PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material

Karl Pearson

Phil. Trans. R. Soc. Lond. A 1895 **186**, doi: 10.1098/rsta.1895.0010, published 1 January 1895

X. *Contributions to the Mathematical Theory of Evolution.*—II. *Skew Variation in Homogeneous Material.*

By KARL PEARSON, *University College, London.*

Communicated by Professor HENRICI, *F.R.S.*

Received December 19, 1894,—Read January 24, 1895.

PLATES 7–16.

CONTENTS.

PART I.—THEORETICAL.

16.7.95.

PART II.—PRACTICAL.

Statistical Examples.

PART I.—THEORETICAL.

Asymmetrical Frequency Curves.

(1.) AN asymmetrical frequency curve may arise from two quite distinct classes of causes. In the first place the material measured may be heterogeneous and may consist of a mixture of two or more homogeneous materials. Such frequency curves, for example, arise when we have a mixed population of two different races, a homogeneous population with a sprinkling of diseased or deformed members, a curve for the frequency of matrimony covering more than one class of the population, or in economics a frequency of interest curve for securities of different types of stability—railways and government stocks mixed with mining and financial companies. The treatment of this class of frequency curves requires us to break up the original curve into component parts, or simple frequency curves. This branch of the subject (for the special case of the compound being the sum of two normal curves) has been treated in a paper presented to the Royal Society by the author, on October 18, 1893.

The second class of frequency curves arises in the case of homogeneous material when the tendency to deviation on one side of the mean is unequal to the tendency to deviation on the other side. Such curves arise in many physical, economic and biological investigations, for example, in frequency curves for the height of the barometer, in those for prices and for rates of interest of securities of the same class, in mortality curves, especially the percentage of deaths to cases in all kinds of

fevers, in income tax and house duty returns, and in various types of anthropological measurements. It is this class of curves, which are dealt with in the present paper. The general type of this class of frequency curve will be found to vary (see Plate 7, fig. 1) through all phases from the form close to the negative exponential curve :

$$y = Ce^{-px},$$

to a form close to the normal frequency curve

$$y = Ce^{-px^2},$$

where C and p are constants.

Hence any theory which is to cover the whole series of these curves must give a curve capable of varying from one to another of these types, *i.e.*, from a type in which the maximum* practically coincides with the extreme ordinate, to a type in which it coincides with the central ordinate as in the normal frequency curve.

It is well known that the points given by the point-binomial $(\frac{1}{2} + \frac{1}{2})^n$ coincide very closely with the contour of a normal frequency curve when n is only moderately large. For example, the 21 points of $(\frac{1}{2} + \frac{1}{2})^{20}$ lie most closely on a normal frequency curve, and the author has devised a probability machine, which by continually bisecting streams of sand or rape seed for 20 successive falls gives a good normal frequency curve by the heights of the resulting 21 columns. Set to any other ratio $p : q$ of division other than bisection, the machine gives the binomial $(p + q)^{20}$, or indeed any less power and thus a wide range of asymmetrical point-binomials. Plate 7, fig. 2, represents, diagramatically, a 14-power binomial machine.

Just as the normal frequency curve may be obtained by running a continuous curve through the point-binomial $(\frac{1}{2} + \frac{1}{2})^n$ when n is fairly large, so a more general form of the probability curve may be obtained by running a continuous curve through the general binomial $(p + q)^n$. As the great and only true test of the normal curve is : Does it really fit observations and measurements of a symmetrical kind ? so the best argument for the generalised probability curve deduced in this paper is that it does fit, and fit surprisingly accurately observations of an asymmetrical character. Indeed, there are very few results which have been represented by the normal curve which do not better fit the generalised probability curve,—a slight degree of asymmetry being probably characteristic of nearly all groups of measurements. Before deducing the generalised probability curve, it may be well to show how any asymmetrical curve may be fitted with its closest point-binomial. This will be the topic of the following five articles.

(2.) Consider a series of rectangles on equal base c and whose heights are respectively the successive terms of the binomial $(p+q)^n \times \alpha/c$, where $p+q=1$. Here α is clearly the area of the entire system. Choose as origin a point O distant $\frac{1}{2}c$ from the

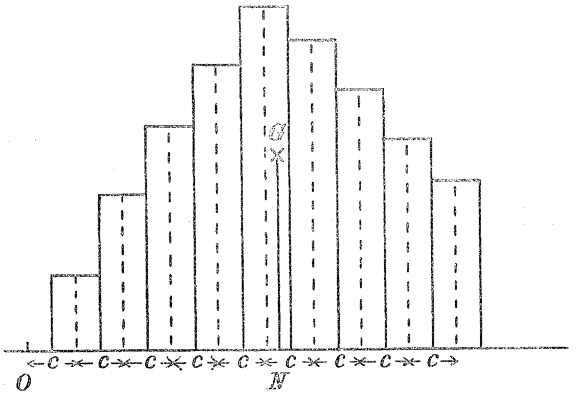
* I have found it convenient to use the term *mode* for the abscissa corresponding to the ordinate of maximum frequency. Thus the "mean," the "mode," and the "median" have all distinct characters important to the statistician.

boundary of the first rectangle, on the line of common bases, and let y_r be the height of the r^{th} rectangle, or

$$y_r = \frac{\alpha}{c} \frac{n(n-1)\dots(n-r+2)}{\lfloor r-1} p^{n-r+1} q^{r-1},$$

while

$$y_1 = \alpha p^n / c .$$



Let us find the values of

$$\Sigma\{y_r c \times (rc)^s\},$$

where s is any integer, for values of s from 0 to 4.

It is easy to see that

$$\Sigma\{y_r c \times (rc)^s\} = \alpha c^s \frac{d}{dq}\left(q \frac{d}{dq}\right)\left(q \frac{d}{dq}\right) \dots q(p+q)^n,$$

where the operation d/dq is repeated s times.

The operations indicated can easily be performed by putting $q = e^u$ when

$$\Sigma\{y_r c \times (rc)^s\} = \frac{\alpha c^s}{q}\left(\frac{d}{du}\right)^s \{e^u (p + e^u)^n\},$$

and the successive values can be found by LEIBNITZ'S theorem. After differentiation we may put $p + q$ or $p + e^u = 1$. There results :

$$\begin{aligned}
\Sigma(y_r c) &= \alpha \\
\Sigma(y_r c \times rc) &= \alpha c\{1 + nq\} \\
\Sigma(y_r c \times (rc)^2) &= \alpha c^2\{1 + 3nq + n(n-1)q^2\} \\
\Sigma(y_r c \times (rc)^3) &= \alpha c^3\{1 + 7nq + 6n(n-1)q^2 + n(n-1)(n-2)q^3\} \\
\Sigma(y_r c \times (rc)^4) &= \alpha c^4\{1 + 15nq + 25n(n-1)q^2 + 10n(n-1)(n-2)q^3 \\
&\qquad + n(n-1)(n-2)(n-3)q^4\}.
\end{aligned}$$

Let NG be the vertical through the centroid of the system of rectangles, then clearly

$$\text{ON} = \Sigma(y_r c \times rc)/\alpha = c\{1 + nq\}.$$

We shall now proceed to find the first four moments of the system of rectangles round GN. *If the inertia of each rectangle might be considered as concentrated along its mid vertical,* we should have for the s^{th} moment round NG, writing $d = c(1 + nq)$,

$$\alpha\mu_s = \Sigma\{y_r c \times (rc - d)^s\}.$$

The resulting values are

$$\mu_2 = npqc^2$$
$$\mu_3 = npq(p - q)c^3$$
$$\mu_4 = npq\{1 + 3(n - 2)pq\}c^4,$$

whence, remembering that $p + q = 1$, we find that p and q are roots of

$$z^2 - z + \frac{(3\mu_2^2 - \mu_4)\mu_2 + \mu_3^2}{4(3\mu_2^2 - \mu_4)\mu_2 + 6\mu_3^2} = 0,$$

$$n = \frac{2\mu_2^3}{(3\mu_2^2 - \mu_4)\mu_2 + \mu_3^2}, \qquad c = \frac{\sqrt{\{2(3\mu_2^2 - \mu_4)\mu_2 + 3\mu_3^2\}}}{\mu_2}.$$

Thus, when μ_2, μ_3, and μ_4 have been calculated for the frequency curve, the elements of the point-binomial are known. These results were given by me in a letter to 'Nature,' October 26, 1893.

They give quite a fair solution so long as n is large and c small, *i.e.*, so long as the asymmetry and the "excess" ('Phil. Trans.,' vol. 185, A, p. 93), measured respectively by μ_3 and $\mu_4 - 3\mu_2^2$ (which vanish for the normal curve) are not considerable.* In many cases, however, they are considerable, and the following solution is perfectly general.

