THE POINT BINOMIAL AND PROBABILITY PAPER

By Frank H. Byron[1]

1. An approximation to the sum of a number of consecutive terms of the point binomial may be found graphically and quite expeditiously by means of so-called "probability paper." This paper is ruled so that the (x, y) graph of the equation of the integral of the normal curve

$$y = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x}{2}} \, dx \tag{1}$$

is a straight line. Let the successive terms of the point binomial be represented as follows:

$$(p + q)^n = u_0 + u_1 + \cdots + u_t + \cdots + u_n , \tag{2}$$

where $u_t = {}_nC_t p^{n-t} q^t$ and $p \geqq q$. Then the (x, y) graph of the equation,

$$y = \sum_{i=0}^{t} u_i , \qquad t + \tfrac{1}{2} = x , \tag{3}$$

i.e., of the sum of first $(t + 1)$ terms of this point binomial, is, in all but extreme cases, a set of points lying on a gently turning curve, so gently that its form may be represented closely by two straight lines, each passing through the median point as will be explained in the next section. As paper of this sort is readily obtainable, and as this method yields as great accuracy as is really useful in many problems, it is suggested that its use ought to be quite general.

2. **Sheppard's Corrections.** The formulae for the moments of the point binomial, mean $= qn$, $\sigma^2 = pqn$, are exact without any corrections such as are used for grouped material. This fact has led us all (apparently) to assume that in fitting the curve to the point binomial one would get a better fit by equating the moments of the curve to the uncorrected moments of the point binomial rather than to the corrected moments. The studies made in connection with the preparation of this paper show that when the purpose is to equate areas to sums of terms the corrected moments should be used. The theoretical basis for this conclusion is as follows:

To simplify the argument let us suppose that one were seeking that curve of Charlier type,

$$F(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots c_4\phi_4(x) , \tag{4}$$

[1] With the assistance of Burton H. Camp.

(where ϕ_0 is the normal curve and ϕ_1, ϕ_2, $\cdots$ its successive derivatives) whose integral would best fit the graph of (3). Since fitting is required only at the isolated points $x = \frac{1}{2}, 1\frac{1}{2}, 2\frac{1}{2}, \cdots$, it is clear that one might obtain this by the two following steps. First let $f(x)$ be any function whose integral meets exactly the requirement at these isolated points. What values this integral has at other points does not for the moment concern us. There are an infinite number of such $f(x)$ curves. Next let the c's of (4) be so chosen that $F(x)$ will fit $f(x)$ as nearly as possible. The ordinary derivation of the c's supposes that the fit between $f(x)$ and $F(x)$ is to be made by least squares, the residuals being weighted by the factor $1/\sqrt{\phi(x)}$. No matter what $f(x)$ is chosen, the c's can be determined so that the weighted integral of $(f(x) - F(x))^2$ will be a minimum, but the value of this minimum will vary from one $f(x)$ to another. We now desire to select that $f(x)$ which will make this minimum value as small as possible, and it is reasonable to suppose that our best selection will be some $f(x)$ which is as kindred to the nature of $F(x)$ as possible. We shall not therefore choose an $f(x)$ which oscillates wildly between the points where perfect fitting is required, (Fig. 1) nor yet an $f(x)$ which is made up of the top bases of the point binomial

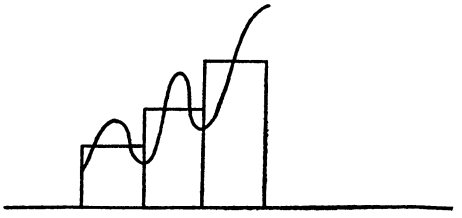

FIG. 1

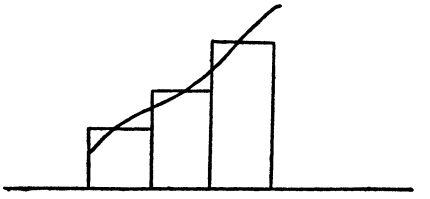

FIG. 2

histogram; we shall prefer a modification (Fig. 2) of that histogram by a smoothing process. Such an $f(x)$ will not have the exact moments of the point binomial, but, more nearly, those moments corrected for grouping. Then the determination of the c's will come out in terms of these corrected moments, not in terms of the uncorrected moments. (In fact the uncorrected moments would be the exact moments of an $f(x)$ having an oscillatory character between the important points.)

