EDITORIAL

THE FUNDAMENTAL NATURE AND PROOF OF SHEPPARD'S ADJUSTMENTS

In the course of our discussion of moment adjustments, we shall have occasion to refer to the following lengthy distribution of discrete variates. By selecting

TABLE 1

*Distribution of the number of items correctly recorded by 244 students in a five minute code transcription test**

Score x	Freq f	Score x	Freq. f	Score x	Freq. f
64	1	94	3	119	1
66	2	95	5	120	2
68	2	96	3	121	6
69	1	97	3	122	2
70	1	98	12	123	3
71	3	99	4	124	2
72	3	100	5	125	6
73	3	101	6	126	3
76	1	102	8	127	4
77	2	103	6	128	2
78	3	104	8	130	2
79	1	105	9	131	1
80	2	106	5	132	5
82	2	107	3	133	1
83	3	108	3	134	1
84	2	109	4	136	1
85	6	110	2	138	1
86	3	111	4	140	1
87	1	112	7	141	1
88	2	113	5	142	2
89	4	114	5	144	2
90	4	115	7	153	1
91	5	116	8	155	1
92	2	117	3		
93	4	118	2	Total	244

* I am indebted to Professor J. A. Gengerelli, of the Department of Psychology of Univ. of California at Los Angeles, for these data.

the provisional mean, $M_0 = 105$, we find that

$$\Sigma x f = -129 \qquad \Sigma x^3 f = -52\ 005$$

$$\Sigma x^2 f = 77\ 591 \qquad \Sigma x^4 f = 69\ 239\ 951.$$

Let us now form the nine possible distributions of grouped-discrete variates that arise from the nine possible "groupings of nine." These are presented in table 2.

TABLE 2

Distributions derived from the data of table 1 by making the nine possible "groupings of nine"

First significant class interval of distribution								
(1) 64–72	(2) 63–71	(3) 62–70	(4) 61–69	(5) 60–68	(6) 59–67	(7) 58–66	(8) 57–65	(9) 56–64
13	10	7	6	5	3	3	1	1
12	15	16	16	14	14	13	15	15
27	23	21	20	22	21	16	14	11
41	41	33	32	30	28	31	29	30
53	54	63	61	55	52	49	45	41
45	45	40	38	42	45	44	48	52
27	27	29	34	36	39	40	42	43
16	19	24	25	23	24	28	30	29
8	6	7	6	10	10	12	11	13
1	2	2	4	5	6	6	7	7
1	2	2	2	2	2	2	2	1
								1

Let us now compute the values of Σxf, Σx^2f, Σx^3f and Σx^4f for each of the distributions of table 2, selecting $M_0 = 105$ in each instance in order to facilitate a comparison of these results with those for table 1. Thus, in spite of what would otherwise be called poor computing technique, we shall use the following class marks as values of x for the first distribution above; $-37, -28, -19, \cdots, 35, 44, 53$. For the second we shall likewise use, $-38, -29, -20, \cdots, 34, 43, 52$, respectively.

TABLE 3

Summations derived from the distributions listed in table 2, using $M_0 = 105$

Dist.	Σxf	Σx^2f	Σx^3f	Σx^4f
(1)	— 181	77 149	— 134 191	69 063 265
(2)	— 218	78 466	— 54 602	74 519 962
(3)	— 111	77 769	2 889	71 465 409

TABLE 3—*Continued*

Dist.	Σxf	Σx^2f	Σx^3f	Σx^4f
(4)	— 139	79 747	— 23 311	74 171 443
(5)	— 104	81 934	19 666	76 143 874
(6)	— 87	80 145	16 551	72 467 541
(7)	— 52	80 302	— 36 118	71 851 930
(8)	— 89	78 553	— 101 357	68 426 497
(9)	— 180	78 894	— 180 792	73 155 150
Average	— 129	79 217⅔	— 54 585	72 362 785⅔

The fact that the average of the values of Σxf appearing in table 3 suggests that no adjustments of the first moment is necessary and that the variations in the nine values for Σxf may be regarded as *accidental errors* and attributed to grouping. An attempt to account for this phenomenon and also for the fact that the averages of the higher order summations of table 3 do not likewise agree with the corresponding summations of table 1 lead us directly to formulae for Sheppard's adjustments.