* If y_0 denote the largest term in $(p + q)^n$ and y_t the tth term beyond it, then an application of STIRLING's theorem—if n be large—shows that

$$y_t/y_0 = \left(1 - \frac{t}{pn}\right)^{t-} \qquad \left(1 + \frac{t}{qn}\right)^{-t-qn-\frac{1}{2}}.$$

Take

$$\log u = (t - pn - \tfrac{1}{2})\log\left(1 - \frac{t}{pn}\right)$$
$$\log v = (-t - qn - \tfrac{1}{2})\log\left(1 + \frac{t}{qn}\right)$$

and expand the right hand side in powers of t, we find

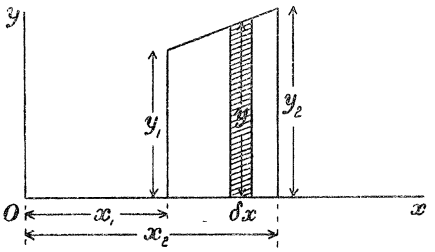
$$\log u = t\left(1 + \frac{1}{2pn}\right) - \frac{t^2}{2pn}\left\{1 - \frac{1}{2pn}\right\} - \frac{t^3}{6p^2n^2}\left(1 - \frac{1}{pn}\right) - \frac{t^4}{12p^3n^3}\left(1 - \frac{3}{2pn}\right) - \text{etc.}$$

Hence, remembering that $p + q = 1$, we have

$$\log uv = -\frac{t(p - q)}{2pqn} - \frac{t^2}{2npq}\left(1 - \frac{1 - 2pq}{2npq}\right) + \frac{t^3(p - q)}{6p^2q^2n^2}\left(1 - \frac{1 - pq}{npq}\right)$$
$$- \frac{t^4}{12p^3q^3n^3}\left(1 - 3pq - \frac{3(1 - 4pq + 2p^2q^2)}{2npq}\right) + \text{etc.}$$

Now, making use of the values given in § 2 for μ_2, μ_3, and μ_4, and writing $t \times c = x$, and $y_t = y$, we find

(3.) To find the nth moment of a trapezium ABCD about a line parallel to its parallel sides, y_1 and y_2 being the lengths of the parallel sides, x_1, x_2, their distances from the moment-axis, and $x_2 - x_1 = c$.



Let M_n be the nth moment. Then

$$M_n = \int_{x_1}^{x_2} yx^n\,dx$$

$$= \frac{y_2 - y_1}{x_2 - x_1}\frac{{x_2}^{n+2} - {x_1}^{n+2}}{n+2} + \frac{y_1x_2 - y_2x_1}{x_2 - x_1}\frac{{x_2}^{n+1} - {x_1}^{n+1}}{n+1}$$

$$= y_2\left(\frac{{x_2}^n c}{\lfloor 2} - \frac{n}{\lfloor 3}{x_2}^{n-1}c^2 + \frac{n(n-1}{\lfloor 4}{x_2}^{n-2}c^3 - \frac{n(n-1)(n-2)}{\lfloor 5}{x_2}^{n-3}c^4 + \ldots\right)$$

$$- y_1\left(\frac{{x_1}^n c}{\lfloor 2} + \frac{n}{\lfloor 3}{x_1}^{n-1}c^2 + \frac{n(n-1)}{\lfloor 4}{x_1}^{n-2}c^3 + \frac{n(n-1)(n-2)}{\lfloor 5}{x_2}^{n-3}c^4 + \ldots\right).$$

(4.) Now consider a curve of observations made up of a series of trapezia on equal bases, as in the accompanying figure:

$$y = y_0\, e^{-\frac{x^2}{2\mu_2}(1-\beta_1-\frac{1}{2}(3-\beta_2))} \times e^{-\frac{\mu_3 x}{2\mu_2^2}} \times e^{\frac{\mu_3 x^3}{6\mu_2^3}(1-\frac{5}{2}\beta_1-\frac{3}{2}(3-\beta_2))} \times e^{-\frac{x^4}{12\mu_2^2}(\frac{3}{4}\beta_1-\frac{1}{2}(3-\beta_2)-\frac{3}{4}(3-\beta_2)^2-\frac{9}{4}\beta_1^2-\frac{9}{2}(3-\beta_2)\beta_1)} \times \text{etc.}$$

where $\beta_1 = {\mu_3}^2/{\mu_2}^3$ and $\beta_2 = \mu_4/{\mu_2}^2$.

This appears to be the more general form of a result given by Professor EDGEWORTH, 'Roy. Soc. Proc.,' vol. 56, p. 271.

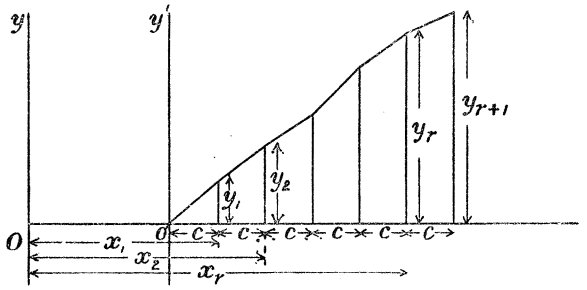
For the normal curve $\mu_3 = 0$, $\mu_4 = 3{\mu_2}^2$; hence, if p does not differ much from q, β_1 and $\beta_2 - 3$ will be small, and we may neglect their products with $x/\sqrt{\mu_2}$. Thus approximately

$$y = y_0\, e^{-\frac{x^2}{2\mu_2}} e^{-\frac{\mu_3}{2\mu_2^2}\left(x - \frac{x^3}{3\mu_2}\right)}.$$

This agrees with Professor EDGEWORTH'S special case if we expand the second exponential. His "negative frequency" is accounted for by the fact that he has only taken the first terms of a long series, *i.e.*,

$$y = y_0\, e^{-x^2/2\mu_2}\left\{1 - \frac{\mu_3}{2{\mu_2}^2}\left(x - \frac{x_3}{3\mu_2}\right)\right\}.$$

I have not considered this form of the skew-curve at length, because it is only a *first* approximation to the more general forms considered in this paper, and further, because it is only applicable in practice within extremely narrow limits.



Here $y_1, y_2, y_5, \ldots y_r$ are the frequencies of deviations falling within the ranges $x_1 \pm \frac{1}{2}c$, $x_2 \pm \frac{1}{2}c$, $x_3 \pm \frac{1}{2}c \ldots x_r \pm \frac{1}{2}c \ldots$, and the tops of the ordinates are joined to form a frequency-curve in the usual manner.

Let M'_n be the nth moment of the system of trapezia about the line Oy, then

$$\mathrm{M}'_n = \mathrm{S}\left\{2y_r\left(\frac{x_r^n c}{\lfloor 2} + \frac{n(n-1)}{\lfloor 4} x_r^{n-2}c^3 + \frac{n(n-1)(n-2)(n-3)}{\lfloor 6} x_r^{n-4}c^5 + \ldots\right)\right\}.$$

In particular, if we take Oy in the position $O'y'$ at distance c from y_1, we have $x_r = rc$, and accordingly,

$$\mathrm{M}'_n = c^{n+1}\left(\mathrm{N}'_n + \frac{n(n-1)}{12}\mathrm{N}'_{n-2} + \frac{n(n-1)(n-2)(n-3)}{360}\mathrm{N}'_{n-4} + \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{20160}\mathrm{N}'_{n-6} + \text{etc.}\right),$$

where $\mathrm{N}'_s = \mathrm{S}(y_r r^s)$.

In particular,

$$\begin{aligned}
\mathrm{M}'_0 &= c\mathrm{N}'_0,\\
\mathrm{M}'_1 &= c^2\mathrm{N}'_1,\\
\mathrm{M}'_2 &= c^3(\mathrm{N}'_2 + \tfrac{1}{6}\mathrm{N}'_0),\\
\mathrm{M}'_3 &= c^4(\mathrm{N}'_3 + \tfrac{1}{2}\mathrm{N}'_1),\\
\mathrm{M}'_4 &= c^5(\mathrm{N}'_4 + \mathrm{N}'_2 + \tfrac{1}{15}\mathrm{N}'_0),\\
\mathrm{M}'_5 &= c^6(\mathrm{N}'_5 + \tfrac{5}{3}\mathrm{N}'_3 + \tfrac{1}{3}\mathrm{N}'_1).
\end{aligned}$$

When we put $\mathrm{M}'_s/\mathrm{M}'_0 = \mu'_s$, and $\mathrm{N}'_s/\mathrm{N}'_0 = \nu'_s$, these reduce to

$$\begin{aligned}
\mu'_1 &= c\nu'_1,\\
\mu'_2 &= c^2(\nu'_2 + \tfrac{1}{6}\nu'_0),\\
\mu'_3 &= c^3(\nu'_3 + \tfrac{1}{2}\nu'_1),\\
\mu'_4 &= c^4(\nu'_4 + \nu'_2 + \tfrac{1}{15}\nu'_0),\\
\mu'_5 &= c^5(\nu'_5 + \tfrac{5}{3}\nu'_3 + \tfrac{1}{3}\nu'_1).
\end{aligned}$$

Now let μ_n be the value of the nth moment of the trapezia system about the vertical through its centroid divided by its area.

