

On Binomial Distribution

By

Jun'ichi Tomita

(Received April 14, 1932)

In this paper, after a brief study of binomial distributions, a new method of graduation is proposed. As a matter of course, the fit is satisfactory when the new method is applied to a theoretical distribution as in example 1. But it should be noticed that the series obtained by this method has, in some practical cases as in example 2, far greater graduating ability than Pearson's frequency curve.

I. Properties of Binomial Distribution

Let p be the a priori probability of occurrence of an event in a single trial and n the total number of trials. When p remains constant throughout the trials, the resulting binomial probability distribution is given by

$$u_i = {}_nC_i p^i q^{n-i}, \quad i=0, 1, 2, \dots, n; \quad (1)$$

where $q=1-p$. (2)

If we denote the abscissa of the mean of a frequency distribution by d and the r th statistical moment of it about the mean by ν_r , $r=2, 3, 4$, using the summation method of Laplace², we have in this case

$$d=np, \quad (3)$$

$$\nu_2=npq, \quad (4)$$

$$\nu_3=npq(q-p), \quad (5)$$

$$\nu_4=npq[1+3(n-2)pq]. \quad (6)$$

Eliminating n , p and q from (4), (5), (6) and (2), we get

$$\nu_4 = 3\nu_2^2 - \frac{1}{2}\nu_2 + \frac{3}{2} \frac{\nu_3^2}{\nu_2};$$

1. Of course, ν_r refers to a total frequency of unity.

2. A. Fisher, The Mathematical Theory of Probabilities, 2nd ed. p. 104.

hence, if we put

$$T = 3\nu_2^2 - \frac{1}{2}\nu_2 + \frac{3}{2} \frac{\nu_3^2}{\nu_2}, \quad (7)$$

we obtain an important property of binomial distributions

$$T - \nu_4 = 0. \quad (8)$$

Also we have

$$\frac{\nu_3}{\nu_2} = q - p, \quad (9)$$

$$\frac{\nu_2}{pq} = n. \quad (10)$$

II. New Method of Graduation

When a given statistical frequency distribution satisfies the conditions

$$T - \nu_4 = 0, \quad (11)$$

$$\left| \frac{\nu_3}{\nu_2} \right| < 1;^1 \quad (12)$$

we can graduate it by the following procedure, assuming that the distribution is binomial.

First, the moments are adjusted as follows:—

$$\begin{aligned} \text{Let } \nu_2 + \xi &= \lambda_2, \\ \nu_3 + \eta &= \lambda_3, \\ \nu_4 + \zeta &= \lambda_4; \end{aligned} \quad (13)$$

and put

$$3\lambda_2^2 - \frac{1}{2}\lambda_2 + \frac{3}{2} \frac{\lambda_3^2}{\lambda_2} = \lambda_4. \quad (14)$$

Neglecting the 2nd and the higher powers of ξ , η , ζ ; we get by expansion

$$\left(6\nu_2 - \frac{1}{2} - \frac{3}{2} \frac{\nu_3^2}{\nu_2^2} \right) \xi + 3 \frac{\nu_3}{\nu_2} \eta - \zeta + (T - \nu_4) = 0. \quad (15)$$

Calculating the value of ξ , η , ζ , which makes $\xi^2 + \eta^2 + \zeta^2$ minimum under the condition (15), we have

1. Strictly speaking $\left| \frac{\lambda_3}{\lambda_2} \right| < 1$; cf. p. 197.

$$\begin{aligned}\xi &= \frac{AD}{E}, \\ \eta &= \frac{BD}{E}, \\ \zeta &= \frac{CD}{E};\end{aligned}\tag{16}$$

$$\text{where } A = 6\nu_2 - \frac{1}{2} - \frac{3}{2} \frac{\nu_3^2}{\nu_2^2},$$

$$B = 3 \frac{\nu_3}{\nu_2},$$

$$C = -1,\tag{17}$$

$$D = -(T - \nu_4),$$

$$E = A^2 + B^2 + C^2.$$

We take $\lambda_2, \lambda_3, \lambda_4$, thus obtained, as the adjusted moments.