For the moment, let us concentrate our intention upon a single variate, x_0, and its associated frequency, f_{x_0}, that are a part of a distribution of discrete variates, such as table 1. Suppose we were to form the k different distributions arising from the k possible "groupings of k." In one of these distributions, x_0 will rest in the first position of a class interval: the limits of this class are x_0 and $(x_0 + k - 1)$ and the class mark is therefore $[x_0 + \frac{1}{2}(k - 1)]$. The contribution of the variate, x_0, to $\Sigma x^s f$ for this particular distribution is therefore

$$[x_0 + \tfrac{1}{2}(k - 1)]^s \cdot f_{x_0}.$$

If x_0 rests in the second position of a class, the limits of this class will be $(x_0 - 1)$ and $(x_0 + k - 2)$ and the corresponding class mark is $[x_0 + \frac{1}{2}(k - 3)]$ and the contribution of x_0 to $\Sigma x^s f$ for this distribution is

$$[x_0 + \tfrac{1}{2}(k - 3)]^s \cdot f_{x_0}.$$

The *expected* value of $\Sigma x^s f$ arising from the k different groupings of variates is therefore,

$$(1) \qquad E(\textstyle\sum x^s f) = \frac{1}{k}\left[\overset{1}{\sum} x^s f + \overset{2}{\sum} x^s f + \cdots + \overset{k}{\sum} x^s f\right]$$

where $\overset{i}{\sum} x^s f$ refers to that distribution in which a specified x_0 rests in the i-th position in the class in which it occurs. The contribution of x_0 to this expected value is therefore

$$(2)\quad \frac{1}{k}\{[x_0+\tfrac{1}{2}(k-1)]^s+[x_0+\tfrac{1}{2}(k-3)]^s+[x_0+\tfrac{1}{2}(k-5)]^s+\cdots\}f_{x_0},$$

this series consisting obviously of k terms.

Expanding each term of (2) by the binomial theorem yields

$$\frac{1}{k}\left[x_0^s-{}_sC_1\,x_0^{s-1}\left(\frac{k-1}{2}\right)+{}_sC_2\,x_0^{s-2}\left(\frac{k-1}{2}\right)^2-{}_sC_3\,x_0^{s-3}\left(\frac{k-1}{2}\right)^3+\cdots\right]$$

$$\frac{1}{k}\left[x_0^s-{}_sC_1\,x_0^{s-1}\left(\frac{k-3}{2}\right)+{}_sC_2\,x_0^{s-2}\left(\frac{k-3}{2}\right)^2-{}_sC_3\,x_0^{s-3}\left(\frac{k-3}{2}\right)^3+\cdots\right]$$

$$\frac{1}{k}\left[x_0^s-{}_sC_1\,x_0^{s-1}\left(\frac{k-5}{2}\right)+{}_sC_2\,x_0^{s-2}\left(\frac{k-5}{2}\right)^2-{}_sC_3\,x_0^{s-3}\left(\frac{k-5}{2}\right)^3+\cdots\right]$$

etc.

Since s is an integer, series (2) may be written as the sum of the $(s+1)$ terms of the series

$$(3)\qquad [x_0^s\,S_0-{}_sC_1\,x_0^{s-1}\,S_1+{}_sC_2\,x_0^{s-2}\,S_2-{}_sC_3\,x_0^{s-3}\,S_3+\cdots]f_{x_0},$$

where

$$S_i=\frac{1}{k}\left[\left(\frac{k-1}{2}\right)^i+\left(\frac{k-3}{2}\right)^i+\left(\frac{k-5}{2}\right)^i+\cdots\text{ to }k\text{ terms}\right].$$

