COMPLETE CORRECTION OF MOMENTS FOR HISTOGRAM BIAS AND FOR LACK OF HIGH CONTACT

Thesis by

Carl H. Savit

(Written in collaboration with Jack H. Irving)

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

California Institute of Technology

Pasadena, California

1943

COMPLETE CORRECTION OF MOMENTS FOR LACK OF HIGH CONTACT AND FOR HISTOGRAM BIAS

There are two types of data which must be handled statistically, that data which by its very nature can take on only discrete values (e.g. the number of teeth in a person's mouth or the electrostatic charge on an oil drop), and that whose values range over a continuum (e.g. the annual rainfall at a given locality or the length of the human tibia.) It is customary to refer to the distribution for the former type as a "point frequency" distribution and to that for the latter as a "histogram" distribution. In the event that a distribution is such that extreme values of the observed data occur infrequently (that is, if when frequency is plotted as ordinate with the observed data as abscissae, the resultant graph is asymptotic to the axis of abscissae at both ends), the distribution is said to possess "high contact".

Assuming that there exists an actual, smooth frequency curve for a given histogram distribution which is known only by the frequency of occurence of data within various ranges, it is clear that moments computed by summing the product of the frequency for each range by the appropriate power of the abscissa of the mid-point of that range will be in error unless the frequency curve is a straight line and only the zeroth and first moments are desired.

In addition, a further error is introduced in computing the moments if the curve does not have high contact at both ends. The former type of error is referred to as histogram bias.

The first practicable correction for histogram bias was devised by W. F. Sheppard (W. F. Sheppard, "Calculation of the Most Probable Values of Frequency Constants", Proceedings of the London Mathematical Society, Vol. XXIX, pp. 353-380; 1898). E. Pairman and Karl Pearson developed a system for applying corrections both for histogram bias and for lack of high contact (E. Pairman & Karl Pearson, "Corrections for Moment Coefficients", Biometrika, Vol. XII, p. 231 et seq.). The methods employed in the above papers are somewhat complex and led to attempts at simplification by Palin Elderton (W. Palin Elderton, Frequency Curves and Correlation, pp. 24-27, Charles and Edwin Layton, London, 2nd Edition, 1927), and Dinsmore Alter (D. Alter, "Correction of Sample Moment Bias Due to Lack of High Contact and to Histogram Grouping", The Annals of Mathematical Statistics, Vol. X, pp. 192-195, 1939). The present paper attempts to combine simplicity of mathematical theory with ease of application.

A set of observations is made which falls within a range of length ℓ on the axis of x. Dividing the range into n equal intervals, we choose a scale such that these intervals are of unit length. Let x_1 be the midpoint of the first unit interval, x_i the midpoint of the ith. Let A_i be the number of experimental observations

falling within the interval whose midpoint is x_i; then $\sum_{i=1}^{n} A_i$ is the total number of observations and is also required to be the area under the frequency curve y(x) from $\alpha = x_1 - \frac{1}{2}$ to $\beta = x_n + \frac{1}{2}$.

From the standard expression for the mth moment

$$\mu'_m = \frac{1}{\sum_{i=1}^{n} A_i} \int_{\alpha}^{\beta} x^m y \, dx \qquad (1)$$

where y is the ordinate of the desired frequency curve.

Since y(x), the desired frequency curve is not known we must approximate to it for purposes of evaluating the integral in (1) by a group of suitable approximation curves, each covering one of the unit intervals of abscissa. Our fundamental assumption for the histogram distribution is that a satisfactory approximation curve for the ith interval is a quartic such that the areas under each of the (i-2)nd, (i-1)th, ith, (i+1)th, and (i+2)nd unit intervals are equal to the experimentally measured A's for these intervals. The practice of using quartic approximations originated with Sheppard and is here retained since in cases where the fourth moment is of the same order of magnitude as the zeroth moment, the correction due to the quartic term is of the order of 3% and even in ordinary cases the correction can be as much as .5% in the fourth moment. Let the equation of the quartic to be used as the approximation curve for the ith unit interval be

$$y_{x_i+t} = a_i t^4 + b_i t^3 + c_i t^2 + d_i t + e_i \qquad (2)$$

The above five conditions on the areas under the unit intervals of the quartic give five conditions from which the coefficients of equation (2) may be determined.

