NOTES

This section is devoted to brief research and expository articles and other short items.

TRANSFORMATIONS RELATED TO THE ANGULAR AND THE SQUARE ROOT

By Murray F. Freeman and John W. Tukey[1]

Princeton University

1. Summary. The use of transformations to stabilize the variance of binomial or Poisson data is familiar (Anscombe [1], Bartlett [2, 3], Curtiss [4], Eisenhart [5]). The comparison of transformed binomial or Poisson data with percentage points of the normal distribution to make approximate significance tests or to set approximate confidence intervals is less familiar. Mosteller and Tukey [6] have recently made a graphical application of a transformation related to the square-root transformation for such purposes, where the use of "binomial probability paper" avoids all computation. We report here on an empirical study of a number of approximations, some intended for significance and confidence work and others for variance stabilization.

For significance testing and the setting of confidence limits, we should like to use the normal deviate K exceeded with the same probability as the number of successes x from n in a binomial distribution with expectation np, which is defined by

$$\frac{1}{2\pi}\int_{-\infty}^{K} e^{-\frac{1}{2}t^2}\,dt = \text{Prob}\,\{x \leq k \mid \text{binomial}, n, p\}.$$

The most useful approximations to K that we can propose here are N (very simple), N^+ (accurate near the usual percentage points), and N^{**} (quite accurate generally), where

$$N = 2\,(\sqrt{(k+1)q} - \sqrt{(n-k)p}).$$

(This is the approximation used with binomial probability paper.)

$$N^+ = N + \frac{N + 2p - 1}{12\sqrt{E}}, \qquad E = \text{lesser of } np \text{ and } nq,$$

$$N^* = N + \frac{(N-2)(N+2)}{12}\left(\frac{1}{\sqrt{np+1}} - \frac{1}{\sqrt{nq+1}}\right),$$

$$N^{**} = N^* + \frac{N^* + 2p - 1}{12\sqrt{E}}. \qquad E = \text{lesser of } np \text{ and } nq.$$

For variance stabilization, the averaged angular transformation

$$\sin^{-1}\sqrt{\frac{x}{n+1}} + \sin^{-1}\sqrt{\frac{x+1}{n+1}}$$

[1] Prepared in connection with research sponsored by the Office of Naval Research.

has variance within $\pm 6\%$ of

$$\frac{1}{n + \frac{1}{2}} \text{ (angles in radians)}, \qquad \frac{821}{n + \frac{1}{2}} \text{ (angles in degrees)},$$

for almost all cases where $np \geq 1$.

In the Poisson case, this simplifies to using

$$\sqrt{x} + \sqrt{x + 1}$$

as having variance 1.

2. Significance testing. In addition to the approximations mentioned above, empirical study was also made of the following

$$L = \frac{x - np}{\sqrt{npq}},$$

$L^* = L$ modified by a term like that in N^*,

$$M = 2\sqrt{n + 1}\left(\sin^{-1}\sqrt{\frac{k + 1}{n + 1}} - \sin^{-1}\sqrt{p}\right),$$

$M^* = M$ modified by a term like that in N^*.

Taking an upper limit of 2.5 or 3.5 on $| K |$ and a lower limit of 0.01, 1, or 4 on np, the greatest observed errors of the approximations were smallest for N^{**}, N^* and M^* and largest for the direct approximations L and L^*. This was true for all six choices of region.

If we exclude the cases $k = 0$ and $k = n$, where the desired probability can be calculated directly, the largest observed errors in the substantial number of cases computed, which are probably representative of the regions where the approximations are worst, were as follows:

$\| K \|$	$E = np$	Largest observed error of							
		N^{**}	M^*	N^*	N^+	N	M	L^*	L
≤ 2.5	≥ 4	.04	.07	.08	.14	.16	.17	.26	.35
	≥ 1	.04	.09	.13	.19	.20	.24	.35	.42
	≥ 0.01	.04	.20	.20	.19	.20	.65	.62	.80
≤ 3.5	≥ 4	.08	.07	.08	.19	.25	.25	.57	.63
	≥ 1	.11	.10	.17	.21	.38	.34	1.51	1.26
	≥ 0.01	.11	.51	.60	.21	.65	.65	5.88	3.42

Within the range of great interest, $| K | \leq 2.5$, that is $.0062 \leq$ probability $\leq .9938$, we have errors of less than 0.04 in N^{**} and less than 0.20 in N.

