A NOTE ON ADJUSTMENTS FOR FIRST AND SECOND MOMENTS IN A GROUPED FREQUENCY DISTRIBUTION SPLIT UP INTO SUB-SECTIONS

By J. M. SEN GUPTA

Statistical Laboratory, Calcutta

Objects of the Study : Sheppard's adjustments for grouping in the moments about the mean for frequency curves with high contact at both the ends are well known. These moments refer to the whole of the frequency distribution. In this note the whole range of the variate has been divided up into several sections, and studies have been made as to how grouping affects the first two moments for each section separately. Incidentally, Sheppard's Correction for the second moment has been verified. The normal frequency distribution with mean zero and unit standard deviation has been taken up for the study and the Normal Probability Table published by Sheppard in Biometrical Tables Part I has been used which gives the accumulated probability for the normal distribution at intervals of ·01 of the deviate.

Experimental Procedure : The normal frequency distribution for N = 100,000 was constructed with the help of Sheppard's Tables, and one half of the curve was used in the computing process. The total range for half the curve, which really extends to positive infinity in terms of the deviate, was assumed to terminate for practical purposes at + 4·8, a point after which the contribution to the total frequency is less than 0·5. The accumulated frequencies read off from Sheppard's Tables at intervals of ·05 were differenced successively, and actual frequencies contained within each class interval of ·05 in the deviate were obtained. The frequencies thus obtained were added up in groups of two's three's and so on, and the derived sub-totals at each stage gave the different frequency distributions with ·05, ·10, ·15, ·20, ·30, ·35, ·40, ·50, ·60, ·80, and 1·20 as class intervals. The sums and sums of squares $S(fx)$ and $S(fx^2)$ with the above class intervals were then calculated directly. The deviate x being measured from the origin, these raw sums and sums of squares required no further correction with reference to the general mean. The sums and sums of squares were however calculated in four sub-totals, corresponding to the four sections into which the entire half-range was arbitrarily split up. Each of these sections represented the portion of the curve intercepted between ordinates at intervals of 1·20 of the deviate. The variation in error, introduced by grouping at different regions in the curve could thus be observed separately. Table 1 gives the sums and sums of squares for one half of the normal curve in four sub-sections for each of the twelve class intervals worked with.

TABLE 1. ABSOLUTE SUMS AND SUMS OF SQUARES FOR ONE HALF OF THE NORMAL CURVE.

Sections	I(·00—1·20) n = 38493		II(1·21—2·40) n = 10687		III(2·41—3·60) n = 804		IV(3·61—4·80) n = 16		Total (·00—4·80) n = 50,000	
h	$S(fx)$	$S(fx^2)$	$S(fx)$	$S(fx^2)$	$S(fx)$	$S(fx^2)$	$S(fx)$	$S(fx^2)$	$S(fx)$	$S(fx^2)$
·05	20480	15189	17182	28263	2179	5901	61	237	39903	50011
·10	20493	15184	17193	28652	2180	5969	62	237	39928	50042
·15	—	—	—	—	—	—	—	—	39969	50093
·20	20544	15104	17236	28769	2196	6097	62	238	40028	50107
·25	—	—	—	—	—	—	—	—	40102	50259
·30	20630	15128	17307	28964	2195	6042	62	239	40194	50373
·35	—	—	—	—	—	—	—	—	40302	50500
·40	20751	15079	17408	29239	2208	6108	62	244	40428	50670
·50	—	—	—	—	—	—	—	—	40729	51039
·60	21099	14925	17694	30035	2242	6290	63	248	41098	51499
·80	—	—	—	—	—	—	—	—	42046	52068
1·20	23096	13857	19237	34026	2412	7236	67	282	44812	56003

Corrections for μ_2 : Table 2 shows the observed μ_2 estimates separately for each of the sections and for the entire curve. The sectional μ_2's however, refer to the general mean and not to the respective sectional means. This has been given for each of the class intervals that have been used.

The true value of μ_2 with an infinitesimal class interval δx, is obviously unity for the entire curve. The true sectional values of μ_2 were available in the Incomplete Normal Moments Tables in Biometric Tables Part 1. The deviations D of the estimated μ_2 values for each of the class intervals by sections and for the total is also being shown in Table 2.

The second differences of $D(h)$ was found to be constant, which at once suggested the form of the fit, as $D(h) = ch^2$. Least square fittings were also tried. The sectional fits were bad, except in the form $D = ch^2$ where c was practically ·0833 for the entire half curve. The graduated deviations accordingly have been shown in Table 2. It will be seen that the total fit is extremely satisfactory, demonstrating that Sheppard's correction term $h^2/12$ for μ_2 is appropriate and exact for all practical purposes. Even for the highest class interval of 1·20, which gives us only 4 classes for half the curve, the agreement is quite good, and we conclude that, with class intervals upto the order of the standard deviation, Sheppard's adjustments are sufficient, so far as a normal frequency distribution is concerned.

It may be noted that the error in the estimates of sectional μ_2, due to grouping changes its sign, being negative within the region 0 — 1·20 and positive for the rest.

