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
© Copyright by Chester H Scott (1949)

Abstract:
The traditional development of Sheppard's formula is as follows. Choose the class interval as a unit of x and let 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 whose class mark is x1 and class width K. And further let ç be the error in using the class mark x1 instead of the variable x. Then (Formula not captured by OCR) were x and ç are independently distributed, x is the variable before grouping and x1 is the variable 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 M. Hust
In Charge of Major Work

Joe J. Lewis
Chairman, Examining Committee

J. J. Nelson
Chairman, Graduate Committee

Bozeman, Montana
June, 1949
TABLE OF CONTENTS

ACKNOWLEDGMENT ............................................................................. 3
ABSTRACT ............................................................................................... 4
INTRODUCTION ....................................................................................... 5
DEVELOPMENT OF PROBLEM ............................................................... 9
CONCLUSIONS ....................................................................................... 20
LITERATURE CITED AND CONSULTED ............................................... 21

LIST OF TABLES

TABLE PAGE
I. Ten Thousand Times the Area under a Pearson Type III Curve (with $x = 0.5$) 12
II. Ten Thousand Times the Area under a Pearson Type III Curve (with $x = 0.5$) 13
III. One Hundred Thousand Times the Area under a Normal Curve 14
IV. Comparison of the Uncorrected, Corrected, Theoretically Expected Moments for Tables I, II, III 15

LIST OF FIGURES

FIGURE PAGE
1. Frequency Distribution 17
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 \( t \)th class, then the approximate moments as directly computed from the observed distribution are defined by

\[
M^i_n = \sum_{t=-\infty}^{\infty} x_t^n \left( \int_{-k/2}^{k/2} g(x_t + h) dh \right),
\]

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

\[
u^i_n = \int_{-\infty}^{\infty} x^n f(x) dx,
\]

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

\[
u^i_n = \sum_{i=0}^{\infty} \frac{k^{2i} n^{2i}}{2^{2i}(2i+1)} u^{i}_{n-2i},
\]

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 \( x_t \) and class width \( k \). And further let \( \varepsilon \) be the error in using the class mark \( x_t \) instead of the variable \( x \). Then \( x_t = x + \varepsilon \), where \( x \) and \( \varepsilon \) are independently distributed, \( x \) is the variable before grouping and \( x_t \) is the variable after grouping. Then

\[
M_{x_t} (\theta) = M_x (\theta) \cdot M_\varepsilon (\theta).
\]

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 \( \int_{a}^{b} f(x) \, dx = P[a \leq x < b] \) where \( a \) and \( b \) are two values of \( x \) with \( a < b \) and \( P[a \leq x < b] \) 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 \( k \)th sample moment about the origin is defined by
\[
m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k f_i,
\]
where \( n \) is the total frequency, \( h \) is the number of intervals, \( x_i \) is the class mark of the \( i \)th class interval, and \( f_i \) is the frequency for the \( i \)th interval. The theoretical \( k \)th 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 \( \overline{x} \) if the data is classified and \( u_1 \) if from a theoretical distribution: \( m_1 = \frac{1}{n} \sum_{i=1}^{n} x_i f_i \) and \( u_1 = \int_a^b x f(x)\,dx \).

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 \( \sqrt{m_2} = \sigma \) is usually used, where \( m_2 \) is defined by
\[
m_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 f_i.
\]
This actually is the second moment about the mean. The quantity \( \sigma \) 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 positive 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 $\alpha_3$ may be defined by $\alpha_3 = \frac{m_3}{s^3} = \frac{u_3}{s^3}$ for grouped data and $\alpha_3 = \frac{u_3}{u_2^{3/2}} = \frac{u_3}{s^2}$ 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 $m_4^2$. If $\alpha_4$ represents this measure of kurtosis, then $\alpha_4 = \frac{m_4}{m_2^2} = \frac{m_4}{s_4^2}$ for classified data and $\alpha_4 = \frac{u_4}{u_2^2} = \frac{u_4}{s_4^2}$ 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.
--9--

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 \(x_0\) be the mid-abscissa of the \(t\)th class, then the approximate moments as directly computed from the observed distribution are defined by

\[ M_n^r = \frac{1}{N} \sum_{i=1}^{k/2} x_n^r g(x_i) \]  

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

\[ u_n^r = \frac{1}{n} \sum_{i=1}^{k/2} x_n^r f(x_i) \]  

where \(f(x)\) is the theoretical relative frequency. The solution to the problem involves a relationship between \(u_n^r\) and \(v_n^r\).