For $1.5 < | K | < 2.5$, the range of greatest interest, the average error of N^+ was less than 0.03 and the maximum was 0.08 (54 cases considered).

Thus, we can recommend

N —as a simple and usually accurate transformation,

N^+ —for rapid significance testing,

N^{**}—for adequate accuracy at all levels.

Figure 1 shows the behavior of the various approximations in the case $n = 50$, $np = 5$. This is roughly typical.

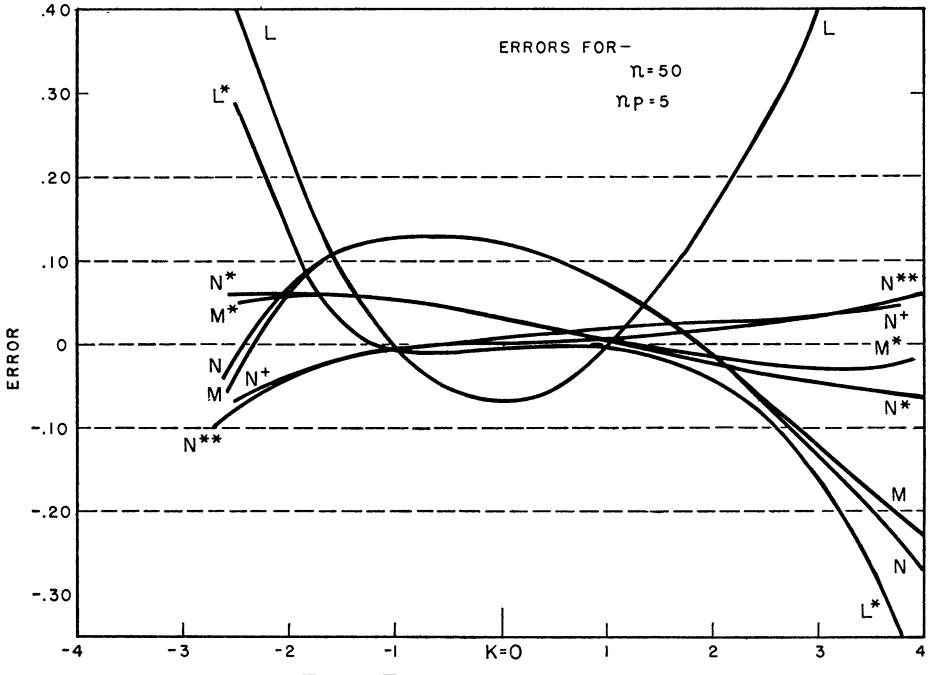

FIG. 1. Errors of approximation.

3. Variance stabilization. The various suggestions for stabilizing the variance of the Poisson are:

$$\sqrt{x + 1/2}, \qquad \text{(Bartlett [2])},$$

$$\sqrt{x + 3/8}, \qquad \text{(Anscombe [1])},$$

$$\sqrt{x} + \sqrt{x + 1}, \qquad \text{(this paper)}.$$

Figure 2 shows the variance of the transformed variate as a function of the Poisson expectation. Clearly $\sqrt{x} + \sqrt{x + 1}$ is the best if small expectations are to be considered. The simplicity with which it can be read from a square-root table, and its unit variance, are also favorable factors.

When an approximation of a given form is to work over as large a range as

possible without the magnitude of its errors exceeding a certain limit, the optimum approximation is almost certain to involve errors of *both* signs. If $\pm 6\%$ variation in variance is permissable, $\sqrt{x} + \sqrt{x+1}$ is usable for expectations of unity or more. It is not surprising that Anscombe's approximation, obtained by eliminating the term in n^{-1}, and dominated by the term in n^{-2}, should only meet the $\pm 6\%$ tolerance for expectations of 2.2 or more.