Corrections for the sectional mean : We shall now consider errors in the estimates of the mean referring to portions of a full curve contained within stated intervals of the deviate. Table 3, analogous to Table 2, is showing the mean values for one half of the normal curve in four sub-sections calculated with different class intervals 'h'. It will be seen, that the means have increasingly been over estimated with increasing 'h'. The true value for these sectional means with infinitesimal class interval can however be theoretically obtained by direct intergration. We find that the deviations of the different estimates for the entire half based on different sizes of class-intervals from the true Mean fits excellently in the form $D = h^2/15$ except for the largest class intervals used. As regards sectional fits we find that a correction term of the form h^2/c (where c depends on the distance from the origin of the region in which the moment is being calculated) gives good fits.

TABLE 2. DEVIATION IN VALUES OF μ_1 BY SECTIONS OF A FULL CURVE (TWO SIDES POOLED).

Section	I(·00—1·20) N = 70086		II(1·21—2·40) N = 21374		III(2·41—3·60) N = 1608		IV(3·61—4·80) N = 32		Total (·00—4·80) N = 100000	
	$\mu_1(0)$ = ·30464		$\mu_1(0)$ = 2·6775		$\mu_1(0)$ = 7·4104		$\mu_1(0)$ = 14750		$\mu_1(0)$ = 1·0000	
h	D	$-h^2/54$	D	$+h^2/27$	D	$+h^2/9$	D		D	$+h^2/15$
·05	−5	−5	8	9	38	28	62		21	21
·10	−18	−19	35	37	139	111	63		84	83
·15									189	188
·20	−70	−74	145	148	486	444	125		335	333
·25									519	521
·30	−163	−167	329	333	1045	1000	188		747	750
·35									1018	1021
·40	−291	−290	584	593	1866	1778	500		1341	1333
·50									2079	2083
·60	−691	−667	1329	1333	4130	4000	750		2997	3000
·80									5336	5333
1·20	−3165	−2667	5625	5333	15890	16000	2875		12003	12000

TABLE 3. ESTIMATED MEANS WITH DIFFERENT CLASS INTERVALS (h).

	N = 38493 I(·00—1·20)		II(1·21—2·40) N = 10687		III(2·41—3·60) N = 804		IV(3·61—4·80) N = 16		Total (·00—4·80) N = 50,000	
	$\mu_1(0)$ = ·53191		$\mu_1(0)$ = 1·6076		$\mu_1(0)$ = 2·7090		$\mu_1(0)$ = 3·8125		$\mu_1(0)$ = ·79789	
h	D	$h^2/22$	D	$h^2/7·5$	D	$h^2/4·5$	D		D	$h^2/15$
·05	·00013	·00011	·0001	·0003	·0012	·0006	·0012		·00016	·00017
·10	47	47	12	13	24	22	313		66	67
·15									150	150
·20	180	182	52	53	99	90	375		268	267
·25									416	417
·30	403	409	118	120	211	200	600		598	600
·35									816	817
·40	718	719	213	213	373	356	875		1068	1067
·50									1669	1667
·60	1622	1636	481	480	790	800	1250		2407	2400
·80									4304	4268
1·20	6810	6545	1024	1920	2010	3200	3875		9834	9600

Table 4 shows the observed and graduated deviations due to grouping for sections starting from the origin and extended up to the limits 1·20σ, 2·40σ, 3·00σ and 4·80σ. It will be seen that the corrections required is maximum for the first and falls rapidly to the limiting value of $h^2/15$ for the entire half.

TABLE 4. DEVIATIONS IN μ_1 UPTO STATED LIMITS FROM THE ORIGIN.

Range	·00-1·20		·00-2·40		·00-3·60		·00-4·80	
h	D	$h^2/22$	D	$h^2/15·23$	D	$h^2/15$	D	$h^2/15$
·05	·00013	11	·00014	16	·00016	17	·00018	17
·10	47	45	63	65	67	67	66	67
·20	180	182	254	262	268	267	268	267
·30	403	409	573	590	599	600	598	600
·40	718	719	1025	1040	1065	1067	1068	1067
·60	1623	1636	2314	2361	2405	2400	2407	2400
1·20	6810	6545	9512	9442	9828	9600	9834	9600

The error due to grouping in the sectional mean as percentage of the true mean of the half-curve, measured from the general mean, may be expressed in the form $100(2k)^2\sigma/15n^2\sqrt{(2/\pi)} = 100(2k)^2\sigma/12n^2$ approximately where the range is confined within Mean $\pm k\sigma$, n being the number of cells used and σ the standard deviation. In routine work it is often a practice to split up the range confined within $\pm 3\sigma$ from the general mean into 12 intervals. The error in the estimate of the Mean of the half-curve in such a case expressed as percentage to the true mean corresponding reduces to $100(6)^2\sigma/(12 \times 12^2)$ or 2·1 per cent with σ equal to unity. In table 3 the p.c. deviation for $h = 1·20$ where the full range is covered by 8 cells with $\sigma = 1$ is = 12 per cent against an observed percentage deviation of ·09834 × 100/·79789 = 12·3 per cent.

Paper received : 6th June, 1943