On the relationship between the approximate moments of an observed distribution and its functional moments

by Chester H Scott

A THESIS Submitted to the Graduate Committee in partial fulfillment of the requirements for the degree of Master of Science in Mathematics
Montana State University

Abstract:
The traditional development of Sheppard's formula is as fol- lows. Choose the class interval as a unit of x and lat xt be the mid-abscissa of the tth class, then the approximate moments as directly computed from the observed distribution are defined by (Formula not captured by OCR) where g(x) is the observed frequency, N is the total frequency, and K is the class width. The functional moments are (Formula not captured by OCR) where f (x) is the theoretical relative frequency. The solution involves a relationship between v'n and u'n which is (Formula not captured by OCR) SInce this approach involved relatively complex mathematics, it seemed desirable to simplify the process. The writer did this by use of the moment-generating function concept. Consider a frequency distribution f (x) divided into any convenient number of classes whoe class mark is x1 and class width K. And further let ς be the error in using the class mark x1 instead of the var- iable x. Then (Formula not captured by OCR) were x and ς are independently dis- tributed, x is the variable before grouping and x1 is the vari- able after grouping. Then (Formula not captured by OCR) After expending and simplifying the above reduces to results identical to that which Sheppard obtained.

ON THE RELATIONSHIP BETWEEN THE APPROXIMATE MOMENTS OF AN OBSERVED DISTRIBUTION AND ITS FUNCTIONAL MOMENTS

by

CHESTER H. SCOTT

A THESIS

Submitted to the Graduate Committee

in

partial fulfillment of the requirements

for the degree of

Master of Science in Mathematics

at

Montana State College

Approved:

John W. Kust
In Charge of Major Work

Joe J. Livers
Chairman, Examining Committee

J. J. Nelson
Chairman, Graduate Committee

Bozeman, Montana
June, 1949

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

It gives me a great deal of pleasure to acknowledge the assistance rendered by my many friends in the preparation of this paper. I wish particularly to thank Dr. J. J. Livers, Montana State College, for his encouragement and assistance.

ABSTRACT

The traditional development of Sheppard's formulae is as follows. Choose the class interval as a unit of x and let x_t be the mid-abscissa of the tth class, then the approximate moments as directly computed from the observed distribution are defined by

$$Nv'_n = \sum_{t=-\infty}^{\infty} x_t^n \left\{ \frac{1}{k} \int_{-k/2}^{k/2} g(x_t + h)dh \right\},$$

where g(x) is the observed frequency, N is the total frequency, and k is the class width. The functional moments are

$$u'_n = \int_{-\infty}^{\infty} x^n f(x)dx,$$

where f(x) is the theoretical relative frequency. The solution involves a relationship between v'_n and u'_n which is

$$v'_n = \sum_{i=0}^{\infty} \frac{k^{2i}\, {}_nC_{2i}}{2^{2i}(2i+1)}\, u'_{n-2i}.$$

Since this approach involved relatively complex mathematics, it seemed desirable to simplify the process. The writer did this by use of the moment-generating function concept. Consider a frequency distribution f(x) divided into any convenient number of classes whose class mark is x_i and class width k. And further let ϵ be the error in using the class mark x_i instead of the variable x. Then $x_i = x + \epsilon$, where x and ϵ are independently distributed, x is the variable before grouping and x_i is the variable after grouping. Then

$$M_{x_i}(\theta) = M_x(\theta) \cdot M_\epsilon(\theta).$$

After expanding and simplifying the above reduces to results identical to that which Sheppard obtained.

INTRODUCTION

In the general process of interpreting the data of a sample, one finds it desirable to obtain some of its statistical moments. When dealing with large quantities of data, it is often convenient to arrange this data into frequency tables with a number of equal class intervals. (This is done to make the calculation of moments less laborious.) It is the practice of statisticians to regard each element of a class as having the value of the mid-abscissa of that class. The procedure involved introduces some error in the final results. It is the writer's objective to develop formulae to correct this error.