$$\begin{aligned}
A_{i-2} &= \int_{-2\frac{1}{2}}^{-1\frac{1}{2}} y_{x_i+t}\,dt = \tfrac{2882}{160} a_i - \tfrac{544}{64} b_i + \tfrac{98}{24} c_i - 2d_i + e_i \\
A_{i-1} &= \int_{-1\frac{1}{2}}^{-\frac{1}{2}} y_{x_i+t}\,dt = \tfrac{242}{160} a_i - \tfrac{80}{64} b_i + \tfrac{26}{24} c_i - d_i + e_i \\
A_i &= \int_{-\frac{1}{2}}^{\frac{1}{2}} y_{x_i+t}\,dt = \tfrac{2}{160} a_i \qquad + \tfrac{2}{24} c_i \qquad + e_i \\
A_{i+1} &= \int_{\frac{1}{2}}^{1\frac{1}{2}} y_{x_i+t}\,dt = \tfrac{242}{160} a_i + \tfrac{80}{64} b_i + \tfrac{26}{24} c_i + d_i + e_i \\
A_{i+2} &= \int_{1\frac{1}{2}}^{2\frac{1}{2}} y_{x_i+t}\,dt = \tfrac{2882}{160} a_i + \tfrac{544}{64} b_i + \tfrac{98}{24} c_i + 2d_i + e_i
\end{aligned} \qquad (3)$$

Solving (3) for a_i, b_i, etc. and substituting into (2) we obtain:

$$\begin{aligned}
y_{x_i+t} = \tfrac{1}{1920} \big[& A_i (2134 - 2640t^2 + 480t^4) \\
& + (A_{i+1} + A_{i-1})(-116 + 1440t^2 - 320t^4) \\
& + (A_{i+1} - A_{i-1})(1360t - 320t^3) \\
& + (A_{i+2} + A_{i-2})(9 - 120t^2 + 80t^4) \\
& + (A_{i+2} - A_{i-2})(-200t + 160t^3) \big]
\end{aligned} \qquad (4)$$

If the data is such that it may by an a priori method be extrapolated, such extrapolated values may be used for A_{-1}, A_0, A_{n+1}, and A_{n+2}. Extrapolation for these A's will, however, not be necessary for the purposes of this paper.

It is to be remembered that although we have forced five areas under y_{x_i+t} to the observed values, only the middle unit interval is used in obtaining the moments, thus increasing the validity of the representation. It follows that (1) becomes

$$\mu'_m = \frac{1}{\sum_{i=1}^{n} A_i} \sum_{i=1}^{n} \int_{-1/2}^{1/2} (x_i+t)^m y_{x_i+t}\, dt \tag{5}$$

Using $(x_i+t)^m = x_i^m + m x_i^{m-1} t + \frac{m(m-1)}{2} x_i^{m-2} t^2 + \cdots$

we obtain, upon integration,

$$\begin{aligned}\int_{-1/2}^{1/2} (x_i+t)^m y_{x_i+t}\, dt = {} & x_i^m A_i + \frac{m x_i^{m-1}}{1440}\left[-11(A_{i+2}-A_{i-2}) + 82(A_{i+1}-A_{i-1})\right] \\ & + \frac{m(m-1)}{2} x_i^{m-2}\left\{\frac{1}{10080}\left[-3(A_{i+2}+A_{i-2}) + 766 A_i + 40(A_{i+1}+A_{i-1})\right]\right\} \\ & + \frac{m(m-1)(m-2)}{6} x_i^{m-3}\left\{\frac{1}{4480}\left[-5(A_{i+2}-A_{i-2}) + 38(A_{i+1}-A_{i-1})\right]\right\} \\ & + \frac{m(m-1)(m-2)(m-3)}{24} x_i^{m-4}\left\{\frac{1}{302400}\left[-19(A_{i+2}+A_{i-2}) + 3306 A_i \right.\right. \\ & \left.\left. + 256(A_{i+1}+A_{i-1})\right]\right\}\end{aligned} \tag{6}$$

Terms containing a factor of degree less than m-4 can be neglected since moments higher than the fourth are rarely, if ever, desired.

