SYNOPSIS OF ELEMENTARY MATHEMATICAL STATISTICS*

By

B. L. SHOOK

SECTION I. ELEMENTARY STATISTICAL FUNCTIONS

1. *Variates.* Practically all statistical data is obtained as the result of observations that endeavor to establish the magnitudes of certain variables. The individual magnitudes that are recorded are termed variates. Thus in computing the average annual rainfall of a region, the variable is rainfall, and the amount of rainfall for any single year is a variate. Likewise, if the bank clearings for the City of New York be under consideration, then the variable is bank clearings, and the clearings for any specified interval is a variate.

2. The *arithmetic mean* of a series of variates is equal to the sum of the variates divided by the number of variates in the series. If M_v designates the arithmetic mean of the N variates $v_1, v_2, v_3, \cdots, v_N$,

$$(1) \qquad M_v = \frac{1}{N}(v_1+v_2+\cdots+v_N) = \frac{1}{N}\sum v$$

3. The nth moment of a series of variates is defined as the arithmetic mean of the nth powers of these variates and is represented by the symbol $\mu'_{n:v}$. Thus,

$$(2) \qquad \mu_{n:v} = \frac{1}{N}(v_1^n+v_2^n+v_3^n+\cdots+v_N^n) = \frac{1}{N}\sum v^n$$

That is

$$\mu'_{1:v} = \frac{1}{N}\sum v$$

$$\mu'_{2:v} = \frac{1}{N}\sum v^2$$

$$\mu'_{3:v} = \frac{1}{N}\sum v^3$$

---

* An abstract of a series of lectures on elementary statistics given by the mathematical statistical staff at the University of Michigan.

1. Observe that the number of variates in a series is denoted by N, whereas the smaller italic n is employed as an ordinal number.

Obviously, by definitions (1) and (2)

$$\mu'_{1:v} = M_v \tag{3}$$

4. *The deviation of a variate from the arithmetic mean* will be designated by the symbol $\bar{v}$, i. e.

$$\bar{v}_i = v_i - M_v \tag{4}$$

5. The *nth moment about the mean** is defined as the arithmetic mean of the nth powers of the deviations of the variates from the mean, and is represented symbolically by $\mu_{n:v}$. Thus

$$\mu_{n:v} = \frac{1}{N}\sum \bar{v}^n \qquad \text{so that} \tag{5}$$

$$\mu_{1:v} = \frac{1}{N}\sum \bar{v} = 0 \tag{5a}$$

$$\mu_{2:v} = \frac{1}{N}\sum \bar{v}^2 \tag{5b}$$

$$\mu_{3:v} = \frac{1}{N}\sum \bar{v}^3 \tag{5c}$$

The fact that $\mu_{1:v} = 0$, is demonstrated as follows:

$$\begin{aligned} \bar{v}_1 &= v_1 - M_v \\ \bar{v}_2 &= v_2 - M_v \\ &\cdots \\ \bar{v}_N &= v_N - M_v \\ \hline \sum \bar{v} &= \sum v - NM_v \end{aligned}$$

$$\mu_{1:v} = \frac{\sum \bar{v}}{N} = \frac{\sum v}{N} - M_v = M_v - M_v = 0 \qquad Q.\ E.\ D.$$

The numerical example of Table I illustrates the definitions of the preceding paragraphs. The data consists of thirteen variates, which represent the number of even numbers found in consecutive blocks of 100 numbers, drawn to determine the order of call for draft-

* For convenience the arithmetic mean is frequently referred to as *the* mean. When referring to geometric or harmonic means, the adjectives geometric or harmonic must therefore be specified.

ing United States soldiers in 1918. These variates were obtained from the *first* 1300 drawings made.

The most obvious conclusion to be drawn from Table I is that the use of fractions in determining the values of $\mu_{n:v}$ is cumbersome. if M_v is a whole number, then the values of $\bar{v}$, $\bar{v}^2$ and $\bar{v}^3$ are integers, and the procedure is simple. Generally, however, M_v will be fractional, and consequently awkward expressions for $\bar{v}$, $\bar{v}^2$ and $\bar{v}^3$ will result. On the other hand, the computation of values of $\mu'_{n:v}$ is relatively easy, and hence it is expedient to express $\mu_{2:v}$ and $\mu_{3:v}$ in terms of the moments $\mu'_{n:v}$. This may be done as follows:

Since by definition,

$$\bar{v}_i = v_i - M_v, \qquad \text{it follows that}$$

$$\bar{v}_i^2 = v_i^2 - 2v_i M_v + M_v^2 \qquad , \text{ and}$$

$$\bar{v}_i^3 = v_i^3 - 3v_i^2 M_v + 3v_i M_v^2 - M_v^3$$

Consequently

$$\begin{aligned}
\bar{v}_1^2 &= v_1^2 - 2M_v v_1 + M_v^2 & \bar{v}_1^3 &= v_1^3 - 3v_1^2 M_v + 3v_1 M_v^2 - M_v^3 \\
\bar{v}_2^2 &= v_2^2 - 2M_v v_2 + M_v^2 & \bar{v}_2^3 &= v_2^3 - 3v_2^2 M_v + 3v_2 M_v^2 - M_v^3 \\
\bar{v}_3^2 &= v_3^2 - 2M_v v_3 + M_v^2 & \bar{v}_3^3 &= v_3^3 - 3v_3^2 M_v + 3v_3 M_v^2 - M_v^3 \\
&\cdots & &\cdots \\
v_n^2 &= v_N^2 - 2M_v v_N + M_v^2 & \bar{v}_N^3 &= v_N^3 - 3v_N^2 M_v + 3v_N M_v^2 - M_v^3 \\
\hline
\sum \bar{v}^2 &= \sum v^2 - 2M_v \sum v + NM_v^2 & \sum \bar{v}^3 &= \sum v^3 - 3M_v \sum v^2 + 3M_v^2 \sum v - NM_v^3
\end{aligned}$$