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

\[ g(x_0 + h) = \sum_{i=0}^{\infty} \frac{h^i}{i!} g^{(i)}(x_0) \]


then
\[ \frac{k}{x} \int_{-k/2}^{k/2} g(x_i + h) \, dh = \int_{-k/2}^{k/2} \frac{1}{k} \sum_{i=1}^{\infty} \frac{h_i}{i!} g(i) \, (x_i) \, dh \]
and
\[ \frac{k}{x} \int_{-k/2}^{k/2} g(x_i + h) \, dh = \sum_{i=0}^{\infty} \frac{k^{2i}}{2^{2i}(2i+1)!} g(2i) \, (x_i) \]
and therefore
\[ \mathcal{I}_{\nu_n} = \sum_{t=-\infty}^{\infty} x_i \left\{ \sum_{i=0}^{\infty} \frac{k^{2i}}{2^{2i}(2i+1)!} g(2i) \, (x_i) \right\}^{3} \]

If \( g(x) \) is such that it and all its derivatives vanish at \( x = \pm \infty \), then by the Maclaurin Sum formula,
\[ \mathcal{I}_{\nu_n} = \sum_{i=0}^{\infty} \frac{k^{2i}}{2^{2i}(2i+1)!} \left\{ \int_{-\infty}^{\infty} x^n g(x) \, (x) \, dx \right\}. \]

By successive integration by parts
\[ \int_{-\infty}^{\infty} x^n g(x) \, (x) \, dx = k^{2i} n(n-1)(n-2) \cdots (n-2i-1) \int_{-\infty}^{\infty} x^{n-2i} g(x) \, (x) \, dx \]
\[ = k^{2i} (2i)! \, n_{2i} \int_{-\infty}^{\infty} x^{n-2i} g(x) \, (x) \, dx \]
\[ v_{n-i} = \sum_{i=0}^{\infty} \frac{k^{2i} n_{2i}}{2^{2i}(2i+1)!} \]
\[ = v_{n-i} + \frac{k^{2i} n_{2i}}{2^{2i}} v_{n-2i} + \frac{k^{2i} n_{2i}}{2^{2i}} \cdots \]
\[ v_{Q} = v_{0} = 1 \text{ which is the area under } f(x) \]
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

\[ u_1 = v_1 = 0 \]
\[ u_2 = v_2 - \frac{k^2}{12} \]
\[ u_3 = v_3 \]
\[ u_4 = v_4 - \frac{u_2 k^2}{2} - \frac{k^2}{80} \]
\[ u_5 = v_5 - \frac{5u_3 k^2}{6} \]
\[ u_6 = v_6 - \frac{5u_4 k^2}{4} - \frac{3u_2 k^4}{16} - \frac{k^6}{48} \]

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 $\alpha_3 = 0.5$)

<table>
<thead>
<tr>
<th>$u_1$</th>
<th>$f_1$</th>
<th>$u_1^2f_1$</th>
<th>$u_1^3f_1$</th>
<th>$u_1^4f_1$</th>
<th>$u_1^5f_1$</th>
<th>$u_1^6f_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-12</td>
<td>1</td>
<td>-12</td>
<td>-1728</td>
<td>20736</td>
<td>-243832</td>
<td>298592</td>
</tr>
<tr>
<td>-10</td>
<td>23</td>
<td>-230</td>
<td>-23000</td>
<td>230000</td>
<td>-2300000</td>
<td>23000000</td>
</tr>
<tr>
<td>-5</td>
<td>196</td>
<td>-1568</td>
<td>-12544</td>
<td>-100352</td>
<td>802816</td>
<td>-6422528</td>
</tr>
<tr>
<td>-6</td>
<td>706</td>
<td>-4236</td>
<td>-25416</td>
<td>-152496</td>
<td>914976</td>
<td>-5485865</td>
</tr>
<tr>
<td>-4</td>
<td>1438</td>
<td>-5732</td>
<td>-23008</td>
<td>-92032</td>
<td>363128</td>
<td>-1472512</td>
</tr>
<tr>
<td>-2</td>
<td>1955</td>
<td>-3910</td>
<td>-7820</td>
<td>-15640</td>
<td>31280</td>
<td>62550</td>
</tr>
<tr>
<td>0</td>
<td>1966</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1567</td>
<td>3134</td>
<td>6268</td>
<td>12536</td>
<td>25072</td>
<td>50144</td>
</tr>
<tr>
<td>4</td>
<td>1027</td>
<td>4148</td>
<td>16592</td>
<td>66369</td>
<td>265472</td>
<td>1061888</td>
</tr>
<tr>
<td>6</td>
<td>591</td>
<td>3546</td>
<td>21276</td>
<td>127656</td>
<td>765936</td>
<td>4595616</td>
</tr>
<tr>
<td>8</td>
<td>296</td>
<td>2268</td>
<td>18944</td>
<td>151552</td>
<td>1212416</td>
<td>9699328</td>
</tr>
<tr>
<td>10</td>
<td>136</td>
<td>1360</td>
<td>13600</td>
<td>136000</td>
<td>1360000</td>
<td>13600000</td>
</tr>
<tr>
<td>12</td>
<td>55</td>
<td>660</td>
<td>7920</td>
<td>95040</td>
<td>1104800</td>
<td>1368760</td>
</tr>
<tr>
<td>14</td>
<td>42</td>
<td>308</td>
<td>4312</td>
<td>60368</td>
<td>845152</td>
<td>11832128</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>112</td>
<td>1792</td>
<td>28672</td>
<td>458752</td>
<td>7340032</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>36</td>
<td>648</td>
<td>11664</td>
<td>209952</td>
<td>3779136</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>40</td>
<td>800</td>
<td>16000</td>
<td>320000</td>
<td>6400000</td>
</tr>
<tr>
<td>Total</td>
<td>10,000</td>
<td>4</td>
<td>163,384</td>
<td>320,608</td>
<td>8,971,168</td>
<td>56,047,744</td>
</tr>
</tbody>
</table>

