Original

5 63 1

ASTRONOMICAL STATISTICS

by

Dr. E. Öpik

Prepared for Astronomy 11[b]

Academic Year 1930-1931

HARVARD COLLEGE OBSERVATORY

E. Öpik

SOME PROBLEMS OF ASTRONOMICAL STATISTICS

I Methods of Treating Frequency Tables etc.

1) **Counted numbers, frequency-tables and frequency-functions.** The primary subject of statistics is a counted number. The counted elements may be chosen either by a qualitative criterion (e.g. number of animals of a certain species), or by a quantitative one (number of stars within definite limits of magnitude). In astronomical statistics only the second kind of counting is of importance (even in counting, e.g. all stars visible on a plate, or within an area in the sky, without taking into account their magnitudes, the limits of magnitude are set implicite by the effective limiting magnitude of the count on the side of low luminosity, and by infinite luminosity on the other side).

By counting the number of things within consecutive adjacent limits of the measured quantity x we obtain a frequency-table. The number within the limits of x from $x - \Delta x/2$ to $x + \Delta x/2$ we denote by Δn_x. x we call the argument of the frequency-table, Δx the tabular interval; it is advisable to have a constant tabular interval over the whole range of x.

The frequency-table is exhaustive with respect of x if containing data for all possible values of x. (The frequency-table of stellar magnitudes cannot be exhaustive; the frequency-table of color-indices may be).

Representing the frequency-table by a smooth curve we obtain the frequency-function F(x), defined by

$$dn = F(x)\,dx \qquad \ldots\ldots\ldots\ldots (1)$$

where dn is the number counted within the limits x and x+dx. F(x) is an idealization of the frequency-table, useful for the purpose of mathematical treatment and for the derivation of statistical laws. The frequency-function is primitive if the condition is fulfilled

$$\Delta n_x = \int_{x-\Delta x/2}^{x+\Delta x/2} F(x)\,dx \qquad \ldots\ldots\ldots\ldots (2)$$

for each tabular interval; i.e., when within each tabular interval the area of the curve $y = F(x)$ equals the counted number. However, in most cases one has to smooth out some irregularities in the table which may be attributed to accidental uncertainty; in this case we obtain an idealized frequency-function, subject to the condition

$$\text{total counted number} = \sum_{x_1}^{x_2} \Delta n_x = \int_{x_1}^{x_2} F(x)\,dx \qquad \ldots\ldots (3)$$

without fulfilling in individual cases (2); x_1 and x_2 denote the smallest and largest values of x (not necessarily $-\infty$; 0 or $+\infty$ respectively).

Frequency-tables and functions may sometimes be discontinuous functions of x, e.g. when x can assume only integer values. In this case the preceding integrals are to be replaced by sums.

2) CLASSIFICATION OF ERRORS. All errors affecting some observed frequency may be divided into two principal groups:
a) Errors independent of the observer, like the natural uncertainty of a counted number, which has an absolute character. The cosmical error depending upon the properties of the portion of the universe we are studying (e.g. in the evaluation of stellar parallaxes from the observed proper motions, where the cosmical error is due to the spread in the true motions of the stars).
b) Observational errors (including also those of computation) which are under control of the observer, and depend upon the degree of perfection of the means of investigation.

3) NATURAL UNCERTAINTY OF A COUNTED NUMBER. Let p denote the probability of an event to happen, 1-p being thus the probability for it not to occur at a certain elementary trial. Let us make an experiment consisting of n trials, among which our event has occurred r times, in other words we have r "positive events." According to the Theory of Probability we have for the probability of the experiment n,r the expression

$$P_{n,r} = C_n^r \, p^r (1-p)^{n-r} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where

$$C_n^r = n!/r!(n-r)! \quad . \quad . \quad . \quad . \quad . \quad (5)$$

is the number of combinations of n things in groups of r things, and

$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots \cdot n$ is the factorial.

(Formula (4) is easily derived, if we take into account that the probability of the simultaneous appearance of a number of events arranged in a given order is equal to the product of their probabilities, and that the product $p^r(1-p)^{n-r}$ so obtained is to be multiplied by the number of different experiments with the same n and r but arranged in all possible combinations.
Formula (4) is a binomial member of $[p + (1-p)]^n$, and

$$\sum_{r=0}^{r=n} P_{n,r} = [p+(1-p)]^n = 1.$$

The factorial r! for r = 0 has no meaning; in all formulas of probabilities it is to be substituted in this case by 1, as it corresponds to the first member of the above mentioned binomial formula. Likewise we must put $C_n^0 = 1$.)

We may regard now r as our counted number. Repeating our experiment N times (n same in all), we shall find in each experiment different values of r, showing an accidental spread around a certain average value,

$$r_0 = pn \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6),$$

the accidental value of the difference $r - r_0$ characterizing the natural uncertainty in r.

(The average, or expected value $\bar{x}$ of a variable x whose probability is p_x is given by

$$\bar{x} = \sum xp_x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

According to this we have

$$\bar{r} = r_0 = \sum_{r=0}^{r=n} r\, C_n^r\, p^r (1-p)^{n-r} = pn \sum_{r-1=0}^{r-1=n-1} C_{n-1}^{r-1}\, p^{n-1} (1-p)^{(n-1)-(r-1)}$$

$$= pn \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8) \quad)$$

The characteristical measure of the size of accidental errors is ordinarily the dispersion, or the mean square deviation from the average, s. We have generally

$$s = \pm \sqrt{\frac{\sum (r-r_0)^2}{N}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

The probable frequency of $r-r_0$ is given by $NP_{n,r}$ (see formula (4)). Substituting this into (9), we have exactly

$$s = \pm \sqrt{pn(1-p)} = \pm \sqrt{r_0(1-p)} \quad . \quad . \quad (10)$$

$$\frac{\sum (r-r_0)^2}{N} = \frac{\sum_{r=0}^{r=n} (r-r_0)^2 NP_{n,r}}{N} = \sum_{r=0}^{r=n} (r-r_0)^2 P_{n,r}$$

$$= \sum_0^n r^2 C_n^r\, p^r (1-p)^{n-r} - 2pn \sum_0^n r\, C_n^r\, p^r (1-p)^{n-r} + p^2 n^2 \cdot 1$$

Taking into account (8), and transforming

$$\sum_0^n r^2\, C_n^r\, p^r (1-p)^{n-r} = np \sum_{r-1=0}^{r-1=n-1} (r-1+1) C_{n-1}^{r-1}\, p^{n-1}\, (1-p)^{(n-1)-(r-1)}$$

$$= np\left[(n-1)p+1\right]$$

we find $$\sum (r-r_0)^2/N = np(1-p)$$

Formula (10) is general, equally valid for large and small numbers. It gives the expression for the natural uncertainty of the counted number, the hypothetical number r_0 being given.

In the case of $p \to 0$, thus $n \to \infty$ with $np \to r_0$, we have

$$s = \pm \sqrt{r_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

(10) may be regarded as the expression of the relative uncertainty, (11) as the absolute uncertainty of the counted number. In dealing with the shape of frequency-functions, formula (10) is to be used, where r_0 is to be put equal to the expected number within the tabular interval, n to the total number counted (formula (3)). For great numbers approximatively $r_0 = \Delta n_x$ and $p = \Delta n_x / \sum \Delta n_x$. In dealing with the

absolute abundancy of the objects counted, the uncertainty is determined by (11); in this case the counted number is to be regarded as a finite sample taken from the infinite (very great) number of similar objects in the universe. We may remark that for most frequency-tables with a tabular interval small enough (10) becomes practically identical with (11).

The frequency-function of the deviation $r-r_0$ is identical with the frequency-function of r and is generally determined by (4). It is an asymmetrical curve, negative deviations being the more froquent, and at the same time smaller on the average than positive deviations.

(The computations according to (4) may be performed by using tables of factorials, or with the aid of Stirling's formula

$$n! = n^n e^{-n}\sqrt{2\pi n}\,(1 + 1/12n + 1/288n^2 + \ldots) \quad . \quad . \quad (12)$$

The greater r_0 is, the more symmetrical is the shape of the curve. There are special cases, when n or both r and n are great, and when formula (4) may be given an approximation more easy for computations.

The first case, of the classical theory of errors, is when n and r are large, p remaining finite with $n\to\infty$. In this case (4) may be substituted for by a Gaussian symmetrical error-function

$$\mathcal{P}_{n,r} = \frac{1}{\sqrt{2\pi r_0(1-p)}}\, e^{-\frac{(r-r_0)^2}{2r_0(1-p)}} \quad . \quad . \quad (13)$$

This is still a formula for the probability of an integer r; but it may be replaced by a mathematical fiction of a continuously changing error; in this case we have the Gaussion frequency-function of errors

$$F(t)dt = \frac{1}{\sqrt{\pi}}\, e^{-t^2} dt \quad . \quad . \quad . \quad . \quad (14)$$

where

$$t = r-r_0/\, s\sqrt{2} \quad . \quad . \quad . \quad . \quad (15)$$

This function is generally supposed to hold for most kinds of observational errors, s being the dispersion, emperically determinable.

The probability of the error falling within certain limits t_1 and t_2 is

$$\Theta_{t_1,t_2} = \frac{1}{\sqrt{\pi}} \int_{t_1}^{t_2} e^{-t^2} dt \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

The values of Θ for $t_1 = 0$ may be found in special tables.

Instead of the dispersion, the average spread of observational errors may be characterized by other quantities, such as the probable error, defined by the condition that one-half of the errors will exceed it, and the mean arithmetical deviation. In the case of a Gaussian they are given by

prob. error = 0.845(mean arithmetical) = 0.674(dispersion).

In special problems of astronomical statistics the condition of r_0, or r being large is not fulfilled; at the same time n is, or may be assumed to be large (e.g. equal to the total number of objects in the universe). This permits us to transform (4) into the following formula, valid for any values of r or r_0, supposing n is great in comparison with r: (Poisson's form)

$$P_{n,r} = \frac{r_0^r}{r!} e^{-r_0} \qquad \ldots\ldots\ (17)$$

Here p does not occur, as it has in this case no definite meaning.

(Substituting into (4) $p = r_0/n$, we have

$$P_{n,r} = \frac{n!}{r!(n-r)!} \frac{r_0^r}{n^r} \cdot (1-\frac{r_0}{n})^{n-r} = \frac{(n-r+1)}{n} \cdot \frac{(n-r+2)}{n} \cdot \ldots$$

$$\ldots \frac{(n-r+r)}{n} \cdot \frac{(1-\frac{r_0}{n})^n}{(1-\frac{r_0}{n})^r} \cdot \frac{r_0^r}{r!}$$

$\frac{(n-r+1)}{n} \cdot \frac{(n-r+2)}{n} \ldots\ldots \frac{(n-r+r)}{n}$ finite number or factors, tends to 1.

*** $(1 - r_0/n)^n \to e^{-r_0}$ for $n \to \infty$. This yields (17))

Formula (17) is indispensable for the computation of the probabilities of small values of r; it is a much better approximation than (13) as it takes full account of the asymmetry in the distribution of the deviations. It represents the error-function of small numbers; the dispersion remains exactly equal to $\sqrt{r_0}$.

$$s = \sqrt{\sum (r-r_0)^2 P_{n,r}} \ ;$$

$$\sum (r-r_0)^2 P_{n,r} = \sum_{r=0}^{r=n} r^2 \frac{r_0^r}{r!} e^{-r_0} - 2r_0 \sum_0^n \frac{r r_0^r}{r!} e^{-r_0} + r_0^2 \sum_0^n \frac{r_0^r}{r!} e^{-r_0}$$

For $r = 1,2,\ldots\ldots n$, we have

$$\sum \frac{r_0^r}{r!} e^{-r_0} = e^{-r_0}(1 + r_0/1! + r_0^2/2! + \ldots.) = e^{-r_0} e^{r_0} = 1;$$

$$\sum r \frac{r_0^r}{r!} e^{-r_0} = r_0 e^{-r_0} \sum_0^n r \frac{r_0^{r-1}}{r!} = r_0 e^{-r_0}(0 + \sum_1^n \frac{r_0^{r-1}}{(r-1)!}) = r_0$$

$$\sum r^2 \frac{r_0^r}{r!} e^{-r_0} = e^{-r_0} \cdot e^{r_0} \cdot r_0(r_0+1) = r_0(r_0+1)$$

$$s = \sqrt{r_0(r_0+1) - 2r_0^2 + r_0^2} = \sqrt{r_0}$$

*** Insert: Now, $(1-\frac{r_0}{n})^r$ as well as the product of the first r factors tend to 1, and $(1-\frac{r_0}{n})^n \to$, etc.

Hitherto we considered the case of the known probability p, or known average frequency, r_0. In practice we encounter mostly the inverse problem; given the observed number r, or the observed relative frequency, $p' = r/n$, to determine the probabilities of r_0, or p to be included within certain limits; we must make hypotheses, and estimate the chances in favor of each hypothesis concerning the unknown law symbolized in p.

The hypothetical probability p of the positive event in one elementary trial may be assigned any value from 0 to 1. Let π denote the probability a-priori for the hypothesis being confined to the limits p and p + dp; according to the theorem of the coincidence of two independent events, the probability of the experiment n,r to take place through the mediation of the given hypothesis is the product of the respective probabilities,

$$P_{n,r} \cdot \pi dp,$$

where $P_{n,r}$ is given by (4).

According to the theorem of the probability of a hypothesis a posteriori (after an experiment has been made), the probability that p is confined to p and p + dp is given by

$$Q_{p,r,n}\, dp = P_{n,r} \cdot \pi\, dp \Big/ \int_{p=0}^{p=1} P_{n,r}\pi dp \quad \ldots\ldots (18)$$

About the probability of p a priori (i.e. before observations were made) we in most cases do not know anything; in this case we cannot do better than to assume that all values of p are equally probable a priori, or to put $\pi = 1$. There may, however, exist cases where we know more about π, and where other assumptions are to be made.

($\int_0^1 \pi dp = 1$ according to the definition of probability; π = const. gives at once $\pi = 1$.)

This gives us, after substituting (4) and performing the integration

$$Q_{p,r,n}\, dp = (n+1)C_n^r\, p^r(1-p)^{n-r} dp \quad \ldots\ldots (19)$$

($Q = p^r(1-p)^{n-r}\, dp \Big/ \int_0^1 p^r(1-p)^{n-r} dp$. The Gamma-function

$$\Gamma(m,n) = \int_0^1 x^m(1-x)^n\, dx$$ for integer m and n

is easily found by partial integration:

$$\Gamma(m,n) = n!/(m+1)(m+2)\ldots(m+n+1) = m!n!/(m+n+1)! = 1/(m+n+1)C_{m+n}^m$$

Substituting m = r, (n) = n - r, we obtain (19)

Original

The probability of p to be found between p and p+dp is the same as for $r_0 = pn$ to be contained between $\bar{r}_0 = np$ and $\bar{r}_0 + d\bar{r}_0 = np + ndp$, and is thus given by (19), where $dp = dr_0/n$:

$$S_{r_0,r,n}\, dr_0 = \frac{n+1}{n} C_n^r p^r (1-p)^{n-r}\, dr_0 \quad \ldots\ldots\ldots (20)$$

and $p = r_0/n$.

For great values of n, $n+1/n \to 1$, and (20) becomes similar to (4), though in no way identical with it, because (20) is a <u>continuous</u> function of the variable r_0, (4) is a <u>discontinuous</u> one of the integer r.

The <u>most probable</u> hypothesis, giving a maximum of the expression in (19), is easily found equal to $p_0 = r/n = p'$.
(According to the rule of finding maxima, we put the derivative of (19) with respect to p equal to zero and solve the equation for p, knowing that p is not 0 or 1)

The average expected hypothesis, however, is not p'; attributing to each hypothesis a weight proportional to its probability, we have

$$\bar{p} = \int_0^1 p\, Q_{p,r,n}\, dp = r+1/n+2 \quad \ldots\ldots (21)$$

(The integral is a gamma-function of the same kind as discussed above).

This gives

$$\bar{r}_0 = (r+1)n/n+2 \quad \ldots\ldots\ldots\ldots (22)$$

For n large we have $\bar{r}_0 = r+1$; thus the <u>average hypothetical frequency</u> exceeds the observed frequency and the most probable one by 1.
(The average hypothetical frequency $\bar{t}_0$ of the negative event in n trials we have, substituting in (22) r by n-r; this gives $t_0 = (n-r+1)n/n+2$ (23) and the check is

$$\bar{r}_0 + \bar{t}_0 = n)$$

Formula (22) refers to the ~~absolute uncertainty of~~ average frequency derived from a single counted number ($n \to \infty$), or ~~to the relative uncertainty in~~ from a table consisting only of two sections (+event and − event). For the members of a frequency-table with many sections it is not valid. Generally, in frequency-tables we have to assume not the average, but the most probable value of the hypothetical frequency, $r_0 = r$.

The uncertainty of the hypothetical frequency may be measured again by the mean square deviation, or the dispersion.