We have :

$$\mu_n = \mu'_n - n\mu'_1\mu'_{n-1} + \frac{n(n-1)}{1.2}\mu'^2_1\mu'_{n-2} - \frac{n(n-1)(n-2)}{1.2.3}\mu'^3_1\mu'_{n-3} + \text{etc.}$$

Thus we find :

$$\mu_1 = 0,$$
$$\mu_2 = c^2(\nu'_2 - \nu'^2_1 + \{\tfrac{1}{6}\}),$$
$$\mu_3 = c^3(\nu'_3 - 3\nu'_1\nu'_2 + 2\nu'^3_1),$$
$$\mu_4 = c_4(\nu'_4 - 4\nu'_1\nu'_3 + 6\nu'^2_1\nu'_2 - 3\nu'^4_1 + \{\nu'_2 - \nu'^2_1 + \tfrac{1}{15}\}),$$
$$\mu_5 = c^5(\nu'_5 - 5\nu'_1\nu'_4 + 10\nu'^2_1\nu'_3 - 10\nu'^3_1\nu'_2 + 4\nu'^5_1 + \{\tfrac{5}{3}\nu'_3 - 5\nu'_1\nu'_2 + \tfrac{10}{3}\nu'^3_1\}).$$

Comparing these results with those given in the 'Phil. Trans.,' vol. 185, p. 79, Eq. (4), we see that treating the curve as built-up of trapezia instead of loaded ordinates introduces the parts into the values of the μ's enclosed in curled brackets. These additions are small, but in many cases quite sensible. Since the series of trapezia gives in general a closer approach than the series of loaded ordinates to the frequency curve, and, further, since the calculation of these additional terms is not very laborious, it will be better for the future to calculate the moments of any frequency curve from the above modified formulæ.

(5.) Returning now to the point-binomial, we have :

$$\nu'_1 = 1 + nq,$$
$$\nu'_2 = 1 + 3nq + n(n-1)q^2,$$
$$\nu'_3 = 1 + 7nq + 6n(n-1)q^2 + n(n-1)(n-2)q^3,$$
$$\nu'_4 = 1 + 15nq + 25n(n-1)q^2 + 10n(n-1)(n-2)q^3 + n(n-1)(n-2)(n-3)q^4.$$

Thus :

$$\mu_2 = c^2(npq + \tfrac{1}{6}),$$
$$\mu_3 = -c^3 npq(q-p),$$
$$\mu_4 = c^4(\tfrac{1}{15} + npq(2 + 3(n-2)pq)).$$

If, instead of taking trapezia, we had taken a series of rectangles, but not, as in § 2, concentrated their areas along their axes, we should have found the following system :

$$\mu_2 = c^2(npq + \tfrac{1}{12}),$$
$$\mu_3 = -c^3 npq(q-p),$$
$$\mu_4 = c^4(\tfrac{1}{80} + npq(\tfrac{3}{2} + 3(n-2)pq)).$$

Hence if we write :

$$\mu_2 = c^2 (npq + \epsilon_1),^*$$
$$\mu_3 = -c^3 npq (q - p),$$
$$\mu_4 = c^4 (\epsilon_2 + npq (\epsilon_3 + 3 (n - 2) pq)),$$

we have :

For trapezia :	$\epsilon_1 = \frac{1}{6}$,	$\epsilon_2 = \frac{1}{15}$,	$\epsilon_3 = 2$,
For rectangles :	$\epsilon_1 = \frac{1}{12}$,	$\epsilon_2 = \frac{1}{80}$,	$\epsilon_3 = 1{\cdot}5$,
For loaded ordinates :	$\epsilon_1 = 0$,	$\epsilon_2 = 0$,	$\epsilon_3 = 1$,

and the above general system may be applied to all cases.

Writing

$$z = npq, \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \text{and } \beta_2 = \frac{\mu_4}{\mu_2^2},$$

we have by elimination the cubic for z :

$$z^3 (6 + 3\beta_1 - 2\beta_2) + z^2 (2\epsilon_3 - 3 + 9\beta_1\epsilon_1 - 4\beta_2\epsilon_1) + z (2\epsilon_2 + 9\beta_1\epsilon_1^2 - 2\beta_2\epsilon_1^2) + 3\beta_1\epsilon_1^3 = 0.$$

The remaining constants of the binomial are :

$$n = \frac{4z}{1 - \beta_1 z (1 + \epsilon_1/z)^3},$$
$$pq = \tfrac{1}{4} (1 - \beta_1 z (1 + \epsilon_1/z)^3),$$

and

$$c = \sqrt{\frac{\mu_2}{z + \epsilon_1}}.^\dagger$$

(6.) Let us illustrate these results by a numerical example. Plate 8 gives Dr. VENN'S curve for 4857 barometric heights. Along the horizontal, 1 cm. equals ·1″ of height of barometer, and the scale of frequency is 1 sq. cm. = 28·304 observations. The centroid vertical and the second, third, and fourth moments about it were found for me‡ by the graphical process described, 'Phil. Trans.,' vol. 185, p. 79. We have the following results :—

* This result seems of considerable importance, and I do not believe it has yet been noticed. It gives the mean square error for any binomial distribution, and we see that for most practical purposes it is identical with the value $\sqrt{npq}$, hitherto deduced as an *approximate* result, by assuming the binomial to be approximately a normal curve.

† If we take $z + \epsilon_1 = \chi$ the fundamental cubic reduces to

$$(6 + 3\beta_1 - 2\beta_2) \chi^3 - (2 - \tfrac{3}{5}\beta_2) \chi^2 + \tfrac{3}{10} \chi - \tfrac{1}{45} = 0,$$

a form in which the coefficients are easily calculated and the nature of the roots discriminated.

‡ By Mr. G. U. YULE, who has given me very great assistance in the laborious calculations required in the reduction of frequency curves. We have used, with much economy of time, the "Brunsviga" calculator.

$$\alpha = 171{\cdot}6, \qquad \mu_2 = 10{\cdot}14,$$
$$\mu_3 = 15{\cdot}95, \qquad \mu_4 = 326{\cdot}34,$$

all in centimetre units.

These give

$$\beta_1 = {\cdot}24401, \ \beta_2 = 3{\cdot}1739.$$

Hence for trapezia,

$${\cdot}3842z^3 - {\cdot}749917z^2 + {\cdot}018008z + {\cdot}003389 = 0,$$

and for rectangles,

$${\cdot}3842z^3 - {\cdot}87496z^2 - {\cdot}003832z + {\cdot}000424 = 0.$$

These give the following solutions :—

	Trapezia.	Rectangles.	Lines.
z	1·92516	2·28034	2·6028
n	19·379	23·983	28·5293
p	·8881	·8936	·89985
q	·1119	·1064	·10015
c	2·2017	2·0712	1·974
α/c	77·94	82·85	86·93
d	6·976	7·3562	7·614

Here $d = c\,(1 + nq)$ gives the distance of the start of the point-binomial from the centroid vertical. The three point-binomials are therefore

$$77{\cdot}94\,({\cdot}8881 + {\cdot}1119)^{19{\cdot}379},$$
$$82{\cdot}85\,({\cdot}8936 + {\cdot}1064)^{23{\cdot}983},$$
$$86{\cdot}93\,({\cdot}89985 + {\cdot}10015)^{28{\cdot}5293},$$

respectively.

These three point-binomials are represented in Plate 8, fig. 3. It will be noticed that they all lie very close to the barometric curve; they would be still closer if that curve were a real curve and not a polygonal line. The total areas between binomial-polygons and observation curves, treating all parts as positive, are for the three cases, 10·3, 10·5, 11·0 sq. centims. respectively, or taking the base range to be 23 centims., we have mean deviations from the observation curve of ·448, ·457, ·478 in the three cases respectively. Thus the method of trapezia gives slightly the best result; the method of concentrating along ordinates the worst result. The total area of the curve being 171·6, we have from another standpoint, mean percentage errors* in the ordinates of about 6·03, 6·06, and 6·3, respectively. The generalised probability curve, if fitted to the same observations, gives an areal deviation of 7 sq. centims., or a percentage error of about 4. Thus it is very nearly one-third as close again as the point-binomials.

* The "percentage error" in ordinate is, of course, only a rough test of the goodness of fit, but I have used it in default of a better.