By the Euler-Maclaurin Sum Formula we have

$$\sum_{x=a}^{b}x^p=\frac{1}{p+1}(b^{p+1}-a^{p+1})+\tfrac{1}{2}(b^p+a^p)+\frac{B_1}{2!}p\,(b^{p-1}-a^{p-1})$$
$$-\frac{B_3}{4!}p^{(3)}(b^{p-3}-a^{p-3})+\frac{B_5}{6!}p^{(5)}(b^{p-5}-a^{p-5})+\cdots,$$

where $p^{(i)}=p(p-1)(p-2)(p-3)\cdots$ to i factors. In our expression for S_i, $a=\frac{1}{2}(k-1)=-b$, and therefore S_i equals zero when i is an odd integer. For even values of i,

$$(4)\qquad S_i=\frac{2}{k}\left\{\frac{(k-1)^i(k+i)}{2^{i+1}(1+i)}+\frac{B_1}{2!}i\left(\frac{k-1}{2}\right)^i-\frac{B_3}{4!}i^{(3)}\left(\frac{k-1}{2}\right)^{i-3}+\frac{B_5}{6!}i^{(5)}\left(\frac{k-1}{2}\right)^{i-5}-\cdots\right\}$$

so that

$$S_0 = 1$$

$$S_2 = \frac{1}{12}(k^2 - 1)$$

$$S_4 = \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$S_6 = \frac{1}{1344}(k^2 - 1)(3k^4 - 18k^2 + 31)$$

etc.

Since expression (3) represents the contribution of any variate, x_0, to the expected value defined by (1), we may obtain by summation

$$(5)\quad E(\sum x^s f) = \sum x^s f + {}_sC_2 \cdot S_2 \cdot \sum x^{s-2} f + {}_sC_4 \cdot S_4 \sum x^{s-4} f + \cdots .$$

To illustrate: if we desire to shorten the distribution of table 1 by forming class intervals of dimension 9,

$$S_2 = \frac{1}{12}(9^2 - 1) = \frac{20}{3}, \qquad S_4 = \frac{1}{240}(9^2 - 1)(3 \cdot 9^2 - 7) = \frac{236}{3},$$

and by formula (5),

$$E(\sum x\ f) = \sum xf = -129$$

$$E(\sum x^2 f) = \sum x^2 f + {}_2C_2 \cdot S_2 \cdot \sum f = 77591 + \frac{20}{3} \cdot 244 = 79217^{2/3}$$

$$E(\sum x^3 f) = \sum x^3 f + {}_3C_2 \cdot S_2 \cdot \sum xf = -52005 + 3 \cdot \frac{20}{3}(-129) = -54585$$

$$E(\sum x^4 f) = \sum x^4 f + {}_4C_2 \cdot S_2 \cdot \sum x^2 f + {}_4C_4 \cdot S_4 \cdot \sum f$$

$$= 69239951 + 6 \cdot \frac{20}{3} \cdot 77591 + \frac{236}{3} \cdot 244 = 72362785^{2/3}.$$

Since these expected values are identical with those computed directly in table 3, we see that formula (5) provides the adjustments necessary to eliminate the effect of the systematic errors caused by grouping.

Dividing both sides of (5) by Σf yields

$$(6)\quad E(\mu_s') = \mu_s' + {}_sC_2 \cdot S_2 \cdot \mu_{s-2}' + {}_sC_4 \cdot S_4 \mu_{s-4}' + {}_sC_6 \cdot S_6 \cdot \mu_{s-6}' + \cdots ,$$

that is

$$E(\mu_1') = \mu_1'$$

$$E(\mu_2') = \mu_2' + \frac{1}{12}(k^2 - 1)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}(k^2 - 1)\,\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}(k^2 - 1)\,\mu_2' + \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$E(\mu_5') = \mu_5' + \frac{10}{12}(k^2 - 1)\,\mu_3' + \frac{5}{240}(k^2 - 1)(3k^2 - 7)\,\mu_1'$$

etc.

In numerical computations we generally prefer to select the class interval as the unit of x and in this case we have

$$E(\mu_1') = \mu_1'$$

$$E(\mu_2') = \mu_2' + \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}\left(1 - \frac{1}{k^2}\right)\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}\left(1 - \frac{1}{k^2}\right)\mu_2' + \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(3 - \frac{7}{k^2}\right)$$

etc.