Dividing both sides of these equations through by N yields, respectively

$$\frac{\sum \bar{v}^2}{N} = \frac{\sum v^2}{N} - 2M_v \cdot M_v + M_v^2$$

$$\frac{\sum \bar{v}^3}{N} = \frac{\sum v^3}{N} - 3M_v \frac{\sum v^2}{N} + 3M_v^2 \cdot M_v - M_v^3$$

Hence

$$(6) \qquad \begin{cases} \mu_{2:v} = \mu'_{2:v} - M_v^2 \\ \mu_{3:v} = \mu'_{3:v} - 3M_v \mu'_{2:v} + 2M_v^3 \end{cases}$$

TABLE I

i	v_i	v_i^2	v_i^3	$\bar{v}_i = v_i - M_v$	$\bar{v}_i^4$	$\bar{v}_i^5$
1	51	2601	132651	− 3/13	9/169	− 27/2197
2	49	2401	117649	− 29/13	841/169	− 24389/2197
3	60	3600	216000	114/13	12996/169	1481544/2197
4	53	2809	148877	23/13	529/169	12167/2197
5	48	2304	110592	− 42/13	1764/169	− 74088/2197
6	51	2601	132651	− 3/13	9/169	− 27/2197
7	42	1764	74088	− 120/13	14400/169	− 1728000/2197
8	50	2500	125000	− 16/13	256/169	− 4096/2197
9	51	2601	132651	− 3/13	9/169	− 27/2197
10	52	2704	140608	10/13	100/169	1000/2197
11	54	2916	157464	36/13	1296/169	46656/2197
12	53	2809	148877	23/13	529/169	12167/2197
13	52	2704	140608	10/13	100/169	1000/2197
Total	666	34314	1777716	0	32838/169	− 276120/2197

$$M_v = \mu'_{1:v} = \frac{666}{13}$$

$$\mu'_{2:v} = \frac{34314}{13}$$

$$\mu'_{3:v} = \frac{177716}{13}$$

$$\mu_{1:v} = 0$$

$$\mu_{2:v} = \frac{32838}{13 \cdot 169} = \frac{2526}{169}$$

$$\mu_{3:v} = \frac{-276120}{13 \cdot 2197} = \frac{-21240}{2197}$$

These formulae are perhaps the most important in our work, since they enable us to obtain the moments about the mean without requiring that we actually determine the deviations. Applying these formulae to the numerical example of Table I,

$$\mu_{2:v} = \frac{34314}{13} - \left(\frac{666}{13}\right)^2 = \frac{2526}{169}$$

$$\mu_{3:v} = \frac{177716}{13} - 3\left(\frac{34314}{13}\right)\left(\frac{666}{13}\right) + 2\left(\frac{666}{13}\right)^3 = -\frac{21240}{2197}$$

The results thus obtained by this *indirect* method are identical with the results obtained in Table I by employing the *direct* method.

7. *Standard Deviation.* The second moment about the mean, $\mu_{2:v}$, is a function of the variability of the data, since its essential elements are the deviations of the variates from the mean. But if the original variates happen to be measured in *inches*, then since $\mu_{2:v}$ is the average of the squares of the deviations, it follows that the unit of $\mu_{2:v}$ is square inch. Nevertheless, by extracting the square root of $\mu_{2:v}$ we would obtain a function which would in general measure the variability of, and possess the same unit as the original data. This function is known as the *standard deviation* and is denoted by the symbol σ_v. Thus

$$\sigma_v = \sqrt{\mu_{2:v}} \tag{7}$$

Verbally we may say that the standard deviation is defined as the square root of the mean of the squared deviations of the variates from their mean.

Actually σ_v is rarely computed directly from the squared deviations, but rather by employing the relationship given in formula (6). For the data of Table I

$$\sigma_v = \sqrt{\frac{2526}{169}} = \frac{50.2593}{13} = 3.78918$$

8. *Standard Units.* If we assume that the arithmetic mean and the standard deviation of the weights of adult males are 150 lbs. and 20 lbs. respectively, then we may say that a man weighing 190 lbs. is

40 lbs. or 2 *standard units* above the average in weight. Likewise an individual weighing 120 lbs. may be considered as being 30 lbs. or 1.5 *standard units* under average weight. Conversely, if the arithmetic mean and the standard deviation for heights be 67 inches and 2.5 inches respectively, then an individual who is 2 standard units above the average height must be five inches above the average stature, or in other words must be 72 inches tall. The magnitude of an observation expressed in standard units is therefore defined as follows:

$$t_i = \frac{v_i - M_v}{\sigma_v} = \frac{\bar{v}_i}{\sigma_v} \tag{8}$$

It will be observed that these *standard* variates, t_i, are abstract numbers. For example, if the original variates be expressed in the unit inch then the unit of M_v, $\bar{v}$ and σ_v is also inch, and it follows that if both the numerator and denominator of a fraction be expressed as *inches* the quotient must be an abstract number, *independent of the unit employed in the measurements.* For instance, one series of variates would result if the height of each of a group of individuals were recorded in *inches.* However, if their heights had been recorded in *centimeters,* each of the resulting set of variates would be numerically about 2.54 times as large as the corresponding variate expressed in inches. Nevertheless, the *standard variates* obtained by both methods would agree in the case of each individual. Thus, if