As typical samples of mean percentage errors considered by various statisticians to give good results, I may note the following, the frequency being about 1,000 or upwards :—AIRY, 9 ; MERRIMAN, 13·5 ; GALTON (Anthropometric), 7 to 15 ; WELDON (Crabs), 6·7, (Shrimps), 8·8; STIEDA (Skulls), 7·6; PORTER (School Girls), 7·7; PEROZZO (Recruits), 6·8 ; BRADLEY'S observations, 5·85 ; PEARSON (Lottery), 6·7, (Tossing), 6·6.

It is therefore clear that our point-binomials and generalized curve may be considered to give good results.* It will be noticed, however, that a little difference in the method of calculating the point-binomials leads, without much alteration of the percentage error, to a considerable change in their centroid-positions and the magnitude of their constants.† Generally speaking we may conclude that in round numbers the barometric frequency corresponds to the binomial $(\cdot 9 + \cdot 1)^{20}$, or to the distribution of zeros when 20 ten-sided teetotums, marked 0, 1 . . . 9, are spun together. There is an apparent upper limit to the height of the barometer, and its deviation below the mean can be much greater than its deviation above. At the same time within the narrower range round the mean, the frequency of a high barometer is greater than the frequency of a low barometer; the odds against a "contributory cause" tending to a low barometer being about 9 to 1. I propose to investigate a wider series of barometric observations, in order to test how far the conclusions which may be drawn from Dr. VENN'S statistics are general.‡

A rather interesting point may be considered at this stage. Is it always possible to fit a point-binomial to a series of observations with a chance frequency? Can we better the normal curve by a point-binomial? The answer is Yes, if the fundamental cubic in χ (second footnote, p. 351), has a real positive root. Now for the normal curve $2(3\mu_2^2 - \mu_4)\mu_2 + 3\mu_3^2$, or $6 + 3\beta_1 - 2\beta_2$ is zero. For the loaded ordinates c will only be real if this expression be positive. It may, however, take small negative values for the trapezia, in which case χ itself will be small and only within narrow limits give suitable values for n.

Hence, for real values of n, p and q, it is impossible to fit a point-binomial to a series of observations for which $6 + 3\beta_1 - 2\beta_2$ has a large negative value. The normal curve, for which $\mu_4 = 3\mu_2^2$, is nearer to any such observations than a point-binomial.

For example, by aid of the modified expressions given in this paper, p. 350, we have

* As another manner of testing, compare the ten-points of the point-binomial for lines with observations :—

Theory	5·6	15·9	21·8	19	11·9	5·7	2·1	·7	·2	·03
Observation . . .	5·7	15·8	22·1	18·8	12	5·8	2·3	1·1	·2	·00

† A curve drawn through the 30 points of the three point-binomials would be very close to the observations. As a matter of fact, the skew probability curve passes very near to all 30 points.

‡ [Miss A. Lee has since calculated the constants of three years of Eastbourne barometric observations for me. While n and c differ widely from the Cambridge values, she finds $p = \cdot 89375$, $q = \cdot 10625$, a striking and suggestive agreement.]

for the data given for Professor WELDON's Crab Measurements, No. 4, 'Phil. Trans.,' A, vol. 185, p. 96.

$$\mu_2 = 7{\cdot}6759, \quad \mu_3 = 3{\cdot}4751, \quad \mu_4 = 184{\cdot}3039.$$

Hence,

$$\beta_1 = \mu_3^2/\mu_2^3 = \ {\cdot}0267022,$$
$$\beta_2 = \ \mu_4/\mu_2^2 = 3{\cdot}12807.$$

Thus $6 + 3\beta_1 - 2\beta_2$ is positive, and accordingly no rational point-binomial is likely to fit as well as the normal curve. As a matter of fact the fundamental cubic is now

$${\cdot}17603z^3 + 1{\cdot}045327z^2 + {\cdot}033773z - {\cdot}0003709 = 0.$$

The two negative roots of this equation give imaginary value for p and q. The small positive root gives p greater than unity and q negative, n is also negative. Although I can give no interpretation to these results, it seemed well to complete in the latter case the solution and test how near the resulting point-binomial fitted the curves. I found

$$z = {\cdot}00866, \quad p = 1{\cdot}19268, \quad q = -{\cdot}19268.$$
$$n = -{\cdot}037685, \quad c = 6{\cdot}61662, \quad d = 6{\cdot}6645.$$

These give for the binomial

$$150{\cdot}0983\ (1{\cdot}19268 - {\cdot}19268)^{-{\cdot}037685},$$

or,

$$151{\cdot}89\ (1 - {\cdot}161552)^{-{\cdot}037685},$$

or,

$$151{\cdot}89 + {\cdot}92532 + {\cdot}07756 + \&c.$$

Thus the sensible part of the binomial to the scale of our figure is a *triangle*. I have drawn this binomial, see Plate 8, fig. 4. The reader will mark a fit very close on the whole to the observations. We have the following percentage mean errors of the ordinates :—

Normal curve	6·7,
Skew probability curve.	4·4,
Binomial	10·5.

We may conclude, therefore, that even if our binomial constants have unintelligible values, yet our method will give, in many cases, a closely-fitting polygonal figure. This remark should be read in connection with Professor EDGEWORTH'S somewhat divergent views* on fitting chance distributions with curves other than the normal error curve. It is possible in almost every case to find simple combinations of lines,

* See 'Phil. Mag.,' vol. 334, p. 24, *et seq.*, 1887.

circles, or parabolas of various degrees which give results extremely close to any given set of observations.

For example, taking the range of frequency to be sensibly π times the standard deviation, we have the following close expression for the error function by harmonic analysis

$$y = y_0 \left\{ \cdot 399 + \cdot 482 \cos \frac{x}{\sigma} + \cdot 109 \cos \frac{2x}{\sigma} + \cdot 009 \cos \frac{3x}{\sigma} \right\}.$$

Here y_0 is the maximum ordinate, x any deviation, and σ the standard deviation. A couple of wave curves* will thus very frequently give us a close approximation to a set of statistical measurements, quite as close as statistical practice shows the error curve to be.

The above expression further allows the normal curve to be constructed by aid of scale and compasses—*geometrically*, or its ordinates calculated from a table of cosines.

Another example of the fitting of a point-binomial will be found in Part 2, § 34, *Pauper Percentages*.

(7.) Consider the point-binomial $e \times (\frac{1}{2} + \frac{1}{2})^n$, where e is any constant, and suppose a polygon formed by plotting up the terms of the binomial at distance c from each other.

Then, corresponding to $x_r = rc$, we have

$$y_r = e \frac{n(n-1)(n-2)\ldots(n-r+2)}{\underline{|r-1}} (\tfrac{1}{2})^n$$

and

$$\frac{y_{r+1} - y_r}{\frac{1}{2}(y_{r+1} + y_r) \times c} = \frac{c(n+2) - (x_r + x_{r+1})}{\frac{1}{2}(n+1)c^2} = -\frac{(x'_r + x'_{r+1})}{\frac{1}{2}(n+1)c^2},$$

if $x'_r = x_r - \frac{1}{2}c(n+2)$.

Now $(y_{r+1} - y_r)/c$ is the slope of the polygon corresponding to the mean ordinate $\frac{1}{2}(y_{r+1} + y_r)$, or, writing† $\sigma^2 = \frac{1}{2} \times \frac{1}{2}(n+1)c^2$,

$$\frac{\text{slope of polygon}}{\text{mean ordinate}} = -\frac{2 \times \text{mean abscissa}}{2\sigma^2}.$$

* It is often sufficient to take

$$y = y_0 \left(\tfrac{2}{5} + \tfrac{1}{2} \cos \frac{x}{\sigma} + \tfrac{1}{9} \cos \frac{2x}{\sigma} \right).$$

† The divergence of this value of σ^2 from the ordinary value $\frac{1}{2} \times \frac{1}{2} \times n$ is to be noted. The two agree sensibly if n be great. [Drawing on a large scale, however, the point-binomial $(\frac{1}{2} + \frac{1}{2})^{10}$ and the two normal curves with standard deviations of 1·5811 and 1·6533, I find that the latter has a mean percentage error of only 1·76 as compared with 5·1 of the former. Thus it would appear that the normal curve corresponding to $\sqrt{(n+1)pq}$ fits the point-binomial closer than one with the standard deviation $\sqrt{npq}$ usually adopted.]