For most data arranged in frequency tables experience indicates that one should use from 10 to 20 classes. Any fewer would lead to a loss of accuracy; whereas more would only tend to unnecessarily increase the arithmetic involved.

It can be said that any group of data one considers is just a portion of a very large population, called the parent population. The exact characteristics of this larger group may not be known, but it can sometimes be represented by a theoretical frequency distribution of a continuous variable x and be denoted by f(x). It is defined as that function for which $\int_a^b f(x)dx = P[a<x<b]$ where a and b are two values of x with a<b and $P[a<x<b]$ represents the probability or theoretical relative frequency with which x will fall between a and b, providing the data or samples have been chosen at random. The total area under the graph of the distribu-

tion function f(x) and above the x-axis is defined as unity. Therefore many functions could not serve as mathematical models for an observed distribution.

For classified data, the kth sample moment about the origin is defined by $m'_k = \frac{1}{n}\sum_{i=1}^{h} x_i^k f_i$, where n is the total frequency, h is the number of intervals, x_i is the class mark of the ith class interval, and f_i is the frequency for the ith interval. The theoretical kth moment about the origin, denoted by u'_k is defined by $u'_k = \int_a^b x^k f(x)dx$, where f(x) is the theoretical relative frequency and is defined over the interval (a,b).

The first moment, a measure of central tendency, is called the mean and may be denoted by m'_1 or $\bar{x}$ if the data is classified and u'_1 if from a theoretical distribution: m'_1 or $\bar{x} = \frac{1}{n}\sum_{i=1}^{h} x_i f_i$ and $u'_1 = \int_a^b xf(x)dx$.

The second moment is a measure of variation. In the latter case it is often convenient to reduce this measure to the same unit as that of the data. Therefore $\sqrt{m_2} = s$ is usually used, where m_2 is defined by $m_2 = \frac{1}{n}\sum_{i=1}^{h} (x_i-\bar{x})^2 f_i$. This actually is the second moment about the mean. The quantity s is called the standard deviation. The second theoretical moment about the mean is defined by $u_2 = \int_a^b (x-u'_1)^2 f(x)dx$. The standard deviation applied here is $\sigma = \sqrt{u_2}$.

The third moment about the mean may be used as a measure of skewness of the distribution. For a symmetrical distribution this moment is zero, because, for each deviation from the right of the

mean there is an equal deviation from the left of the mean, so that when these deviations are cubed and multiplied by the class frequency they will cancel each other in the summation. However, if the distribution has a long right tail (this is said to be skewness to the right), the third moment will be positive because these large ppsitive deviations when cubed and multiplied by their class frequencies will more than overbalance the relatively smaller deviations on the left. In order that this measure be independent of the unit of x and also independent of the choice of origin skewness or α_3 may be defined by $\alpha_3 = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{s^3}$ for grouped data and $\alpha_3 = \frac{u_3}{u_2^{3/2}} = \frac{u_3}{\sigma^3}$ for the model $f(x)$.

Quite often the fourth moment about the mean is used to determine the extent of peakedness of the graph. If two distributions have the same standard deviation but one of them has a very large percentage of its data concentrated about the mean, then that one will usually have considerably longer tails to compensate for the concentration of data about the mean. Because of the heavy contribution of the fourth powers of these tails in the fourth moment, the peaked distribution with long tails will tend to have a relatively larger fourth moment about the mean. It does not necessarily follow that this is always true. Distribution functions can be constructed for which this is not true. However, the fourth moment is useful in the great majority of cases to determine a measure of peakedness. In order to obtain a measure of peakedness, often spoken of as kurtosis, which is independent of the considered unit

of measurement, one divides this fourth moment by m_2^2. If α_4 represents this measure of kurtosis, then $\alpha_4 = \frac{m_4}{m_2^2} = \frac{a_4}{s^4}$ for classified data and $\alpha_4 = \frac{u_4}{u_2^2} = \frac{u_4}{\sigma_4}$ for the theoretical frequency curve. In most elementary treatments of the subject writers make no concrete interpretations of moments beyond the fourth.