In order to obtain μ'_m it remains merely to sum equation (6). We notice that

$$\sum_{i=1}^{n} x_i^{m-1} A_{i+1} = \sum_{i=1}^{n} (x_i - 1)^{m-1} A_i - x_0^{m-1} A_1 + x_n^{m-1} A_{n+1}$$

and

$$\sum_{i=1}^{n} x_i^{m-1} A_{i-1} = \sum_{i=1}^{n} (x_i + 1)^{m-1} A_i + x_1^{m-1} A_0 - x_{n+1}^{m-1} A_n \qquad \text{etc.}$$

therefore

$$\begin{aligned} -\sum_{i=1}^{n} x_i^{m-1} (A_{i+1} - A_{i-1}) &= \sum_{i=1}^{n} A_i \left[(x_i + 1)^{m-1} - (x_i - 1)^{m-1} \right] + (\alpha + \tfrac{1}{2})^{m-1} A_0 \\ &\quad + (\alpha - \tfrac{1}{2})^{m-1} A_1 - (\beta + \tfrac{1}{2})^{m-1} A_n - (\beta - \tfrac{1}{2})^{m-1} A_{n+1} \\ &= 2 \sum_{i=1}^{n} A_i \left[(m-1) x_i^{m-2} + \frac{(m-1)(m-2)(m-3)}{6} x_i^{m-4} \right] \\ &\quad + (A_0 + A_1) \left[\alpha^{m-1} + \frac{(m-1)(m-2)}{2} \frac{\alpha^{m-3}}{4} \right] \\ &\quad + (A_0 - A_1) \left[(m-1) \frac{\alpha^{m-2}}{2} + \frac{(m-1)(m-2)(m-3)}{6} \frac{\alpha^{m-4}}{8} \right] \end{aligned}$$

minus similar terms in β from the upper end of the distribution.

The summation of the remaining terms which appear in (6) may be accomplished in an analogous manner. When combined, they yield the fundamental formula

$$\mu'_m \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} x_i^m A_i - \frac{m(m-1)}{24} \sum_{i=1}^{n} x_i^{m-2} A_i + \frac{7m(m-1)(m-2)(m-3)}{5760} \sum_{i=1}^{n} x^{m-4} A_i$$

$$+ m\alpha^{m-1} \left\{ \frac{1}{1440} \left[11 (A_{-1} + A_2) - 71 (A_0 + A_1) \right] \right\}$$

$$+ \frac{m(m-1)}{2} \alpha^{m-2} \left\{ \frac{1}{5040} \left[37 (A_{-1} - A_2) - 153 (A_0 - A_1) \right] \right\}$$

$$+ \frac{m(m-1)(m-2)}{6} \alpha^{m-3} \left\{ \frac{1}{6720} \left[43 (A_{-1} + A_2) + 41 (A_0 + A_1) \right] \right\}$$

$$+ \frac{m(m-1)(m-2)(m-3)}{24} \alpha^{m-4} \left\{ \frac{1}{75600} \left[419 (A_{-1} - A_2) + 5073 (A_0 - A_1) \right] \right\}$$

$$- m\beta^{m-1} \left\{ \frac{1}{1440} \left[11 (A_{n-1} + A_{n+2}) - 71 (A_n + A_{n+1}) \right] \right\}$$

$$+ \frac{m(m-1)}{2} \beta^{m-2} \left\{ \frac{1}{5040} \left[37 (-A_{n-1} + A_{n+2}) - 153 (-A_n + A_{n+1}) \right] \right\} \qquad (7)$$

$$- \frac{m(m-1)(m-2)}{6} \beta^{m-3} \left\{ \frac{1}{6720} \left[43 (A_{n-1} + A_{n+2}) + 41 (A_n + A_{n+1}) \right] \right\}$$

$$+ \frac{m(m-1)(m-2)(m-3)}{24} \beta^{m-4} \left\{ \frac{1}{75600} \left[419 (-A_{n-1} + A_{n+2}) + 5073 (-A_n + A_{n+1}) \right] \right\}$$

We observe that in equation (7) the first term is the uncorrected "raw" moment; the second and third terms constitute the well known Sheppard's correction for histogram grouping; the terms involving α comprise the correction for lack of high contact at the lower end of the distribution; and the terms involving β comprise the correction for lack of high contact at the upper end of the distribution. At this point the formula may

be applied using any or all of the corrections independently.