The <u>exact</u> formulae are: for the dispersion s_1 of r_0 around its most probable value $r_0 = r$,

$$s_1 = \pm \sqrt{\int_0^1 (pn-r)^2\, Q_{p,r,n}\, dp} = \pm \sqrt{\frac{n^2(r+2)-nr(r+6)+6r^2}{(n+3)(n+2)}} \quad \ldots (23)$$

symmetrical

$s_1(r) = s_1(n-r)$

For n great in comparison with r this becomes

$$s_1 = \pm \sqrt{r+2} \quad \ldots\ldots\ldots\ldots (23')$$

The dispersion s_2 of r_0 around its average expected value, $\bar{r}_0$, is

$$s_2 = \pm \sqrt{\int_c^1 \left[pn - \frac{(r+1)n}{(n+2)}\right]^2 Q_{p,r,n}\, dp} = \pm \sqrt{(r+1)\cdot\frac{n^2(n-r+1)}{(n+2)^2(n+3)}} \quad \ldots\ldots (24)$$

For n great we have

$$s_2 = \pm \sqrt{r+1} \quad \ldots\ldots\ldots\ldots (24')$$

~~We see that the uncertainty in the hypothetical frequency is always greater, than the uncertainty in the observed frequency of the same amount.~~ We must also remember that the distribution of the deviations for small values of r is asymmetrical.

The approximations to (19) and (20) in special cases are very similar to those of formulas (13) and (17). For both n and r large we have a Gaussian with the dispersion $\sqrt{r}$ smaller than the dispersion of the exact formula. ($\sqrt{r}$ differs little from $\sqrt{r+1}$ if r is large).

For n large, r small we have

$$S_{r_0,r}\, dr_0 = (r_0^r\, e^{-r_0}/r!)dr_0 \quad \ldots\ldots\ldots (25)$$

which is a good approximation to the exact formula, giving an average $r_0 = r+1$ with a dispersion $\pm\sqrt{r+1} = \pm\sqrt{\bar{r}_0}$ (this may be easily tested, according to the procedure many times applied).

4) RELATIVE UNCERTAINTY IN COUNTED NUMBERS OR PROPORTIONS. For a known hypothesis, p, the relative uncertainty in r is generally (exact formula):

$$\frac{\bar{s}}{r} = \pm\sqrt{\frac{1-p}{r}} \quad \ldots\ldots\ldots\ldots (26)$$

and for $p \to 0$, $n \to \infty$ $\quad = \pm\sqrt{1/r}$

The spread in the observed proportion, $p' = r/n$, is given by

$$\frac{s}{n} = \pm\sqrt{\frac{p(1-p)}{n}} \quad \ldots\ldots\ldots\ldots (27)$$

This formula leads to the fundamental theorem of the theory of errors, that the mean error of an average observed quantity (here--the ratio p' approaching p) decreases as the square root of the number of observations.

Similar formulae for the relative uncertainty of the hypothesis may be derived from the formulae of section 3.

5) INTERRELATED ELEMENTS. The preceding formulae are valid only in the case when all counted individuals are independent of one another. However, in nature we find many cases when the appearance of one element may have as a consequence the observability of another similar element which, if judged from its physical properties only, cannot be distinguished from any other independent element, and is thus counted as an independent individual. In such a case of interrelated elements the formulae of probabilities and errors are applicable only with respect to the number of independent groups of elements. As an example let us imagine that all stars are observable visual pairs; if there is found a counted number of 10 stars upon a certain area = 1, the number of independent groups observed is $r = 10/2 = 5$, the average hypothetical density of pairs per unit area is $\bar{r}_0 = 6 \pm \sqrt{6} = 6 \pm 2.5$, and the average density of individual stars is 12 ± 5. The relative error in this case is greater than for an equal observed number of independent elements. In practical problems the character of interrelation is ordinarily much more complicated than in our example; e.g., the number of elements in a group may itself represent a certain frequency-function, and in the above mentioned case of the double stars their distribution of distances is of great importance. The special cases of interrelation will be analyzed in the chapters dealing with the individual problems of astronomical statistics. From the discordance between the predicted and the observed frequencies it is possible to study the character of interrelation that disturbs the statistical data.

6) CORRECTION OF FREQUENCY-FUNCTIONS FOR NATURAL UNCERTAINTY IN THE COUNTED NUMBERS is attained by smoothing out all unevennesses which are within the limits of the natural error. Figure 1 is an illustration of the practical proceeding. The rectangles represent the numbers (r) counted within successive tabular intervals; the arrows indicate the size of the natural dispersion $= \pm\sqrt{r+2}$. The area of the curve must be equal to the sum of the rectangular areas. In drawing the curve it may be held in view that, from considerations of probabilities, one third of the deviations are likely to exceed the dispersion, and one twentieth to exceed the double dispersion, and that

Original

for small r positive deviations are on the average greater than negative deviations; further, negative values of F(x) are, of course, excluded.

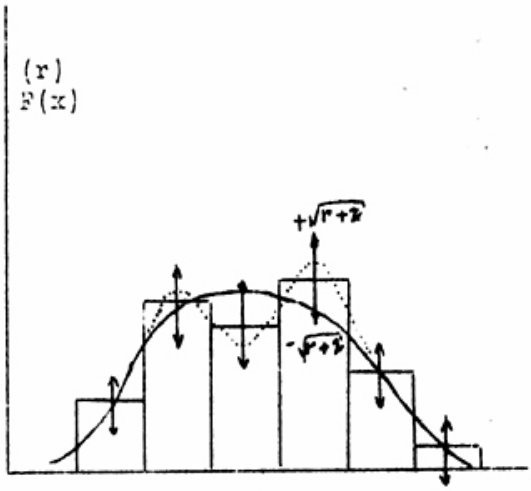

Figure 1.
Full curve= smoothed freq-function.
Dotted curve= primitive freq-function.

In certain cases we know the general character of F(x), and the smoothing is then confined to the determination of the parameters of this function, from a system of equations of the form

$$r = F(x)\Delta x \quad . \quad . \quad . \quad (28)$$

supposing the tabular interval Δx is not too large. In solving system (28), by least squares or some other analogous procedure, the weights of the individual equations are to be assumed proportional to $1/r+2$. Moreover, an additional equation

$$\sum r = \sum F(x)\, dx = N \quad (28')$$

must be introduced as the condition determining the area of the curve.

The smoothing for natural uncertainty must be made before the correction for accidental errors in x (see following paragraphs), because otherwise all accidental irregularities in the curve will be accentuated, and spurious maxima and minima will appear. From the same standpoint it is necessary to make the smoothed curve as simple as possible, in other words--to make the average radius of curvature as great, or the average curvature as small, as compatible with the limits allowed by the natural uncertainty. Doubtful waves in the primitive frequency-function (Fig. 1) must be avoided. (Mathematically the measure of "complexity" of the curve within the limits x_1 and x_2 may be assumed equal to

$$\overline{(1/\rho^2)} = \int_{x_1}^{x_2} \left(\frac{d^2y}{dx^2}\right)^2 \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{-3} dx \quad . \quad . \quad . \quad . \quad . \quad (29)$$

Correction

or to the mean square of the inverse radius of curvature; the greater this quantity, the more "complicated" is the curve.)

7) DECIMAL EQUATION is an error caused by the preference given by the observer, or the computer to certain figures, chiefly in the last place. The most common kind of this error reveals itself in the last decimal, whence its name. In rough observations, e.g. in estimates of magnitudes of meteors which are noted at the best only to one half a magnitude, the "decimal" equation consists in a more frequent record of integer numbers as compared with the fractional ones or vice versa. In observations it depends highly upon the individuality of the observer, and can be determined only by counting the relative frequency of the different "decimals" In computations sometimes its character may be pre-

dicted from the rule used in rounding off the last place, though it is advisable even in this case to use the empirical method, because all circumstances cannot be foreseen.

Let $\bar{\nu}$ denote the average frequency of all decimals, ν_i the observed frequency of the i^{th}decimal; we will call the <u>decimal excess</u> the quantity

$$\delta_i = \nu_i/\bar{\nu} \qquad \ldots\ldots\ldots\ldots (30)$$

In the case when F(x) does not change too much within one tabular interval, a frequency-table may be corrected for decimal equation in the following way. Let $\bar{\delta}_i$ be the average decimal excess, r'_i -- the counted frequency within one tabular interval; the corrected frequency will be

$$r_i = r'_i/\bar{\delta}_i \qquad \ldots\ldots\ldots (31)$$

after this a slight adjustment must be made, to fulfill the condition

$$\sum(r_i - r'_i) = 0 \qquad \ldots\ldots\ldots\ldots (32)$$

In the case when in adjacent tabular intervals δ_i changes from > 1 to < 1 consecutively, formula (31) may be used only for the intervals where $\delta_i > 1$, and the residual frequency $r'_i - r_i$ may be then distributed between the adjacent intervals $(i-1)^{st}$ and $(i+1)^{st}$, either in equal parts, or proportionally to their counted frequencies; this procedure saves any further adjustment.

The relative amount of the decimal equation does not depend upon the absolute frequency, and though ordinarily small, in certain cases it may exceed the natural dispersion. If data for the decimal correction are available it is advisable to introduce this correction before the smoothing of the curve.

8) SYSTEMATICAL CORRECTIONS OF THE ARGUMENT affect the frequency-function in the same way as a <u>change of the variable</u>; let t denote the new argument, f(t) its frequency-function. We have

$$dn = f(t)\,dt = F(x)\,dx$$

whence

$$f(t) = F(x)/\frac{dt}{dx} \qquad \ldots\ldots\ldots\ldots (33)$$

E.g., for $t = \log nx$, we have $f(t) = x\,F(x)$. If $u(x)$ is the systematical correction, $t = x+u$, and formula (33) is transformed into

$$f(x+u) = \frac{F(x)}{1 + du/dx} \qquad \ldots\ldots\ldots (33')$$

In numerical tables, for a given tabular interval Δx we may assume

$$\frac{du}{dx} = \frac{u(x + \Delta x/2) - u(x - \Delta x/2)}{\Delta x} \qquad \ldots\ldots (34)$$

9) INFLUENCE OF ACCIDENTAL ERRORS IN THE ARGUMENT UPON FREQUENCY FUNCTIONS. (The conclusions of this and subsequent paragraphs may be applied equally in certain problems other than the above mentioned, generally in cases where a distribution function is changed by some kind of spread in the argument, not necessarily accidental; e.g., the correction of observed contours of spectral lines for the effect of imperfect definition is mathematically the same problem.)

Let x denote the true, ξ the observed argument, and $Z = \xi - x$ -- the accidental error (no matter whether observational or cosmical).

Let $\chi(x,Z)$ denote the frequency-function of Z, generally assumed to depend upon x also (a generalization very important in practice; an assumption of the same error-curve, e.g. a Gaussian with constant dispersion all over the range of x corresponds seldom to reality); χ must vanish for $Z = -\infty$ and $+\infty$, though it may vanish outside of certain finite limits of Z. The condition

$$\int_{-\alpha}^{+\alpha} \chi(x,Z)dZ = 1 \qquad \ldots\ldots\ldots (35)$$

gives to χ the meaning of a probability. χ need not be a continuous function throughout and the integrals in the regions of discontinuity are thought to be replaced by the corresponding sums.

Of the number $dn = F(x)dx$, within x and x+dx, the fraction $\chi(x,Z)dZ$ will have the error between Z and Z+dZ, the total number within these limits of x and Z being thus

$$F(x)dx \cdot \chi(x,Z)dZ = F(\xi - Z)\chi(\xi - Z, Z)dx\,dZ \ldots\ldots (36)$$

For a given x, and Z between Z and Z+dZ, the average value of ξ is x+Z+dZ/2; similarly, the average value of $\xi + d\xi$, corresponding to x+dx, is x+dx+Z+dZ/2. Hence the average or effective value of $d\xi$ is $d\xi = dx$; the total frequency found within the limits ξ and $\xi + d\xi$ we find by taking the sum of (36) over the entire range of Z, i.e.

$$\varphi(\xi)d\xi = d\xi \int_{-\alpha}^{+\alpha} F(\xi - Z)\chi(\xi - Z, Z)dZ \ldots\ldots\ldots (37)$$

where $\varphi(\xi)$ denotes the observed frequency-function. Equation (37), which may be called the equation of diffusion, is of fundamental importance in this and similar problems. Given the error-function, χ, and the true frequency-function, F, the observed frequency-function is easily computed according to (37).

The inverse problem--to determine F, the true frequency-function, if φ and χ are given, is more difficult, and involves the solution of the integral equation (37). However, though a difficult mathematical problem even in restricted cases, the practical solution is easily found by successive approximations in the most general case, on the condition of a certain, though remote resemblance between F and φ; in other words, the error-dispersion must not be strong enough to change F altogether, so that its main features must remain recognizable in φ; this condition will be defined more precisely later on. It is easily conceivable

that only such a case is of practical interest; if the error-dispersion is so strong that it conceals all the original features of the curve, the correction for accidental errors will be extremely uncertain and of no real meaning.

Thus, in this case, in practice the only important one, we take as a first approximation $F = \varphi$, and compute

$$\varphi_1(\xi) = \int_{-\infty}^{+\infty} \varphi(\xi - z)\, \chi(\xi - z, z)\, dz \quad \ldots\ldots\ldots \quad (38)$$

Comparing (38) and (37), we find that φ_1 refers to φ as φ does with respect to F. For the second approximation we assume

$$F_1 = \varphi + (\varphi - \varphi_1) = 2\varphi - \varphi_1 \quad \ldots\ldots\ldots\ldots \quad (39)$$

Similarly

$$\varphi_2(\xi) = \int_{-\infty}^{+\infty} F_1(\xi - z)\, \chi(\xi - z, z)\, dz \quad \ldots\ldots\ldots \quad (40)$$

and the third approximation for F

$$F_2 = \varphi + (F_1 - \varphi_2) \quad \ldots\ldots\ldots\ldots \quad (41)$$

From one to three approximations ordinarily lead to a good solution which is to be checked by substituting into (37).

All this is easily performed in a purely numerical way. Both, the observed frequency-function φ, and the error-function χ we represent by frequency-tables with the same size of the tabular interval in each, $(\Delta x) = \delta\xi = \delta Z$. Let $\Delta\chi_k$ denote the fraction of Δn_i (see page 1), the true number, in the i^{th} tabular interval, which by error-dispersion is spread into the $i+k^{th}$ interval; let $i+k = m$, or $i = m-k$, and $\Delta\nu_m$ denote the observed frequency in the m^{th} interval; then

$$\Delta\nu_m = \sum_{k_1}^{k_2} \Delta n_{m-k} \cdot \chi_k \quad \ldots\ldots \quad (42)$$

where the sum is to be taken over all important values of k, i.e. such that χ_k is not too small as compared with the accuracy of computations. Formula (42) represents the numerical equivalent of (37). The successive approximations may be found in a manner described analytically above; however, in the numerical proceeding a still simpler method may be used, giving at once the approximations F_1, F_2 etc. The method consists in inverting the process of spreading in formula (38) and substituting it by a process of concentration; all amounts that correspond to (42) (with an approximation for Δn) must go from the interval m-k to m, we make go from m to m-k. This gives for the first approximation

$$\Delta n_i' = \Delta\nu_i + \Delta\nu_i \sum_{\substack{k=\ldots-2,-1,+1,+2\ldots \\ k\neq 0}} \chi_k - \sum_{k\neq 0} \Delta\nu_{i+k}\, \chi_k \quad \ldots \quad (43)$$

The second approximation is

$$\Delta n''_i = \Delta\nu_i + \Delta n'_i \sum_{k\neq 0} \chi_k - \sum_{k\neq 0} \Delta n'_{i+k} \chi_k \quad . . (43')$$

(Note that the first member, $\Delta\nu_i$, is the same in all approximations). The corrections, e.g. in (43'), consist of positive and negative members, but only the positive members are to be computed, because the term $\Delta n_{i+k}\chi_k$, negative in the computation of $\Delta n''_i$, is a positive one in the computation of $\Delta n''_{i+k}$.

Remark. If, in the course of making the approximations, negative values of Δn are found, it means that probably the error-dispersion is over-estimated, and the computation of further approximations becomes useless.

10

GENERAL THEOREM OF DISPERSIONS. The spread of an arbitrary function may be measured just in the same manner as the error-dispersion, i.e. by the mean square deviation, s_x, from the arithmetical mean, x_0.

$$x_0 = \Sigma x/n \quad . . . . . (44)$$

$$s_x = \pm\sqrt{\frac{\Sigma(x-x_0)^2}{n}} \quad . . . . . (45)$$

Let s_z denote the error-dispersion of the accidental quantity Z, such that the expectation, or average probable value of $Z_0 = \Sigma Z/n$ is zero, not only for the total range of x, but also for every selected value of x. In other words, Z is supposed to be free of systematical errors (which are treated in a different manner, section 8). (The observed value of Z_0 may, and will differ from zero, on account of its accidental character). And let s_ξ denote the dispersion of the observed frequency-function, defined in a similar manner relative to the observed mean value ξ_0.

