The Use of the Central Limit Theorem for Interpolating in Tables of Probability Distribution Functions

By Gerard Salton

In using tables of probability density and distribution functions, the difficulty of interpolating for functional values which are not directly tabulated constitutes a major problem. In particular, when the variance of a random variable becomes small, the corresponding density and distribution functions approach, respectively, a delta-function and a step-function. As a result, for small variations in the variables, there occur very large variations in the corresponding functional values, and interpolation by difference methods will give very poor approximations. Moreover, space limitations in a volume of tables often make it impractical to tabulate a given function on a mesh fine enough to allow for accurate interpolation. This often increases the difficulty of interpolating for a given probability distribution.

By a fundamental limit theorem of probability theory it is known that, under rather general conditions, the sum of a set of n independent random variables appropriately standardized by the mean and the standard deviation is asymptotically normal with mean 0 and variance 1, as n tends to infinity [1]. A method which makes use of this property to improve the accuracy of interpolating in tables of probability distribution functions was suggested by Rossow [2]. The binomial probability distribution is chosen here for purposes of illustration. The method described is however applicable to any set of random variables which obeys the central limit theorem.

The cumulative binomial probability distribution may be denoted by

$$E(n, r, p) = \sum_{i=r}^{n} C_i^n (1 - p)^{n-i} p^i$$

where $0 \leqq p \leqq 1$ and $0 \leqq r \leqq n$. If one considers a series of n independent repetitions of some random experiment, then $E(n, r, p)$ represents the probability of at least r successes in n repetitions of the experiment, p being the probability of success for each experiment. By the theorem previously stated, it is known that for a given value of p, and for increasing values of n

$$E(n, r, p) \approx \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt = 1 - N(x)$$

where $x = (r - np - \frac{1}{2})/\sqrt{np(1 - p)}$, and $N(x)$ denotes the cumulative standard normal distribution. The effectiveness of the approximation increases with increasing values of $\sqrt{np(1 - p)}$, so that for a given value of n, the approximation is closest for $p = \frac{1}{2}$. The argument x, corresponding to a given value of $N(x)$, is often called the normal deviate corresponding to the total frequency N.

Consider now the problem of interpolating in a table of the binomial probability

Received April 9, 1958.

distribution. As an example, let it be desired to find some value $E(n_0, r_0, p_0 + \Delta p)$, where the function is not tabulated for the argument $p_0 + \Delta p$. The standard procedure consists in taking the values of $E(n, r, p)$ corresponding to the arguments $n_0r_0p_0, n_0r_0p_1, \cdots, n_0r_0p_{k+1}$ and in constructing a $k + 1^{\text{th}}$ degree polynomial which takes on the given functional values at the $k + 2$ given points. The value of the polynomial at the point $n_0r_0p_0 + \Delta p$ is then calculated and taken to be equal to $E(n_0, r_0, p_0 + \Delta p)$.

It is proposed to modify this procedure by making use of the fact that the binomial distribution is often much better approximated by the corresponding normal distribution than by the interpolation polynomial which is constructed in the standard method. To this effect, it is first assumed that $E(n_0, r_0, p_0) = 1 - N(x_0)$, $E(n_0, r_0, p_1) = 1 - N(x_1), \cdots, E(n_0, r_0, p_{k+1}) = 1 - N(x_{k+1})$. The $k + 1$ normal deviates corresponding to the $k + 1$ values of $1 - N$ are then found in a table of the normal distribution. Thereafter the interpolation is performed in the normal deviates, that is, from the values $-x_i$ corresponding to the arguments p_i, a value $-x_p$ is determined which corresponds to the argument $p_0 + \Delta p$. The value of $1 - N(x_p)$ is then looked up in a table of the normal distribution, and $E(n_0, r_0, p_0 + \Delta p)$ is approximated by $1 - N(x_p)$.* The procedure is outlined in the following diagram, where the arrows denote a table look-up and the brace stands for the evaluation of an ordinary interpolation polynomial:

$$\left.\begin{array}{lllll} p_0 \rightarrow E(n_0, r_0, p_0) & = & 1 - N(x_0) & \rightarrow & -x_0 \\ p_1 \rightarrow E(n_0, r_0, p_1) & = & 1 - N(x_1) & \rightarrow & -x_1 \\ \vdots \qquad\quad \vdots & & \vdots & & \vdots \\ p_{k+1} \rightarrow E(n_0, r_0, p_{k+1}) & = & 1 - N(x_{k+1}) & \rightarrow & -x_{k+1} \end{array}\right\} -x(p_0 + \Delta p) \rightarrow 1 - N(x_p)$$

$$= E(n_0, r_0, p_0 + \Delta p).$$

The procedure is easier to perform if tables are available which tabulate the normal deviates as a function of the cumulative normal probabilities [3]. However, tables which give $N(x)$ as a function of x can also be used [4]. In either case, the accuracy of the method depends on the availability of accurate values of the normal deviates; this is especially important for cumulative probabilities close to 0 and close to 1 where the deviates change rapidly for small changes in the probabilities. It should be noted that the additional labor required by the proposed method is restricted to some table look-up operations which can be performed rapidly.