Next, put

$$\frac{\lambda_3}{\lambda_2} = q - p,\tag{18}$$

taking account of (2), find the value of p and q , and put

$$\frac{\lambda_2}{pq} = n.\tag{19}$$

Then compute $\log(Nq^n)$, where N is the observed total frequency. Take an integer n_0 such that

$$n - \frac{1}{2} < n_0 \leq n + \frac{1}{2},\tag{20}$$

and adding to $\log(Nq^n)$ successively $\log\left(\frac{n-r+1}{r} \frac{p}{q}\right)$, $r=1, 2, \dots, n_0$, we get

$$\log(N_n C_r p^r q^{n-r}), \quad r = 0, 1, 2, \dots, n_0.\tag{21}$$

When, for some values of r , $\log(N_n C_r p^r q^{n-r})$ is smaller than -2 , it is generally useless to calculate the value.

It is necessary that $\left|\frac{\lambda_3}{\lambda_2}\right| < 1$, because p and q must both be positive. This is the reason why we have put the approximate condition (12).

Finally, superpose the theoretical range upon the observed so that both means coincide.

Using the method of interpolation in the series of ordinates $(21)^1$, find the ordinate t at the point corresponding to the mid point of a class of the observed range. Then 10^t or the nearest integral value to it is the graduated frequency of the class.

As for the interpolation, Everett's formula² is used, differences of higher than the 5th order being neglected, and x being put equal to the distance between the integral point of the theoretical range and the nearest (to the right of it) mid point of the class of the observed range.

Thus the graduation finishes.

III. Numerical Examples

Let us apply this new method to frequency distributions, and see whether the fit is good or not.

Example 1. The following table shows a theoretical example, the given frequency f being the terms of the binomial series $1000 \times (0.7 + 0.3)^{20}$. (Figures given to the nearest unit.)³

Given frequency f	Entry $\log(N, C_r p^r q^{n-r})$ $r=0, 1, 2, \dots, 15$	Interpolated value when $x=0.1076$	Graduated frequency f'	$\frac{(f-f')^2}{f'}$
1	1.81810	1.92601	1	0
7	0.76759	0.84799	7	0
28	1.39306	1.44710	28	0
72	1.81817	1.85431	71 ⁴	0.014
130	2.09263	2.11440	130	0
179	2.24285	2.25234	179	0
192	2.28472	2.28323	192	0
164	2.22838	2.21676	165	0.006
114	2.08035	2.05915	115	0.009
65	1.84464	1.81414	65	0
31	1.52329	1.48363	30	0.033
12	1.11662	1.06773	12	0
4	0.62329	0.56493	4	0
1	0.04013	1.97185	1	0
1000	1.36169	1.28066	1000	$\chi^2=0.062$
	2.57945			

1. We take this for the entry, because, if we put $x=i-\frac{n}{2}$, $\log u_i$ is expansible in a convergent power series of x , provided $|x| < \frac{n}{2}$.

2. H. I. Rietz, Handbook of Mathematical Statistics, p. 40.

3. G. U. Yule, An Introduction to the Theory of Statistics, 6th ed. p. 294.

4. The actual value is 71.50; we take 71 as the nearest integral value, so that the total frequency shall amount to 1000.

The constants are as follows :

mean = 5.9940	$\xi = -0.01788$	$\frac{\lambda_3}{\lambda_2} = 0.37083$
$v_2 = 4.19996$	$\eta = -0.00081$	$q = 0.68542$
$v_3 = 1.55160$	$\zeta = 0.00073$	$p = 0.31458$
$v_4 = 51.24033$	$\lambda_2 = 4.18208$	$n = 19.3960$
$T = 51.67982$	$\lambda_3 = 1.55079$	theoretical mean
$T - v_4 = 0.43949$	$\lambda_4 = 51.24106$	$np = 6.1016$

To find the value of x in Everett's formula, let

M be the point where the means coincide,

P_i be the point on the graduated (or theoretical) range whose abscissa is i , referred to the point where the frequency Nq^n stands as origin and