$$M_v = 67 \text{ ins.} = 67(2.54) \text{ cms.},$$

and

$$\sigma_v = 2.5 \text{ ins.} = 2.5(2.54) \text{ cms.},$$

then for an individual 6 feet tall

$$v = 72 \text{ ins.} = 72(2.54) \text{ cms.},$$

$$\bar{v} = 5 \text{ ins.} = 5(2.54) \text{ cms.},$$

$$t = \frac{5 \text{ ins.}}{2.5 \text{ ins.}} = \frac{5(2.54) \text{ cms.}}{2.5(2.54) \text{ cms.}}, \text{ or}$$

$$t = 2 = 2.$$

With the aid of a computing machine, the series of standard variates corresponding to any observed series of variates may be completed very rapidly by means of a so-called continuous process. To

illustrate, we found that for the data of Table I, page 17,

$$M_v = 51.230769$$

$$\sigma_v = 3.86610$$

By formula (8), then

$$t_i = \frac{v_i - 51.230769}{3.86610} = -13.2513 + .258659\ v_i$$

In using this equation one should first subtract out 13.2513 from the machine, and then set up .258659 as a multiplier. The product of this multiplier by 51 will cause the value $t = -.059691$ to appear on the machine. By merely subtracting the multiplier two times, the value $t = -.577009$, corresponding to $t = 49$, appears. Continuing this "build-over" method, the following set of standard variates is readily obtained:

TABLE II

i	v_i	t_i
1	51	- .06
2	49	- .58
3	60	2.27
4	53	.46
5	48	- .84
6	51	- .06
7	42	- 2.39
8	50	- .32
9	51	- .06
10	52	.20
11	54	.72
12	53	.46
13	52	.20
Total	666	0.00

It is scarcely an exaggeration to state that the theory of mathematical statistics hinges on standard units. Although in many problems this might not appear on the surface, yet we shall see that the fact is nevertheless true.

9. The properties of the *moments* of standard variates are both

interesting and important. Thus

(9) $$\mu_{1:t} = M_t = 0$$

since

$$M_t = \frac{\Sigma t}{N} = \frac{1}{N}\sum \frac{v_i - M_v}{\sigma_v} = \frac{1}{N\sigma_v}\sum \bar{v}_i = 0$$

(see formula 5a)

Referring to formula (6) we see that

$$\mu_2 = \mu_2' - M^2$$

$$\mu_3 = \mu_3' - 3\mu_2' M + 2M^3$$

But since M_t has already been proven equal to 0,

$$\mu'_{2:t} = \mu_{2:t}$$

$$\mu'_{3:t} = \mu_{3:t}$$

Which is an important simplification in the moments of the standard variates.

(10) $$\mu_{2:t} = 1$$

for $$\mu_{2:t} = \frac{\Sigma t^2}{N} = \frac{1}{N}\sum \left(\frac{v_i - M_v}{\sigma_v}\right)^2 = \frac{1}{\sigma_v^2}\sum \frac{\bar{v}^2}{N}$$

$$= \frac{\mu_{2:v}}{\sigma_v^2} = 1$$ (see formula 7)

(11) $$\mu_{3:t} = \frac{\mu_{3:v}}{\sigma_v^3} = \frac{\mu_{3:v}}{\mu_{2:v}\sigma_v}$$

for

$$\mu_{3:t} = \frac{\Sigma t^3}{N} = \frac{1}{N}\sum \left(\frac{v_i - M_v}{\sigma_v}\right)^3 = \frac{1}{\sigma_v^3}\sum \frac{\bar{v}^3}{N} = \frac{\mu_{3:v}}{\sigma_v^3}$$

We see, therefore, that although the values of $\mu_{1:t}$ and $\mu_{2:t}$ are always 0 and 1 respectively, the value of $\mu_{3:t}$ will possess an abstract value depending, nevertheless, upon the variates themselves. The expression, $\mu_{3:t}$, is known as the coefficient of *skewness* and is denoted

by the symbol $\alpha_{3:\nu}$, i. e.

$$(12) \qquad \alpha_{3:\nu} = \frac{\mu_{3:\nu}}{\sigma_\nu^3} = \frac{\mu_{3:\nu}}{\mu_{2:\nu}\,\sigma_\nu}$$

Summary of Section I. From the viewpoint of *Elementary Mathematical Statistics,* we characterize a series of variates by its

(a) number, N,
(b) mean, M_ν,
(c) standard deviation, σ_ν, and
(d) skewness, $\alpha_{3:\nu}$

The *moments about the mean,* $\mu_{n\nu}$, are introduced solely to facilitate the determination of σ_ν and $\alpha_{3:\nu}$. Other moments, $\mu'_{n:\nu}$, are used to simplify the numerical calculation of the moments about the mean, $\mu_{n:\nu}$.

Verbally, we may state that the mean serves as a convenient average, and the standard deviation measures the concentration of the variates about their mean.

A thorough discussion of the significance of the coefficient of skewness must be slightly deferred. We may say at this time merely that the value of $\alpha_{3:\nu}$ depends obviously upon the value of $\mu_{3:\nu}$ and that a glance at the last column of Table I will lend weight to the statement that a positive or negative skewness indicates a weighted preponderance of those variates which are considerably greater than, or less than the mean, respectively.

Finally, the operations of mathematical statistics, and even certain comparisons in descriptive statistics, require that we introduce the notion of a standard variate, defined as follows:

$$t_i = \frac{v_i - M_\nu}{\sigma_\nu}$$

SECTION II.