We have:

$$\xi = x + Z, \qquad \xi_0 = x_0 + Z_0 = x_0 \quad . . . . . (46)$$

$$\Sigma(\xi - \xi_0)^2 = \Sigma(x-x_0)^2 + \Sigma Z^2 + \Sigma 2(x-x_0)Z \quad , \text{ or, dividing by } n$$

$$s_\xi^2 = s_x^2 + s_z^2 + \frac{2\Sigma(x-x_0)Z}{n} \quad . . . . . (47)$$

Now, the expectation of the third term in (47) is 0 (because for each given $x-x_0$ the expectation of Z is 0, and the expectation of the product is also 0); hence,

$$s_\xi^2 = s_x^2 + s_z^2 \ .$$

or the expected value of the observed dispersion is equal to the geometrical sum of the true dispersion and the error-dispersion.

In real cases the third term of (47) will not be zero; it gives us a means for evaluating the statistical uncertainty of our theorem. We find finally,

$$s_\xi^2 = s_x^2 + s_Z^2 \pm \frac{2 s_x \cdot s_Z}{\sqrt{n}} \quad \ldots\ldots (48)$$

(If we have a sum of n numbers, each of which contains an accidental error h, the accidental error of the sum is $\sum_n h_i$; the expected dispersion of the sum is

$$s^2 = (\sum h_i)^2 = \sum h_i^2 + 2\sum_{i \neq k} h_i h_k = \sum h_i^2 = nh^2$$

where h denotes the mean dispersion for all members.

$$s = \pm h\sqrt{n} \quad \ldots\ldots (49)$$

The only assumption of this otherwise purely algebraic theorem is that the expectation of $\sum h_i = 0$. (49) is the famous formula, upon which is based the theory of observational errors. Applying this, we have

$$\sum (x - x_0) Z = \sum (x - x_0) \cdot (\pm s_g) = s_Z \sum \pm (x - x_0) = \pm s_Z s_x \sqrt{n}$$

Substituting this into (47), we obtain (48).

Formula (48) is valid for arbitrary frequency-functions; thus, one of the most important characteristics of the true frequency-function, its dispersion, may be found directly, without the complicated methods described in the preceding and following sections. The generality of the error-function holds also for (48) except that the expression for the uncertainty is derived on the assumption of a constant s_Z, independent of x; in the more general case of variable s_Z, the substitution of an average value for it serves the practical purpose perfectly, because what we need is only a guess of the size of the error, not its exact analytical expression.

In the case of variable error-dispersion, $s_Z = f(x)$, the value of s_Z in (48) must be computed from (50) in conformance with our presumption:

$$s_Z^2 = \frac{\sum z_i^2}{n} = \frac{\int s_Z^2 F(x) dx}{n} \quad \ldots\ldots (50)$$

A rough approximation for F(x), e.g. $F(x) = \varphi(x)$, is ordinarily more than wanted.

From (48) we may judge how reliable the correction for error-dispersion is. As a limit of reliability we may put

$$\frac{2 s_x \cdot s_Z}{\sqrt{n}} < s_x^2/2 \ ,$$

Original

or,
$$\frac{s_x}{s_z} > \sqrt{\frac{16}{n}} \quad \ldots\ldots\ldots\ldots \quad (51)$$

$$\frac{s_\xi}{s_z} > \sqrt{1+\frac{16}{n}} \quad \ldots\ldots\ldots\ldots \quad (52)$$

For different values of n we have:

TABLE I.

Limits of Reliability for s_x

n=	4	10	50	100	1000	~~10000~~ 10 000
$s_x/s_z >$	2	1.27	0.57	0.40	(0,13)	(0.04)
$s_\xi/s_z >$	2.2	1.61	1.15	1.08	(1,008)	(1.0008)

These limits of reliability take into account only a kind of natural uncertainty depending chiefly upon the total number of individuals. The uncertainty e in the adopted value of s_z makes the apparent gain in the limit of reliability, illustrated by the preceding Table, illusory. The uncertainty of (48) will be given more generally by

$$S = \pm 2s_x s_z \sqrt{\frac{1}{n} + \frac{e^2}{s_z^2}} \quad \ldots\ldots\ldots \quad (53)$$

For s_x^2 one may put $s_\xi^2 - s_z^2$.

$$(s_z' \pm e)^2 = s_z'^2 + e^2 \pm 2es_z' = s_z^2 \pm 2es_z$$

because the constant e^2 is already included in the apparent error-dispersion s_z^2. According to the theorem of summation of accidental errors, the total dispersion will be

$$\pm \left[(2es_z)^2 + (2s_x s_z/\sqrt{n})^2\right]$$ which gives (53) ~~etc~~

The conditions set above determine when s_x is probably real; of course, having a reliable value for s_x, it is not certain yet that the true shape of the frequency-function may be determined with success; for this purpose a much greater n is required.

We may ask now, when does the correction for error-dispersion become too small or doubtful, so that its determination will not be worthy of the labor of computation, especially if the risk of changing our data by arbitrary assumptions is taken into account. Using still expression (48), the limit of reality of the correction, defined by the condition that the uncertainty does not exceed s_z^2, is given by

$$\frac{2s_x s_z}{\sqrt{n}} < s_z^2 \quad \text{or,} \quad \frac{s_x}{s_z} < \sqrt{\frac{n}{4}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (54)$$

$$\frac{s_\xi}{s_z} < \sqrt{1+\frac{n}{4}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (55)$$

For small values of n all reliable cases may be unreal and vice versa. The limiting number n is defined by

$$\sqrt{\frac{n}{4}} = \sqrt{\frac{16}{n}} \qquad \text{or} \quad n = 8.$$

For a number of individuals less than this, corrections for error-dispersion possess no real weight.

TABLE II

Limit of Reality of the Correction for Error Dispersion

$n =$	4	10	50	100	1000	10,000
$s_x/s_z <$	1.0	1.58	3.5	5.0	(15.8)	(50)
$s_\xi/s_z <$	1.4	1.87	3.7	5.1	(15.8)	(50)

From Tables I and II we infer, that for n = 50, the correction for error-dispersion is of practical value for $s_\xi/s_z > 3.7$; $s_\xi/s_z > 1.15$. If s_ξ/s_z is less than 1.15, the result for s_x is extremely uncertain; if s_ξ/s_z exceeds 3.7, the correction is so small that it will be within the limits of natural uncertainty, and we may safely put $s_x = s_\xi$. A special uncertainty in s_z, of course, changes these conclusions, because then (53) must be taken as the measure of uncertainty.

The above results refer formally to the actual apparent spread as defined by (45), relative to the observed mean x_0. If there is supposed to exist a certain true mean value, $\bar{x}_0$, from which x_0 may differ, the expected true dispersion relative to the unknown x_0 is given by

$$\bar{s}_x = \pm\sqrt{\frac{\Sigma(x-\bar{x}_0)^2}{n-1}} = \pm\sqrt{\frac{n}{n-1}}\, s_x \qquad \ldots\ldots (56).$$

The change is in a constant ratio, thus our conclusions are valid also in the case of the true dispersion.

(Let h_i denote true deviations from an unknown mean value, h their mean dispersion

$$\Delta_i = h_i - \frac{\sum h_i}{n}$$

their observed, or apparent deviations, and

$$s^2 = \frac{\sum \Delta_i^2}{n}$$

the apparent dispersion.
We have

$$\sum \Delta_i^2 = \sum h_i^2 - 2\sum\left(\frac{h_i \sum h_i}{n}\right) + \sum\left[\frac{(\sum h_i)^2}{n^2}\right]$$

but $\sum h_i = \text{const.}$; $(\sum h_i)^2 = \sum h_i^2 + 2\sum h_i h_k = \sum h_i^2 = nh^2$

We find

$$\sum \Delta_i^2 = nh^2 - 2nh^2/n + n \cdot nh^2/n^2 = (n-1)h^2, \text{ or}$$

or

$$h = \pm\sqrt{\frac{\sum \Delta_i^2}{n-1}}$$

11) SYSTEMATIC CHANGES PRODUCED BY ACCIDENTAL ERRORS. For a given x (notations as before) the mean value of the accidental error, $\bar{Z} = 0$, and the mean observed value $\bar{\xi} = x$. The reverse, however, is not true except in the very special case of $F(x) = \text{const.}$ Let $\eta = \eta(\xi)$ denote the mean true value of the argument x for a given ξ, $\bar{Z}$ the average error, so that

$$\eta = \xi - \bar{Z} \quad \ldots\ldots\ldots\ldots (57)$$

We have

$$\bar{Z} = \frac{\int_{-\infty}^{+\infty} Z\, F(\xi - Z)\, \chi(\xi - Z, Z)\, dZ}{\int_{-\infty}^{+\infty} F(\xi - Z)\, \chi(\xi - Z, Z)\, dZ} \quad \ldots\ldots (58)$$

The divisor is nothing else than $\varphi(\xi)$. The value of Z will fall on the side of greater weight, i.e. of greater F; thus the η will be displaced relative to ξ in the direction of the rise of the frequency-curve. This is a very important, frequently troublesome source of error, consisting in an apparent "attraction" of the observed values by the maxima, or the higher parts of the frequency-curve; in a set of measures otherwise free of systematic errors there may arise thus a systematic error depending upon the statistical distribution of the material; frequently this circumstance may make the observed range in x more or less illusory, a circumstance important not only in statistical but also in other investigations, and often overlooked.

Original

The function

$$\psi(\xi,\bar{z}) = F(\xi-z)\,\chi(\xi-z,z)\Big/\varphi(\xi) \quad \ldots\ldots\ldots (59)$$

represents the frequency-function of errors for the observed value, ξ; the function conforms not with our definition of accidental errors, as its expectation (58) is not zero. The distribution of

$$v = \bar{z} - \bar{\bar{z}} = x - \eta \quad \ldots (59')$$

of the deviation of x from η, is again "accidental", and its frequency function is given by

$$\mu(\eta,v) = \frac{F(\eta+v)\,\chi(\eta+v,\bar{z}+v)}{\int_{-\infty}^{+\infty} F(\eta+v)\,\chi(\eta+v,\bar{z}+v)\,dv} \quad \ldots\ldots\ldots (60)$$

There are cases in observational practice where directly η, and not ξ is determined; this is likely to occur in indirect observations, e.g., in spectroscopic determinations of absolute magnitudes, when the argument itself is not observable, and when its mean true value, η, is found in some empirical way with the aid of other criteria. In this case we obtain a special frequency-function of the mean true argument,

$$\tau(\eta) = \frac{\psi(\eta+\bar{z})}{1 - d\bar{z}/d\xi} = \frac{\varphi(\xi)}{\frac{\partial\eta}{\partial\xi}} \quad \ldots\ldots (61) \qquad (\text{form. (33)})$$

different from F as well as from φ.

Let the dispersion of the true values of x around η be s_v (this is the mean square of v, computed for given η from

$$s_v^2 = \frac{\int v^2 \mu(v,\eta)\,dv}{\int \mu(v,\eta)\,dv} = \frac{\int v^2 \mu(v,\eta)\,dv}{\tau(\eta)}$$

and averaged over the whole range like (50)); taking into account that the expectation of v is zero, we apply the theorem of section 10 and have

$$s_x^2 = s_\eta^2 + s_v^2 \pm \frac{2 s_\eta s_v}{\sqrt{n}} \quad \ldots\ldots (62)$$

Here s_η is the dispersion of $\tau(\eta)$. Comparing this with (48), we have

$$s_\xi^2 = s_x^2 + s_z^2 \quad \ldots\ldots\ldots (63)$$

and

$$s_\eta^2 = s_x^2 - s_v^2$$

Thus the frequency-function of the "mean true argument", $\tau(\eta)$, shows a smaller spread than the true frequency-function, and deviates from it in the opposite direction as compared with $\varphi(\xi)$.

The greater s_v is, i.e. the less accurate are the observations, the smaller is the apparent dispersion s_η ; therefore sets of observations of this kind are always suspicious if the dispersion in the measured quantities turns out to be small. No weight can be attributed to the internal agreement of observations of η, because the agreement may be produced equally by very low and very high accuracy.

12) SPECIAL CASE OF BOTH F AND χ GAUSSIANS. For the sake of simplicity we assume $x_0 = 0$ for the maximum of the frequency-function. Then

$$F(x)dx = \frac{n}{s_x\sqrt{2\pi}} e^{-\frac{x^2}{2s_x^2}} dx \quad \ldots\ldots\ldots (64)$$

and $\chi(x,z)$ independent of x and given by the Gaussian of constant dispersion

$$\chi(z)dz = \frac{1}{s_z\sqrt{2\pi}} e^{-\frac{z^2}{2s_z^2}} dz \quad \ldots\ldots\ldots (65)$$

The observed frequency-function is in this case also a Gaussian with the maximum at $\xi = 0 = x_0$, as may be verified by integration,

$$\varphi(\xi)d\xi = \frac{n}{\sqrt{2\pi(s_x^2+s_z^2)}} e^{-\frac{\xi^2}{2(s_x^2+s_z^2)}} d\xi \quad \ldots (66)$$

The dispersion being

$$s_\xi = \sqrt{s_x^2 + s_z^2}$$

in agreement with(63).

Likewise the inverse is true; if φ and χ are Gaussians, the latter with constant dispersion as before, the true frequency-function is also a Gaussian. The correction for error dispersion consists in this case in simply computing the dispersion s_x of the original frequency-function according to the rules of section 10.

The average "systematic" error is found according to (58)

$$\bar{z} = \frac{s_z^2}{(s_x^2 + s_z^2)}\xi \quad \ldots\ldots\ldots\ldots (67)$$

The mean true argument is

$$\eta = \xi - \bar{z} = \frac{s_x^2}{s_x^2 + s_z^2}\xi \quad \ldots\ldots\ldots (68)$$

and

$$\bar{z} = (s_z^2/s_x^2)\eta \quad \ldots\ldots\ldots\ldots (69)$$

Thus the systematical error produced by the accidental errors consists in a change of the scale in the constant ratio $s_x^2/(s_x^2+s_z^2)$, the mean true scale being shorter than the apparent scale.

The frequency function of the mean true argument is also a Gaussian (equ. (61)):

$$\tau(\eta)d\eta = \frac{n}{s_\eta\sqrt{2\pi}}\, e^{-\frac{\eta^2}{2s_\eta^2}}\, d\eta \quad \ldots\ldots(70)$$

where

$$s_\eta = \frac{s_x^2}{\sqrt{s_x^2 + s_z^2}} = \frac{s_\xi^2 - s_z^2}{s_\xi} = \frac{s_x^2}{s_\xi} \quad (71)$$

In other words, we have $s_\xi s_\eta = s_x^2$, or s_x is the geometrical mean of s_ξ and s_η.

The spread of x for a given η is now easily found from (63) as

$$s_v = \frac{s_x s_z}{\sqrt{s_x^2 + s_z^2}} = \frac{\sqrt{s_\xi^2 - s_z^2}}{s_\xi} \cdot s_z \quad \ldots\ldots(72)$$

or

$$s_v = s_z \Big/ \left(1 + \frac{s_z^2}{s_x^2}\right)^{1/2} \quad \ldots\ldots\ldots (72')$$

Equation (72') tells us that the apparent error in η is smaller than the true error is$_z$, and decreases with increasing s_z/s_x, for $s_x = 0$ we have $s_v = 0$; this in our particular case of Gaussian distribution, in the most inaccurate series of observations we may find a very small empirical spread s_v around the mean value, due simply to the fact that the true spread of x (s_x) is small as compared with the true error of observation (s_z). What is found to hold for a Gaussian, is qualitatively valid also for any other form of F(x) having one simple maximum, and even for more complicated curves--in the case of detached, or very pronounced maxima. (e.g. distribution of absolute magnitudes of late type stars).

13) KAPTEYN'S METHOD of finding the true frequency function refers to the special case when the observations yield η, the "mean true argument", and when in some empirical way it is possible to determine s_v, the spread of the true argument, x, and $\mu(v,\eta)$, the law of this spread, for a given η. The true dispersion of x is then given by

$$s_x^2 = s_\eta^2 + s_v^2 \qquad \text{(section 11)}$$

and the correction for error dispersion is mathematically equivalent to the addition of the error dispersion to the observed frequency function, $\tau(\eta)$, and is attained by simple integration (mechanical quadratures) ($\eta = x - v$):

$$F(x)dx = dx \int_{-\infty}^{+\infty} \tau(x-v)\,\mu(x-v,v)\, dv \quad \ldots\ldots(73)$$

In his researches, Kapteyn confined himself to the assumption of μ depending on v only, but if the material is large enough one may study and take into account the change of μ with η. (73) is from this

standpoint quite general, and the empirical way of attacking the problem is most free of hypothetical elements.