Now compare this property of the polygon with that of the curve:

$$y = y_0 e^{-x^2/2\sigma^2}.$$

We have by differentiation:

$$\frac{\text{slope of curve}}{\text{ordinate}} = -\frac{2\ \text{abscissa}}{2\sigma^2}.$$

Hence: *this binomial polygon and the normal curve of frequency have a very close relation to each other, of a geometrical nature, which is quite independent of the magnitude of* n. In short their slopes are given by an identical relation. By a proper choice of σ and y_0, we can get the normal curve to fit closely the point-binomial, owing to this slope property, *without any assumption as to the indefinitely great value of* n. It is this geometrical property which is largely the justification for the manner in which statisticians apply, and apply with success, the normal curve to cases in which n is undoubtedly small. No stress seems hitherto to have been laid upon the fact that the normal curve of errors besides being the limit of a symmetrical point-binomial has also this intimate geometrical relationship with it.*

(8.) Now let us deal with the skew point-binomial in precisely the same manner as we have dealt with the symmetrical binomial. Taking its form to be $c\,(p+q)^n$, we have, if $x_r = r \times c$ and $\lambda = q/p$:

$$\frac{y_{r+1} - y_r}{\frac{1}{2}(y_{r+1} + y_r)\,c} = \frac{2}{c}\,\frac{(n-r+1)\,\lambda/r - 1}{(n-r+1)\,\lambda/r + 1} = \frac{2\,(\lambda\,(n+1) - \lambda\,(\lambda+1))}{c\,(\lambda\,(n+1) + r\,(1-\lambda))}.$$

Let us write $\Delta_r y = y_{r+1} - y_r$, $\Delta x = c$.

$$\mathrm{Y}_{r+\frac{1}{2}} = \tfrac{1}{2}\,(y_{r+1} + y_r),\quad \mathrm{X}_{r+\frac{1}{2}} = \tfrac{1}{2}\,(x_{r+1} + x_r).$$

Then $\mathrm{X}_{r+\frac{1}{2}}/c = r + \frac{1}{2}$, and:

* The following table shows the closeness of frequency within a given range as determined by the binomials:—

Range of deviation.	Frequency per cent.		Normal curve.
	$(1+1)^{10}$.	$(1+1)^{20}$.	
3	24	23	24
5	37	37	38
7	50	52	52
11	71	73	73
15	87	87	87
21	96	96	96
33	100	100	100

Here the distribution of 100 groups each of 100 events is seen to be practically the same whether we take $n = 10$ or $n = \infty$.

$$\frac{\Delta_r y}{\Delta x \times y_{r+\frac{1}{2}}} = \frac{2}{c} \frac{\lambda(n+1) - (1+\lambda)\left(\frac{X_{r+\frac{1}{2}}}{c} - \frac{1}{2}\right)}{\lambda(n+1) + (1-\lambda)\left(\frac{X_{r+\frac{1}{2}}}{c} - \frac{1}{2}\right)},$$

or, if $X'_{r+\frac{1}{2}} = X_{r+\frac{1}{2}} - c\left(\frac{1}{2} + q(n+1)\right)$,

$$= \frac{-X'_{r+\frac{1}{2}}}{pq(n+1)c^2 + (p-q)\frac{c}{2}X'_{r+\frac{1}{2}}}$$

$$= \frac{-\gamma X'_{r+\frac{1}{2}}}{\alpha + X'_{r+}},$$

if $\gamma = \frac{2}{(p-q)c}$ and $\alpha = \frac{2pq(n+1)c}{p-q}$.

The curve which has the same law of slope as this skew binomial is:

$$y = y_0 (1 + x/\alpha)^{\gamma\alpha} e^{-\gamma x}.$$

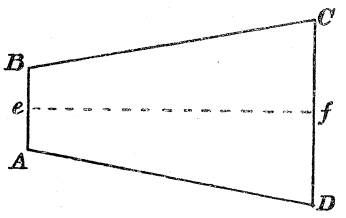
(9.) This curve accordingly stands in the same relationship to the skew binomial as the normal curve to the symmetrical binomial.* There are several points, however, to be considered with regard to it. In the first place it is usually assumed that n is indefinitely great and c indefinitely small, and then it is supposed that we may neglect $(p-q)cX'_{r+\frac{1}{2}}$ as compared with $pq(n+1)c^2$, and so we deduce the normal error curve whether p be equal to q or not. But I contend that this is unjustifiable except for very small values of $X'_{r+\frac{1}{2}}$. When the deviation X' is considerable and c vanishingly small, X' will be an indefinitely great multiple of c; c must be in fact the unit in which X' is measured and unless $p = q$, the ordinary normal curve is only an approximation, even if n be large, near the maximum frequency. In the next place, when we speak of n being large, are we quite clear as to what we mean in the case of physical or biological frequency curves? We speak of a multiplicity of small "causes" determining the actual dimensions of an organ, or the size of a physical error, or the height of the barometer. But it is less clear why this multiplicity should be identified with the infinite greatness of n. If we take Dr. VENN'S frequency curve for barometric height, we see that the closest point-binomial is by no means consistent with either $p = q$, or with n being indefinitely great. Further, many statistical results in games of chance are given with great exactness by the normal curve, although we are then able to show that n is quite moderate.

Now, it is true that the biological and physical statistics to which we are referring, give essentially continuous curves, but it does not seem to follow of necessity that n must be infinite; while their frequent skewness sufficiently indicates that the neglect

* Note again the deviation of the constant $pq(n+1)c^2$ from its usually adopted value $pqnc^2$.

of $X'_{r+\frac{1}{2}}$ as compared with a is unjustifiable. Thus, the maximum of a fever mortality curve cannot be an infinite distance from birth, which limits the curve in one direction, nor an age-at-marriage curve have a maximum frequency infinitely distant from the age of puberty, nor a frequency of interest curve separate its maximum, between 3 or 4 per cent., by an infinite distance from 0 per cent. It is clear, therefore, that if such frequency curves as those referred to are to be treated as chance distributions at all, it would be idle to compare them to the limit of a symmetrical binomial. We are really quite ignorant as to the nature of the contributory "causes" in biological, physical, or economic frequency curves. The continuity of such frequency curves may depend upon other features than the magnitude of n. If I toss twenty coins, a discrete series of 0, 1, 2, 3, . . . 20, heads is the only possible range of results. Each individual coin, here representing a "contributory cause" can only give head or tail, and so many whole coins must give head, so many tail. If I want to make any ratio of head to tail, I have to take an indefinitely great number of coins, for each "contributory cause" must give a unit to the total. But it may possibly be that continuity in biological or physical frequency curves may arise from a limited number of "contributory causes" with a power of *fractionizing* the result. We cannot conceive on the tossing of 20 coins that 13·5 will give heads and 6·5 will give tails, we are obliged to deal with 200 coins, 135 giving heads and 65 tails. Yet the two things are not identical. The former corresponds to a value intermediate between two ordinates of $(\frac{1}{2}+\frac{1}{2})^{20}$, and the latter to a definite ordinate of $(\frac{1}{2}+\frac{1}{2})^{200}$. So long as we remain in ignorance of the nature and number of "contributory causes" in physics and biology, so long as we do find markedly skew distributions, it seems to me that we must seek more general results than flow from the assumption that $p = q$ and $n = \infty$. The form of curve given in § 8 above is suggested as a possible form for skew frequency curves. Its justification lies essentially, like that of the normal curve, in its capacity to express statistical observations.

(10.) But it must be noted that the generalised probability curve in § 8, although it contains the normal curve as a special case, is not sufficiently general. It is limited in *one* direction, indefinitely extended in the other. This limitation at one end only, corresponds *theoretically* to many cases in economics, physics, and biology. But there are a great variety of cases in which there is theoretical limitation at both



ends; that is to say, there is a limited *range* of possible deviations. For example, let a trapezium, ABCD, of white paper be pasted on a cylinder of black surface with

ef, the axis of symmetry parallel to the axis of the cylinder. Then, if the cylinder be rotated, we shall have a series of grey tints from a darkish *e* to a lighter *f*. Now, if we ask several hundred persons to select a tint which would result from mixing the tints at *e* and *f*, we shall obtain a continuous frequency curve, falling, however, entirely within the *range* *e* to *f*. Or, again suppose a frequency curve obtained by plotting up the frequency of a given ratio of leg-length to total body-length, or of carapace to body-length. Here the range must lie between 0 and 1. It is not that other values are excessively improbable, they are by the conditions of the problem absolutely impossible. Hence, it is clear that the curves obtained by Professor WELDON and Mr. H. THOMPSON in the case of shrimps, crabs, and prawns, can only be approximately normal curves, even if it were possible for the ratios to run from 0 to 1. But as a matter of fact, the possible range is very much smaller. We may not be able to assert, *à priori*, what it is, but for an adult prawn to have a carapace $\frac{2}{3}$ or $\frac{1}{1000}$ of its body-length, or a man a leg $\frac{2}{3}$ or $\frac{1}{20}$ of his body-length, may be regarded as impossibilities; they are abnormalities, which could hardly survive to the adult condition. Precisely the same remarks apply to skull indices, and probably to the relative size of all sorts of organs in the adult condition. We may not know the range, *à priori*, but we are quite certain that one exists, and it is a quantity to be determined—just as the mean or the standard deviation—from our measurements themselves. We may take it that in most biological measurements of adults there is a range of stability, so to speak, organs not falling within this range are inconsistent with the continued existence of the individual, with the assumption that he has lived to be an adult.* Nor is this question of range confined to biological statistics. A barometric frequency curve must show the same peculiarity; there are excessively low and excessively high barometric heights which would be not only inconsistent with the survival of any meteorological observer, but also with the existing features of physical nature on this earth. In vital statistics we find precisely the same thing, a curve of percentages of mothers of different ages for the children born during any year in a country would be definitely limited by the ages of puberty and the climacteric, which cannot be pushed indefinitely towards childhood and senility respectively. Again in disease and mortality curves, while the lower limit of life is clear, it is highly probable that an upper limit exists, if we can only fix it by investigation of our statistics themselves. A man of the present day, as now organised, may be able to live 120 years, perhaps, but we have exceeded his vital possibilities if we take, say, 200 years.