Let us assume that the data being considered was arranged into frequency tables. In such a case the range of variation of x would be divided into a number of equal class intervals. For simplicity of calculations of moments it is the practice to regard each frequency of the class as having the value of the class mark of the interval in which it falls. This method introduces some error in the final results. It is the purpose of this thesis to develop formulae to correct this error.

There has been considerable work done on the subject, but in most instances the development of these formulae has involved relatively complex mathematics. It is the writer's intention to devise a simpler and more understandable approach.

DEVELOPMENT OF PROBLEM

"This subject, which is of prime importance to an understanding of the theory of frequency distributions has been treated in considerable detail in papers[1] by Pearson, Filon, and Sheppard."[2] In the literature relationships between the corrected and uncorrected moments of a grouped distribution are known as "Sheppard's corrections" after W. F. Sheppard. His approach to the problem is roughly as follows. Let us choose the class interval as a unit of x and let x_t be the mid-abscissa of the tth class, then the approximate moments as directly computed from the observed distribution are defined by $Nv'_n = \sum_{t=-\infty}^{\infty} x_t^n \int_{-k/2}^{k/2} g(x+h)dh$, where g(x) is the observed frequency, N is the total frequency and k is the class width. The functional moments are $u'_n = \int_{-\infty}^{\infty} x^n f(x)dx$, where f(x) is the theoretical relative frequency. The solution to the problem involves a relationship between u'_n and v'_n.

If g(x) can be expanded by Taylor's theorem:

$$g(x_t+h) = \sum_{i=0}^{\infty} \frac{h^i}{i!} g^{(i)}(x_t),$$

---

1. Pearson, K., "On the Probable Errors of Frequency Constants," Biometrika, vol. 2 (1903), pp. 273-81.

   Pearson, K., and Filon, L. N. G., "On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation," Phil. Trans. A., vol. 191 (1898), pp. 229-311.

   Sheppard, W.F., "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlations," Phil. Trans. A., vol. 192 (1898), pp. 101-67.

2. Rietz, H.L., Handbook of Mathematical Statistics, p. 95, Houghton Mifflin Co., New York, 1924.

then

$$\frac{1}{k}\int_{-k/2}^{k/2} g(x_t+h)dh = \int_{-k/2}^{k/2} \frac{1}{k}\sum_{i=0}^{\infty} \frac{h^i}{i!} g^{(i)}(x_t)dh$$

and

$$\frac{1}{k}\int_{-k/2}^{k/2} g(x_t+h)dh = \sum_{i=0}^{\infty} \frac{k^{2i}}{s^{2i}(2i+1)!} g^{(2i)}(x_t)$$

and therefore

$$Mv_n' = \sum_{t=-\infty}^{\infty} x_t^n \left\{ \sum_{i=0}^{\infty} \frac{k^{2i}}{s^{2i}(2i+1)!} g^{(2i)}(x_t) \right.$$

$$= \sum_{i=0}^{\infty} \frac{k^{2i}}{s^{2i}(2i+1)!} \left\{ \sum_{t=-\infty}^{\infty} x_t^n g^{(2i)}(x_t) \right\}.^{3}$$

If $g(x)$ is such that it and all its derivatives vanish at $x = \pm\infty$, then by the Maclaurin Sum formula,

$$Mv_n' = \sum_{i=0}^{\infty} \frac{k^{2i}}{s^{2i}(2i+1)!} \left\{ \int_{-\infty}^{\infty} x^n g^{(2i)}(x)dx \right\}.$$

By successive integration by parts

$$\int_{-\infty}^{\infty} x^n g^{(2i)}(x)dx = k^{2i} n(n-1)(n-2) \cdots (n-2i-1) \int_{-\infty}^{\infty} x^{n-2i} g(x)dx$$

$$= k^{2i}(2i)!\, {}_nC_{2i} \int_{-\infty}^{\infty} x^{n-2i} g(x)dx$$

$$v_n' = \sum_{i=0}^{\infty} \frac{k^{2i}\, {}_nC_{2i}}{s^{2i}(2i+1)} u_{n-2i}'$$

$$= u_n' + \frac{k^2\, {}_nC_2}{3 \cdot s^2} u_{n-2}' + \frac{k^4\, {}_nC_4}{5 \cdot s^4} + \cdots$$

$v_0' = u_0' = 1$ which is the area under $f(x)$

and above the x-axis.