INDIRECT METHOD OF OBTAINING ELEMENTARY FUNCTIONS

10. One of the fundamental theorems of moments states that if a constant be added to, or subtracted from each variate of a series, the moments computed about the mean for the revised series will be

identical with the corresponding moments of the original series. By way of a simple example:

The mean of the following five variates is 138, consequently the values of $\bar{v}$ are as given below:

i	v_i	$\bar{v}_i$
1	133	−5
2	142	4
3	138	0
4	141	3
5	136	−2
Total	690	0

If we subtract, say, 130 from each of the variates, then for the revised series x_1, x_2, x_3, x_4 and x_5,

i	$M_o = 130$ x_i	$\bar{x}_i = x_i - M_x$
1	3	− 5
2	12	4
3	8	0
4	11	3
5	6	− 2
Total	40	0

$$M_x = \frac{40}{5} = 8, \quad M_v = 130 + 8 = 138$$

The value subtracted, 130, is termed the *provisional mean*, and in general is designated by the symbol, M_o. It follows, therefore, that

(13) $$x_i = v_i - M_o$$

(14) $$M_v = M_o + M_x$$

(15) $$\mu'_{n:x} = \frac{\Sigma x^n}{N}$$

(16) $$\mu_{n:v} = \mu_{n:x}$$

It is understood that the functions of x are defined in precisely the same manner as corresponding functions of v, that is

$$M_x = \frac{\sum x}{N}$$
$$\bar{x}_i = x_i - M_x$$
$$\mu'_{n:x} = \frac{\sum x^n}{N}$$
$$\mu_{n:\bar{x}} = \frac{\sum \bar{x}^n}{N}$$

etc.

11. Formula (13) follows from definition, although (14)—seemingly self-evident—needs proof. Thus by (13)

$$\begin{aligned} v_1 &= M_0 + x_1 \\ v_2 &= M_0 + x_2 \\ v_3 &= M_0 + x_3 \\ &\vdots \\ v_N &= M_0 + x_N \\ \hline \sum v &= NM_0 + \sum x \end{aligned}$$

Dividing both sides through by N yields, by definition,

$$M_v = M_0 + M_x \qquad Q.\ E.\ D.$$

Formula (15) is proved by means of (13) and (14) as follows:

$$\begin{aligned} \bar{x}_i &= x_i - M_x && \text{(Definition)} \\ &= (v_i - M_0) - (M_v - M_0) && \text{(Formulae 13 and 14)} \\ &= v_i - M_v \\ &= \bar{v}_i && Q.\ E.\ D. \end{aligned}$$

Since

$$\mu_{n:v} = \frac{\sum \bar{v}^n}{N} \quad \text{and} \quad \mu_{n:x} = \frac{\sum \bar{x}^n}{N}$$

and we have just shown that always for corresponding values

$$\bar{V}_i = \bar{x}_i$$

the truth of (16) is apparent.

12. A comparison of tables III and I will reveal an advantage of the indirect over the direct method of calculation.

TABLE III

i	v_i	$M_0=50$ x_i	x_i^2	x_i^3
1	51	1	1	1
2	49	− 1	1	− 1
3	60	10	100	1000
4	53	3	9	27
5	48	− 2	4	− 8
6	51	1	1	1
7	42	− 8	64	− 512
8	50	0	0	0
9	51	1	1	1
10	52	2	4	8
11	54	4	16	64
12	53	3	9	27
13	52	2	4	8
Total		16	214	616

$$M_x = \frac{16}{13}$$

$$\mu'_{2:x} = \frac{214}{13} \qquad \mu_{2:x} = \mu'_{2:x} - M_x^2 = \frac{2526}{13^2}$$

$$\mu'_{3:x} = \frac{616}{13} \qquad \mu_{3:x} = \mu'_{3:x} - 3\mu_{2:x}M_x + 2M_x^3 = -\frac{21240}{13^3}$$

$$\sigma_x = \sqrt{\frac{2526}{13^2}} = 3.78918$$

$$\alpha_{3:x} = \frac{\mu_{3:x}}{\sigma_x \mu_{2:x}} = -\frac{21240}{2526\sqrt{2526}} = -.167303$$

$$M_v = 50 + \frac{16}{13} = 51\ 3/13$$

$$\sigma_v = \sigma_x = 3.78918$$

$$\alpha_{3:v} = \alpha_{3:x} = -.167303$$

It will be observed that the values

$$\mu_{2:x} = \frac{2526}{13^2} \quad \text{and} \quad \mu_{3:x} = \frac{-21240}{13^3}$$

agree exactly with those of Table I, namely

$$\mu_{2:v} = \frac{2526}{169} \quad \text{and} \quad \mu_{3:v} = \frac{-21240}{2197}$$

The following will illustrate an important advantage of the indirect method of determining the moments, $\mu_{n:v}$. Let us suppose that after computing the values of M_v, σ_v and $\alpha_{3:v}$ for the 13 variates of Table I we desire to delete the 13th variate, $v_{13} = 52$, and compute the values of M, σ and α_3 for the remaining twelve variates.

By the direct method of Table I, the revision would be quite laborious, but by the indirect method of Table III, revisions are made easily, as follows:

$$N = 13 - 1 = 12, \quad \sum x = 16 - 2 = 14, \quad \sum x^2 = 214 - 4 = 210;$$
$$\sum x^3 = 616 - 8 = 608$$

Consequently

$$M_x = \frac{14}{12}$$

$$\mu'_{2:x} = \frac{210}{12} \qquad \mu_{2:x} = \mu'_{2:x} - M_x^2 = \frac{581}{6^2}$$

$$\mu'_{3:x} = \frac{608}{12} \qquad \mu_{3:x} = \mu'_{3:x} - 3\mu'_{2:x}M_x + 2M_x^3 = -\frac{1600}{6^3}$$

$$\sigma_x = \frac{1}{6}\sqrt{581} = 4.01732$$

$$\alpha_{3:x} = \frac{-1600}{581\sqrt{581}} = -.114250$$

$$M_v = 50 + \frac{14}{12} = 51\ 1/6$$

$$\sigma_v = \sigma_x = 4.01732$$

$$\alpha_{3:v} = \alpha_{3:x} = -.114250$$

13. In a word, revisions of series arising from

(a) increasing or decreasing the number of variates,
(b) combining two or more series, or
(c) correcting the original variates

together with the resulting smaller numbers that result by employing the indirect method, lead us ordinarily to avoid using the direct method of section I in computing the fundamental functions, ***mean, standard deviation*** and *skewness*.

