mu_{20}t^2 + 2\mu_{11}t\tau + \mu_{02}\tau^2}{2!} + \cdots.$$

Differentiating both sides with respect to t and replacing the exponential factor by the right member gives an identity which is convenient for evaluating the μ_{i0}. Differentiating both sides with respect to τ and making the same replacement gives an identity which is convenient for evaluating the μ_{ij} for $j > 0$.

These identities express μ_{ij} as a sum of products of κ's and μ's, each such product being of total degree i and j in its subscripts. By repeated substitution, μ_{ij} can be expressed as a sum of products of κ's only. From Fisher's formulas each such semi-invariant, κ_{rs}, can be expressed as a sum of products of semi-invariants of the basic distribution, each term of which sum is of order $N^{-(r+s-1)}$ in N. Hence it follows that the lowest order term, or at least one of the lowest order terms, in N in the expression for μ_{ij} will be a term with the maximum number of κ factors. Since the κ_{rs} of lowest degree in subscripts are κ_{10} and κ_{01}, the term with the maximum number of κ factors will be the term in $\kappa_{10}^{i}\kappa_{01}^{j}$. However, since $w = k_1 - \kappa_1$ has a zero mean value, $\mu_{10} = \kappa_{10} = 0$; consequently the lowest degree term involving the subscript $i > 0$ is κ_{20} or κ_{11}. As a result, the maximum number of κ factors will be found in the term containing $\kappa_{20}^{\frac{1}{2}i}\kappa_{01}^{j}$ fòr i even and $\kappa_{20}^{\frac{1}{2}(i-1)}\kappa_{01}^{j-1}\kappa_{11}$ for i odd. These terms are of order $N^{-\frac{1}{2}i}$ and $N^{-\frac{1}{2}(i+1)}$ respectively. Since it is desired to obtain accuracy of order N^{-3}, it therefore will suffice to evaluate μ_{ij} for $i \leq 6$.

The validity of the expansions used in arriving at (1) could now be shown by writing them as partial sums with remainder terms and then showing that the remainder terms are of higher order than N^{-3}.

Neglecting terms of higher order than N^{-3}, the above identities give the following expressions for μ_{ij} for $j = 0, 1, 2$ and $i = 0, 1, \cdots, 6$, with slightly longer expressions for $j = 3$ and 4.

$$\begin{array}{ll}
\mu_{10} = 0 & \mu_{01} = \kappa_{01} \\
\mu_{20} = \kappa_{20} & \mu_{11} = \kappa_{11} \\
\mu_{30} = \kappa_{30} & \mu_{21} = \kappa_{21} + \kappa_{01}\mu_{20} \\
\mu_{40} = \kappa_{40} + 3\kappa_{20}\mu_{20} & \mu_{31} = \kappa_{31} + 3\kappa_{11}\mu_{20} + \kappa_{01}\mu_{30} \\
\mu_{50} = 6\kappa_{30}\mu_{20} + 4\kappa_{20}\mu_{30} & \mu_{41} = 6\kappa_{21}\mu_{20} + 4\kappa_{11}\mu_{30} + \kappa_{01}\mu_{40} \\
\mu_{60} = 5\kappa_{20}\mu_{40} & \mu_{51} = 5\kappa_{11}\mu_{40} + \kappa_{01}\mu_{50} \\
 & \mu_{61} = \kappa_{01}\mu_{60}
\end{array}$$

$$\begin{aligned}
\mu_{02} &= \kappa_{02} + \kappa_{01}\mu_{01} \\
\mu_{12} &= \kappa_{12} + \kappa_{11}\mu_{01} + \kappa_{01}\mu_{11} \\
\mu_{22} &= \kappa_{22} + \kappa_{21}\mu_{01} + \kappa_{02}\mu_{20} + 2\kappa_{11}\mu_{11} + \kappa_{01}\mu_{21} \\
\mu_{32} &= \kappa_{31}\mu_{01} + 3\kappa_{12}\mu_{20} + 3\kappa_{21}\mu_{11} + \kappa_{02}\mu_{30} + 3\kappa_{11}\mu_{21} + \kappa_{01}\mu_{31} \\
\mu_{42} &= 6\kappa_{21}\mu_{21} + \kappa_{02}\mu_{40} + 4\kappa_{11}\mu_{31} + \kappa_{01}\mu_{41} \\
\mu_{52} &= 5\kappa_{11}\mu_{41} + \kappa_{01}\mu_{51} \\
\mu_{62} &= \kappa_{01}\mu_{61} .
\end{aligned}$$