Thus the problem of *range* seems a very important one, it *theoretically* excludes the use of the normal curve in many classes of statistics; it is quite true that, for many practical purposes, frequency curves of limited range may be sensibly identical either with unlimited curves, or even with normal curves, but, in other cases, this

* Absolute malformations, congenital, or due to post-natal accident are excluded. Abortions or amputations would be naturally excluded from our measurements.

is not so, and under any circumstances the limited curve may actually give information as to the possible range—the "limits of stability"—which is itself of great value.

We have, then, reached this point: *that to deal effectively with statistics we require generalised probability curves which include the factors of skewness and range.* The generalised curve we have already reached, possesses skewness, but its range is limited in one direction only.

Accordingly, we require the following types of frequency curves:—

Type I.—Limited range in both directions, and skewness.

Type II.—Limited range and symmetry.

Type III.—Limited range in one direction only and skewness.

Type IV.—Unlimited range in both directions and skewness.

Type V.—Unlimited range in both directions and symmetry.

Type V. is the normal curve; Type IV., with slight skewness, has been dealt with by POISSON in the form of an approximative series.* Type III. has been given above, it was first published by me without discussion in 'Roy. Soc. Proc.,' vol. 54, p. 331.

We can now turn to the general problem.

(11.) A very simple example will illustrate how a frequency curve, with limited range and skewness, may be considered to arise. Take n balls in a bag, of which pn are black, and qn are white, and let r balls be drawn and the number of black be recorded. If $r > pn$, the range of black balls will lie between o and pn; the resulting frequency polygon will be skew and limited in range. This polygon, which is given by a hypergeometrical series, leads us to generalised probability curves, in the same manner as the symmetrical and skew binomials lead us to special cases of such curves. If we consider our balls to become fine shot, or ultimately sand, and suppose each individual grain to have an equal chance of being drawn, we obtain a continuous curve.† It is not, however, impossible that, could we measure with sufficient accuracy, many physical as well as biological statistics might be found to proceed by units, much as in certain types of economic statistics we are not troubled with fractions of a penny. For this reason we shall keep our results in the most general form, and obtain a curve approximating to the hypergeometrical series referred to without any assumptions as to the relative magnitude of the quantities involved.

We easily obtain for the series giving the chances of r, $r - 1$, $r - 2 \ldots 0$, black balls being drawn out of a bag containing pn, black, and qn, white, the expression

* "Sur la Probabilité des Jugements," chapter 3.

† p pints of red sand and q pints of white sand are put into a vessel, and r pints are withdrawn. We have if $r > p$, a perfectly continuous frequency curve for red sand withdrawn ranging between o and p pints. We are here supposing no "perfect mixture" of the two kinds of sand, but theoretical equality of chances for each grain.

$$\frac{pn\,(pn-1)\,(pn-2)\ldots(pn-r+1)}{n\,(n-1)\,(n-2)\ldots(n-r+1)}$$
$$\times\left(1+r\,\frac{qn}{pn-r+1}+\frac{r\,(r-1)}{1.2}\,\frac{qn\,(qn-1)}{(pn-r+1)\,(pn-r+2)}\right.$$
$$\left.+\frac{r\,(r-1)\,(r-2)}{1.2.3}\,\frac{qn\,(qn-1)\,(qn-2)}{(pn-r+1)\,(pn-r+2)\,(pn-r+3)}+\&c.\right).$$

If y_s be the s^{th} ordinate of this polygon, and we suppose these ordinates plotted up at distances c apart, we have

$$\frac{y_{s+1}}{y_s}=\frac{r-s+1}{s}\,\frac{qn-s+1}{pn-r+s},$$
$$x_s=s\times c,\qquad x_{s+1}=(s+1)\,c,$$
$$\mathrm{X}_{s+\frac{1}{2}}=c\,(s+\tfrac{1}{2}).$$

Thus

$$\frac{y_{s+1}-y_s}{\frac{1}{2}\,(y_{s+1}+y_s)\times c}=\frac{2}{c}\,\frac{(r+1)\,(1+qn)-s\,(n+2)}{(r+1)\,(1+qn)-s\{2\,(r+1)+n\,(q-p)\}+2s^2}$$
$$=\frac{2}{c}\,\frac{(r+1)\,(1+qn)-\left(\frac{\mathrm{X}_{s+\frac{1}{2}}}{c}-\frac{1}{2}\right)(n+2)}{(r+1)\,(1+qn)-\left(\frac{\mathrm{X}_{s+\frac{1}{2}}}{c}-\frac{1}{2}\right)\{2\,(r+1)+n\,(q-p)\}+2\left(\frac{\mathrm{X}_{s+\frac{1}{2}}}{c}-\frac{1}{2}\right)^2}.$$

Write

$$\mathrm{X}'_{s+\frac{1}{2}}=\mathrm{X}_{s+\frac{1}{2}}-c\left(\frac{1}{2}+\frac{(r+1)\,(1+qn)}{n+2}\right),$$

and we find with our previous notation

$$\frac{\Delta_s y}{\Delta x}\,\frac{1}{\mathrm{Y}_{s+\frac{1}{2}}}=\frac{-\mathrm{X}'_{s+\frac{1}{2}}}{\beta_1+\beta_2\mathrm{X}'_{s+\frac{1}{2}}+\beta_3\mathrm{X}'^2_{s+\frac{1}{2}}}\qquad\ldots\ldots\qquad(\epsilon),$$

where

$$\beta_1=\frac{c^2\,(r+1)\,(n-r+1)\,(1+qn)\,(1+pn)}{(n+2)^3},$$
$$\beta_2=\frac{cn\,(n-2r)\,(p-q)}{2\,(n+2)^2},\qquad\beta_3=\frac{1}{n+2}.$$

Now, if we attempt to find the curve which has the same geometrical relation for the slope as the above hypergeometrical polygon, we see that it will change its type according to the sign of $\beta_2^2-4\beta_1\beta_3$.

After some reductions we have

$$\sqrt{\{\beta_2^2-4\beta_1\beta_3\}}$$
$$=\frac{cn}{n+2}\left\{\left(\tfrac{1}{2}-\sqrt{\left(p+\frac{1}{n}\right)\left(q+\frac{1}{n}\right)}-\frac{r}{n}\right)\left(\tfrac{1}{2}+\sqrt{\left(p+\frac{1}{n}\right)\left(q+\frac{1}{n}\right)}-\frac{r}{n}\right)\right\}^{\frac{1}{2}}.$$

Hence $\sqrt{\{\beta_2^2 - 4\beta_1\beta_3\}}$ will be real or imaginary, according as r/n lies outside or between the limits

$$\tfrac{1}{2} \pm \sqrt{\left\{\left(p + \frac{1}{n}\right)\left(q + \frac{1}{n}\right)\right\}}.$$

If r/n lies outside these limits, then the integral of the right-hand side of equation (ϵ) is purely logarithmic; if it lies between these limits, the integral is in part trigonometrical.

Since r must be less than n, it follows that the integral must be trigonometrical if these limits are respectively $= <0$ and $= >1$, *i.e.*, if

$$(p + 1/n)(q + 1/n) = \text{or} > \tfrac{1}{4},$$

or p must lie between $\tfrac{1}{2} \pm \sqrt{\left\{\frac{1}{n}\left(1 + \frac{1}{n}\right)\right\}}$.

For example, if $n = 100$, then, if p lies between ·6005 and ·3995, the integral must be trigonometrical. If p lies outside these limits, say $= ·7$ for example, then the integral will be logarithmic if r/n does not lie between ·04 and ·96, *i.e.*, if we draw a small or large proportion of the total contents.

Let us treat the trigonometrical and logarithmic cases separately.

(12.) Case I. $\beta_2^2 < 4\beta_1\beta_3$.