Ordinarily we are interested in estimating the values of the moments that would have been obtained if we had not used the time-saving device of grouping the variates and therefore we solve the previous set of equations for the moments of the ungrouped distribution and obtain

$$(7)\qquad \begin{cases} \mu_1' = E(\mu_1') \\ \mu_2' = E(\mu_2') - \frac{1}{12}\left(1 - \frac{1}{k^2}\right) \\ \mu_3' = E(\mu_3') - \frac{3}{12}\left(1 - \frac{1}{k^2}\right)E(\mu_1') \\ \mu_4' = E(\mu_4') - \frac{6}{12}\left(1 - \frac{1}{k^2}\right)E(\mu_2') + \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(7 - \frac{3}{k^2}\right) \\ \qquad\qquad \text{etc.} \end{cases}$$

In general we may write, corresponding to formula (6),

$$(8)\qquad \mu_s' = E(\mu_s') - {}_sC_2 \cdot P_2 \cdot E(\mu_{s-2}') + {}_sC_4 \cdot P_4 \cdot E(\mu_{s-4}') - \cdots$$

where

$$P_2 = \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$P_4 = \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(7 - \frac{3}{k^2}\right)$$

$$P_6 = \frac{1}{1344}\left(1 - \frac{1}{k^2}\right)\left(31 - \frac{18}{k^2} + \frac{3}{k^4}\right)$$

$$P_8 = \frac{1}{11520}\left(1 - \frac{1}{k^2}\right)\left(381 - \frac{239}{k^2} + \frac{55}{k^4} - \frac{5}{k^6}\right)$$

$$P_{10} = \frac{1}{33792}\left(1 - \frac{1}{k^2}\right)\left(2555 - \frac{1636}{k^2} + \frac{410}{k^4} - \frac{52}{k^6} + \frac{3}{k^8}\right)$$

$$P_{12} = \frac{1}{5591040}\left(1 - \frac{1}{k^2}\right)\left(1414477 - \frac{910573}{k^2} + \frac{233570}{k^4} - \frac{32410}{k^6} + \frac{2625}{k^8} - \frac{105}{k^{10}}\right).$$

In actual problems we do not know the exact values of the expectations involved in formulae (7) and (8), and are forced to obtain mere approximations by utilizing in their stead the corresponding moments computed from the single chance grouped distribution. These approximations correspond to those employed in the theory of probable error, namely, substitutions of the moments derived from a single sample for the corresponding expected moments of the parent population.

The adjustments so far considered may properly be referred to as *Sheppard's adjustments about a fixed point.* At first thought it might appear that we might obtain corresponding formulae for the expectations of moments *about the mean* by merely dropping the primes in formula (6) and obtain, for example,

$$\mu_2 = E(\mu_2) - \frac{1}{12}(k^2 - 1),$$

but unfortunately this is not true. For example, the exact value for the variance of the distribution of table 1 is $18915563/244^2$. Using the summations of table 3 and computing the variance for each of the nine groupings yields

$$E(\mu_2) = \frac{1}{9.244^2}[18791595 + 19098180 + 18963315 + 19438947 + 19981080 + 19547811 + 19590984 + 19159011 + 19217736] \tag{9}$$
$$= 19309851/244^2.$$

Since $\frac{1}{12}(k^2 - 1) = \frac{1}{12}(9^2 - 1) = 20/3$ we see that

$$\mu_2 < E(\mu_2) - \frac{1}{12}(k^2 - 1).$$

In the theory of sampling we differentiate between the standard errors of moments about a fixed point and the standard error of moments about the mean

of the sample. Apparently writers on the subject of Sheppard's adjustments have overlooked the case of adjustments about the mean, although the solution for the second moment is readily obtained as follows:

$$E(\mu_2) = E(\mu_2' - M^2) = E(\mu_2') - E(M^2)$$
$$= \mu_2' + \frac{1}{12}(k^2 - 1) - \frac{1}{k}(M_1^2 + M_2^2 + \cdots + M_k^2),$$

where M_i represents the mean of the i-th of the k different grouped distributions. Since

$$\mu_2 = \mu_2' - M^2 = \mu_2' - \frac{1}{k}(M_1 + M_2 + \cdots + M_k),$$

$$E(\mu_2) = \mu_2 + \frac{1}{12}(k^2 - 1)$$
$$- \left[\frac{M_1^2 + M_2^2 + \cdots + M_k^2}{k} - \left(\frac{M_1 + M_2 + \cdots + M_k}{k}\right)^2\right].$$