---

3. With sufficient restrictions on the functions involved one can change the order of summation.

The primed moments are about any arbitrary origin. If the mean is chosen as that origin, the primes may be dropped. Moments up to the sixth are

$$u_1 = v_1 = 0$$

$$u_2 = v_2 - \frac{k^2}{12}$$

$$u_3 = v_3$$

$$u_4 = v_4 - \frac{u_2 k^2}{2} - \frac{k^2}{80}$$

$$u_5 = v_5 - \frac{5u_3 k^2}{6}$$

$$u_6 = v_6 - \frac{5u_4 k^2}{4} - \frac{3u_2 k^4}{16} - \frac{k^6}{448}$$

Certainly it would be desirable to know whether these so called "corrections" corrected and if so, how well. Tables I, II, III and IV will tend to illustrate this. Table IV shows the results of applying the "corrections" to the moments of the frequency distributions in tables I, II and III.

TABLE I

TEN THOUSAND TIMES THE AREA UNDER A PEARSON TYPE III CURVE (WITH $\alpha_3 = 0.5$)[4]

u_i	f_i	$u_i f_i$	$u_i^2 f_i$	$u_i^3 f_i$	$u_i^4 f_i$	$u_i^5 f_i$	$u_i^6 f_i$
-12	1	- 12	144	- 1728	20736	- 248832	2985984
-10	23	- 230	2300	- 23000	230000	-2300000	23000000
- 8	196	-1568	12544	-100352	802816	-6422528	51380224
- 6	706	-4236	25416	-152496	914976	-5489856	32939136
- 4	1438	-5752	23008	- 92032	368128	-1472512	5890048
- 2	1955	-3910	7820	-15640	31280	- 62560	125120
0	1966	0	0	0	0	0	0
2	1567	3134	6268	12536	25072	50144	100288
4	1037	4148	16592	66369	265472	1061888	4247552
6	591	3546	21276	127656	765936	4595616	27573696
8	296	2368	18944	151552	1212416	9699328	77594624
10	136	1360	13600	136000	1360000	13600000	136000000
12	55	660	7920	95040	1140480	13685760	164229120
14	22	308	4312	60368	845152	11832128	165649792
16	7	112	1792	28672	458752	7340032	117440512
18	2	36	648	11664	209952	3779136	68024448
20	2	40	800	16000	320000	6400000	128000000
Total	10,000	4	163,384	320,608	8,971,168	56,047,744	1,005,180,544

4. This table was condensed from "Pearson's Type III Function" by R. L. Salvosa, Annals of Mathematical Statistics, vol. 1, No. 2, May 1930.

TABLE II

TEN THOUSAND TIMES THE AREA UNDER A PEARSON TYPE III CURVE (WITH $\alpha_3 = 0.5$) [5]