---

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>
<thead>
<tr>
<th>$u_1$</th>
<th>$f_1$</th>
<th>$u_1^2f_1$</th>
<th>$u_1^3f_1$</th>
<th>$u_1^4f_1$</th>
<th>$u_1^5f_1$</th>
<th>$u_1^6f_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-11</td>
<td>5</td>
<td>55</td>
<td>-6655</td>
<td>73205</td>
<td>-805255</td>
<td>8857805</td>
</tr>
<tr>
<td>-9</td>
<td>97</td>
<td>-693</td>
<td>6237</td>
<td>-56223</td>
<td>505197</td>
<td>-4546773</td>
</tr>
<tr>
<td>-7</td>
<td>405</td>
<td>-2825</td>
<td>19845</td>
<td>-138915</td>
<td>972405</td>
<td>-6806835</td>
</tr>
<tr>
<td>-5</td>
<td>1069</td>
<td>-5369</td>
<td>26725</td>
<td>-133625</td>
<td>668125</td>
<td>-3340625</td>
</tr>
<tr>
<td>-3</td>
<td>1750</td>
<td>-5250</td>
<td>15750</td>
<td>-47250</td>
<td>14750</td>
<td>-425250</td>
</tr>
<tr>
<td>-1</td>
<td>2027</td>
<td>-2027</td>
<td>2027</td>
<td>-2027</td>
<td>2027</td>
<td>2027</td>
</tr>
<tr>
<td>1</td>
<td>1800</td>
<td>1800</td>
<td>1800</td>
<td>1800</td>
<td>1800</td>
<td>1800</td>
</tr>
<tr>
<td>3</td>
<td>1302</td>
<td>3936</td>
<td>11718</td>
<td>35134</td>
<td>105282</td>
<td>316986</td>
</tr>
<tr>
<td>5</td>
<td>736</td>
<td>3936</td>
<td>19900</td>
<td>99500</td>
<td>497500</td>
<td>2487500</td>
</tr>
<tr>
<td>7</td>
<td>425</td>
<td>2375</td>
<td>20825</td>
<td>145775</td>
<td>1020425</td>
<td>7129755</td>
</tr>
<tr>
<td>9</td>
<td>202</td>
<td>1815</td>
<td>16365</td>
<td>147258</td>
<td>1325322</td>
<td>11927898</td>
</tr>
<tr>
<td>11</td>
<td>89</td>
<td>965</td>
<td>10648</td>
<td>117126</td>
<td>1286408</td>
<td>14172428</td>
</tr>
<tr>
<td>13</td>
<td>35</td>
<td>455</td>
<td>5915</td>
<td>76895</td>
<td>999635</td>
<td>12995255</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>180</td>
<td>2700</td>
<td>40500</td>
<td>607500</td>
<td>9112500</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>68</td>
<td>1156</td>
<td>19652</td>
<td>334084</td>
<td>5679428</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>57</td>
<td>1063</td>
<td>20377</td>
<td>390956</td>
<td>7232977</td>
</tr>
<tr>
<td>Total</td>
<td>10000</td>
<td>2</td>
<td>163,296</td>
<td>319,624</td>
<td>8,933,808</td>
<td>55,337,762</td>
</tr>
</tbody>
</table>