Of course, when n is large, the difference is too small to be noticed and the use of Sheppard's corrections is not worth while, and since n usually is large when approximations of this sort are needed, the point is not usually important. It was important in the computation of the tables of §4. Moreover, the use of Sheppard's corrections does not invariably yield better results, the gain being masked sometimes by other effects to be considered in §3. An excellent illustration of uniformly better results is in fitting $(\frac{1}{2} + \frac{1}{2})^9$ by a curve of Type 4. The errors in the sums as derived from (4) with and without the corrections, is given on the following page.

t	0	1	2	3	4	5	6	7	8	9
With Corrections	.0002	.0001	−.0003	−.0001	.0000	.0001	.0003	−.0001	−.0002	.0000
Without Corrections	.0007	.0022	.0039	.0036	.0000	−.0036	−.0039	−.0022	−.0007	−.0001

3. **The Stubby End.** The other effects which mask this improvement are especially noticeable at the stubby end of a point binomial. We have to keep in mind here that the approximating curve (such as (4)), is required to turn a sharp corner, for, due to the least square method of fitting, it is just as important that it be close to zero when t is negative, as it is that it be close to u_0, u_1, $\cdots$ when t is positive. Therefore, in order to turn this corner it has to dip below the x-axis in the neighborhood of $t = -\frac{1}{2}$. This makes the approximating curve too low just to the right of $t = -\frac{1}{2}$, unless the whole curve be arbitrarily widened. This arbitrary widening is customarily performed by not using Sheppard's correction for σ, and the result is a betterment of the fit at these points but a corresponding loss over the rest of the infinite interval. A good example[2] is $(\frac{2}{3} + \frac{1}{3})^{25}$. The fit is worse at the left end when Sheppard's corrections are used but better over the rest of the interval.

The same difficulty arises in another connection. If we compare the closeness of fit to a point binomial made by $F(x)$ as written in (4) and by $F(x)$ as it would be written if c_4 were zero, it often happens (as is well known) that the latter is actually slightly better on the average. How can this be true if the c's are chosen by the method of least squares and the best choice as thus indicated makes c_4 different from zero? The answer is that the c's are chosen so that the fit is best over the infinite interval, not merely over the interval from $t = -\frac{1}{2}$ to $t = n + \frac{1}{2}$, and that furthermore the distant points are weighted more heavily than those near the center. Thus it might happen that a choice, other than the least square choice, and one in which c_4 would be zero, might be better for the restricted interval covered by the point binomial. This does happen especially when due to the abruptness of the stubby end of a very skew binomial, the curve has to dip below the axis in order to get by a sharp corner. A good example is the problem considered by Fry:[3] $(\frac{9}{10} + \frac{1}{10})^{100}$. All the effects mentioned are present here. The fit is on the average a little worse if c_4 is not equal to zero over the point binomial interval, a little better over the infinite interval.

4. For graphical purposes a sufficiently good approximation to the median of $(p + q)^n$, is given by

$$M = nq - (p - q)/6.$$

[2] The true values are given on page 220 of Mathematical Part of Elementary Statistics, by Camp, D. C. Heath and Company, 1931.

[3] T. C. Fry, Probability and its Engineering Uses, p. 258, Van Nostrand, 1928.