14) EDDINGTON'S FORMULA. For the case of a Gaussian error function with constant dispersion and an arbitrary continuous observed frequency-function $\varphi(\xi)$ Eddington has given a general solution in the form of a series. The solution is highly interesting from the mathematical standpoint, though its practical value is not so great. The numerical method described in section 9, involving less complicated computational work, at the same time being more general with respect to the basic assumptions, and more free of errors of computation, is in most cases to be preferred.

(The derivation of Eddington's formula is based upon the properties of the symbol

$$\exp.\left(a\frac{d^n}{dx^n}\right).u(x) = e^{a\frac{d^n}{dx^n}}.u(x) = u + \frac{a}{1!}\frac{d^n u}{dx^n} + \frac{a^2}{2!}\frac{d^{2n}u}{dx^{2n}} + \ldots$$

$$+ \frac{a^k}{k!}\frac{d^{kn}u}{dx^{kn}} + \ldots \qquad \ldots\ldots (74)$$

This series has certain properties relative to non-symbolical operations of the symbolic function on the left-hand side, regarding the symbolic exponent $a\,d^n/dx^n$ as independent of x, a being a real, d^n/dx^n a symbolic factor. Some of the properties may be directly tested, e.g. differentiating (74) with respect to x or to a. A general proof may be given by considering at first n = 1, when

$$e^{a\frac{d}{dx}}.u(x) = u(x+a),$$

and when the properties are easily tested; the series (74) is such that what is true for n = 1, is true for any n.

One of the properties we shall make use of is the following. If

$$v(x) = \exp\left(a\frac{d^n}{dx^n}\right).u(x)$$

then

$$u(x) = \exp\left(-a\frac{d^n}{dx^n}\right).v(x)$$

$$= v(x) - a\frac{d^n v}{dx^n} + \frac{a^2}{2!}\frac{d^{2n}v}{dx^{2n}} - \ldots\ldots \qquad \ldots (75)$$

The equation of diffusion in Eddington's problem is

$$\varphi(\xi) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} F(\xi - z)\,e^{-\frac{z^2}{2s_z^2}}\,dz$$

Substituting

$$F(\xi - z) = F(\xi) - z\frac{\partial F}{\partial \xi} + \frac{z^2}{2!}\frac{\partial^2 F}{\partial \xi^2} - \ldots\ldots$$

and taking into account that

$$\theta_{2k} = \int_{-\infty}^{+\infty} t^{2k}\, e^{-t^2}\, dt = \frac{1\cdot 3\cdot 5\cdot \ldots (2k-1)\sqrt{\pi}}{2^k} \qquad 15$$

and $$\theta_{2k+1} = 0$$

we find

$$\varphi(\xi) = \exp.\left(\frac{s_z^2}{2}\frac{d^2}{d\xi^2}\right)\cdot F(\xi),$$

whence

$$F(x) = \exp.\left(-\frac{s_z^2}{2}\frac{d^2}{dx^2}\right)\cdot \varphi(x)$$

$$= \varphi(x) - \frac{s_z^2}{2}\frac{d^2\varphi}{dx^2} + \frac{s_z^4}{2^2 2!}\frac{d^4\varphi}{dx^4} - \ldots \qquad (76) \qquad 4$$

This is the general solution of Eddington's problem.

For functions given in tabular form the derivatives are

$$\frac{d^k\varphi}{dx^k} = \frac{\Delta^{(k)}_n}{\Delta x^k} \qquad \ldots\ldots \qquad (77) \qquad \Delta$$

where $\Delta^{(k)}_n$ denotes the k^{th} tabular difference, and Δx as before is the tabular interval. This gives the formula for tabular use

$$\Delta n = \Delta\nu - \left(\frac{s_z}{\sqrt{2}\,\Delta x}\right)^2 \Delta^{(2)}\nu + \left(\frac{s_z}{\sqrt{2}\,\Delta x}\right)^4 \frac{\Delta^{(4)}\nu}{2!} - \ldots\ldots \qquad (78)$$

Notice that $s_z/\sqrt{2} = 1.046$ x(probable error).

The series in (78) is convergent under all circumstances,

(taking into account that $\Delta\nu > 0$, and assuming in the following for $\Delta\nu$ $\Delta\nu$

its greatest value, we have $|\Delta^{(2)}\nu| < \Delta\nu$, $|\Delta^{(4)}\nu| < 2|\Delta^{(4-1)}\nu|$ and generally $|\Delta^{(k)}\nu| < 2^k\Delta\nu$; thus the absolute value of the sum of the series in (78) will be smaller than the sum of the following series

$$\Delta\nu + a^2\cdot 2^2\Delta\nu + \frac{a^4\cdot 2^4}{2!}\Delta\nu + \ldots + \frac{a^{2k}\cdot 2^{2k}}{k!}\Delta\nu + \ldots$$

where $a = s_z/\sqrt{2}\sigma$;

this series is convergent, because the ratio of two adjacent terms, equal to $4a^2/k+1$, tends to zero with $k \to \infty$. Hence (78) is convergent, too.)

Thus (78) gives always a definite result, though not always of practical value; first, because sometimes too many terms are to be taken into account (the case when only the first correction,--

$$-\left(\frac{s_z}{\sqrt{2}\,\Delta x}\right)^2 \Delta^{(2)}\nu'$$

is to be applied, occurs seldom in practice, and then ordinarily it is so small that it might have been neglected as well as the subsequent terms); then, negative values of n are likely to be found; further and this is the most important circumstance, fluctuations having the character of a "false decimal equation" are likely to be introduced into the final result from small unevennesses in the adopted $\varphi(\xi)$, as may be seen in the following example.

$\Delta\nu$	0	1	0	1	0	1	0	1
$\Delta^{(1)}\nu$		+1	−1	+1	−1	+1	−1	+1
$\Delta^{(2)}\nu$			−2	+2	−2	+2	−2	+2
$\Delta^{(3)}\nu$				+4	−4	+4	−4	+4
$\frac{1}{4}\Delta^{(4)}\nu$					−8	+8	−8	+8

These fluctuations have a cumulative effect; as the consecutive differences of even order change sign, e.g. from $\Delta^{(1)}\nu$ to $\Delta^{(3)}\nu$, in (78) they enter in all terms with the same sign. To avoid this spurious "decimal equation", one has to smooth out not only $\Delta\nu'$, but also consecutively all the tabular differences, which introduces too much arbitrariness and extra work; some real unsteadiness of the differences may as well be blotted out in this procedure, an unsteadiness quite common for functions given in tabular form. No such effect can occur in applying the method of section 9. Only by smoothing out the observed $\Delta\nu$ with the aid of some analytical expression and making the computations more accurately than wanted, this disadvantage may be eliminated, and Eddington's formula will give excellent results.

15) SPECIAL DETAILS IN CORRECTING FREQUENCY FUNCTIONS FOR ACCIDENTAL ERRORS. When negative values of n are found, it may mean that the error-dispersion is overestimated. A single negative value, not too deep, in the neighborhood of some very steep maximum may be the result of the "natural uncertainty" or may be due to the non-homogeneity of the error dispersion; it is advisable to "fill up" the negative minimum, assuming the frequency to be equal to zero, and to subtract the same amount from the adjacent maximum, according to the principle shown in Figure 2. The two portions of the curve where the compensation is made must not be too far apart, so that an interaction of error-dispersion of the given amount appears possible. One may also try, by changing the error-dispersion at the particular portion of the curve, to make the negative values disappear; this may be attained by increasing the error-dispersion, s_z, if the distance A-B is less than s_z, and by decreasing s_z, if A-B exceeds s_z.

The error-table, χ_k (see section 9) may be constructed in different ways, according to the degree of approximation required. A satisfactory approximation that works well when the tabular interval (Δx) is of the order of the probable error, or less, is given by

$$\chi_k = \int_{(k-\frac{1}{2})\Delta x}^{(k+\frac{1}{2})\Delta x} \chi(x,z)\,dz \qquad \ldots\ldots\ldots (79)$$

Here it is assumed that all the counted amount Δn is concentrated in the middle of the given interval. A better approximation, approaching the assumption of a uniform distribution of Δn over x, is furnished by

$$\chi_k = \frac{1}{2}\int_{(k-\frac{1}{2})\Delta x}^{(k+\frac{1}{2})\Delta x} \chi\, dz + \frac{1}{4}\int_{(k-1)\Delta x}^{(k+1)\Delta x} \chi dz \quad \ldots\ldots (80)$$

The more general formula, valid under all conditions, is

$$\chi_k = \int_{y=0}^{y=\Delta x} F(x-\tfrac{1}{2}\Delta x+y)dy \int_{z=(k-\frac{1}{2})\Delta x+y}^{z=(k+\frac{1}{2})\Delta x+y} \chi dz \quad \ldots\ldots\ldots (81)$$

(K+1) - K -

Here the change of the frequency-function within one tabular interval is taken into account. In most cases we may substitute φ for F. If great refinement is desirable, one may work by successive approximations; with some preliminary table of χ_k determine F, then according to (81) find the second approximation for χ_k, which now yields the final result for F. There are cases where χ_k may be found directly in some empirical way.

It may be noted, that even when $\chi(x,z)$, the true error function, does not depend upon x, the resulting error-table, according to (81), will generally depend upon x, the more the greater the ratio $(\Delta x/s_z)$; of course, the character of F is important in this case, e.g. when F is a Gaussian, χ_k will be independent of x if Δx is not too large.

16) AN EXACT NUMERICAL METHOD OF SOLVING THE EQUATION OF DIFFUSION, where the result is found directly, without successive approximations, may be theoretically applied in the case of any tabular data, though in practice the method is of use only when the total number of tabular intervals is small. Equations (42) of section 9,

$$\Delta\nu_m = \sum \Delta n_{m-k} \cdot \chi_k$$

written for all tabular intervals, give a system for the determination of the unknowns, Δn; the number of unknowns is smaller than the number of equations, because in all cases where $\Delta\nu$ is zero, we have to put Δn equal to zero (on account of the more concentrated character of F relative to φ); at the same time the cases $\Delta\nu=0$ represent additional equations. One may apply least squares solutions, though generally it is too cumbersome. A reduction of the number of equations by making it equal to the number of unknowns (leaving only the more important ones and rejecting those of the extremities of the curve with $\Delta\nu=0$ or smaller) works practically as well and is more convenient.

A combination of the method of successive approximations, which gives the unimportant values of Δn at the extremities and reduces thus the number of unknowns, with the method here described for the determination of the Δn in the main portion of the curve, may prove valuable in certain cases.

Here it is assumed that all the counted amount Δn is concentrated in the middle of the given interval. A better approximation, assuming the assumption of a uniform distribution of Δn over x, is furnished by

$$\chi_k = \frac{1}{2}\int_{(k-\frac{1}{2})\Delta x}^{(k+\frac{1}{2})\Delta x} \chi\, dz + \frac{1}{4}\int_{(k-1)\Delta x}^{(k+1)\Delta x} \chi dz \quad \ldots\ldots (80)$$

The more general formula, valid under all conditions, is

$$\chi_k = \int_{y=0}^{y=\Delta x} F(x-\tfrac{1}{2}\Delta x+y)\,dy \int_{z=(k-1)\Delta x+y}^{z=(k+1)\Delta x-y} \chi dz \quad \ldots\ldots\ldots (81)$$

Here the change of the frequency-function within one tabular interval is taken into account. In most cases we may substitute φ for F. If great refinement is desirable, one may work by successive approximations; with some preliminary table of χ_k determine F, then according to (81) find the second approximation for χ_k, which now yields the final result for F. There are cases where χ_k may be found directly in some empirical way.

It may be noted, that even when $\chi(x,z)$, the true error function, does not depend upon x, the resulting error-table, according to (81), will generally depend upon x, the more the greater the ratio $(\Delta x/s_z)$; of course, the character of F is important in this case, e.g. when F is a Gaussian, χ_k will be independent of x if Δx is not too large.

16) AN EXACT NUMERICAL METHOD OF SOLVING THE EQUATION OF DIFFUSION, where the result is found directly, without successive approximations, may be theoretically applied in the case of any tabular data, though in practice the method is of use only when the total number of tabular intervals in small. Equations (42) of section 9,

$$\Delta\nu_m = \sum \Delta n_{m-k} \cdot \chi_k$$

written for all tabular intervals, give a system for the determination of the unknowns, Δn; the number of unknowns is smaller than the number of equations, because in all cases where $\Delta\nu$ is zero, we have to put Δn equal to zero (on account of the more concentrated character of F relative to φ); at the same time the cases $\Delta\nu=0$ represent additional equations. One may apply least squares solutions, though generally it is too cumbersome. A reduction of the number of equations by making it equal to the number of unknowns (leaving only the more important ones and rejecting those of the extremities of the curve with $\Delta\nu=0$ or smaller) works practically as well and is more convenient.

A combination of the method of successive approximations, which gives the unimportant values of Δn at the extremities and reduces thus the number of unknowns, with the method here described for the determination of the Δn in the main portion of the curve, may prove valuable in certain cases.

17) COMPUTATION OF $\bar{z}, h$ AND THE DISTRIBUTION OF x and z FOR A GIVEN ξ. After we have found by any of the preceding methods the true frequency function, we make a control computation according to (42) (where unlike (43) the term k=0 also enters); the single terms $\Delta n_{m-k} \cdot \chi_k$ represent the frequency of occurrence of $z = k\Delta x$, or of $x = \xi - k\Delta x$, within the given m^{th} observed interval; dividing the frequencies by $\Delta \nu_m$, we obtain the relative frequencies; $\bar{z}$ is found as the mean z weighted by the frequency; thus this problem is easily solved in a purely arithmetical way.

It may be remarked that for F of very small curvature, and χ a Gaussian, the systematic error in ξ is given by

$$\bar{z} = - s_z^2 \frac{\partial \log_e F(x)}{\partial x} \quad \ldots\ldots\ldots (82)$$

Thus the systematic error in this case increases as the square of the error dispersion.

18) DETERMINATION OF $\varphi(\xi)$ FROM GIVEN $\tau(\eta)$ AND $\mu(v)$. (Compare section 11). Except in special cases like the case treated in section 12, the problem is solved by successive approximations. Taking into account that F(x) is easily found (section 13), and assuming in (58) μ for χ as a first approximation, and writing x for ξ except in μ where x stands for $\xi - z$, we have the first approximation of $\bar{z}$, ξ and $\varphi(\xi)$

$$\bar{z}_1(x) = \frac{\int_{-\infty}^{+\infty} z\, F(x-z)\, \mu(x,z)\, dz}{\int_{-\infty}^{+\infty} F(x-z)\, \mu(x,z)\, dz} \quad \ldots\ldots (83) \qquad (x)$$

$$\xi_1 = \eta + \bar{z}_1(\eta) \quad \ldots\ldots (84) \qquad (\eta)$$

$$\varphi_1(\xi) = \tau(\eta) \Big/ \left(\frac{\partial \xi_1}{\partial \eta}\right) = \tau(\xi_1 - \bar{z}_1) \Big/ \left(1 + \frac{\partial \bar{z}_1}{\partial \eta}\right) \quad \ldots (85)$$

From (59) we have the second approximation for χ (taking into account that ψ and μ are displaced by $\bar{z}$ and otherwise identical)

$$\chi_2(x,z) = \frac{\mu(x, \bar{z}-z)\, \varphi(x)}{F(x-z)} \cdot \text{CONST.} \quad \ldots (86)$$

The constant is chosen so as to satisfy condition (35). (This formula we have by substituting in (60) x for $\eta + \bar{z}$ except in μ where x stands for $\xi - z$; this latter circumstance makes it necessary to introduce the constant factor in (86))

Substituting χ_2 instead of μ into (83), we find the second approximation of $\bar{z}$ etc.

For data given in tabular form all these operations are performed numerically in a much simpler way.

19) DECOMPOSITION INTO GAUSSIAN COMPONENTS. Another method of performing the correction for accidental errors consists in representing $\varphi(\xi)$ by a sum of Gaussians and correcting each of them separately according to section 12. The method is applicable when none of the component Gaussian curves has a dispersion less than s_z. The decomposition into Gaussian components, however, involves a great deal of work (even using the simplified method of G. Doetsch, Zeits. f. Physik, 49, 705) and cannot be regarded as convenient except in particular cases when the components are easily recognizable in the original curve, i.e. when the decomposition must have some real meaning, being thus more than a procedure of mere mathematical convention.

20) DETERMINATION OF THE ERROR DISPERSION AND THE ERROR FUNCTION. The error dispersion is easily found from a set of observations of the same object; if the set is great enough, an approximation to the true shape of the error function is obtained by tabulating the apparent deviations Δ_i (section 10), and by assuming that the distribution of the true deviations $\Delta_i \sqrt{n/n-1}$ follows the same law.

Ordinarily it is found, that an error function, even when symmetrical, deviates from a Gaussian with equal dispersion, and shows a positive excess: very small, and very great errors are more frequent, than in the Gaussian. Ogorodnikoff suggested that this may be explained by a changing accuracy of the observations when making them (unsteady conditions.)