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

**One Hundred Thousand Times the Area Under a Normal Curve**

<table>
<thead>
<tr>
<th>$t_1$</th>
<th>$f_1$</th>
<th>$t_1 f_1$</th>
<th>$t_1^2 f_1$</th>
<th>$t_1^3 f_1$</th>
<th>$t_1^4 f_1$</th>
<th>$t_1^5 f_1$</th>
<th>$t_1^6 f_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.0</td>
<td>9</td>
<td>36.0</td>
<td>144.0</td>
<td>576.0</td>
<td>2304.0</td>
<td>-9216.0</td>
<td>36864.0</td>
</tr>
<tr>
<td>-3.5</td>
<td>49</td>
<td>-171.5</td>
<td>600.25</td>
<td>-2100.88</td>
<td>7353.06</td>
<td>-25735.72</td>
<td>90075.02</td>
</tr>
<tr>
<td>-3.0</td>
<td>240</td>
<td>-720.0</td>
<td>2160.00</td>
<td>-6480.00</td>
<td>19440.00</td>
<td>-58320.00</td>
<td>174960.00</td>
</tr>
<tr>
<td>-2.5</td>
<td>924</td>
<td>-2310.0</td>
<td>5775.00</td>
<td>-14375.00</td>
<td>36093.75</td>
<td>-90234.38</td>
<td>225585.94</td>
</tr>
<tr>
<td>-2.0</td>
<td>2784</td>
<td>-5568.0</td>
<td>11136.00</td>
<td>-22272.00</td>
<td>44544.00</td>
<td>-80983.00</td>
<td>176176.00</td>
</tr>
<tr>
<td>-1.5</td>
<td>6559</td>
<td>-9838.5</td>
<td>14757.75</td>
<td>-22136.62</td>
<td>33204.94</td>
<td>-49807.41</td>
<td>74711.11</td>
</tr>
<tr>
<td>-1.0</td>
<td>12098</td>
<td>-12098</td>
<td>12098.00</td>
<td>-12098.00</td>
<td>12098.00</td>
<td>12098.00</td>
<td>12098.00</td>
</tr>
<tr>
<td>-0.5</td>
<td>17466</td>
<td>-8733.0</td>
<td>4366.50</td>
<td>-2183.25</td>
<td>1091.62</td>
<td>-545.81</td>
<td>272.91</td>
</tr>
<tr>
<td>0.0</td>
<td>19742</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.5</td>
<td>17466</td>
<td>8733.0</td>
<td>4366.50</td>
<td>2183.25</td>
<td>1091.62</td>
<td>545.81</td>
<td>272.91</td>
</tr>
<tr>
<td>1.0</td>
<td>12098</td>
<td>12098.0</td>
<td>12098.00</td>
<td>12098.00</td>
<td>12098.00</td>
<td>12098.00</td>
<td>12098.00</td>
</tr>
<tr>
<td>1.5</td>
<td>6559</td>
<td>9838.5</td>
<td>14757.75</td>
<td>22136.62</td>
<td>33204.94</td>
<td>49807.41</td>
<td>74711.11</td>
</tr>
<tr>
<td>2.0</td>
<td>2784</td>
<td>5568.0</td>
<td>11136.00</td>
<td>22272.00</td>
<td>44544.00</td>
<td>80983.00</td>
<td>176176.00</td>
</tr>
<tr>
<td>2.5</td>
<td>924</td>
<td>2310.0</td>
<td>5775.00</td>
<td>14375.00</td>
<td>36093.75</td>
<td>90234.38</td>
<td>225585.94</td>
</tr>
<tr>
<td>3.0</td>
<td>240</td>
<td>720.0</td>
<td>2160.00</td>
<td>6480.00</td>
<td>19440.00</td>
<td>58320.00</td>
<td>174960.00</td>
</tr>
<tr>
<td>3.5</td>
<td>49</td>
<td>171.5</td>
<td>600.25</td>
<td>2106.88</td>
<td>7533.06</td>
<td>25735.72</td>
<td>90075.02</td>
</tr>
<tr>
<td>4.0</td>
<td>9</td>
<td>36.0</td>
<td>144.0</td>
<td>576.00</td>
<td>2304.00</td>
<td>9216.00</td>
<td>36864.00</td>
</tr>
<tr>
<td>Total</td>
<td>100,000</td>
<td>0.0</td>
<td>102,075.09</td>
<td>0.00</td>
<td>312,258.74</td>
<td>0.00</td>
<td>1,585,485.94</td>
</tr>
<tr>
<td>n</td>
<td>( v_n )</td>
<td>( v_{on} )</td>
<td>( u_n )</td>
<td>( v_n )</td>
<td>( v_{on} )</td>
<td>( u_n )</td>
<td>( v_n )</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>16.3384</td>
<td>16.005</td>
<td>16</td>
<td>16.3296</td>
<td>16.00</td>
<td>16</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>32.0608</td>
<td>32.06</td>
<td>32</td>
<td>31.9634</td>
<td>31.96</td>
<td>32</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>697.1158</td>
<td>664.91</td>
<td>664</td>
<td>893.3803</td>
<td>861.18</td>
<td>864</td>
<td>3.00</td>
</tr>
<tr>
<td>5</td>
<td>5604.7744</td>
<td>5497.90</td>
<td>5504</td>
<td>5539.7762</td>
<td>5527.29</td>
<td>5504</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>1001235.0544</td>
<td>98,140.33</td>
<td>98,460</td>
<td>98,535.8976</td>
<td>94,181.86</td>
<td>96,640</td>
<td>14.00</td>
</tr>
</tbody>
</table>