In practice, one continually faces the problem of revision. Thus, in business statistics, publications serving as sources of data frequently are obliged to present revisions for estimates made in previous issues. Moreover, monthly and annual endeavors to bring statistics up to date require the addition of variates to series. In problems arising in the field of psychology and education, it may develop after preliminary calculations have been made that one or more observations of the original series must be deleted due to the presence of factors such as unusual physical or mental impairment at the time of examination, cheating, etc. Again, we may desire to combine the statistics for several distinct intervals, for several classes, or for various schools of a city or state, etc.

In the numerical examples above, calculations were made in terms of fractions, rather than decimals, in order to emphasize the fact that the direct and indirect methods will yield identical results. Ordinarily, decimals are employed, and the results will consequently differ slightly.

Section III

Frequency Distributions

14. In dealing with *large* groups of quantitative data, the computation of the elementary statistical functions and an appreciation of the variation in the magnitudes of the series of measurements is greatly facilitated by systematically presenting the data in the form of a ***frequency distribution***. Such a distribution may present in tabular form

(a) each *different* variate observed. and
(b) the number of times that each different variate was observed in the investigation.

It is evident at the very outset, therefore, that if a frequency distribution merely reproduces precisely the same data that might otherwise have been listed serially, the values of M, σ and α_3 computed from such a frequency distribution must correspond exactly with the values of M, σ and α_3 that would have been obtained by the serial method. This *serial method* has been considered in the two preceding sections.

15. As an illustration, suppose that we consider the complete table from which the 13 variates, used in earlier computations, were taken. Since, according to the regulations, 17,000 numbers were withdrawn, we shall have 170 groups of one hundred numbers each, consequently 170 variates. These are listed below.

We shall see that one can compute the fundamental functions from the frequency distribution more readily than from Table IV. Again, certain phenomena are apparent at a glance at Table V, though by no means evident from a short inspection of Table IV. Thus the *range* of the variates is immediately observed in Table V, and the degree of symmetry in the distribution can be guessed rather accurately by one accustomed to computing the coefficient of skewness from distributions.

TABLE IV

Number of even numbers in 170 samples of 100 numbers each.

U. S. Order of Call, 1918

51	42	49	53	49	46	47	51	57	48
49	51	55	50	46	53	46	47	46	54
60	59	42	42	58	43	53	49	54	53
53	46	47	50	55	50	48	47	44	51
48	57	49	52	57	56	45	64	37	58
51	53	51	49	39	54	51	56	44	41
42	46	50	56	42	54	50	45	47	58
50	52	53	55	52	48	50	53	45	48
51	55	47	45	55	51	47	54	48	46
52	60	52	53	49	52	46	62	43	48
54	50	51	50	50	53	44	54	51	45
53	47	44	48	55	45	55	45	55	50
52	55	54	56	42	49	45	55	45	55
44	37	44	53	52	50	51	47	56	44
54	56	50	53	49	52	60	48	50	51
56	45	50	51	53	44	47	54	46	54
42	44	49	43	57	46	48	48	49	48

The frequency distribution for Table IV may be obtained readily by means of the "cross-five" method as follows:

TABLE V

Frequency Distribution for Data of Table IV

v	Tabulation	f
37	||	2
38		0
39	|	1
40		0
41	|	1
42	𝍸 ||	7
43	|||	3
44	𝍸 ||||	9
45	𝍸 𝍸	10
46	𝍸 𝍸	10
47	𝍸 𝍸	10
48	𝍸 𝍸 ||	12
49	𝍸 𝍸 |	11
50	𝍸 𝍸 𝍸	15
51	𝍸 𝍸 ||||	14
52	𝍸 ||||	9
53	𝍸 𝍸 ||||	14
54	𝍸 𝍸 |	11
55	𝍸 𝍸 |	11
56	𝍸 ||	7
57	||||	4
58	|||	3
59	|	1
60	|||	3
61		0
62	|	1
63		0
64	|	1
Total		170

16. The above type of distribution should be differentiated from others in which it has been found advantageous to combine the variates

into *classes* and likewise to group together the corresponding frequencies. A distribution of grades will serve to illustrate this second type of distribution.

TABLE VI

Distribution of Examination Grades of 168 Students

Class	Frequency
0- 10	0
11- 20	2
21- 30	3
31- 40	5
41- 50	7
51- 60	16
61- 70	39
71- 80	45
81- 90	41
91-100	10
Total	168

Such a table does not represent *exactly* the original data in which the grades were recorded for each student as an integral number of per cents; nevertheless, it gives a very good idea of the general form of the distribution and enables us to compute the fundamental functions with a considerable degree of accuracy.

17. *Discrete Variates.* The distribution of Table V is obviously one in which the variates can, from their very nature, be expressed only as integers. A distribution of this type is termed one of *discrete variates*, or one of a *discrete variable*. Common illustrations of this type are to be found in distributions of the number of individuals in a family, the number of petals on a flower, the number of coins turning up heads, etc.