Systematic errors depending upon the properties of the individual object cannot be determined in this way; such systematic errors, having character of accidental errors when different objects are considered, may render the determination of the error function according to the above illusory mentioned method.

To allow for this circumstance, or in the case when the object can be observed only once (meteors), the data for the error function may be found from parallel, independent series of observations, e.g. from simultaneous observations made by several observers.

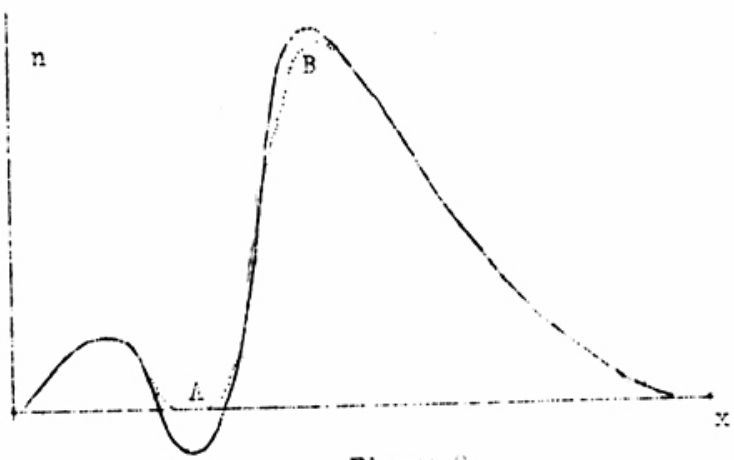

Figure 3.

Full curve, computed F(x)

Dotted curve, corrected for negative values.

(The figure refers to page 24.)

Let us consider two series of observations referring to common objects; let the individual error in those series be z and v, the error function $\chi(z)$ and H(v), the dispersions s_z and s_v; the relative error, u, is given by

$$u = v - z \quad \ldots\ldots\ldots\ldots (87)$$

the relative dispersion is (section 10)

$$s_u^2 = s_z^2 + s_v^2 \quad \ldots\ldots\ldots (88)$$

and the relative error function is

$$\omega(u) = \int_{-\infty}^{+\infty} H(u+z)\chi(z)dz \quad \ldots\ldots\ldots (89)$$

As before, the limits of integration are written here in the conventional manner, and in certain cases may be assumed to have finite values. We obtain (87) from (37), assuming F = H and $\xi = u$, from (37) v = u+z. When $\chi(z)$ is known and $\omega(u)$ is given by observation, we can find H(v) using the methods of solving the equation of diffusion. The problem is more complicated in the case when both error functions are unknown. The problem is generally indefinite and may be solved only upon making some assumptions as to the relative character of H and χ.

Let us first consider the case of two similar series, where *a priori* we may expect identical error functions. We assume $H = \chi$, $s_v = s_z$; if in reality the assumption is not fulfilled, our solution will refer to the *average* errorfunction of the two series. We have then

$$s_z^2 = s_u^2/2 \quad \ldots\ldots\ldots\ldots (90)$$

and

$$\omega(u) = \int_{-\infty}^{+\infty} \chi(u+z)\cdot\chi(z)dz \quad \ldots\ldots\ldots (91)$$

(91) may be solved by successive approximations in the following manner. As a first approximation for $\chi(z)$ we assume a Gaussian,

$$\chi_1(z) = \frac{1}{s_z\sqrt{2\pi}} \; e^{-\frac{z^2}{2s_z^2}}$$

where s_z is given by (90). With this we solve the following equation of diffusion.

$$\omega(u) = \int_{-\infty}^{+\infty} \chi_2(u+z)\cdot\chi_1(z)dz$$

whence χ_2 is found. The next approximation to the error function is given by

$$\chi_3(z) = \frac{\chi_1(z) + \chi_2(z)}{2} \quad \ldots\ldots (92)$$

With this we find χ_4 from

$$\omega(u) = \int_{-\infty}^{+\infty} \chi_4(u+z)\cdot\chi_3(z)dz$$

and have a still closer approximation

$$\chi_5(z) = \frac{\chi_3(z) + \chi_4(z)}{2} \quad \ldots\ldots (92') \text{ etc.}$$

When the true error function resembles roughly a Gaussian, χ_3 will practically give the final approximation. The check is that the difference $\chi_2 - \chi_1$ is not considerable. Otherwise we have to proceed with the approximations until $\chi_{n+1} - \chi_n$ is small enough.

It may be remarked that with tabular data the most convenient method of solving the equation of diffusion in this case is the method described in section 16, because the number of tabular intervals in an error table is ordinarily small.

For two equal error functions; $\omega(u)$ is always symmetrical, though χ may be unsymmetrical; this is easily proved mathematically, though it is obvious without any special proofs, because the two sets of observations are identical, and interchanging them the distribution of relative deviations cannot change, i.e. $\omega(u) = \omega(-u)$. If $\omega(u)$ is found to be unsymmetrical, this means that H and χ cannot be identical. For a symmetrical $\omega(u)$, when the first approximation, χ_1, is assumed symmetrical, χ_2 will also be found symmetrical, and the final result for χ will also be symmetrical; thus, such a result never can be regarded as a proof that χ really is symmetrical; all depends upon the character of the first approximation, χ_1. If we know that χ resembles a certain non-Gaussian function, we must assume for χ_1 this function taking the dispersion according to (90), and proceed as described above. It may be noted that in numerical computations a different method of approximations may be applied.

The procedure for tabular data may be illustrated by the following example, which refers to a very small number of tabular intervals, but which may be easily extended to any number. Let $\Delta x = 1$, and let χ show such a low spread that only for $k = -1$, 0 and +1 χ assumes important values, being practically zero outside of those limits. The equations will be as follows:

$$\left.\begin{aligned}
\chi_{+1}\,\chi_{-1} &= \omega_{+2} && \ldots (a)\\
\chi_0\,\chi_{-1} + \chi_{+1}\,\chi_0 &= \omega_{+1} && \ldots (b)\\
\chi_{-1}^2 + \chi_0^2 + \chi_{+1}^2 &= \omega_0 && \ldots (c)\\
\chi_{-1}\,\chi_0 + \chi_0\,\chi_{+1} &= \omega_{-1} && \ldots (d)\\
\chi_{-1}\,\chi_{+1} &= \omega_{-2} && \ldots (e)
\end{aligned}\right\} \ldots (93)$$

Equations (a) and (e), (d) and (b) are identical, whence $\omega_{-k} = \omega_{+k}$, in conformance with the required symmetry of ω. Thus we have only 3 independent equations with three unknowns: χ_{-1}, χ_0, χ_{+1}, and theoretically we have one and only one solution.
This apparently contradicts our former conclusion as to the general indefiniteness of our problem; however, this is only an apparent contradiction. It is true that for some peculiar tabular data the problem

may have a definite solution. But in the general case of an error table, χ_1 and χ_{-1} are expected to be small in comparison with χ_0--otherwise we have no right to assume that χ_{+2} vanishes, and at the left hand side of equation (8) will be found a small quantity of second order relative to χ_{+1}, the equation approaching practically $0 = 0$, and being of no use in the solution.
Thus in the general case we have practically 2 equations with 3 unknowns. The successive approximations may be avoided now, making simply an assumption about the ratio χ_{+1}/χ_{-1}, which for symmetrical functions is 1. This at once reduces the number of unknowns and makes a solution possible.

When the dispersions are unequal, and when of the two unknown error functions we know which has the smaller dispersion, we assume for it a Gaussian with some guessed value of the dispersion and solve (39) with respect to the other function having the larger dispersion.

When we have three or more independent sets of observations relating to common objects, the error functions may be determined practically without any ambiguity; the best determination is made when the sets are approximately of equal accuracy; in unequal sets the error function of the most accurate set is the least accurately determinable. The procedure will be described here for three sets. For a greater number it is easy to extend the method by analogy.

Let ξ_1, ξ_2 and ξ_3 represent three independent measures of the same quantity x; z_1, z_2 and z_3 are the corresponding errors, s_1, s_2, s_3--the error dispersions. Let us form the following differences

$$\left.\begin{aligned} u_1 &= \xi_1 - \frac{\xi_2 + \xi_3}{2} \\ u_2 &= \xi_2 - \frac{\xi_1 + \xi_3}{2} \\ u_3 &= \xi_3 - \frac{\xi_1 + \xi_2}{2} \end{aligned}\right\} \quad \ldots(94)$$

Let the dispersions of u_1, u_2 and u_3 be σ_1, σ_2, and σ_3. According to the general theorem of section 10, we have

$$\left.\begin{aligned} \sigma_1^2 &= s_1^2 + \tfrac{1}{4}s_2^2 + \tfrac{1}{4}s_3^2 \\ \sigma_2^2 &= \tfrac{1}{4}s_1^2 + s_2^2 + \tfrac{1}{4}s_3^2 \\ \sigma_3^2 &= \tfrac{1}{4}s_1^2 + \tfrac{1}{4}s_2^2 + s_3^2 \end{aligned}\right\} \quad \ldots(95)$$

$\tfrac{1}{4}s_2^2$ etc

whence

$$\left.\begin{aligned} s_1^2 &= \frac{10}{9}\sigma_1^2 - \frac{2}{9}\sigma_2^2 - \frac{2}{9}\sigma_3^2 \\ s_2^2 &= \frac{10}{9}\sigma_2^2 - \frac{2}{9}\sigma_1^2 - \frac{2}{9}\sigma_3^2 \\ s_3^2 &= \frac{10}{9}\sigma_3^2 - \frac{2}{9}\sigma_1^2 - \frac{2}{9}\sigma_2^2 \end{aligned}\right\} \quad \ldots(96)$$

Knowing the dispersions, the shape of the individual error functions may be found by successive approximations, the uncertainty depending upon the character of the first approximation being much less than in the case of only two sets.

We may consider $\xi' = (\xi_2+\xi_3)/2$ as a single set of observations, giving with respect to ξ_1 a relative deviation u_1, the law of distribution of which, $\omega(u_1)$, is known from observations; The error dispersion of ξ' is

$$s_{2,3} = \frac{\sqrt{s_2^2+s_3^2}}{2} \quad \ldots\ldots (97)$$

In the case when it is less than s_1, or when

$$s_2^2+s_3^2 < 2s_1^2 \quad \ldots\ldots\ldots\ldots (98)$$

(for equal sets we have $s_2^2+s_3^2 = 2s_1^2$) we may assume for the distribution of ξ' a Gaussian and solve the equation (89) assuming

$$\chi_\varphi(z) = \frac{1}{s_{2,3}\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2s_{2,3}^2}}$$

1

with the purpose of finding H_1, the error function of ξ_1. In the same manner we find the error functions of the two other sets. We may enter a second approximation with

H

$$L(y) = \int_{-\infty}^{+\infty} H_2(y-v)\cdot H_3(v)dv \quad \ldots\ldots\ldots (99)$$

2

assuming: $\chi(z/2) = L(z)$ or $\chi(z) = L(2z)$

χ

where H_2 and H_3 are the error functions of ξ_2 and ξ_3 found in the first approximation.

(The integral gives the error function $L(z)$ of the sum $\xi_2+\xi_3$; the error function of the average, $(\xi_2+\xi_3)/2$, will be the same except that the argument must be divided by 2. The integral does not change when H_2 and H_3 change places, which is easily demonstrated by substituting $y-v = u$)

If condition (98) is not fulfilled, i.e. when ξ_1 is more accurate than the average of all the other sets, the solution becomes indefinite with respect to H_1. (practically)

practically

21) SELECTION. In many, perhaps in most cases not all required objects, but only a certain fraction q of them can be counted; q is called the coefficient, or factor of selection. Generally q may be regarded as a function of the argument. The true frequency function, $F_0(x)$, and the selective frequency function, $F(x)$, are related by the equation

$$F(x) = q\,F_0(x)$$

or

$$F_0(x) = F(x)/q \quad \ldots\ldots\ldots (100)$$

This equation represents the general form of the correction for selection.

When $q = q(x)$ depends directly upon the true argument, and is not influenced by observational errors, the correction for selection must be applied to the true frequency function, $F_0(x)$; it may be applied to $\tau(\eta)$ also, assuming for q a mean value

$$\bar{q}(\eta) = \frac{\int q\,\mu(v)dv}{\int \mu(v)dv} \qquad \ldots\ldots\ldots\ldots (101)$$

notation as in section 11; and $v = x - \eta$.

The same average value, $\bar{q}(\eta) = \bar{q}(\xi - \bar{z}) = \bar{q}'(\xi)$, may be used in correcting $\varphi(\xi)$, though generally such corrections of τ or φ are to be made only when for some reason it is difficult to derive $q(x)$, whereas observations yield $\bar{q}'(\xi)$ or $\bar{q}(\eta)$.

When the apparent argument, ξ, influences the selection without regard to what the value of the true argument is, the observed frequency function must be corrected for selection

$$\varphi_0(\xi) = \frac{\varphi(\xi)}{q(\xi)} \qquad \ldots\ldots\ldots\ldots (102)$$

22) DETERMINATION OF THE COEFFICIENTS OF SELECTION. There may exist two different kinds of selection; cosmical selection, and observational, or subjective selection. An example of the first kind is given by the apparent distribution of stellar luminosities in the sky, which is a selection depending upon the true luminosity, the coefficient of selection being proportional to the observable volume of space. Cosmical selection may be determined theoretically, on the basis of general considerations, according to the circumstances of the problem.

An example of observational selection is given by star counts, where some stars may be omitted, the percentage of omissions increasing as the limiting magnitude is approached. The coefficient of selection may be called in this case the coefficient of perception.

From two, or more independent sets of observations, relating to the same (data simultaneous) in time and space, the coefficients of perception may be determined. Such sets we will call double-count observations. Let us first consider the case of two independent sets, and let us limit ourselves by a certain range of the argument, homogeneous with respect to selection, i.e. within which q may be assumed to be constant. Let the coefficients of selection be q_1 and q_2; the numbers counted in the two sets be n_1 and n_2, the number of objects found in common between the two sets be $m_{1,2}$, and the probable value of the true number be N. We have

$$\begin{aligned} n_1 &= q_1 N \\ n_2 &= q_2 N \end{aligned} \qquad \ldots\ldots\ldots\ldots (103)$$

on account of the independence of the two sets we have also

$$m_{1,2} = q_1 n_2 = q_2 n_1 \qquad \ldots\ldots\ldots (104)$$

or

$$q_1 = \frac{m_{1,2}}{n_2} \pm \sqrt{\frac{q_1(1-q_1)}{n_2}}$$
$$q_2 = \frac{m_{1,2}}{n_1} \pm \sqrt{\frac{q_2(1-q_2)}{n_1}} \quad \ldots\ldots\ldots (104')$$

Hence we find

$$N = \frac{n_1}{q_1} = \frac{n_2}{q_2} = \frac{n_1 n_2}{m_{1,2}} \quad \ldots\ldots\ldots (105)$$

If S denotes the number of <u>different</u> individuals in both sets together, we have

$$S = n_1 + n_2 - m_{1,2} \quad \ldots\ldots (106)$$

From other considerations we find

$$N = S/[1-(1-q_1)(1-q_2)] \pm N/\sqrt{S} \quad \ldots (107)$$

which after substituting (106) and (104') turns out to be identical with (105).

When there are more than two independent sets of observations, our formulae are transformed into the following:

$$q_1 = \frac{\sum m_{1,h}}{\sum n_h} \pm \frac{q_1(1-q_1)}{n_h} \qquad (h = 2,3, \ldots)(108)$$

$$N = S/P \pm N/\sqrt{S} \quad \ldots (109)$$

where $$P = 1-(1-q_1)(1-q_2)(1-q_3) \quad \ldots \quad (110)$$

As everywhere before, the mean errors, or dispersions are given. The probable errors are found by multiplying by 0.674.

Ordinarily we obtain from observations directly the coefficient of perception, $q'(\xi)$, as a function of ξ, the measured argument. If q is supposed to be a direct function of x, we have

$$q'(\xi) = \frac{\int_{-\Phi}^{+\Phi} q(\xi-z)\cdot F(\xi-z)\cdot \chi(\xi-z,z)dz}{\varphi(\xi)} \quad \ldots (111)$$

Knowing $q'(\xi)$ from observations, and F, χ and φ given, we may find the function q by successive approximations.

The coefficient of perception may depend upon several arguments, $\xi_1, \xi_2, \xi_3 \ldots\ldots$; the theory of this section applies to the case when the given group of counted individuals is homogeneous with respect to all the arguments. The greater the number of arguments, the more detailed must be the subdivision into groups, and consequently the smaller will be the number within one group. This puts a limit to the fineness of detail; when the total number of individuals is small, we must neglect the influence of some arguments upon q and take into account only the most important ones.