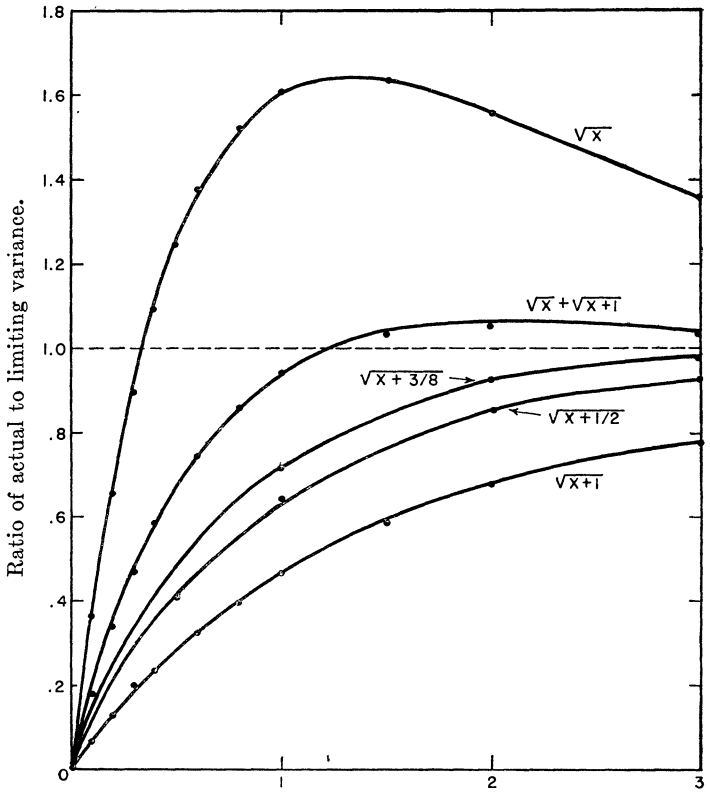

FIG. 2. Stabilization of Poisson variance.

4. Scope. Values of K, and with some occasional exceptions, of L, L^*, M^*, N, N^+, N^* and N^{**} were calculated for

$$n = 2, 5, 10, 20, 100,$$

$$p = 1\%, 2\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%,$$

$$k \quad \text{giving } K < 4.5,$$

and similar computations were made for the Poisson case with expectations

$$1/100, 1/50, 1/20, 1/10, 1/5, 1/2, 1, 2, 4, 8, 16, 32, 64.$$

These computations were made to only two decimal places, so that the final results may easily err by 1, 2, or 3 in the second decimal place.

A more complete discussion of the problem, the origin of the approximations, and tables showing a representative collection of actual values can be found in Memorandum Report 24 of the Statistical Research Group, Princeton University, which bears the same title as this note. Copies may be obtained from its Secretary, Box 708, Princeton, N. J.

REFERENCES

[1] F. J. ANSCOMBE, "The transformation of Poisson, binomial, and negative binomial data", *Biometrika*, Vol. 35 (1948), pp. 246–254.

[2] M. S. BARTLETT, "The square root transformation in the analysis of variance", *Jour. Roy. Stat. Soc., Suppl.*, Vol. 3 (1936), pp. 68–78.

[3] M. S. BARTLETT, "The use of transformations", *Biometrics*, Vol. 3 (1947), pp. 39–51.

[4] J. H. CURTISS, "On transformations used in the analysis of variance", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 107–122.

[5] CHURCHILL EISENHART, "The assumptions underlying the analysis of variance", *Biometrics*, Vol. 3 (1947), pp. 1–21.

[6] FREDERICK MOSTELLER AND JOHN W. TUKEY, "The uses and usefulness of binomial probability paper", *Jour. Am. Stat. Assn.*, Vol. 44 (1949), pp. 174–212.

REMARK ON THE ARTICLE "ON A CLASS OF DISTRIBUTIONS THAT APPROACH THE NORMAL DISTRIBUTION FUNCTION" BY GEORGE B. DANTZIG[1]

BY T. N. E. GREVILLE

Federal Security Agency

In this interesting and valuable article, Dr. Dantzig showed that, under certain conditions, a sequence of frequency distributions connected by a linear recurrence formula converges to the normal distribution. Among several applications of his results which are discussed, the author mentions their relation to certain types of smoothing formulas, and has shown that if a linear smoothing formula and the data to which it is applied satisfy certain conditions, the iteration of the smoothing process produces a sequence of smoothed distributions which, upon normalization, approaches the normal frequency curve.

In a summary paragraph at the end of the article, it is stated that "successive application of one or many such linear formulas will usually smooth *any* set of values to the normal curve of error." The entire article was concerned with frequency distributions, and a careful reading makes it clear that the author intended the quoted statement to apply only to data in this form. However, its rather general wording seems to have led a number of readers to interpret it as being applicable to other types of data, such as time series, which frequently may not satisfy the conditions assumed. Moreover, it is easy to overlook the

[1] *Annals of Math. Stat.*, Vol. 10 (1939), pp. 247–253.