The curve having the same geometrical slope relation is

$$\log y = \text{constant} - \frac{1}{2\beta_3}\log(\beta_1 + \beta_2 x + \beta_3 x^2) - \frac{\beta_2}{2\beta_3}\frac{2}{\sqrt{\{4\beta_1\beta_3 - \beta_2^2\}}}\tan^{-1}\frac{2\beta_3 x + \beta_2}{\sqrt{\{4\beta_1\beta_3 - \beta_2^2\}}}.$$

Write x for $x + \beta_2/2\beta_3$, changing the origin; further put a for $\sqrt{\{4\beta_1\beta_3 - \beta_2^2\}}/(2\beta_3)$, m for $1/(2\beta_3)$, and ν for $\dfrac{\beta_2}{\beta_3\sqrt{\{4\beta_1\beta_3 - \beta_2^2\}}}$, then we have, y_0 being a constant of integration,

$$y = \frac{y_0}{(1 + x^2/a^2)^m} e^{-\nu \tan^{-1}(x/a)}.$$

This frequency curve is asymmetrical and has an unlimited range on either side of the origin. It corresponds accordingly to the curve required as Type IV.

Here

$$a = \tfrac{1}{4}c\sqrt{\{4(1 + pn)(1 + qn) - (n - 2r)^2\}},$$

$$\nu = \frac{n(n - 2r)(p - q)}{\sqrt{\{4(1 + pn)(p + qn) - (n - 2r)^2\}}}$$

$$m = \tfrac{1}{2}(n + 2).$$

Special cases. (i.) Suppose $r/n = \chi$, and n very large, then

$$m/a^2 = \frac{2}{c^2 n\,(pq - (\frac{1}{2} - \chi)^2)} = \alpha_1, \text{ say,}$$

$$\nu/a = \frac{(1 - 2\chi)\,(p - q)}{c\,(pq - (\frac{1}{2} - \chi)^2)} = \alpha_2, \text{ say.}$$

Thus we have

$$y = y_0 e^{-\alpha_1 x^2 - \alpha_2 x},$$

which reduces to the normal type by a change of origin. It is important to notice, however, that the standard deviation of this normal type

$$= \sqrt{(1/2\alpha_1)} = \tfrac{1}{2} c \sqrt{\{n\,(pq - (\tfrac{1}{2} - \chi)^2)\}},$$

and is *very different* from the value $c\sqrt{\{(r + 1)\,pq\}} = \frac{1}{2}c\sqrt{(npq \times 4\chi)}$, nearly, which is the usual form. Only when we put $p = q = \frac{1}{2}$ and make χ small do they agree. We thus conclude: *That the normal form may fit a chance distribution, but it does not follow that the standard deviation is of the binomial type generally assumed.*

(ii.) Suppose $\chi = \frac{1}{2}$, corresponding to the withdrawal of one-half of the contents of a vessel, then

$$y = y_0\,(1 + x^2/a_0^2)^{-m},$$

where

$$a_0 = \tfrac{1}{2} c \sqrt{\{(1 + pn)\,(1 + qn)\}}.$$

This is an unlimited and symmetrical frequency curve approaching more and more nearly to the normal form as we increase n. It has, however, a standard deviation $= \frac{1}{2}c\sqrt{(npq)}$, while the normal curve would give $\frac{1}{2}c\sqrt{(npq \times 2)}$.

(iii.) Suppose $p = q = \frac{1}{2}$, we again reach the form

$$y = y_0\,(1 + x^2/a_0^2)^{-m},$$

where

$$a_0 = \tfrac{1}{4} c\,(n + 2) \sqrt{\left\{1 - \left(\frac{n - 2r}{n + 2}\right)^2\right\}}.$$

Make n infinite and we have again the normal type, but a standard deviation of the form $\frac{1}{2}c\sqrt{\{n\chi\,(1 - \chi)\}}$, only approaching the usual value when χ is small.

We postpone until we have discussed the remaining types the problem of fitting a curve of Type IV. to a series of observations.

(13.) Case II. $\beta_2^2 > 4\beta_1\beta_3$.

Let α_1 and α_2 be the roots of $\beta_1 + \beta_2 x + \beta_3 x^2 = 0$. Then the curve having the same geometrical relation for its slope is

$$\frac{d(\log y)}{dx} = -\frac{x}{\beta_3\,(x - \alpha_1)\,(x - \alpha_2)}$$

$$= -\frac{1}{\beta_3\,(\alpha_1 - \alpha_2)} \frac{d}{dx} \{\alpha_1 \log (x - \alpha_1) - \alpha_2 \log (x - \alpha_2)\},$$

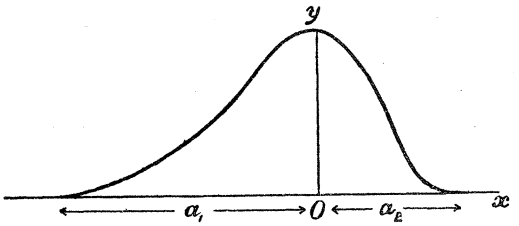
or, if

$$1/\nu = \beta_3 (a_1 - a_2),$$
$$y = y'_0 (x - a_1)^{-\nu a_1} (x - a_2)^{\nu a_2}$$
$$= y_0 (1 - x/a_1)^{-\nu a_1} (1 - x/a_2)^{\nu a_2}$$

by changing constants.

Assuming that y_0, ν, a_1 and a_2 can take any sign whatever, we see that there are three fundamental subtypes of this frequency curve,

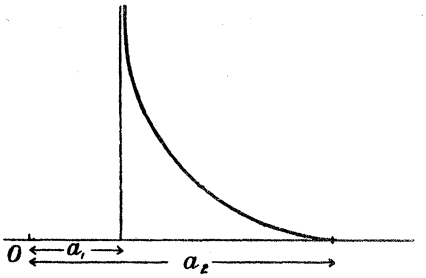
(i.) $y = y_0 (1 + x/a_1)^{\nu a_1} (1 - x/a_2)^{\nu a_2}$.



This is an asymmetrical curve with limited range and maximum towards mediocrity.

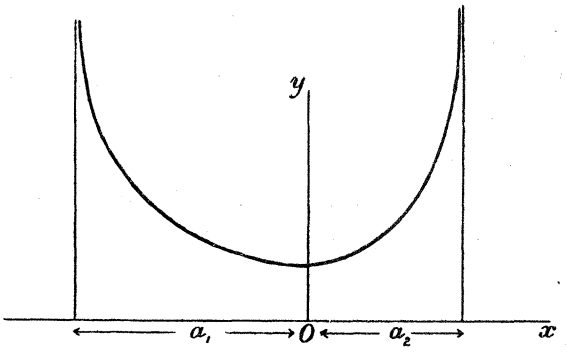
As a rule νa_1 and νa_2 are fractional and the curve becomes imaginary beyond the limits $x = -a_1$ and $x = a_2$.

(ii.) $y = y_0 (x/a_1 - 1)^{-\nu a_1} (1 - x/a_2)^{\nu a_2}$.



Here the ordinate between $x = a_1$ and $x = a_2$ varies from infinity to zero, and resembles the frequency curves given by "wealth" distribution or infant mortality.

(iii.) $y = y_0 (1 - x/a_1)^{-\nu a_1} (1 + x/a_2)^{-\nu a_2}$.



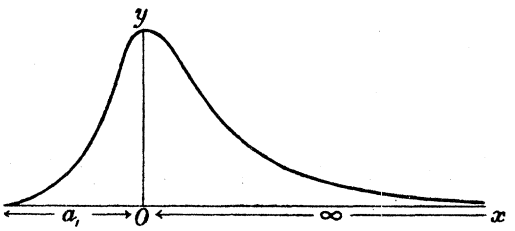
This is an asymmetrical curve with limited range, mediocrity being in a minimum. The disappearance of mediocrity is not a very uncommon feature of statistics; the

"prevalence of extremes" may appear not only in meteorological phenomena but in competitive examinations, where the mediocre have occasionally sufficient wisdom to refrain from entering. The type is that of Mr. F. GALTON's curve of "consumptivity."*

The curve contains an interesting number of less fundamental subtypes.

(iv.) Make $a_2 = \infty$ in (i.),

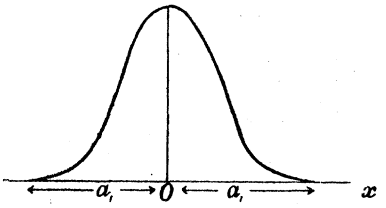
$$y = y_0 (1 + x/a_1)^{\nu a_1} e^{-\nu x}.$$



This is the limit to the asymmetrical binomial, which has been already referred to in § 8.