Note: The \( v_n \) represents uncorrected moments, \( v_{on} \) represents corrected moments, and \( u_n \) represents theoretically expected moments.
From the previous demonstration of the derivation of Shepard's corrections it would seem desirable to have a less complex approach to the problem. To aid in this approach the moment-generating function concept will be introduced here. As the name implies, it is a function which generates moments and is defined by \( M_x(\theta) = \int_a^b e^{\theta x} f(x) dx \). Expanding \( e^{\theta x} \) into a power series one obtains

\[
e^{\theta x} = 1 + \theta x + \frac{\theta^2 x^2}{2!} + \frac{\theta^3 x^3}{3!} + \cdots
\]

therefore

\[
M_x(\theta) = \int_a^b \left[ 1 + \theta x + \frac{\theta^2 x^2}{2!} + \frac{\theta^3 x^3}{3!} + \cdots \right] f(x) dx
\]

Integrating term by term one gets

\[
M_x(\theta) = \int_a^b f(x) dx + \theta \int_a^b x f(x) dx + \frac{\theta^2}{2!} \int a^b x^2 f(x) dx + \cdots
\]

\[
= u_0 + u_1 \theta + u_2 \frac{\theta^2}{2!} + u_3 \frac{\theta^3}{3!} + \cdots
\]

Observe that the coefficient of \( \frac{\theta^k}{k!} \) is the \( k \)th moment about the origin. Also useful in this approach is the function

\( g(x_1, x_2) = x_1 + x_2 \) where \( x_1 \) and \( x_2 \) are independently distributed. Independence implies that \( f(x_1, x_2) = f(x_1) \cdot f(x_2) \). One can generalize the moment-generating function definition to hold for the variable \( g(x) \), therefore

\[
M_g(\theta) = \int_a^b e^{\theta g(x)} f(x) dx
\]

By extending this to the case where \( g \) is a function of two variables one has

\[
M_g(\theta) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} e^{\theta g(x_1, x_2)} f(x_1, x_2) dx_1 dx_2.
\]
Then
\[
M_{x_1 + x_2}(\theta) = \int_{b_2}^{b_1} \int_{a_2}^{a_1} e^{\theta(x_1 + x_2)} f(x_1)f(x_2) dx_1 dx_2
\]
\[
= \int_{b_1}^{b_1} \int_{a_1}^{a_2} e^{\theta x_1} f(x_1) dx_1 \int_{b_2}^{b_2} e^{\theta x_2} f(x_2) dx_2
\]
Since each of the integrals on the right is the moment-generating function of the indicated variable, then
\[
M_{x_1 + x_2}(\theta) = M_{x_1}(\theta)M_{x_2}(\theta).
\]

Consider a frequency distribution \( f(x) \) divided into any convenient number of classes whose class mark is \( x_1 \) and class width \( k \). And further let \(\varepsilon\) be the error in using the class mark \( x_1 \) instead of the variable \( x \). See Fig. 1.