The following tables enable us to find the first quartile Q_1, and the ninth decile D_9. The accuracy to which they can be plotted is only about one-tenth that to which they are given here. Therefore accurate interpolation is seldom necessary. The values of S_{t+1} are to be read from the graph at the points $t + \frac{1}{2}$, as indicated in the directions preceding the tables. The graphical method will be found efficient if one uses common sense in the computation. Numbers which are to be plotted should not be computed to a higher degree of accuracy than can be used graphically. In reading the values of S_{t+1} it is well to remember that the true values lie on a curve, and that outside the interval from Q_1 to D_9, they are slightly less than those given by the straight line. Once the graph has

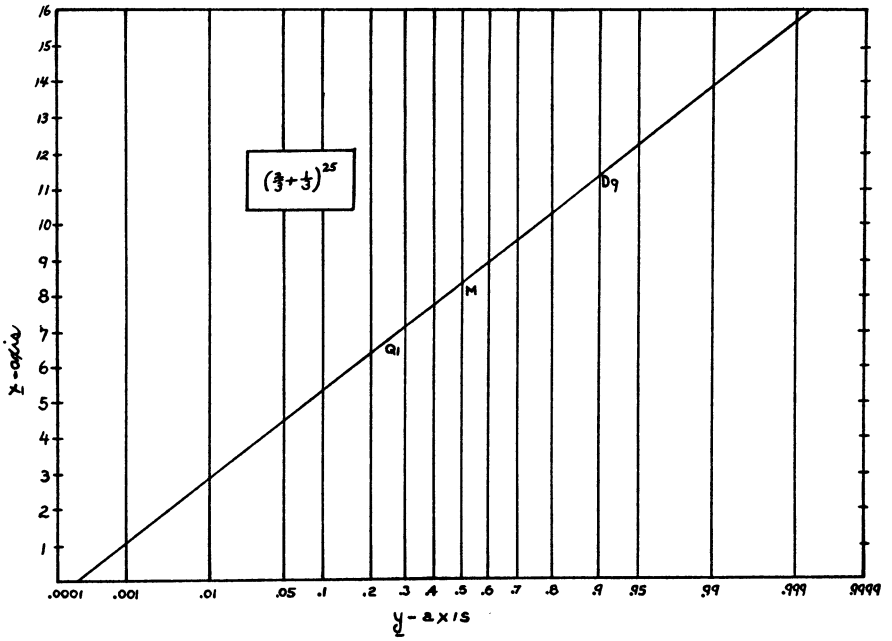

FIG. 3

been made, all the values of S_{t+1} can be read quickly; it is not necessary to make a separate computation for each t. This method is therefore specially advantageous when one wishes to find several sums of this sort for the same point binomial. It should also be noticed that one can tell from the appearance of the graph about how far the true sum would be from the two straight lines and so estimate the error to which his reading is liable.

5. **Illustration.** Find the sum of the first 7 terms of $(\frac{2}{3} + \frac{1}{3})^{25}$.

Here $t = 6$, $M = 8.278$, $Q_1 = 6.726$, $D_9 = 11.369$. The graph shows that $\sum_0^t = 0.224$. The true value is 0.222. So the error is 0.002.

An idea of the accuracy of the method is given by the errors (out of two places) that would be obtained for this point binomial for various values of t, as follows:

t	2	4	6	8	10	12	14	16
Errors	.00	.01	.00	.00	.00	.00	.00	.00

DIRECTIONS FOR USE OF THE TABLES: Let $p = q$, $M = nq - (p - q)/6$, $Q_1 = x_1 + qn$, $D_9 = x_2 + qn$. On the graph draw the lines MQ_1 and MD_9. Read S_{t+1} at $t + \frac{1}{2}$.