Q_i be the point on the given (or observed) range whose abscissa is t , referred to the mid point of the first class as origin ;

then we have

$$\overrightarrow{P_6 M} = 6.1016 - 6 = 0.1016$$

$$\text{and } \overrightarrow{M Q_6} = 6 - 5.9940 = 0.0060,$$

$$\therefore \overrightarrow{P_6 Q_6} = 0.1016 + 0.0060 = 0.1076,$$

as Q_6 is the nearest mid point to P_6 (on the right of it)

$$x = 0.1076.$$

Completing the graduation and finding the value²

$$\chi^2 = \sum_{i=0}^n \left\{ \frac{(f_i - f'_i)^2}{f'_i} \right\}$$

to be 0.062 ; we see that the fit is, in this case, satisfactorily good.

Example 2. For the following actual data and its graduation by Pearson's normal curve of error, readers are referred to the Treatise of W. P. Elderton, Frequency Curves and Correlation, 2nd ed. p. 79.

1. $\overrightarrow{P_6 M}$ means the distance between the two points P_6 and M .
2. H. I. Rietz, loc. cit. p. 78.

Observation		Calculation	
Central Age for groups of 5 years of birth	Sum Assured and Bonuses $\div$ 1000 f	Entry $\log (N_n C_r p^r q^{n-r})$ $r=0, 1, 2, \dots, 13$	Interpolated value when $x=0.7998$
17	11	0.22054	0.98769
22	48	1.15960	1.66290
27	124	1.76341	2.08991
32	213	2.15398	2.34581
37	281	2.37899	2.45620
42	295	2.46226	2.43554
47	185	2.41636	2.29070
52	104	2.24701	2.02322
57	40	1.95471	1.62952
62	15	1.53484	1.10009
67	3	0.97641	0.41633
	<u>1319</u>	0.25842	1.54408
		1.33977	
		2.12274	

Straightforward calculation gives :

mean = 4.4405	$\xi = 0.00944$	$\frac{\lambda_3}{\lambda_2} = 0.20590$
$v_2 = 3.15023$	$\eta = 0.00032$	$q = 0.60295$
$v_3 = 0.65026$	$\zeta = -0.00051$	$p = 0.39705$
$v_4 = 28.57146$	$\lambda_2 = 3.15967$	$n = 13.1977$
$T = 28.39772$	$\lambda_3 = 0.65058$	theoretical mean
$T - v_4 = -0.17374$	$\lambda_4 = 28.57095$	$np = 5.2403$

For the value of x , we have

$\overrightarrow{P_5 M} = 5.2403 - 5 = 0.2403$

and $\overrightarrow{MQ_5} = 5 - 4.4405 = 0.5595,$

$\therefore \overrightarrow{P_5 Q_5} = 0.2403 + 0.5595 = 0.7998,$

as Q_5 is the nearest mid point to P_5 (on the right of it)

$x=0.7998.$

Thus the following table is obtained :

	Our Method Binomial Distribution		Pearson's Method Normal Curve of Error	
	Graduated frequency f'	$\frac{(f-f')^2}{f'}$	Graduated frequency f''	$\frac{(f-f'')^2}{f''}$
17	10	0.100	13	0.308
22	46	0.087	40	1.600
27	123	0.008	104	3.846
32	222	0.365	202	0.599
37	286	0.087	282	0.004
42	273	1.773	288	0.170
47	195	0.513	214	3.930
52	105	0.010	116	1.241
57	43	0.209	44	0.364
62	13	0.308	13	0.308
67	3	0.000	3	0.000
	1319	$\chi^2_P=3.460$	1319	$\chi^2_P=12.370$

Turning to Pearson's "Tables for Statisticians and Biometricians (p. 26)", when $n'=11$, we find

$$P_T=0.966 \quad \text{for } \chi^2_T=3.460$$

$$\text{and } P_P=0.262 \quad \text{for } \chi^2_P=12.370.$$

Thus, we conclude that the graduation by our new method is, in this case, much more satisfactory than that by Pearson's.

The author wishes here to acknowledge his debt of gratitude to Professor T. Wada for his kindness in reading the original manuscript and for valuable criticisms.

College of Niigata