A study was made of the efficiency of the method, based on a recent tabulation of the cumulative binomial probability distribution [5]. It is stated in that volume that interpolation by standard difference methods gives poor results for certain ranges of the arguments. In particular, p-wise interpolation is poor for $n > 100$ and p small, r-wise interpolation is poor for $n < 100$ and p small, and n-wise interpolation is inaccurate for $n > 100$ and p close to $\frac{1}{2}$.

The normal deviate method fills the need for both p-wise and n-wise interpola-

* Since the arguments of $1 - N(x)$ are the negatives of the corresponding arguments of $N(x)$, the method can also be carried out by interpolating in the values x_i, rather than in the values $-x_i$. Such an interpolation will result in the determination of the function $N(x_p)$ corresponding to $1 - E(n_0, r_0, p_0 + \Delta p)$.

tion, since it is most accurate for large n where the standard method fails. It can also be used for interpolation in r when n is not too small. When interpolating for p and for n, a four-point interpolation in the normal deviates was found to result in final probabilities whose error did not exceed 5.10^{-5}, except for very small p ($p \leqq 0.03$) where the error could be as high as 5.10^{-3}. Four-point interpolation in r for $n > 30$ resulted in errors not exceeding 5.10^{-4}.

A different method which also improves the interpolation in tables of probability distribution functions consists in approximating the given distribution by the logistic function [6]

$$L(x) = \frac{1}{1 + e^{-\beta x}}$$

in lieu of the cumulative standard normal distribution. In fact it is known that

$$L(x) \approx N(x),$$

the two functions differing by less than 0.01 when β is close to 1.7. The interpolation proceeds exactly as before except that $L(x)$ is used for $N(x)$. The arguments x corresponding to various values of $L(x)$ are sometimes called logits. By inverting the expression for $L(x)$ it is seen that

$$x = \frac{1}{\beta} \ln \frac{L}{1 - L}.$$

Tables of the natural logarithms and exponential tables can therefore be used, if logit tables are not available [6].

Linear interpolation in the logits will sometimes give slightly better values than the corresponding interpolation in the normal deviates. However, a higher order interpolation formula brings much less improvement in the logits than it does in the normal deviates. A four-point interpolation in the normal deviates is usually to be preferred to the corresponding four-point formula in logits. It is again important to use accurate logit tables, especially for values of L close to 0 and 1. Since tables of the logistic function may not be generally available, interpolation in the normal deviates is usually easier to perform and will moreover give more accurate results for higher order interpolation formulas.

The following examples give a good idea of the improvement which can be

Comparison of Interpolation Methods

	E (600,225,0.375) p-wise, $\Delta p=0.5$	E (1000,53,0.0625) p-wise, $\Delta p=0.25$	E (39,6,0.08) r-wise, $\Delta r=1$	E (689,285,0.42) n-wise, $\Delta n=39$
Correct Values	.51541	.90685	.08786	.64612
Ordinary 2-pt.	.51475 (−66)	.87776 (−2909)	.11597 (+2811)	.63732 (−880)
Ordinary 4-pt.	.51531 (−10)	.90178 (−507)	.09194 (+408)	.64157 (−455)
Normal Deviates 2-pt.	.51524 (−17)	.90497 (−188)	.09004 (+218)	.64263 (−349)
Normal Deviates 4-pt.	.51542 (+1)	.90685 (0)	.08770 (−16)	.64616 (+4)
Logits 2-pt.	.51537 (−4)	.91249 (+564)	.08456 (−330)	.64401 (−211)
Logits 4-pt.	.51543 (+2)	.90973 (+288)	.08726 (−60)	.64870 (+258)

achieved over the standard interpolation method by using either normal deviates or logits. In each case, the error in the value obtained is shown in parentheses.

Computation Laboratory, Harvard University,
Cambridge, Massachusetts

1. HARALD CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1951, p. 213–220.
2. E. ROSSOW, Technische Universität, Berlin, private communication.
3. R. A. FISHER & F. YATES, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd Ltd., Edinburgh, 1948.
4. HARVARD UNIVERSITY, COMPUTATION LABORATORY, *Annals*, v. 23: *Tables of the Error Function and of its First Twenty Derivatives*, Harvard University Press, Cambridge, Mass., 1952.
5. HARVARD UNIVERSITY, COMPUTATION LABORATORY, *Annals*, v. 35: *Tables of the Cumulative Binomial Distribution*, Harvard University Press, Cambridge, Mass., 1955.
6. JOSEPH BERKSON, "A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function," *Journal of the American Statistical Association*, v. 48, No. 263, September, 1953.