The next step is to apply Fisher's formulas expressing the κ_{rs} in terms of the semi-invariants of the basic variable distribution, which in this case is the binomial distribution. In Fisher's notation κ_{rs} would be written as $\kappa(1^r2^s)$, since the variables w and y are respectively k_1, measured from its expected value, and k_2. Applying such formulas, the following expressions for the μ_{i1} and μ_{i2} are obtained, with somewhat longer expressions for the μ_{i3} and μ_{i4}.

$$\mu_{01} = \kappa_2, \qquad \mu_{11} = \frac{\kappa_3}{N}, \qquad \mu_{21} = \frac{\kappa_4}{N^2} + \frac{\kappa_2^2}{N},$$

$$\mu_{31} = \frac{\kappa_5}{N^3} + \frac{4\kappa_3\kappa_2}{N^2}, \qquad \mu_{41} = \frac{7\kappa_4\kappa_2}{N^3} + \frac{4\kappa_3^2}{N^3} + \frac{3\kappa_2^3}{N^2},$$

$$\mu_{51} = \frac{25\kappa_3\kappa_2^2}{N^3}, \qquad \mu_{61} = \frac{15\kappa_2^4}{N^3}$$

$$\begin{aligned}
\mu_{02} &= \frac{\kappa_4}{N} + \kappa_2^2\left[\frac{2}{N-1} + 1\right] \\
\mu_{12} &= \frac{\kappa_5}{N^2} + \frac{2\kappa_3\kappa_2}{N}\left[\frac{2}{N-1} + 1\right] \qquad (2) \\
\mu_{22} &= \frac{\kappa_6}{N^3} + \frac{\kappa_4\kappa_2}{N^2}\left[\frac{4}{N-1} + 3\right] + \frac{2\kappa_3^2}{N^2}\left[\frac{2}{N-1} + 1\right] + \frac{\kappa_2^3}{N}\left[\frac{2}{N-1} + 1\right] \\
\mu_{32} &= \frac{5\kappa_5\kappa_2}{N^3} + \frac{7\kappa_4\kappa_3}{N^3} + \frac{7\kappa_3\kappa_2^2}{N^2}\left[\frac{2}{N-1} + 1\right] \\
\mu_{42} &= \frac{16\kappa_4\kappa_2^2}{N^3} + \frac{20\kappa_3^2\kappa_2}{N^3} + \frac{3\kappa_2^4}{N^2}\left[\frac{2}{N-1} + 1\right] \\
\mu_{52} &= \frac{40\kappa_3\kappa_2^3}{N^3} \\
\mu_{62} &= \frac{15\kappa_2^5}{N^3}.
\end{aligned}$$

It is necessary to express these κ's in terms of the parameters of the binomial distribution. Here the κ's are defined by the following formal identity in θ,

$$e^{\kappa_1\theta + \kappa_2\frac{\theta^2}{2!} + \kappa_3\frac{\theta^3}{3!} + \cdots} = (q + pe^{\theta})^n.$$

Taking logarithms, expanding in powers of θ, and equating coefficients of powers of θ, the following expressions are obtained:

$$\begin{aligned}
\kappa_1 &= m \\
\kappa_2 &= mq \\
\kappa_3 &= mq(q - p) \\
\kappa_4 &= mq(1 - 6pq) \\
\kappa_5 &= mq(q - p)(1 - 12pq) \\
\kappa_6 &= mq(1 - 30pq + 120p^2q^2) \\
\kappa_7 &= mq(q - p)(1 - 60pq + 360p^2q^2) \\
\kappa_8 &= mq(1 - 126pq + 1680p^2q^2 - 5040p^3q^3).
\end{aligned}$$

These values of the κ's are inserted in (2) to give the following expressions for the μ_{i1} and μ_{i2}, with considerably longer expressions for the μ_{i3} and μ_{i4}:

$$\mu_{01} = mq$$

$$\mu_{11} = mq\,(q - p)\,\frac{1}{N}$$

$$\mu_{21} = mq\left(\frac{1 - 6pq}{N^2} + \frac{mq}{N}\right)$$

$$\mu_{31} = mq(q - p)\left(\frac{1 - 12pq}{N^3} + \frac{4mq}{N^2}\right)$$

$$\mu_{41} = m^2q^2\left(\frac{11 - 58pq}{N^3} + \frac{3mq}{N^2}\right)$$

$$\mu_{51} = m^3q^3(q - p)\,\frac{25}{N^3}$$

$$\mu_{61} = m^4q^4\,\frac{15}{N^3}$$

$$\mu_{02} = mq\left(\frac{1 - 6pq}{N} + \frac{2mq}{N - 1} + mq\right)$$

$$\mu_{12} = mq(q - p)\left(\frac{1 - 12pq}{N^2} + \frac{4mq}{N(N - 1)} + \frac{2mq}{N}\right)$$