Hypotheses may be made in advance about the probable character of the dependence of q upon the different arguments. One of the most useful hypotheses, without which we cannot do when the total number of observations is not very great, is to assume that the influence of the different arguments is independent of one another, so that

$$q = q'_1(\xi_1)\cdot q'_2(\xi_2)\cdot q'_3(\xi_3)\ldots \qquad (112)$$

according to the theorem of the product of probabilities. In this way all the equations of the form (108), written for the separate homogeneous groups, will form one single system and may be solved together.

23) CORRELATION. Two measurable quantities, x and y, relating to the same object, may depend upon one another:

$$y = f(x) \qquad \ldots\ldots\ldots (113)$$

This is a functional relationship when (113) is fulfilled exactly, and a statistical relationship, or statistical correlation when (113) involves a certain accidental error, μ,

$$y = f(x) + \mu \qquad \ldots\ldots\ldots (114)$$

For a given value of x, the average value, or the expectation of y satisfies equation (113);

$$\overline{y} = f(x) \qquad \ldots\ldots\ldots\ldots (114')$$

(μ cancels out of the mean).

The theory of correlation deals with the properties of expressions like (114'); this equation is called the regression curve of y upon x. It must be emphasized that the inverse regression curve, i.e. of x upon y, cannot be found by solving (114') with respect to x (only in the case of functional relationship may this procedure be applied); indeed, (114') does not contain $\overline{x}$ nor y, thus it can never give us an equation of the form

$$\overline{x} = f_1(y) \qquad \ldots\ldots\ldots\ldots (114'')$$

except when $\overline{x} = x$, $\overline{y} = y$, i.e. when there is no spread, $\mu = 0$, in other words, when the relationship is a functional one. Measured quantities, any kind of observational data may be only in statistical correlation, on account of the errors of measurement; attributing all the disagreement to these errors, we assume that at the base there exists a true functional dependence between the quantities, though we are never able to prove this. It is by such a process of abstraction that the laws of nature are derived, the correlation in this case we assume to be apparently statistical, but intrinsically functional. When μ is partly, or wholly cosmical, the correlation is intrinsically statistical. From the philosophical standpoint, we may affirm that only statistical correlations really exist, the so-called exact laws of nature only appearing so because of the imperfection of our measurements. In practice, however, we will call a correlation functional when the errors are negligible.

In the case of an intrinsically statistical correlation (e.g. between the distance and the proper motion of a star) we nevertheless may introduce a certain fictitious functional correlation,

$$y = f_0(x) \qquad \ldots\ldots\ldots (115)$$

from which the regression-curves (114') and (114") originate through the mediation of the accidental errors; (115) we will call the ideal correlation; it is to be expected that the laws of nature are revealed by the ideal correlation, whereas calculations of the most probable value of $\bar{y}$ must be made according to (114').

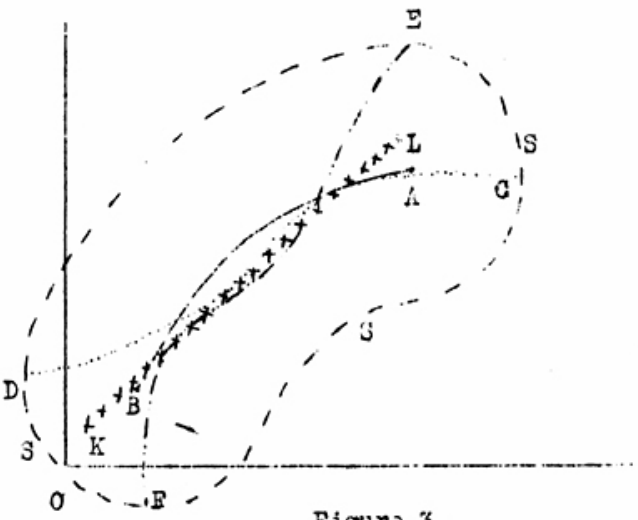

Figure 3.

AB = ideal correlation $y = f_0(x)$
3S = contour of observed spread, assuming equal errors in x and y.
CD = regression curve $\bar{y} = f(x)$
EF = regression curve $\bar{x} = f_1(y)$
KL = straight line median between the two regression curves

Figure 3 gives a schematical representation of the different curves of correlation; SS is the effective contour of the error spread, assumed equal in both axes, so that SS is the geometrical envelope of circles, having their centers on AB, of radius equal to the effective mean error. Assuming schematically that the area within SS is uniformly covered by the observed points, (x,y), we find that the regression curve of y upon x, CD, is the geometrical locus bisecting the chords parallel to the y-axis, and vice versa. Though in reality a uniform density of the observed points over the contour SS is not very likely to occur, qualitatively the phenomenon will be the same as in the schematical case. We see that both regression curves may have very little resemblance with the ideal (true) correlation and that some average (intermediate) curve between CD and EF will not give much better agreement. With decreasing error-dispersion, or increasing length of the ideal curve, the regression-curves will come closer to the ideal curve.

It may be noted that $\bar{y} = f(x)$ covers a range in x greater than the true one, whereas the range in $\bar{y}$ equals the true range in y; hence a flattening of the ends of this regression curve arises, the ends tending to assume a small angle with the x-axis; similarly the ends of the $\bar{x} = f_1(y)$ stretch out at small angles with the y-axis; these diverging "horns" formed by the two regression curves are one of the most characteristic features of statistical correlation; the amount of the divergence of the ends is at the same time a measure of the distortion of the original curve produced by the accidental errors.

When x is regarded as an accurate quantity, i.e. when its error-dispersion is negligible, the regression curve of y upon x coincides with the ideal correlation,

$$\bar{y} = f(x) = f_0(x),$$

as illustrated in Figure 4. Generally, the regression curve having as

independent variable the variable of smaller accidental error, will approach more closely the ideal correlation. When the mean error in x is smaller than in y, $y = f(x)$ represents a better approximation to reality than $x = f_1(y)$.

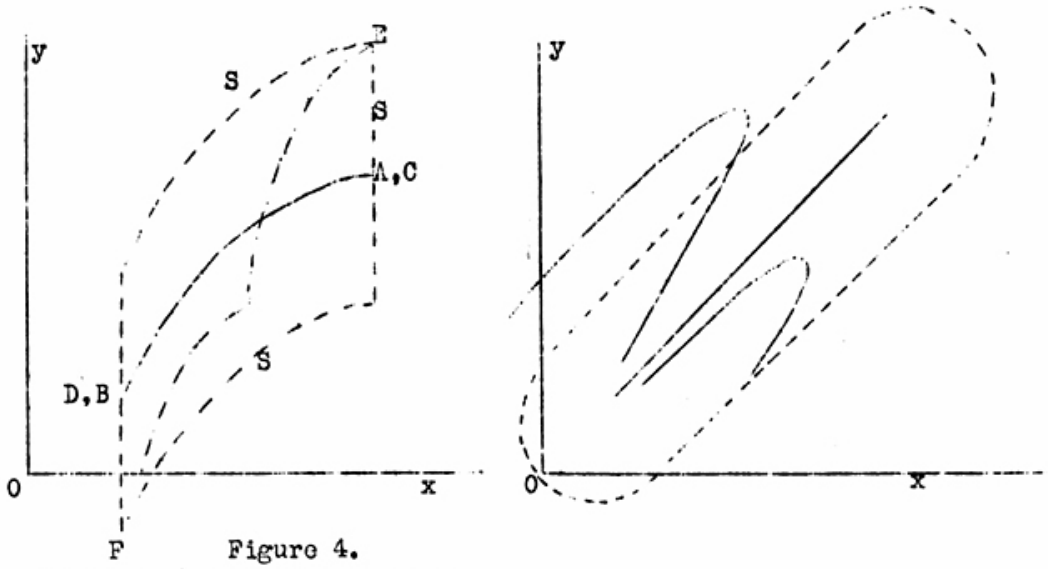

Figure 4.
Case of zero dispersion in x.
Notation as in Figure 3.

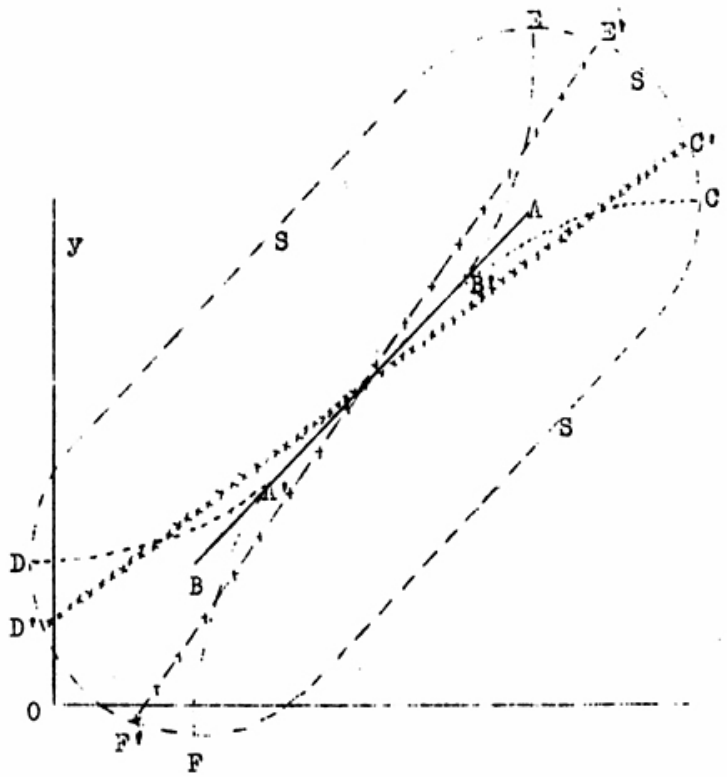

Figure 5

Case when ideal correlation is linear (AB); equal error dispersion in x and y. CD, EF--true regression curves C'D', E'F'--Pearson's regression lines.

$$\bar{y} = f(x) = \frac{\int_{-\infty}^{+\infty} y\omega(x,y)dy}{\int_{-\infty}^{+\infty} \omega(x,y)dy} \quad \ldots\ldots\ldots (117)$$

Similarly

$$\bar{x} = f_1(y) = \frac{\int_{-\infty}^{+\infty} x\omega(x,y)dx}{\int_{-\infty}^{+\infty} \omega(x,y)dx} \quad \ldots\ldots\ldots (118)$$

Instead of infinite limits one may assume, according to circumstances, finite limits of integration. The substitution of x,y for ξ, η is purely a matter of convention. As ω is directly given by the observations, the problem of finding the regression curves from a given observational material is always definite; in practice the integration is made arithmetically as explained in a following section.

The problem of finding the true correlation, $y = f_0(x)$, is more complicated, and even to some extent indefinite.

Let F(x) be the frequency function of x; because of the functional relationship between y and x, F(x) determines the original distribution without ambiguity. Let $\chi(z)$ and H(v) be the error functions, supposed here to depend upon the errors themselves, though it is easy to generalize our deductions for the case when the error functions depend upon the coordinates and the error in the other coordinate also(generalized error function $\psi(x,y,z,v)dzdv$). The number of points between x and x+dx, z and z+dz v and v+dv is evidently

$$F(x)\, H(v)\, \chi(z)\, dxdvdz$$

For given ξ and η we have $x = \xi - z$, $v = \eta - y = \eta - f_0(x) = \eta - f_0(\xi - z)$; for given v ($dv = 0$), $d\eta = dy$ and likely, for given z ($dz = 0$) $dx = d\xi$. Thus we reduce our expression to a single arbitrary variable, z, and integrating over the extreme limits of z we obtain the observed density of population (dividing by $d\xi\, d\eta$)

$$\omega(\xi,\eta) = \int_{-\infty}^{+\infty} F(\xi - z)H\left[\eta - f_0(\xi - z)\right]\chi(z)dz \quad \ldots\ldots (119)$$

From this integral equation f_0 may be determined by successive approximations; the procedure is more complicated than in the one-dimensional equation of diffusion (37). The direct observational data consist in ω and F, the latter must be determined from (37). We must know also both error functions. Taking as a first approximation $f_0(x) = f(x)$ from (117), we compute an auxiliary density function, $\omega'(\xi,\eta)$ from (119) and with this auxiliary function we compute a regression curve $f_{(1)}(x)$ from (117); adding the difference $f(x) - f_{(1)}(x)$ to f(x), we obtain the second approximation to f_0. We may proceed more quickly by taking as a first approximation of f_0 not one of the regression curves, but a certain median between them (Figure 3, KL); the median curve may be found by attributing to each of the regression curves certain relative weights, by analogy with linear correlation as explained in one of the next sections.

In Figure 5 we have the case when the ideal correlation, AB, is a straight line; the resulting regression curves are in this case never straight lines, except in the portion A'B' where they coincide with the ideal curve. Pearson assumes his regression curves to be straight lines, C'D' and E'F' in Figure 5; we see that they nowhere coincide with the true regression curves, and that they really cannot be regarded as regression curves at all. Pearson's regression lines are mathematical fictions, useful when applied with full realization of their meaning, but frequently misleading when one forgets what they are. In using Pearson's straight regression lines we loose the advantage of finding directly a portion (A'B') of the true correlation; the extent of this portion is the greater the longer is AB as compared with the error dispersion.

Generally, when the error dispersion is not negligible, the true regression curves cannot both be straight lines; we may find a certain case where one of the regression lines is approximately straight, produced by a peculiarly curved ideal correlation. However, the other regression curve will deviate the more from a straight line. In practice, when the number of observations is small, or the error dispersion considerable, we may be unable to find the true shape of the regression curves. According to the principle of minimum hypothesis, to avoid arbitrariness of judgment, we assume the regression lines to be straight; this is Pearson's case. All the curves in Figures 3, 4 and 5 may be substituted by straight lines, and the general trend of variations revealed by the curves will still be reflected by the straight lines; this happens because the curvature of the curves is not great. However, for a strongly curved correlation Pearson's regression lines may be misleading and must not be applied; Pearson's linear correlation may be applied with success when we know, from experience or general considerations, that dy/dx does not change sign, or even does not vary considerably over the whole range covered by the observations,--i.e. when the correlation possesses a more or less steady slope.

24) GENERAL THEORY OF CORRELATION; CASE WHEN THERE EXISTS A TRUE CORRELATION, $f_0(x)$. Our preceding qualitative analysis was bound to the scheme of a contour SS covered uniformly by the observed points, or to a contour of constant density of population; in reality the density of population may not be uniform--it depends upon the true initial distribution of the objects according to x or y, or both, and upon the error spread in both coordinates which changes the original distribution; also there may not exist a limited contour containing all points, but the density of the points may gradually fade away with increasing distance from the original curve. All these circumstances may influence quantitatively the results, though qualitatively they stand as exposed in the preceding section.

Let x and y be the true coordinates, ξ and η --the measured ones (η has here and in the following another meaning than before), z and v--the errors, so that

$$\xi = x+z, \qquad \eta = y+v.$$

Let further $dn = \omega(\xi,\eta)d\xi d\eta$ be the number of observed points comprised between ξ and $\xi+d\xi$, η and $\eta+d\eta$, i.e. within the element of area $d\xi d\eta$; ω is evidently the observed density of population at the point (ξ,η); it is a two-dimensional frequency function which may be represented in three-dimensional space as a surface,

$$\zeta = \omega(\xi,\eta) \quad \ldots\ldots\ldots\ldots\ldots (116).$$

The regression curve of y upon x is defined as a function representing the average value of η for constant ξ. Thus we have

Original

The indefiniteness of the problem consists in the following. After attaining the final approximation of f_0 and substituting it into (119), and performing the integration, we may find that the resulting density-function does not coincide with the observed one, $\omega(\xi,\eta)$. This may be attributed partly to the accidental character of the errors which in a real case need not follow the exact theoretical laws of distribution, especially when the total number is not large; and partly to the uncertainty in the adopted functions H and χ; if this is the case, one may try to change H and χ in the sense indicated by the difference between ω observed and computed, and then enter for a new solution of f_0, etc. Such a refined procedure is worth while only when the total number of points is great and when the extent of the curve exceeds considerably the average error dispersion. As a criterion for the latter circumstance the relative divergence of the two regression curves may be used; if in their main portion they are similar, or coincide, we may hope that the detailed treatment will yield real results. Otherwise it is safe to content oneself with a rough similarity of ω and ω', and not to try any further approximations.

25) CASE WHEN NO TRUE FUNCTIONAL RELATIONSHIP EXISTS. When x and y involve cosmical errors of an accidental character, we may proceed by two different methods. Including the combined effect of observational and cosmical errors into our error functions, we may try to find the fictitious ideal correlation, $y = f_0(x)$, according to the method of the preceding section. Or, knowing the laws of observational errors, we may seek for the density function of the true arguments, determined by the spread of the cosmical errors. This latter procedure is ordinarily to be preferred, because we may at once get important information of the character of the cosmical errors which may reflect certain laws of nature as well as the ideal correlation itself. Thus, even having for the final aim the determination of the ideal correlation, it is advisable to perform it in two steps, by freeing first from purely observational errors, and then looking for $f_0(x)$ as described in section 24; especially, when the laws of observational and cosmical errors are so different that it is inconvenient to join them together in a single error function.