For purposes of computation it is possible to assume that that quartic which was used as an approximation for y in the third interval of x is valid as an approximation for y in the first and second intervals. A similar assumption may be made at the upper end of the distribution. To adapt (7) to these assumptions it is simplest merely to evaluate A_0, A_{-1}, A_{n+1}, A_{n+2} by means of the quartics for the third interval of x and for the (n-2)th interval of x and to substitute the values so obtained into (7). The resulting expression for the moments is:

$$
\begin{aligned}
\mu'_m \sum_{i=1}^{n} A_i = {} & \sum_{i=1}^{n} x_i^m A_i - \frac{m(m-1)}{24} \sum_{i=1}^{n} x_i^{m-2} A_i + \frac{7m(m-1)(m-2)(m-3)}{5760} \sum_{i=1}^{n} x_i^{m-4} A_i \\
& + m\alpha^{m-1}\left[\frac{1}{1440}(-261A_1 + 281A_2 - 215A_3 + 91A_4 - 16A_5)\right] \\
& + \frac{m(m-1)}{2}\alpha^{m-2}\left[\frac{1}{5040}(-57A_1 + 13A_2 + 135A_3 - 123A_4 + 32A_5)\right] \\
& + \frac{m(m-1)(m-2)}{6}\alpha^{m-3}\left[\frac{1}{6720}(891A_1 - 2087A_2 + 2345A_3 - 1237A_4 + 256A_5)\right] \\
& + \frac{m(m-1)(m-2)(m-3)}{24}\alpha^{m-4}\left[\frac{1}{75600}(26577A_1 - 67909A_2 + 69585A_3 \right. \\
& \qquad \left. - 35421A_4 + 7168A_5)\right] \qquad (8) \\
& + m\beta^{m-1}\left[\frac{1}{1440}(16A_{n-4} - 91A_{n-3} + 215A_{n-2} - 281A_{n-1} + 261A_n)\right] \\
& + \frac{m(m-1)}{2}\beta^{m-2}\left[\frac{1}{5040}(-32A_{n-4} + 123A_{n-3} - 135A_{n-2} - 13A_{n-1} + 57A_n)\right] \\
& + \frac{m(m-1)(m-2)}{6}\beta^{m-3}\left[\frac{1}{6720}(-256A_{n-4} + 1237A_{n-3} - 2345A_{n-2} + 2087A_{n-1} - 891A_n)\right] \\
& + \frac{m(m-1)(m-2)(m-3)}{24}\beta^{m-4}\left[\frac{1}{75600}(-7168A_{n-4} + 35421A_{n-3} - 69585A_{n-2} \right. \\
& \qquad \left. + 67909A_{n-1} - 26577A_n)\right]
\end{aligned}
$$

The curve $y = \sqrt{x}$ has been used to demonstrate the various corrections mentioned above. Table I gives the values of the first four moments of $y = \sqrt{x}$ from $x = 0$ to $x = 1$ using ten divisions of the abscissa to form ten A_i's; uncorrected (a), with Sheppard's corrections only (b), with Pairman-Pearson corrections (c), with Alter's corrections (d), using formula (7) and graphical extrapolation (e), using tables prepared from formula (8) (f).

Table II gives the values prepared from formula (8) for use as coefficients of the A_i's in computing the first four moments. The coefficients are included only for the case of $n = 10$.

Table I

m	a	b	c	d	e	f	True Value
1	5.9883	5.9883	5.9994	5.9996	5.9997	5.9997	6.0000
2	42.6900	42.6067	42.8570	42.8576	42.8571	42.8569	42.8571
3	331.0854	329.5884	333.3349	333.3387	333.3335	333.3328	333.3333
4	2698.7740	2677.4585	2727.2757	2727.3555	2727.2706	2727.1399	2727.2727

Table II

i	$1440\,E_{1i}$
1	459
2	2441
3	3385
4	5131
5	6464
6	7936
7	9269
8	11015
9	11959
10	13941

i	$5040\,E_{2i}$
1	783
2	10933
3	31215
4	61197
5	101672
6	153192
7	206027
8	298265
9	344063
10	472653

i	$6720\,E_{3i}$
1	891
2	18073
3	103145
4	281003
5	605056
6	1132224
7	1703477
8	3126455
9	3721847
10	6107829

i	$75600\,E_{4i}$
1	24057
2	61871
3	2788665
4	10848459
5	30244648
6	71577448
7	113660109
8	282458415
9	333886121
10	666280107

$$\mu_m \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} E_{mi} A_i$$