Values of x_1

p \ n	2000	1000	750	500	400	300	200	100	75	50	25
.99	−.693	−.701	−.704	−.710	−.714	−.720	−.728	−.747	−.756	−.771	−.804
.98	−.688	−.693	−.696	−.700	−.703	−.707	−.714	−.728	−.735	−.746	−.770
.97	−.685	−.690	−.692	−.696	−.698	−.701	−.707	−.718	−.724	−.734	−.784
.96	−.684	−.687	−.689	−.693	−.695	−.697	−.702	−.712	−.718	−.726	−.744
.95	−.683	−.686	−.688	−.691	−.692	−.695	−.699	−.708	−.713	−.721	−.737
.94	−.682	−.685	−.686	−.689	−.691	−.693	−.697	−.705	−.709	−.717	−.732
.93	−.681	−.684	−.685	−.688	−.689	−.691	−.695	−.703	−.707	−.713	−.727
.92	−.681	−.683	−.685	−.687	−.688	−.690	−.693	−.701	−.704	−.710	−.723
.91	−.680	−.683	−.684	−.686	−.687	−.689	−.692	−.699	−.702	−.708	−.720
.90	−.680	−.682	−.683	−.685	−.686	−.688	−.690	−.697	−.700	−.704	−.717
.88	−.679	−.681	−.682	−.684	−.685	−.686	−.689	−.695	−.697	−.702	−.713
.85	−.679	−.680	−.681	−.682	−.683	−.685	−.687	−.691	−.694	−.698	−.707
.80	−.677	−.679	−.679	−.681	−.681	−.682	−.684	−.688	−.690	−.693	−.700
.75	−.677	−.678	−.678	−.679	−.680	−.681	−.682	−.685	−.686	−.689	−.694
.70	−.676	−.677	−.677	−.678	−.679	−.679	−.680	−.682	−.683	−.685	−.690
.65	−.676	−.676	−.677	−.677	−.677	−.678	−.678	−.680	−.681	−.682	−.686
.60	−.675	−.676	−.676	−.676	−.676	−.677	−.677	−.678	−.679	−.680	−.682
.50	−.675	−.675	−.675	−.675	−.675	−.675	−.675	−.675	−.675	−.675	−.675

ERRATA

THE ANNALS OF MATHEMATICAL STATISTICS

Volume VI, No. 1, March, 1935

On page 25, in Directions for Use of the Tables, $p = q$ should read $p \gtrless q$, $Q_1 = x_1 + qn$ should read $Q_1 = x_1\sigma + qn$, $D_9 = x_2 + qn$ should read $D_9 = x_2\sigma + qn$. In the tables of values of x under $p = .97$, $n = 25$, instead of $-.784$ the number should be $-.754$.

Values of x_2

p \ n	2000	1000	750	500	400	300	200	100	75	50	25
.99	1.307	1.318	1.325	1.336	1.344	1.356	1.378	1.439	1.481	See Auxiliary	
.98	299	307	311	318	323	330	343	376	396	Tables	
.97	295	301	304	310	314	319	329	353	367		
.96	293	298	301	306	309	313	321	341	352		
.95	292	296	299	303	305	309	316	332	342		
.94	291	295	297	300	303	306	312	327	335		
.93	290	293	295	298	301	304	309	322	329	1.342	1.374
.92	289	292	294	297	299	302	307	318	325	336	365
.91	289	292	293	296	298	300	305	315	321	331	357
.90	288	291	292	295	296	299	303	313	318	325	351
.88	287	290	291	293	295	297	300	309	313	321	341
.85	286	288	289	291	292	294	297	304	308	314	330
.80	285	287	288	289	290	291	293	298	301	306	317
.75	284	285	286	287	288	289	291	294	297	300	308
.70	284	285	285	286	286	287	288	291	293	295	301
.65	283	284	284	285	285	286	286	288	290	292	296
.60	283	283	283	284	284	284	285	286	287	288	291
.50	282	282	282	282	282	282	282	282	282	282	282

Auxiliary Table

p \ n	60	50	40	35	30	25	20
.99	1.525	1.575	1.663	1.740	1.871	2.149	3.209
.98	416	435	455	488	520	1.568	1.652
.97	381	394	413	433	445	472	514
.96	362	372	387	397	410	428	457
.95	350	359	370	378	389	405	425
.94	336	349	359	366	375	387	405