Also the ideal cosmical correlation is frequently known in advance from general considerations, in which case equation (119) may be solved with respect to one of the cosmical error functions when the other is known.

Let $\omega_0(x,y)$ denote the true, or cosmical density function, and let the other notations be as before. The observed density function is then given by the two-dimensional equation of diffusion:

$$\omega(\xi,\eta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \omega_0(\xi-z,\eta-v)H(v)\chi(z)\,dv\,dz \quad . \quad (120)$$

The solution of this equation with respect to ω_0 is very inconvenient in the general case, though for data given in tabular form (table with two entries, ξ and η) a solution without any successive approximations analogous to the one described in section 16 gives a definite answer; the inconvenience arises from the great number of equations, equal to tt_1, where t and t_1 are the number of tabular intervals in ξ and η respectively

Similarly, the method of successive approximations may be applied, using ω as a first approximation for ω_0 and proceeding exactly as described in section 9. Simplifications due to some special properties of the functions involved may be introduced.

A quite general simplification of the two-dimensional case may, however, be used. According to the rules of section 9-16 we may find the separate frequency functions of x and y, F(x) and G(y) respectively. On the other hand, those are related to the true density function in the following manner:

$$F(x) = \int_{-\infty}^{+\infty} \omega_0(x,y)dy \qquad \cdots\cdots$$

$$G(y) = \int_{-\infty}^{+\infty} \omega_0(x,y)dx \qquad \cdots\cdots \qquad (121)$$

In the tabular data the number of equations here is $(t+t_1)$. When this number is less than tt_1, no advantage is presented, because then the number of different values of ω_0 determinable is also smaller.

Generally, in (121) the number of unknowns, tt_1, is greater than the number of equations, $t+t_1$; solutions, however, are made possible by applying the principle of smoothness. This is most easily shown by an example: Figure 6 is a scheme of a table with five entries, giving a total of 5+5=10 equations of the form (121); the number of unknowns is 25; of them we choose only 9 (those in the circles) as independent unknowns, and for the other unknowns we assume intermediate values, determined according to a certain rule of interpolation (the formula of interpolation may be chosen according to the expected character of ω_0). For linear interpolation we have, e.g.

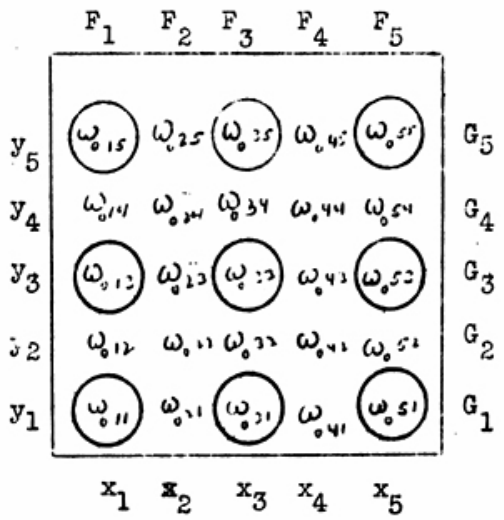

Figure 6.

$$\omega_{12} = \frac{\omega_{13} + \omega_{11}}{2}$$

$$\omega_{23} = \frac{\omega_{13} + \omega_{33}}{2}$$

$$\omega_{22} = \frac{\omega_{11} + \omega_{31} + \omega_{13} + \omega_{33}}{4}$$

etc.

The advantage of the simplified solution (121) is that it does no more contain the error functions; its disadvantage consists in a less detailed solution for ω_0.

Original

26) NUMERICAL DETERMINATION OF THE REGRESSION CURVES. The regression curve of y upon x is found in the following way. Dividing the whole range of x into sections (AB, BC, etc.) not necessarily equal, we compute the mean values of y and x within each section. We thus obtain normal points of the curve $\overline{y} = f(x)$; the mean error of y is found from

$$m.e. = \pm\sqrt{\frac{\sum(y-\overline{y})^2}{n(n-1)}}$$

for each section separately, or, on the assumption of constant dispersion, from all sections,

$$m.e. = \pm\sqrt{\frac{\sum \Delta y^2}{n(N-m)}} \quad \ldots\ldots\ldots (122)$$

where Δy is the deviation $(y-\overline{y})$, n the number of points within one section, $N = \sum n$ the total number of points, m the number of sections. The normal points with arrows indicating the size of the mean errors are represented in Figure 7. A smooth curve is drawn through the points, according to the rules of section 6. In a like manner the regression curve of x upon y may be constructed, dividing the area into sections parallel to the x-axis. If we try to do this in Figure 7, we may, however, arrive at a result which, though formally justified, has no real meaning: it is not permissible to join together data from the two different branches of the curve, ~~which are clearly separated from each other. This is because the circumstance that they fall within the same limits of~~ wide range in y is certainly not due to the effect of accidental errors. In such cases a modified method of treatment may be applied; though in cases like Figure 7

~~which~~ are clearly separated from each other. This ← which are The wide range in

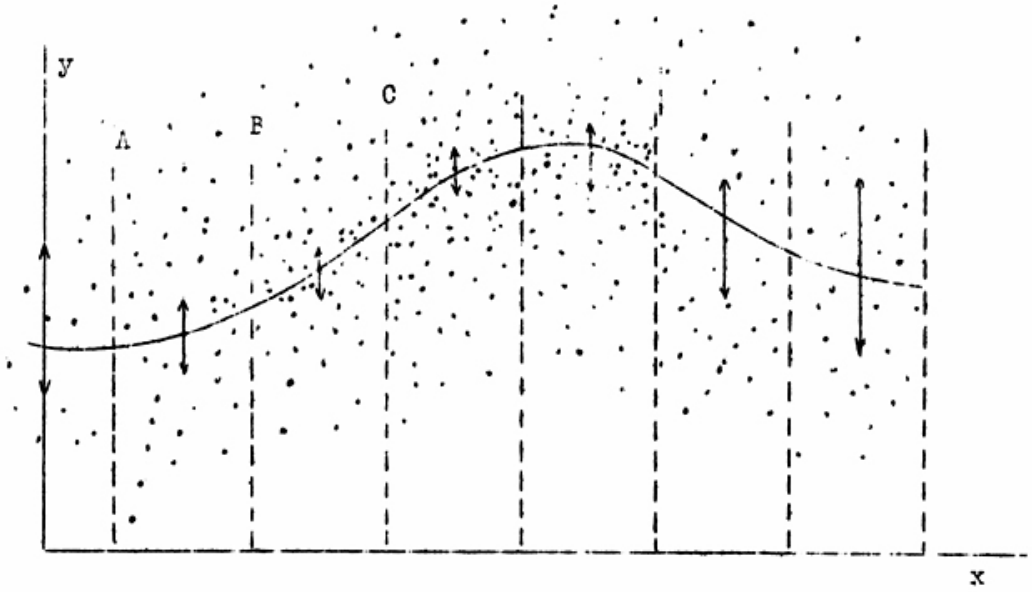

Figure 7.

the regression curve of x upon y cannot be constructed with any degree of precision.

Figure 8 illustrates the case of a certainly bifurcated regression curve. It is difficult to decide whether the bifurcation continues on the left-hand side (broken lines). The three isolated points encircled may be divided between the two branches proportionally to their mathematical expectation from each of the branches. Assuming a Gaussian spread of the points on both sides of the branches, the mathematical expectation is proportional to

$$n\alpha \; e^{-\frac{(y-\bar{y})^2}{2s_y^2}} \frac{dy}{s_y},$$

where n is the number of points belonging to one branch in the section KL, $(y-\bar{y})$ the deviation of the point from , and s_y (the dispersion around the given branch; the factor dy being constant may be rejected.

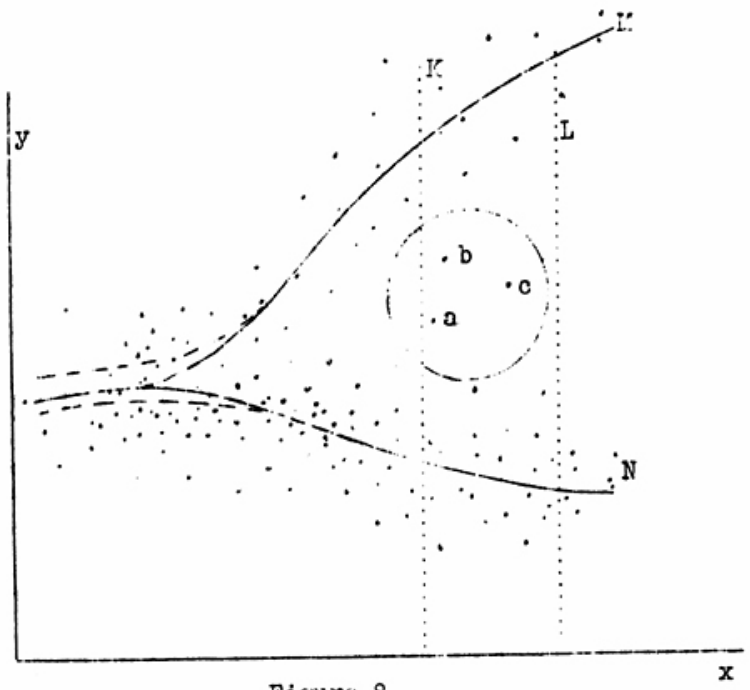

Figure 8.

Double-valued regression curve $\bar{y} = f(x)$.

E.g. let for a given point, b,

$$\frac{1}{s_y} e^{-\frac{(y-\bar{y})^2}{2s_y^2}}$$

be equal to 0.1 for branch N and 0.5 for branch M, $n_N = 20$, $n_M = 8$; hence the ratio of mathematical expectations N:M

$$N:M = \frac{(0.1)(20)}{(0.5)(8)} = \frac{2}{4} = \frac{1}{2}$$

and 1/(1+2) = 1/3 of the point is assumed to belong to N, 2/3 to M.

In more complicated cases, when the separation of the two branches is not clear enough, one may apply a more rigorous method, which involves, however, more work. Instead of computing the mean y, we form the frequency function of y separately for each section (interval A-B, B-C, etc. in Fig. 7 or K-L in Figure 8). Decomposing this frequency curve into two components (Figure 9), the ordinates $\bar{y}$ of the two branches are found as the median values of the argument, $\bar{y}_M$ and $\bar{y}_N$ (Corresponding to the maxima in the case of Gaussians). This method is the only one which may help to settle doubtful cases like the left-hand portion in Figure 8. Of course, the number of points must be great enough to be adequate to the method.

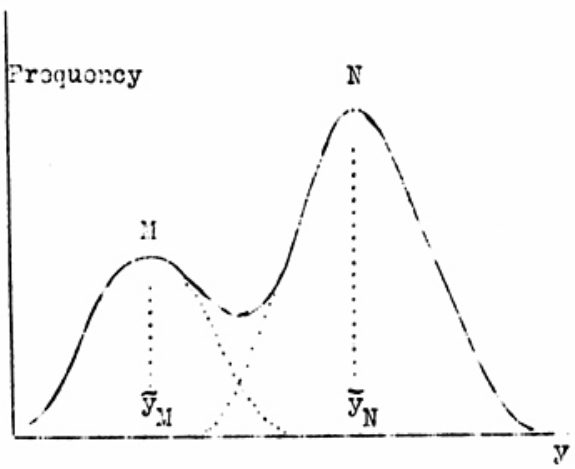

Figure 9.

There exists another frequently used method of constructing regression curves, which is believed to give "smooth" results. Instead of forming mean values for different intervals of x, one displaces a given interval along the x-axis, and obtains thus a great number of normal points, two neighboring normal points being based partly on common observed points. This method we will call the method of displaced normal points. Such a method is at least useless, as it gives no more information than the standard method described above, and involves a great deal of extra work; in most cases it may be harmful, leading to self-deception about the reality of the "smooth" curve obtained; if ϵ_1 and ϵ_2 are the accidental errors of two adjacent normal points in the standard method, the method of displaced normal points is nothing else than a smooth interpolation between these accidental quantities, a procedure of no real meaning. Only when the accidental errors are small, or negligible, this procedure may be used to avoid subjectiveness in drawing ~~in drawing~~ the curve; but in this case still we gain more by increasing the number of normal points in the standard method, taking the intervals of x more narrow and attaining in this way real detail.

27) LINEAR CORRELATION. Let us now consider Pearson's case, when the true regression curves are substituted by straight lines. Choosing the origin of coordinates in such a manner, that $x = 0$ and $y = 0$ (when this is not the case, we have to substitute in the following $x-x_0$ for x and $y-y_0$ for y, where $x_0 = \sum x/n$, $y_0 = \sum y/n$), we may write the regression lines in the following form:

$$\bar{y} = ax \qquad \ldots\ldots (123)$$

$$\bar{x} = by \qquad \ldots\ldots (124)$$

The coefficients a and b may be calculated according to the rules of least squares; assuming the equations of condition, $y = ax$ to possess equal weight, we find

$$a = \frac{\Sigma xy}{\Sigma x^2} \quad \ldots\ldots\ldots\ldots (125)$$

and

$$b = \frac{\Sigma xy}{\Sigma y^2} \quad \ldots\ldots\ldots\ldots (126)$$

In the case of ideal correlation, $y = a_0 x$ and $x = b_0 y$, equations (123) and (124) should be reversible, and we would have

$$b_0 = 1/a_0 , \quad \text{or} \quad a_0 b_0 = 1$$

In the case of no correlation we have $a = 0$ and taking into account that $\Sigma x^2 \neq 0$ (otherwise all points would be distributed along a straight line, we have $\Sigma xy = 0$, whence $b = 0$ also and $ab = 0$. Hence the quantity $r^2 = ab$ may be regarded as the measure of the apparent intensity of correlation; r, called the coefficient of correlation, is defined by

$$r = \pm\sqrt{ab} \quad \ldots\ldots (127)$$

$$r = + \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} \quad \ldots\ldots (127')$$

The sign of r, depending upon the sign of Σxy, indicates whether y increases with x (+), or decreases (-). $r = \pm 1$ means functional relationship; $r = 0$--no correlation.

The reality of correlation, however, does not directly depend upon the value of r; when r is great, this means real correlation; but when r is small, real correlation may still exist, being only concealed by accidental errors. To this we may add the possibility of curvilinear correlation which renders r less than unity up to 0, even in the case of functional relationship. Thus the absolute value of r is rather a measure of the reliability of linear correlation for the given set of observational data. From the standpoint of reality of some ideal correlation, the relative probable error, not the absolute value of r, is of importance.

The probable errors in a, b and r depend all upon the uncertainty in Σxy. In section 10, for independent x and y, we found the uncertainty equal to $\pm s_x s_y \sqrt{n}$; however, here x and y are not independent. Let us consider at first the average error, s_μ, to y as determined by the equation of condition:

$y = ax + \mu$ (μ = accidental error)

$\mu = y - ax$

$$s_\mu^2 = \frac{\Sigma \mu^2}{n} = \frac{\Sigma y^2 + a^2 \Sigma x^2 - 2a\Sigma xy}{n} = s_y^2 + a^2 s_x^2 - \frac{2a\Sigma xy}{n}$$

(127') may be represented as

$$r = \frac{\Sigma xy}{n s_x s_y} , \quad \text{whence} \quad \Sigma xy = nrs_x s_y ;$$

this gives

$$s_\mu^2 = s_y^2 + a^2 s_x^2 - 2ars_x s_y ;$$

further, we have from (125)

$$a = \frac{\Sigma xy}{ns_x^2} = r\frac{s_y}{s_x} \qquad \ldots\ldots (128)$$

whence

$$s_\mu^2 = s_y^2(1-r^2)$$

or

$$s_\mu = s_y\sqrt{1-r^2} \qquad \ldots\ldots (129)$$

Thus the spread in y relative to the regression line is a fraction $\sqrt{1-r^2}$ of the total spread in y; this gives us another aspect of the significance of the coefficient of correlation.