But since for any set of k variates

$$\sigma_\nu^2 = \frac{\Sigma\nu^2}{k} - \left(\frac{\Sigma\nu}{k}\right)^2,$$

we have that

$$E(\mu_2) = \mu_2 + \frac{1}{12}(k^2 - 1) - \sigma_M^2. \tag{10}$$

Referring back to table 3 we find that

$$\sigma_M^2 = \frac{7856}{3.(244^2)}$$

and the numerical results now satisfy equation (10).

For the benefit of those interested in unsolved problems of mathematical statistics we may say that nothing appears to have been written as yet on the most important problem associated with the systematic errors due to grouping. It is of course desirable to eliminate these systematic errors introduced by grouping, but it is even more important to investigate the distribution of the accidental errors that remain after the systematic errors have been eliminated. For example it is gratifying to know that no systematic errors are present in the Σxf column of table 3 and that equation (6) will enable us to add a constant to each summation of the $\Sigma x^3 f$ column so that the mean of these adjusted values will agree with the value $\Sigma x^3 f = -52005$ obtained in table 1. It is rather disconcerting, however, to realize that in actual practice we *may* in the case of discrete variates and *must* in the case of continuous variates select an arbitrary set of class limits for our recorded data, and that after adjustments for grouping

have been made, our estimates of the true values of the moments of the distribution will—as in table 3—depend so much upon the choice of these limits. Thus, the standard error of the mean attributed to grouping is

$$\sigma_M = \frac{1}{244}\sqrt{\frac{7856}{3}} = 0.21,$$

which is about twenty percent as large as the approximation for the standard error of the mean due to sampling from an infinite parent population, namely,

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = 1.15.$$

If one will take the trouble to compute the values of μ_3 and μ_4 for each of the distributions of table 2, utilizing the summations of table 3, and then compute and compare the values of σ_{μ_3} and σ_{μ_4} due to grouping with the corresponding functions associated with sampling, he will realize the seriousness of the situation.

Summary

The formula for Sheppard's adjustments for distributions of grouped discrete variates was first given without proof in the Editorial of Vol. 1, No. 1 of the *Annals* (page 111). The method used to develop the general formula was extremely laborious and paralleled the method used for the case of continuous variates in the *Handbook of Mathematical Statistics*, Chapter 7, except that the calculus of finite differences was employed. A more satisfactory proof of this formula was presented by Dr. J. R. Abernethy in Vol. 4, No. 4 of the *Annals* in an article entitled *"On the Elimination of Systematic Errors Due to Grouping."* An extremely elegant development of the same formula and an extension to the case of two variables appears elsewhere in this volume by Professor C. C. Craig. From the point of view of expectations, all of these developments are adjustments about a fixed point, although this fixed point may be selected arbitrarily at the mean of the distribution in question. The obtaining of formulae for the adjustments about the mean of each grouping and the distribution of the accidental errors that remain after these systematic errors have been removed has apparently been neglected to date and should interest students of mathematical statistics.

From a mathematical standpoint, the development of this paper is the simplest of all that have appeared to date: the adjustments for the first four moments can be worked out with the aid of the binomial considerations leading to formula (3) and the following well known formulae for the sums of the powers of the first n integers:

$$S_1 = \frac{n(n+1)}{2} \qquad S_3 = \frac{n^2(n+1)^2}{4}$$

$$S_2 = \frac{n(n+1)(2n+1)}{6} \qquad S_4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

One should note that the condition of high contact is not required in this paper or in the developments of Abernethy or Craig. The results of the three preceding papers agree with those obtained about a fixed point in this paper, but fail to hold for the case of expectations about the mean, if we accept the following definition:

$$E(\mu_s) = \frac{1}{k}(\mu_{s:1} + \mu_{s:2} + \cdots + \mu_{s:k}), \qquad (s = 2, 3, \cdots)$$

where $\mu_{s:i}$ designates the s-th moment computed about the mean of the i-th grouped distribution, $(1 \leqq i \leqq k)$.

H. C. CARVER.