18. *Continuous Variates.* In the majority of distributions the variates by their nature may differ by infinitesimals, and the observed values, as recorded, are merely more or less accurate estimates of the *true values*, which never can be established with *absolute* accuracy by any method of measurement. Thus the variates in the case of heights may be correct to the nearest inch, one-hundredth of an inch, or even the one millionth part of an inch, etc., but theoretically it can be shown that the chances that any measurement of a continuous variable is exact is about one in infinity. A frequency table for the distribution of continuous variates must always, therefore, be one of *grouped frequencies.*

19. The fundamental differences between distributions which may be classified as

(a) discrete
(b) grouped discrete, and
(c) continuous

are of vital importance whenever the accurate determination of the mean, standard deviation, or skewness, is concerned. We shall now illustrate in detail and by numerical examples the procedure which should be followed in each case.

20. *Frequency Distributions of Discrete Variates.*

If 180 dice were thrown, and a throw of a six spot counted a success, then the *expected* frequencies of successes that would be obtained in one thousand such trials are as follows:

TABLE VII

v	f	$M_o = 30$ x	x^2	x^3
15	1	−15	225	−3375
16	1	−14	196	−2744
17	2	−13	169	−2197
18	4	−12	144	−1728
19	6	−11	121	−1331
20	10	−10	100	−1000
21	16	− 9	81	− 729
22	23	− 8	64	− 512
23	31	− 7	49	− 343
24	41	− 6	36	− 216
25	51	− 5	25	− 125
26	61	− 4	16	− 64
27	69	− 3	9	− 27
28	75	− 2	4	− 8
29	79	− 1	1	− 1
30	80	0	0	0
31	77	1	1	1
32	72	2	4	8
33	64	3	9	27
34	56	4	16	64
35	46	5	25	125
36	37	6	36	216
37	29	7	49	343
38	22	8	64	512
39	16	9	81	729
40	11	10	100	1000
41	8	11	121	1331
42	5	12	144	1728
43	3	13	169	2197
44	2	14	196	2744
45	1	15	225	3375
46	1	16	256	4096

$\sum f = 1000$ $\quad M_o = 30$

$\sum xf = -27$ $\quad M_x = -.027$

$\sum x^2 f = 24687$ $\quad \mu'_{2:x} = 24.687$

$\sum x^3 f = 11259$ $\quad \mu'_{3:x} = 11.259$

$\mu_{2:x} = 24.6863$ $\quad \sigma_x = 4.96853$

$\mu_{3:x} = 13.2586$ $\quad \sigma_x \mu_{2:x} = 122.655$

$\alpha_{3:x} = .108097$

$M_v = 29.973, \quad \sigma_v = 4.96853, \quad \alpha_{3:v} = .108097$

Explanation. Since this distribution of discrete variates is an exact reproduction of the original data listed serially, we know that the moments obtained by the frequency distribution method must be identical with those which would have resulted had the serial method been employed. In fact

$$(17)\qquad \begin{cases} \sum f = N, \\ \sum xf = \sum x, \\ \sum x^2 f = \sum x^2, \text{ and} \\ \sum x^3 f = \sum x^3 \end{cases}$$

Numerically, $\sum x^n f$ is absolutely equivalent to $\sum x^n$. However, $\sum x^n f$ implies more; it indicates a brief and systematic method of attaining a total in which multiplication replaces repeated additions. Thus, in the serial method the value $x = 5$ would be added 46 times during the numerical determination of $\sum x$. In the frequency distribution method one multiplication, 5×46, represents likewise the contribution of this variate to the total $\sum xf = \sum x$.

If a computing machine be not available, the headings of Table VII should be

v	f	x	xf	x^2f	x^3f

and the totals $\sum x^n f$ obtained by a detailed process. With the aid of a computing machine the values of $\sum x^n f$ may be obtained readily by a continuous process, and it is necessary to record only the totals.

Since

$$(x+1)^3 = x^3 + 3x^2 + 3x + 1$$

it follows that

$$(18)\qquad \sum (x+1)^3 f = \sum x^3 f + 3\sum x^2 f + 3\sum xf + \sum f$$

Formula (18) is known as *Charlier's check.* By associating with each value, f the value of x^3 appearing on the next lower line, the value of $\sum (x+1)^3 f$ may be obtained as readily as that of $\sum x^3 f$. Then if equation (18) be satisfied we may assume with a considerate

degree of confidence that all five summations have been accurately determined.

It follows that we may now write, employing (17),

$$(19) \qquad \begin{cases} \mu_{2:x} = \dfrac{\Sigma x^2 f}{\Sigma f} \\ \mu_{3:x} = \dfrac{\Sigma x^3 f}{\Sigma f} \end{cases}$$

and observe that here, as in the serial method,

$$\mu_{2:x} = \mu'_{2:x} - M_x^2$$

$$\mu_{3:x} = \mu'_{3:x} - 3M_x\mu'_{2:x} + 2M_x^3$$

$$M_v = M_o + M_x$$

$$\mu_{2:v} = \mu_{2:x}$$

etc.

21 *The Grouping of Discrete Variates.* Occasionally frequency distributions of discrete variates contain so many different variates that some sort of grouping must be employed. Thus, the distribution of Table VII and the numerical calculations may be abbreviated as in Table VIII.

Explanation. The *class mark* of a class is defined as the rithmetic mean of the greatest and least variates that can occ· within that class. In Table VIII we might have used the class marks as values of v, but the use of a provisional mean, as has already been demonstrated, saves a large amount of labor.