![Fig. 1. Frequency Distribution](image)

Then one has the relationship \( x_1 = x + \varepsilon \), where \( x \) and \( \varepsilon \) are independently distributed, \( x \) is the variable before grouping and \( x_1 \) is the variable after grouping. \( x \) and \( \varepsilon \) are independently distributed because \( P \left[ x \leq x \leq (x + dx) \right] = f(x)dx \) and
\[
P \left[ \varepsilon \leq \varepsilon \leq (\varepsilon + d\varepsilon) \right] = \frac{1}{k} d\varepsilon. \]
Then \( M_{x_1}(\theta) = M_{x}(\theta) M_{\varepsilon}(\theta) \). It has been proved that \( M_{x}(\theta) = 1 + u_1 \theta + u_2 \frac{\theta^2}{2!} + u_3 \frac{\theta^3}{3!} + \cdots \) and it
follows that $M_{X_1}(\theta) = 1 + v_1 \theta + v_2 \frac{\theta^2}{2!} + v_3 \frac{\theta^3}{3!} + \cdots$. It is then left to find the expansion of $M_{\xi}(\theta)$. The definition of moment-generating function tells one that

$$M_{\xi}(\theta) = \frac{1}{k} \int_{-k/2}^{k/2} e^{\theta \varepsilon} d\varepsilon.$$

Upon integrating

$$M_{\xi}(\theta) = \frac{1}{\theta k} \left[ e^{\theta \varepsilon} \right]_{-k/2}^{k/2} = \frac{1}{\theta k} \left[ e^{\theta k/2} - e^{-\theta k/2} \right] = \frac{2}{\theta k} \sinh \theta k/2.$$

Expanding into a power series and simplifying

$$M_{\xi}(\theta) = \frac{2}{\theta k} \left[ \theta k/2 - \frac{(\theta k/2)^3}{3!} - \frac{(\theta k/2)^5}{5!} - \frac{(\theta k/2)^7}{7!} - \cdots \right]$$

$$= 1 + \frac{k^2}{12} \frac{\theta^2}{2!} + \frac{k^4}{80} \frac{\theta^4}{4!} + \frac{k^6}{448} \frac{\theta^6}{6!} + \cdots.$$

By substitution, multiplication and dropping primes to indicate moments about the mean one obtains

$$M_{X_1}(\theta) = M_{X}(\theta) M_{\xi}(\theta)$$

$$= 1 + u_1 \theta + \left( u_2 - \frac{k^2}{12} \right) \frac{\theta^2}{2!} + \left( u_3 + \frac{u_1 k^2}{4} \right) \frac{\theta^3}{3!} -$$

$$\left( u_4 - \frac{u_2 k^2}{2} - \frac{k^4}{80} \right) \frac{\theta^4}{4!} - \left( u_5 - \frac{5u_2 k^2}{6} - \frac{u_1 k^4}{16} \right) \frac{\theta^5}{5!} +$$

$$\left( u_6 + \frac{5u_2 k^2}{4} + \frac{3u_2 k^4}{16} + \frac{k^6}{448} \right) \frac{\theta^6}{6!} + \cdots.$$
Moments up to the sixth are

\[ u_1 = v_1 = 0 \]

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

\[ u_3 = v_3 \]

\[ u_4 = v_4 - \frac{u_{bk}^2}{2} - \frac{k^2}{80} \]

\[ u_5 = v_5 - \frac{5u_{bk}^2}{6} \]

\[ u_6 = v_6 - \frac{5u_{bk}^2}{4} - \frac{3u_{bk}^4}{16} - \frac{k^6}{448} \]
Illustrations have been given in this paper to demonstrate how "Sheppard's corrections" actually function. Also presented was a simplified development of the formulae for these "corrections" by use of the moment-generating function concept. The writer believes that this approach simplifies the traditional solution.

For the numerical examples chosen in this paper the "corrections" do give very good adjustments to the lower approximate moments. However, they seem to give inadequate corrections to the observed moments higher than the fourth. This implies no basic wrong in the "corrections"; but the methodology employed in grouping the data may have been at fault or the quantity of data may have been insufficient. Table IV indicates that the data here used seems to bear out both of the above suppositions.

Arising out of this paper, then, is the possible problem of further investigating the adjustments to higher moments by varying the method of grouping as well as varying the quantity of data used. Since $\gamma_4$ is not always a reliable index of peakedness, one could study that problem with the purpose of finding a new and valid measure. At present the writer does not have the opportunity for further investigation of the above mentioned problems.

CONCLUSIONS
LITERATURE CITED AND CONSULTED


Scott, Chester H.

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

ISSUED TO
MAY 26 '54
Glenn Ingram
Math Dept.

N378
Sco81o
cop.2

92609