$$\mu_{22} = mq\left(\frac{1 - 30pq + 120p^2q^2}{N^3} + \frac{8mq(1 - 5pq)}{N^2(N - 1)}\right.$$
$$\left. + \frac{mq(5 - 26pq)}{N^2} + \frac{2m^2q^2}{N(N - 1)} + \frac{m^2q^2}{N}\right)$$

$$\mu_{32} = m^2q^2(q - p)\left(\frac{12 - 102pq}{N^3} + \frac{14mq}{N^2(N - 1)} + \frac{7mq}{N^2}\right)$$

$$\mu_{42} = m^3q^3\left(\frac{36 - 176pq}{N^3} + \frac{6mq}{N^2(N - 1)} + \frac{3mq}{N^2}\right)$$

$$\mu_{52} = m^4q^4(q - p)\,\frac{40}{N^3}$$

$$\mu_{62} = m^5q^5\,\frac{15}{N^3}\,.$$

It remains to express the coefficients of (1) in terms of these same parameters. From the definition of c_i, a, and κ_1, it follows that

$$c_i = \frac{\left(\dfrac{1}{a}\right)^{i+1} + (-1)^i\left(\dfrac{1}{\kappa_1}\right)^{i+1}}{\dfrac{1}{a} + \dfrac{1}{\kappa_1}} = \frac{p^{i+1} + (-1)^i q^{i+1}}{m^i q^i}\,.$$

If now the above values of the μ_{ij} and c_i are inserted in the expressions (1), the following final formulas are obtained.

$$
\begin{aligned}
E(z) &= (N-1)\left\{1 + \frac{p}{Nm} + \left(\frac{p}{Nm}\right)^2 + \left(\frac{p}{Nm}\right)^3 + \cdots\right\} \\
E(z^2) &= (N-1)^2\left\{1 + \frac{2}{N-1} - \frac{2(1-6pq)}{(N-1)Nmq} + \frac{pq(2-11pq)}{(Nmq)^2}\right. \\
&\quad \left. - \frac{2(1+2pq-25p^2q^2)}{(N-1)(Nmq)^2} + \frac{2pq(1+3pq-30p^2q^2)}{(Nmq)^3} + \cdots\right\} \\
E(z^3) &= (N-1)^3\left\{1 + \frac{6}{N-1} - \frac{3pq}{Nmq} + \frac{8}{(N-1)^2} - \frac{6(1-3pq)}{(N-1)Nmq}\right. \\
&\quad + \frac{2pq(1-5pq)}{(Nmq)^2} + \frac{4(1-4pq)(N-2)}{(N-1)^2Nmq} - \frac{24(1-5pq)}{(N-1)^2Nmq} \\
&\quad - \frac{6(1-11pq+40p^2q^2}{(N-1)(Nmq)^2} + \frac{6pq(1-16pq+55p^2q^2)}{(Nmq)^3} \\
&\quad \left. + \frac{60pq(1-4pq)(N-2)}{(N-1)^2(Nmq)^2} + \cdots\right\} \\
E(z^4) &= (N-1)^4\left\{1 + \frac{12}{N-1} - \frac{8pq}{Nmq} + \frac{44}{(N-1)^2} - \frac{12(1+2pq)}{(N-1)Nmq}\right. \\
&\quad - \frac{2pq(2-21pq)}{(Nmq)^2} + \frac{16(1-4pq)(N-2)}{(N-1)^2Nmq} + \frac{48}{(N-1)^3} - \frac{8(15-46pq)}{(N-1)^2Nmq} \\
&\quad - \frac{12(3-44pq+138p^2q^2)}{(N-1)(Nmq)^2} + \frac{64pq(1-4pq)(N-2)}{(N-1)^2(Nmq)^2} \\
&\quad + \frac{96(1-4pq)(N-2)}{(N-1)^3Nmq} + \frac{8(1-12pq+36p^2q^2)(4N^2-9N+6)}{(N-1)^3(Nmq)^2} \\
&\quad \left. + \frac{4pq(1-43pq+168p^2q^2)}{(Nmq)^3} + \cdots\right\}.
\end{aligned}
\tag{3}
$$

By considering the formation of terms, it can also be shown that the above expressions are correct to terms of order m^3, m^2, m^1, and m^0, respectively, in the parameter m. If m is large these expressions are considerably more accurate than the order N^{-3} would indicate since the lowest order terms neglected in these expressions are respectively N^4m^4, N^4m^3, N^4m^2, and N^4m.

3. Applications. To compare these moments with those of the chi-square distribution, consider the ratios of corresponding moments, both for the Poisson distribution and for the binomial distribution in the special case of $p = \frac{1}{3}$.