TABLE VIII (Unadjusted)

Class	Class Mark	f	$M_o = 30,\ \lambda = 3$ x
14-16	15	2	- 5
17-19	18	12	- 4
20-22	21	49	- 3
23-25	24	123	- 2
26-28	27	205	- 1
29-31	30	236	0
32-34	33	192	1
35-37	36	112	2
38-40	39	49	3
41-43	42	16	4
44-46	45	4	5

Σf = 1000	M_o = 30, λ = 3
Σxf = - 9	M_x = - .009
$\Sigma x^2 f$ = 2817	$\mu'_{2:x}$ = 2.817
$\Sigma x^3 f$ = 405	$\mu'_{3:x}$ = .405
$\mu_{2:x}$ = 2.81692	σ_x = 1.67837
$\mu_{3:x}$ = .481058	$\sigma_x \mu_{2x}$ = 4.72783

$\alpha_{3:x}$ = .101750

M_v = 29.973, σ_v = 5.03511, $\alpha_{3:v}$ = .101750

The *class interval* is defined as the common difference between two consecutive class marks. In the example of Table VIII, the class interval has been chosen as the *unit* of x, consequently M_x and σ_x are expressed in class units. If λ denotes the class interval for a distribution, then

(20) $$M_v = M_o - \lambda M_x \ , \quad \text{and}$$

(21) $$\sigma_v = \lambda \sigma_x$$

Thus in Table VIII we had

$$M_v = 30 + 3(-.009) = 29.973$$
$$\sigma_v = 3(1.67837) = 5.03511$$

Since the skewness is an abstract number, completely independent of the unit employed

$$\alpha_{3:v} = \alpha_{3:x} \tag{22}$$

22. Table IX shows in the second, third, and fourth columns the values of M_v, σ_v and $\alpha_{3:v}$ which are obtained by various groupings of the data of Table VII. The grouping employed in Table VIII is listed as $D(3:2)$ in Table IX, the 3 denoting the number of different variates in each group, and the 2 designating the position of the first observed variate (i. e. $v = 15$) in the first grouping. Thus the classes of the grouping symbolized by $D(6:4)$ would be

12-17
18-23
24-29
etc.

From Table IX it may be observed that, although all of the values of M_v agree to a rather remarkable extent, nevertheless the unadjusted values of σ_v reveal the fact that an increase in the class interval is as a rule accompanied by an increase in the associated standard deviation and a decrease in the corresponding skewness.

23. In computing the moments $\mu'_{1:x}$, $\mu'_{2:x}$, and $\mu'_{3:x}$ for distributions of grouped frequencies, the assumption is made that each variate in a class may be treated as being numerically equal to the class mark. A mathematical investigation that lies beyond the scope of an elementary course shows that in the computation of M_x and $\mu_{3:x}$ it is entirely legitimate to treat each variate after this manner, but the demonstration also reveals that grouping tends to introduce a systematic error into the value of $\mu_{2:x}$. To eliminate this systematic tendency we find that one should introduce a correction and write

$$\mu_{2:x} = \mu'_{2:x} - M_x - \frac{1 - 1/k^2}{12} \tag{23}$$

where k denotes the *number of different variates* that are grouped together in each class. Thus, in Table VIII we should have introduced as a correction

$$\frac{3^2 - 1}{12 \cdot 3^2} = \frac{2}{27} = .074074$$

TABLE IX

Comparison of Adjusted and Unadjusted Values of σ_v and $\alpha_{3:v}$

(1) Grouping	(2) M_v	(3) Unadjusted σ_v	(4) Unadjusted $\alpha_{3:v}$	(5) Adjusted σ_v	(6) Adjusted $\alpha_{3:v}$
$D(1:1)$	29.973	4.969	.108	4.969	.108
$D(2:1)$	29.972	4.992	.106	4.967	.108
$D(2:2)$	29.974	4.995	.107	4.970	.108
Avg. $D(2)$	29.973	4.935	.106	4.968	.108
$D(3:1)$	29.974	5.030	.109	4.963	.113
$D(3:2)$	29.973	5.035	.102	4.968	.106
$D(3:3)$	29.972	5.041	.101	4.974	.105
Avg. $D(3)$	29.973	5.035	.104	4.968	.108
$D(4:1)$	29.968	5.089	.096	4.964	.103
$D(4:2)$	29.976	5.094	.104	4.970	.112
$D(4:3)$	29.976	5.078	.104	4.970	.112
$D(4:4)$	29.972	5.096	.098	4.970	.105
Avg. $D(4)$	29.973	5.089	.100	4.968	.108
$D(5:1)$	29.975	5.160	.105	4.962	.118
$D(5:2)$	29.975	5.170	.097	4.972	.109
$D(5:3)$	29.970	5.167	.094	4.970	.105
$D(5:4)$	29.970	5.163	.085	4.966	.096
$D(5:5)$	29.975	5.170	.100	4.972	.112
Avg. $D(5)$	29.973	5.166	.096	4.968	.108
$D(6:1)$	29.974	5.247	.107	4.961	.126
$D(6:2)$	29.976	5.256	.099	4.971	.117
$D(6:3)$	29.972	5.259	.087	4.974	.102
$D(6:4)$	29.974	5.250	.085	4.965	.100
$D(6:5)$	29.970	5.251	.080	4.966	.094
$D(6:6)$	29.972	5.259	.091	4.974	.108
Avg. $D(6)$	29.973	5.254	.092	4.968	.108
$D(7:1)$	29.977	5.347	.097	4.959	.121
$D(7:2)$	29.971	5.347	.093	4.958	.117
$D(7:3)$	29.972	5.358	.087	4.971	.109
$D(7:4)$	29.966	5.354	.070	4.966	.087
$D(7:5)$	29.974	5.361	.088	4.974	.110
$D(7:6)$	29.975	5.360	.086	4.973	.108
$D(7:7)$	29.976	5.365	.084	4.978	.105
Avg. $D(7)$	29.973	5.356	.086	4.968	.108