We may now find the uncertainty in Σxy. We have

$$\Sigma xy = \Sigma x(ax \pm \mu) = a\Sigma x^2 + \Sigma(\pm \mu x), \text{ and with (125)}$$

$$\Sigma xy = \Sigma xy + \Sigma(\pm \mu x).$$

The second term represents the uncertainty. We have

$$\Sigma(\pm x) = s_\mu \Sigma(\pm x) = \pm s_\mu s_x \sqrt{n} = \pm s_x s_y \sqrt{n(1-r^2)}$$

Hence it is easy to find the probable errors in a, b and c((s_x and s_y we regard as fixed for the given set of data). Finally we write

$$a = \frac{\Sigma xy}{ns_x^2} \pm \frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n}} \qquad \ldots\ldots (130)$$

or

$$a = \frac{s_y}{s_x}\left(r \pm \sqrt{\frac{1-r^2}{n}}\right) \qquad \ldots\ldots (130')$$

$$r = \frac{\Sigma xy}{ns_x s_y} \pm \sqrt{\frac{1-r^2}{n}} \qquad \ldots\ldots (132)$$

$$b = \frac{s_x}{s_y}\left(r \pm \sqrt{\frac{1-r^2}{n}}\right) \qquad \ldots\ldots (131)$$

The relative probable error in r, a or b is thus

$$\frac{s_a}{a} = \frac{s_b}{b} = \frac{s_r}{r} = \pm\sqrt{\frac{1-r^2}{nr^2}} \qquad \ldots (133)$$

Thus it depends not only upon r, but also upon the number of observations, hence, from a relatively small value of r we may infer the reality of a correlation when the number of observations is sufficiently large. The limit of reality of a correlation, for which the relative uncertainty is assumed equal to 1/2, is found from (133) as

$$|r| > \frac{2}{\sqrt{n+4}} \qquad \ldots\ldots (134)$$

and is contained in Table III.

TABLE III

Limit of Reality of the Coefficient of Correlation.

n =	5	10	20	50	100	500	1000	10,000
$\|r\| \geq$	0.67	0.54	0.41	0.27	0.20	0.09	0.06	0.02

0.45

The coefficient of correlation has the following geometrical significance. We have, Figure 10:

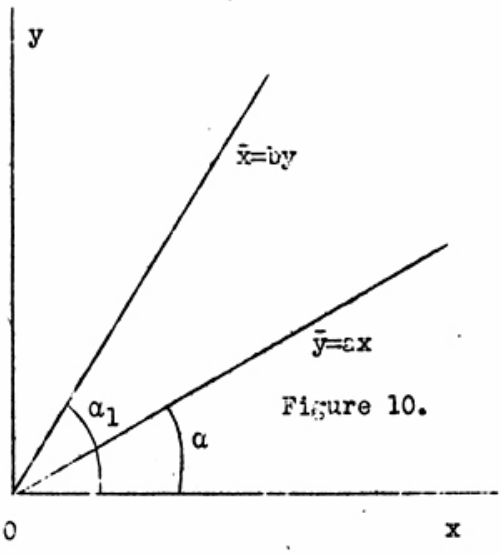

Figure 10.

$$a = \tan \alpha$$
$$1/b = \tan \alpha_1$$

whence

$$\frac{\tan \alpha_1}{\tan \alpha} = \frac{1}{ab} = \frac{1}{r^2} \quad . . \quad (135)$$

Thus the coefficient of correlation determines the relative inclination of the two regression lines; the smaller is r, the more the regression lines differ. We have also

$$\tan (\alpha_1 - \alpha) = \frac{a(1-r^2)}{a^2+r^2} = \frac{b(1-r^2)}{b^2+r^2} \quad \text{...(136)}$$

$$\tan(\alpha_1-\alpha) = \frac{s_x s_y}{(s_x^2+s_y^2)} \cdot \frac{(1-r^2)}{r} \quad \text{.....(136')}$$

This equation determines the angle between the regression lines. For r=0, $\alpha_1 - \alpha = 90°$; for r=1, $\alpha_1 - \alpha = 0$.

Assuming $s_x = s_y$, we have $\tan (\alpha_1 - \alpha) = \frac{1}{2} \frac{(1-r^2)}{r}$; for this case we obtain the following Table.

TABLE IV

Reliability of Correlation for $s_x = s_y$; $\alpha_1 - \alpha$ = angle between Regression Lines

r =	0.00	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\alpha_1 - \alpha$ =	90°	79°	67°	57°	46°	37°	28°	20°	13°	6°	0°

The angle between the regression lines may be regarded as a measure of their reliability: the greater the angle, the worse is the approximation to the true correlation.

When $s_x \neq s_y$, we obtain the same angles as in Table IV by changing the scale of y in the ratio (s_x/s_y), (r does not depend upon the scale).

Let us now find the expression for the regression lines in terms of the true correlation and vice versa. The ideal correlation we assume to be also in the linear form,

$$y = a_0 x \quad . . . . . \quad (137)$$

28) FRACTIONAL INDIVIDUALS. In dealing with individuals belonging to different groups it may happen that the limits of the groups are not sharp, in other words, that the demarkation of the groups is of a statistical character. Such a circumstance arose in section 26, for the double-valued correlation. Since we are not able to decide to which group the individual belongs, we may assign certain probabilities in favor of the different groups. In trying to find the probable number of individuals in a given group, we have to sum up these probabilities, or have to perform a count with fictitious fractional individuals. Let f denote the counted fraction of one individual; the probable number of individuals in a group is

$$n_f = \sum f \pm \sqrt{\sum f^2} \qquad \ldots\ldots\ldots (150)$$

In the simplest case, when f = const., we have

$$n_f = nf \pm f\sqrt{n} \qquad \ldots\ldots\ldots (150')$$

the relative natural uncertainty

$$\pm \frac{f\sqrt{n}}{n \cdot f} = \pm \frac{1}{\sqrt{n}}$$

in this case is exactly the same as for n counted whole individuals, although the apparent counted number, nf, is smaller.

The expression for the uncertainty in (150) takes into account only the natural uncertainty of the counted number; a possible error in the guessed value of f will increase the uncertainty. For f = const., the relative uncertainty is given by

$$\pm \sqrt{\left(\frac{\Delta f}{f}\right)^2 + \frac{1}{n}} \qquad \ldots\ldots\ldots (151)$$

where Δf is the mean error of f. For unequal f we may still use (151), assuming a certain average value of $(\Delta f/f)$ (the arithmetical mean of f, and the mean error of this arithmetical mean, Δf, are to be substituted). mean

29) AVERAGE AND EFFECTIVE QUANTITIES. Let $y = f(x)$ . . . (152) denote a certain functional relationship, and

$$x = f_1(y) \qquad \ldots (153)$$

the corresponding inverse function. The average values of x and y are

$$\bar{x} = \frac{\sum x}{n} \qquad \ldots (154)$$

and

$$\bar{y} = \frac{\sum y}{n} = \frac{\sum f(x)}{n} \qquad \ldots (155)$$

where n is the total number of points.

The <u>effective value of x, found through the relation of $y = f(x)$</u>, is defined as:

$$x_e = f_1(\bar{y}) = f_1\left[\frac{\sum f(x)}{n}\right] \qquad \ldots (156)$$

Original

Generally x_e is not equal to $\bar{x}$; only for x = const., i.e. when there is no spread in x (nor y), $x_e = \bar{x}$. The average value, $\bar{x}$, is evidently a particular case of effective values for f(x) = x. Effective quantities are frequently alluded to as simply mean values of certain functional significance, e.g.

$\bar{x}$ is the direct mean

$x_e =$

$$x_e = \frac{1}{\frac{\sum \frac{1}{x}}{n}}$$ is the harmonic mean $(y = 1/x)$

$$x_e = 10^{\frac{\sum \log x}{n}}$$ is the logarithmic mean $(y = \log x)$

etc.

The effective value depends both upon the frequency function of x, and upon the character of the intermediate function. Two sets of data giving identical effective values for one function, may yield different values for another function.

In practice one ordinarily calls effective quantities for which we have no observational control of the frequency function, and when y is found directly from observation; e.g. the effective temperature through the mediation of Stefan's Law.

$$I = kT^4$$

$$T_e = \sqrt[4]{\frac{\sum I}{kn}} = \sqrt[4]{\frac{\bar{I}}{k}} \quad ;$$

The quantity $\bar{I} = (\sum I/n)$

$$= \frac{\int k T^4 F(T)\, dT}{\int F(T) dT}$$

may be directly measured, though we may have no idea of the real distribution function of the temperatures, F(T). When in possession of full mathematical control over the process of formation of the effective quantity it is mostly called a mean quantity; the difference is, however, rather a psychological one, not a difference of principle.

II METEORS

1) BRIGHTNESS OF METEORS. The internal consistency of estimates of the magnitudes of meteors amounts to ±0.3 to ±0.5 mg (P.e.), according to the skill and experience of the observer. The estimates are thus of the same order of accuracy as absolute (not differential) estimates of magnitudes of fixed stars, provided that the observers bear in mind a definite scale of standard objects for comparison (direct or mental).

The physical meaning of the estimated brightness of meteors is not clear *a priori*. The estimated brightness of fixed stars corresponds to the intensity of radiation per unit of area and time; for moving objects we may expect errors depending upon the apparent (angular) speed; for objects of very short duration, or very great angular speed the total amount of light intercepted by the eye may determine the subjective impression of brightness. In any case, one has to expect a certain effect of motion, such that for low angular speeds the estimated brightness will give the intensity of radiation in the same units as for fixed stars, and with increasing angular speed the apparent brightness will be more and more underestimated.

The general character of the effect of motion may be guessed as follows, though the true circumstances of eye-sight are much more complicated than supposed here, the photographic effect of motion, on the other hand, must be very near to the subsequent scheme.

Let r be the effective radius of the image of a point source as projected upon the sensitive surface (of the eye, or the photographic plate); this quantity is defined so, that for a separation less than 2r, the combined light of two point-sources determines the photometric effect, whereas for a separation exceeding 2r, two point-sources are observed as independent photometric individuals (we neglect the gradual transition from one case to the other). 2r need not be equal to the optical separation; indeed, it appears that the radius of photometric separation of the human eye is greater than the radius of optical separation (double stars).

The optical effect is determined by the intensity i of radiation, and by the effective time of exposure, t; we may assume the following schematical form for the observable optical effect, or subjective brightness S:

$$S = i \cdot \left(\frac{t}{t_0}\right)^n \qquad \ldots\ldots\ldots (1)$$

for $t \leq t_0$; and $S = i$ .... (1') for $t > t_0$. Here t_0 is the limit of over-exposure, above which increase of the time of exposure does not sensibly change the optical impression already gained (save the case of very long exposures, when then rose of solarization, or weariness of the eye appear; The photographic plate possesses also the peculiarity to prove not an increased impression on such prior portions when the directly exposed portion is overexposed, giving thus an increase of impression after overexposure; the eye is free from this effect).

Let φ denote the speed of the moving light-source across the sensitive surface (retina, plate-film); both r and φ may be measured in angular units, e.g. degrees per second. The time of action is given evidently by

$$t = \frac{2r}{\varphi} \qquad \ldots\ldots (2)$$

Original

Hence

$$s = i\left(\frac{2r}{t_0\varphi}\right)^p \quad \ldots\ (3) \quad \text{for } \varphi \geq \frac{2r}{t_0}$$

Denoting by m_s the estimated, by m_i the true stellar magnitude, the effect of motion or the apparent decrease in magnitude from (3) is given by

$$m_s - m_i = \Delta m = 2.5\ p \log \varphi + C \quad \ldots(4)$$

where $C = 2.5\ p \log \frac{2r}{t_0}$.

Some preliminary experiments with stars observed through a moving telescope with power of 500 gave data contained in Table I. The data suggest a value of p for the human eye approaching 2; at the same time (4) does not represent the data well; instead of an abrupt transition to a line of constant slope, the transition is gradual. The effective value of the limiting angular velocity is unexpectedly high, about 300 degrees per second; with $2r = 1/3°$ this gives t_0 the value of about 10^{-3} sec., as compared with the duration of residual effect, after the light exposure is over, of about 10^{-1} sec. One may suppose that the eye follows the meteor if it does not move too fast. In any case, it seems that naked-eye observations of meteors require no correction for the effect of motion, because the angular velocities in this case seldom exceed 30° per second, and probably never reach as high values as 100° per second. The effect of motion begins to be of importance in meteor observations with telescopes of magnifying powers of 10 or more.

TABLE I

Observed Effect of Motion in Eye-estimates of the Brightness of a moving starlike object.*

φ, degrees per sec	100	160	250	400	630	1000	1600	2500	4000
Log φ	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6
Δm, mag.	0.1	0.2	0.3	0.6	1.2	2.1	3.0	4.0	5.0

In telescopic observations with high magnifying power the effect will play an important role in producing a selection of apparently slowly moving objects (near the radiant); this is exactly the fact noted by Denning, though he has tried to give it another explanation (enormous height of 1000 and more kilometers, quite improbable).

* H.C. 355, 1930

Original

EXERCISES

1. There is not a single star brighter than absolute magnitude 0 with a parallax exceeding 0".2. What are the probabilities for the density in space of stars with M<0 to exceed certain given values?

2. The average number (near the sun) of stars brighter than ⊙ is 6 per 1000 cubic parsecs (Kapteyn and Van Rhijn). What are the probabilities to find 0, 1, 2 etc. such stars with a parallax exceeding 0".2; what is the probability to find 20 or more such stars with a parallax >0".2?

3. The average number of stars brighter than 14.0 mag.pg. in galactic latitude 0° is 360 per square degree; what is the probability that there will be not a single star within a square 10' x 10'; and what is the expected number of star-void areas of this size on a Carte-du-Ciel plate (2° 10' x 2° x 10')?

4. There are 4 stars brighter than the sun with a parallax exceeding 0".2; what is the most probable space-density, and its uncertainty; what are the probabilities of different space-densities?

5. Among 21 stars of low luminosity (M≥10.0) with known spectra and parallax exceeding 0".2, there are 5 "collapsed" stars ("white dwarfs") (spectra B9; K0; F;K2;F0), the remaining oneshaving normally late (M) spectra. What is the hypothetical proportion (the most probable and the average) of white dwarfs, and its uncertainty; what are the probabilities of this proportion deviating on both sides from the most probable one by a certain amount?

(Pages 1-8)

Original.

SUPPLEMENTARY NOTE TO SECTION 7, "DECIMAL EQUATION."

The interval in the argument (ξ) after which the decimal equation repeats itself may be called the period of the decimal equation. For a decimal equation in the first decimal the period is 1.0.

When the period of the decimal equation is large in comparison with the range of the frequency table, the decimal equation may be determined in the following way. Before making any corrections for natural uncertainty, we drew a smooth curve through the observed points, taking care only that the total area will remain the same. Let $\Delta\nu$ denote the observed number, $\Delta\nu_0$ the smoothed number; the average decimal excess is then given by

$$\bar{\delta}_i = \frac{\sum \Delta\nu}{\sum \Delta\nu_0} \quad \ldots\ldots\ldots\ldots\ldots (a)$$

where the sum is to be taken over all tabular intervals having the same decimal, or group of decimals. If the total number of different groups of decimals is m, we must have the additional equation

$$\sum \bar{\delta}_i = m \quad \ldots\ldots\ldots\ldots\ldots (b)$$

The simplest case is when m = 2. Then $\delta_1 = 2-\delta_2$; putting

$$\bar{\delta}_1 = 1+\beta \; ; \qquad \bar{\delta}_2 = 1-\beta \quad \ldots\ldots\ldots (c)$$

we have for β a single equation

$$\beta = \frac{\sum_1 (\Delta\nu - \Delta\nu_0) + \sum_2 (\Delta\nu_0 - \Delta\nu)}{n} \quad \ldots\ldots (d)$$

where $\sum_1$ is the sum over the intervals having the 1st decimal, $\sum_2$ the sum over those with the 2nd decimal, and n is the total number.

Original

EXERCISES

9. Determine the general character of the error-functions in estimates of the magnitudes of meteors by different observers from the data of Tartu Publication 25, 4, 1923.

10. Investigate the correlation between spectroscopic absolute magnitudes and those determined from color index, for H.D. spectra Go and later, on the basis of Tartu Publication 27.1, 1929.

11. Same, for B5 and earlier.

12. Determine the coefficients of selection as a function of spectrum and apparent magnitude, for spectroscopic absolute magnitudes in the Tables of Tartu Publication 27.1, 1929.

13. Correlation between true total absorption, and total absorption computed from the approximate formula $A' = w_e(1-r_0)$, where w_e is the effective width corresponding to $\frac{1-r}{1-r_0} = 0.286$, from the data of H.C. 348.

READING PERIOD

May 1931

The Derivation of the Luminosity and General Density Laws of the Stars, with special reference to the effect of selection and error dispersion.

Literature (chief sources):

1) Kapteyn, Kapteyn and Van Rhijn
Groningen Publications 11, 30, 34, 38 (and others ad lib.)
Mt. Wilson Contribution 188.

2) Presad
Monthly Notices 85, 157, 1924.

3) Seares
Mt. Wilson Contribution 273.

4) Öpik
Tartu Publication 26, 4.

5) Furuhjelm, Recherches sur les mouvements Propres des Etoiles dans la Zone Photographique de Helsin[illegible]

Make a brief account of each of these papers, stating the object, methods and material used, and results; the methods are to be described (as far as possible) in terms of Dr. Öpik's lectures*; if possible, natural uncertainty of the results is to be estimated. The size of the accounts must not be large, from half to three or four pages each, according to the paper.

* E.G. methods of correcting for error dispersion and selection: state what kind of quantities is used: x, ξ or η etc.

An Introduction to Practical Statistics with Astronomical Applications.