For the Poisson distribution, these ratios are

$$R_1 = 1$$

$$R_2 = 1 - \frac{1}{Nm} - \frac{1}{(Nm)^2}$$

$$R_3 = 1 + \frac{1}{2m} - \frac{4}{Nm}$$

$$R_4 = 1 + \frac{2}{N+3}\left\{\frac{3}{m} + \frac{1}{3m^2} - \frac{7}{Nm}\right\}.$$

For the binomial distribution with $p = \frac{1}{2}$, these ratios are

$$R_1 = 1 + \frac{1}{Nn} + \frac{1}{(Nn)^2} + \frac{1}{(Nn)^3}$$

$$R_2 = \left(1 - \frac{1}{n}\right)\left(1 + \frac{5}{2Nn} - \frac{7}{4N^2n^2}\right) - \frac{7}{4(Nn)^3}$$

$$R_3 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{7}{4n}\right) + \frac{1}{Nn}\left(4 - \frac{13}{2n} + \frac{5}{2n^2}\right) + \frac{1}{N^2n^2}\left(1 - \frac{5}{n}\right) + \frac{5}{2(Nn)^3}$$

$$R_4 = 1 + \frac{N}{N+3}\left\{\frac{-2}{n} + \frac{1}{n^2} + \frac{1}{Nn}\left(-13 + \frac{37}{2n} - \frac{17}{2n^2}\right)\right.$$
$$\left. + \frac{1}{N^2n}\left(11 - \frac{67}{2n} + \frac{51}{2n^2}\right) + \frac{1}{N^3n^2}\left(\frac{31}{2} - \frac{51}{2n}\right) + \frac{17}{2N^4n^3}\right\}.$$

From these expressions the following table is constructed.

m	n	N	R_1	R_2	R_3	R_4
25	∞	3	1	.99	.97	1.01
25	75	3	1	1	.98	.97
5	∞	5	1	.96	.94	1.08
5	15	5	1.01	.96	.87	.84
2	∞	∞	1	1	1.25	1
2	∞	10	1	.95	1.05	1.19
2	∞	5	1	.89	.85	1.21
2	6	∞	1	.83	.59	.69
2	6	10	1.02	.87	.64	.64
2	6	5	1.03	.90	.69	.62
1	∞	25	1	.96	1.34	1.22
1	∞	10	1	.89	1.10	1.39
1	∞	5	1	.76	.70	1.44
1	3	25	1.01	.69	.31	.41
1	3	10	1.03	.72	.35	.38
1	3	5	1.07	.77	.41	.36

For $m \geq 5$ these ratios are close to unity even for N as small as 5; hence it appears that the chi-square approximation is satisfactory as long as $m \geq 5$.

For $m \leq 2$ most of these ratios differ considerably from unity, particularly for the binomial distribution. Since R_1 is practically constant, the reduction in R_2 here indicates that the chi-square approximation will contain too many extreme values. For the Poisson distribution there is an increase in R_4 to compensate slightly for this decrease in R_2 so that the 5 percent points, for example, would not differ very much. The use of the chi-square approximation would therefore tend to give slightly too few significant results when they exist. For the binomial distribution, however, there is a decrease in both R_2 and R_4, so that the distribution tends toward normality; consequently the chi-square approximation will contain far too many extreme values and the 5 percent point will be much too large. This situation becomes slightly worse with increasing N.

4. Conclusions.

From a consideration of the approximations for the first four moments of the distribution of the index of dispersion, it appears that the chi-square approximation is highly satisfactory provided that $m \geq 5$. For smaller values of m, the approximation is still fairly accurate for the Poisson distribution but not for the binomial distribution. For decreasing small values of m there is an increasing tendency to claim compatibility between data and theory when it does not exist; hence the binomial index must be handled carefully in such situations. These general conclusions are in agreement with the specialized results of Cochran and Sukhatme.

The semi-invariant technique for problems such as this is exceedingly laborious and is of questionable accuracy. The coefficients in Fisher's heavier formulas are so large that increased accuracy comes slowly with increased accuracy of order of terms. In addition, there are numerous typographical mistakes in Fisher's formulas, some of which are not easily detected. The formulas (3) may be used to investigate the accuracy of the chi-square approximation for situations not covered in the numerical table, but they are of questionable accuracy, when m is small, for N as small as 5.

REFERENCES

[1] P. V. Sukhatme, "On the distribution of chi-square in samples of the Poisson series," *Jour. Roy. Stat. Soc.*, Vol. 101 (1938), pp. 75–79.

[2] W. G. Cochran, "The chi-square distribution for the binomial and Poisson series with small expectations," *Annals of Eugenics*, Vol. 7 (1936), pp. 207–17.

[3] R. A. Fisher, "Moments and product moments of sampling distributions," *Proc. London Math. Soc.*, Series 2, Vol. 30 (1930), pp. 199–238.