u_i	f_i	$u_i f_i$	$u_i^2 f_i$	$u_i^3 f_i$	$u_i^4 f_i$	$u_i^5 f_i$	$u_i^6 f_i$
-11	5	- 55	605	- 6655	73205	- 805255	8857805
- 9	77	- 693	6237	- 56133	505197	-4546773	40920957
- 7	405	-2835	19845	-138915	972405	-6806835	47647845
- 5	1069	-5345	26725	-133625	668125	-3340625	16703125
- 3	1750	-5250	15750	- 47250	141750	- 425250	1275750
- 1	2027	-2027	2027	- 2027	2027	2027	2027
1	1800	1800	1800	1800	1800	1800	1800
3	1302	3906	11718	35154	105462	316386	949158
5	796	3980	19900	99500	497500	2487500	12437500
7	425	2975	20825	145775	1020425	7142975	50000825
9	202	1818	16362	147258	1325322	11927898	107351082
11	88	968	10648	117128	1288408	14172488	155897368
13	35	455	5915	76895	999635	12995255	168938315
15	12	180	2700	40500	607500	9112500	136687500
17	4	68	1156	19652	334084	5679428	96550276
19	3	57	1083	20577	390963	7428297	141137643
Total	10,000	2	163,296	319,634	8,933,808	55,337,762	985,358,976

5. This table was condensed from "Pearson's Type III Function" by R. L. Selvoss, Annals of Mathematical Statistics, vol. 1, No. 2, May 1930.

TABLE III

ONE HUNDRED THOUSAND TIMES THE AREA UNDER A NORMAL CURVE

t_i	f_i	$t_i f_i$	$t_i^2 f_i$	$t_i^3 f_i$	$t_i^4 f_i$	$t_i^5 f_i$	$t_i^6 f_i$
-4.0	9	- 36.0	144.00	- 576.00	2304.00	- 9216.00	36864.00
-3.5	49	- 171.5	600.25	- 2100.88	7353.06	-25735.72	90075.02
-3.0	240	- 720.0	2160.00	- 6480.00	19440.00	-58320.00	174960.00
-2.5	924	- 2310.0	5775.00	-14437.50	36093.75	-90234.38	225585.94
-2.0	2784	- 5568.0	11136.00	-22272.00	44544.00	-89088.00	178176.00
-1.5	6559	- 9838.5	14757.75	-22136.62	33204.94	-49807.41	74711.11
-1.0	12098	-12098.0	12098.00	-12098.00	12098.00	-12098.00	12098.00
-0.5	17466	- 8733.0	4366.50	- 2183.25	1091.62	- 545.81	272.91
0.0	19742	0.0	00.00	00.00	00.00	00.00	00.00
0.5	17466	8733.0	4366.50	2183.25	1091.62	545.81	272.91
1.0	12098	12098.0	12098.00	12098.00	12098.00	12098.00	12098.00
1.5	6559	9838.5	14757.75	22136.62	33204.94	49807.41	74711.11
2.0	2784	5568.0	11136.00	22272.00	44544.00	89088.00	178176.00
2.5	924	2310.0	5775.00	14437.50	36093.75	90234.38	225585.94
3.0	240	720.0	2160.00	6480.00	19440.00	58320.00	174960.00
3.5	49	171.5	600.25	2100.88	7353.06	25735.72	90075.02
4.0	9	36.0	144.00	576.00	2304.00	9216.00	36864.00
Total	100,000	0.0	102,075.00	00.00	312,258.74	00.00	1,585,485.94

TABLE IV

COMPARISON OF THE UNCORRECTED, CORRECTED, THEORETICALLY EXPECTED MOMENTS FOR TABLES I, II, III[6]

	Table I			Table II			Table III		
n	v_n	v_{cn}	u_n	v_n	v_{cn}	u_n	v_n	v_{cn}	u_n
0	1	1	1	1	1	1	1	1	1
1	0.0004	0.0004	0	0.0002	0.0002	0	0.00	0.00	0.00
2	16.3384	16.005	16	16.3296	16.00	16	1.0208	1.02	1.00
3	32.0688	32.06	32	31.9634	31.96	32	0.00	0.00	0.00
4	897.1168	864.91	864	893.3808	861.16	864	3.1226	3.00	3.00
5	5604.7744	5497.90	5504	5533.7762	5527.28	5504	0.00	0.00	0.00
6	100,538.0544	96,140.33	96,640	98,535.8976	94,181.96	96,640	15.8548	14.91	15.00

6. The v_n represents uncorrected moments, v_{cn} represents corrected moments, and u_n represents theoretically expected moments.