This would have resulted in the following revision:

$$\mu_{2:x} = 2.74285 \qquad \sigma_x = 1.65616$$
$$\mu_{3:x} = .481058 \qquad \sigma_x \mu_{4:x} = 4.54260$$
$$\alpha_{3:x} = .105899$$

$$M_y = 29.973, \qquad \sigma_y = 4.96848, \qquad \alpha_{3:y} = .105899$$

Again, for $k = 7$ we would use

$$\mu_{2:x} = \mu'_{2:x} - M_x^2 - \frac{7^2 - 1}{12 \cdot 7^2} = \mu'_{2:x} - M_x^2 \frac{4}{49}$$

When the simple adjustment of formula (24) is made, Table IX shows that the *systematic* errors in the values of σ_y and $\alpha_{3:y}$, caused by grouping, are eliminated. Thus in columns 5 and 6 the averages for each group are constant, consequently the errors remaining are accidental variations, which, due to a complete lack of compensation, still remain, but such discrepancies are not serious.

It should be noted that for distributions of discrete variates in which no grouping occurs, as in Table VII, the correction vanishes, since for $k = 1$

$$\frac{1 - 1/k^2}{12} = 0 \tag{24}$$

24. *Frequency Distributions of Continuous Variates.* The following will serve as an illustration of the method of obtaining the fundamental functions for a distribution of continuous variates.

TABLE X

Weights of 1000 Female Students

(Original Measurements Made to Nearest 1/10 lb.)

Class (Pounds)	Class Mark $\lambda = 10$	f	$M_o = 114.95$ x
70- 79.9	74.95	2	−4
80- 89.9	84.95	16	−3
90- 99.9	94.95	82	−2
100-109.9	104.95	231	−1
110-119.9	114.95	248	0
120-129.9	124.95	196	1
130-139.9	134.95	122	2
140-149.9	144.95	63	3
150-159.9	154.95	23	4
160-169.9	164.95	5	5
170-179.9	174.95	7	6
180-189.9	184.95	1	7
190-199.9	194.95	2	8
200-209.9	204.95	1	9
210-219.9	214.95	1	10
Total		1000	

$\sum f$ = 1000	M_o = 114.95
$\sum xf$ = 379	M_x = .379 class units
$\sum x^2 f$ = 3089	$\mu'_{2:x}$ = 3.089
$\sum x^3 f$ = 8131	$\mu'_{3:x}$ = 8.131
$\mu_{2:x}$ = 2.86203	σ_x = 1.69175
$\mu_{3:x}$ = 4.72769	$\sigma_x \mu_{2:x}$ = 4.84184

$\alpha_{3:x}$ = .976424

M_v = 118.74 lbs., σ_v = 16.9175 + lbs., $\alpha_{3:v}$ = .976424

Explanation: The class mark has previously been defined as the mean of the greatest and least variates that can be included in a class. Since the original measuremtns were made to the nearest tenth of a pound, the *true limits* of the 150-159.9 class are 149.95-159.95, and

their mean is 154.95, which accordingly is the class mark in this instance. If the original measurements had been made to the *nearest pound*, then the classes would be written

150.0-159.0
160.0-169.0

and the true limits of the 150.0-159.0 class would be 149.5 and 159.5 pounds respectively, and the corresponding class mark would be 144.5 lbs. It is apparent, therefore, that a table of continuous variates should specify clearly the accuracy with which the original measurements were made, for the values of the class marks and consequently that of the mean, hinges on this point.

It will be noticed that in this example the class interval has again been taken as the unit of x, and this fact must be taken into consideration in determining the value of M_v and σ_v.

Since the assumption is also made that the class mark may represent the magnitudes of all variates occurring in that class, the question of correcting the second moment, $\mu_{2:x}$ again arises. Since in each class of a distribution of continuous variates an infinite number of different variates may occur, the correction is in this case

$$\frac{1 - 1/k^2}{12} = \frac{1}{12}$$

Therefore, corresponding to formula (24), we must write, in order properly to adjust the second moment of a distribution of continuous variates

$$\mu_{2:x} = \mu'_{2:x} - M_x^2 - \frac{1}{12} \tag{25}$$

As before, neither the values of M_x nor $\mu_{3:x}$ require adjustment.

Summary of Section III. The frequency distribution is a device for presenting an extensive series of variates in a systematic and compact form. Not only are the phenomena of aggregation more readily perceptible by this method of presenting the data, but the calculations of the fundamental functions are facilitated.

The formulae for obtaining the mean, standard deviation and skewness are, with the exception of a single adjustment that may

arise, identical with those employed in the serial method. One need only observe that

$$N = \sum f$$
$$\sum x = \sum xf$$
$$\sum x^2 = \sum x^2 f$$
$$\sum x^3 = \sum x^3 f$$

The adjustment referred to is that we should in general regard

$$\mu_{2:x} = \mu'_{2:x} - M_x^2 - \frac{1 - 1/k^2}{12}$$

For ungrouped distributions of discrete variates this correction vanishes, since in this instance $k = 1$. For distributions of continuous variates, since here k would equal infinity, the correction is numerically equal to $1/12$.

These corrections will remove systematic errors in the standard deviation and skewness that arise from the phenomenon of grouping complete frequency distributions.

Editor's Note: This abstract of Elementary Mathematical Statistics will be continued in the May issue of the ANNALS.