(v.) Make $a_1 = a_2$,

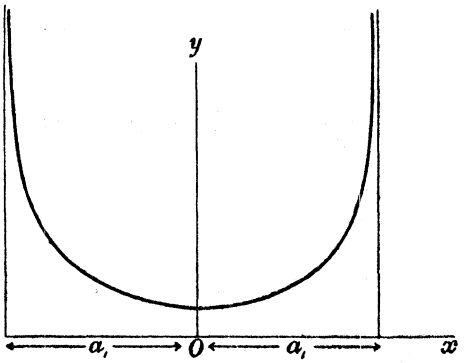
$$y = y_0 (1 - x^2/a_1^2)^{\nu a_1}.$$



This is the symmetrical frequency curve of limited range.

(vi.) Make ν negative in (v.),

$$y = \frac{y_0}{(1 - x^2/a_1^2)^{\nu a_1}}.$$



This is a symmetrical frequency curve, with limited range, and minimum of mediocrity.

(vii.) Put $\nu = pa_1$ in (v.) and make $a_1 = \infty$,

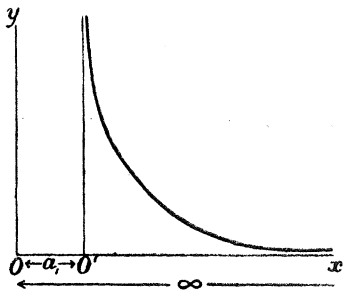
$$y = y_0 e^{-px^2}.$$

This is the normal curve.

* 'Natural Inheritance,' 1889, p. 174.

(viii.) Put $a_2 = \infty$ in (ii.),

$$y = y_0 (x/a_1 - 1)^{-\nu a_1} e^{-\nu x}.$$



This is an asymmetrical frequency curve, with an ordinate varying from a_1 to ∞ along an infinite range.

All eight of the above types are included in the single form

$$y = y_0 (1 + x/a_1)^{\nu a_1} (1 - x/a_2)^{\nu a_2},$$

or

$$y = y_1 x^r (1 - x/c)^s,$$

if we give positive, negative, or limiting values to the constants. But to do this we require to give values to n and r in the expressions for β_1, β_2, and β_3, which are not easily intelligible, if we rigidly adhere to our example of drawing a definite quantity of sand from a limited mixture of two kinds of sand. The last type of curve given is, however, the frequency curve for *à priori* probabilities,* and readily admits of a direct interpretation of the following kind.

Given a line of length l, and suppose $\overline{r + 1}$ points placed on it at random; what is the frequency with which the point pr from one end and qr from the other of the series of $\overline{r + 1}$ points falls on the element δx of the line?

The answer is clearly

$$\frac{\lfloor r}{\lfloor pr \, \lfloor qr} \left(\frac{x}{l}\right)^{pr} \left(1 - \frac{x}{l}\right)^{qr} \frac{\delta x}{l},$$

or, we have a frequency curve of the type

$$y = y_0 x^{pr} (1 - x/l)^{qr}.$$

We may express the problem a little differently. Take $\overline{r + 1}$ cards and slip them at random between the pages of a book, the frequency of the page succeeding the $\overline{pr + 1}^{\text{th}}$ card is given by the above curve.†

* See CROFTON, "Probability," § 17, 'Encycl. Brit.'

† The important point to be noticed here is that we are dealing with a distribution in which contributory causes are inter-dependent.

Until we know very much more definitely than we do at present, how the size of an organ in any individual, say, depends on the sizes of the same organ in its ancestors, or what are the nature of the causes which lead to the determination of prices, or of income, or of mortality at a given age, I do not see that we have any right to select as our sole frequency curve the normal type

$$y = y_0\, c^{-px^2}$$

in preference to the far more general

$$y = y_0\,(1 + x/a_1)^{\nu a_1}\,(1 - x/a_2)^{\nu a_2},$$

which not only includes the former, but supplies the element of skewness which is undoubtedly present in many statistical frequency distributions. As we may look upon the former as a limit to a coin-tossing series, so the latter represents a limit to teetotum-spinning and card-drawing experiments. It is not easy to realise why nature or economics should, from the standpoint of chance, be more akin to tossing than to teetotum-spinning or card-dealing. At any rate, from purely utilitarian and prudent motives, we are justified so long as the analysis is manageable, in using the more general form. It will always give us a measure of the divergence of particular statistics from the normal type, and in many cases of skew frequency, it can be used when it would be the height of absurdity to apply the normal curve at all.

Since Types I., II., III., and V. are all represented by the curve

$$y = y_0\,(1 + x/a_1)^{\nu a_1}\,(1 - x/a_2)^{\nu a_2}$$

and Type IV. by the curve

$$y = y_0\,\frac{1}{(1 + x^2/a^2)^m}\,e^{-\nu \tan^{-1} x/a},$$

we have only to deal with these two cases in general. We shall refer, in the course of our work, to special simplifications arising in particular sub-cases. After a description of the manner in which these generalised probability curves may be fitted to statistics, we shall indicate, by examples, their practical applications.

(14.) *On the Generalised Probability Curve. Type I.*

$$y = y_0\,(1 + x/a_1)^{\nu a_1}\,(1 - x/a_2)^{\nu a_2}.$$

Let the range $a_1 + a_2 = b$; let $m_1 = \nu a_1$, $m_2 = \nu a_2$, $z = (a_1 + x)/(a_1 + a_2)$, whence $x = -a_1$, $z = 0$ and $z = a_2$, $z = 1$.

Further let

$$\eta = y_0\,(a_1 + a_2)^{m_1 + m_2}/a_1^{m_1} a_2^{m_2},$$
$$= y_0\,(m_1 + m_2)^{m_1 + m_2}/m_1^{m_1} m_2^{m_2},$$

thus

$$y = \eta z^{m_1}\,(1 - z)^{m_2}.$$

Let α be the area of the curve between $x = -\alpha_1$ and $x = \alpha_2$, $\alpha\mu'_n$ its n^{th} moment round a parallel to the axis of y through $x = -\alpha_1$, and $\alpha\mu_n$ its n^{th} moment round the centroid vertical.

Then we have

$$\begin{aligned}\alpha\mu'_n &= \int_0^b yx'^n dx,\\ &= b^{n+1}\eta \int_0^1 z^{m_1+n}(1-z)^{m_2} dz,\\ &= b^{n+1}\eta\, \mathrm{B}(m_1+n+1,\ m_2+1),\\ &= b^{n+1}\eta \frac{\Gamma(m_1+n+1)\,\Gamma(m_2+1)}{\Gamma(m_1+m_2+n+2)}.\end{aligned}$$

Thus, by the fundamental property of the Γ function, we have

$$\alpha = b\eta\,\Gamma(m_1+1)\,\Gamma(m_2+1)\,/\,\Gamma(m_1+m_2+2),$$

$$\mu'_1 = \frac{b(m_1+1)}{m_1+m_2+2}, \qquad \mu'_2 = \frac{b^2(m_1+2)(m_1+1)}{(m_1+m_2+3)(m_1+m_2+2)},$$

$$\mu'_3 = \frac{b^3(m_1+3)(m_1+2)(m_1+1)}{(m_1+m_2+4)(m_1+m_2+3)(m_1+m_2+2)},$$

$$\mu'_4 = \frac{b^4(m_1+4)(m_1+3)(m_1+2)(m_1+1)}{(m_1+m_2+5)(m_1+m_2+4)(m_1+m_2+3)(m_1+m_2+2)}.$$

From these we easily deduce by the formulæ connecting μ and μ', if we write for brevity, $m_1+1 = m'_1$, $m_2+1 = m'_2$, and $m'_1+m'_2 = r$:

$$\mu_2 = \frac{b^2 m'_1 m'_2}{r^2(r+1)}, \qquad \mu_3 = \frac{2b^3 m'_1 m'_2 (m'_2 - m'_1)}{r^3(r+1)(r+2)}, \qquad \mu_4 = \frac{3b^4 m'_1 m'_2 (m'_1 m'_2 (r-6) + 2r^2)}{r^4(r+1)(r+2)(r+3)}.$$

Now, α, μ_2, μ_3, and μ_4 are to be found by the methods indicated in Art. 4 from the polygon of observations, and may be supposed known quantities, when we are dealing with the fitting of frequency-curve to observations.

Then, if $\beta_2 = \mu_4/\mu_2^2$, and $\beta_1 = \mu_3^2/\mu_2^3$, $\epsilon = m'_1 m'_2$, we have:

$$\beta_1 = \frac{4(r^2-4\epsilon)(r+1)}{\epsilon(r+2)^2}, \qquad \beta_2 = \frac{3(r+1)(2r^2+\epsilon(r-6))}{\epsilon(r+2)(r+3)}.$$

Thus:

$$\frac{\beta_1(r+2)^2}{4(r+1)} = \frac{r^2}{\epsilon} - 4, \qquad \frac{\beta_2(r+2)(r+3)}{3(r+1)} = \frac{2r^2}{\epsilon} + r - 6,$$

whence, eliminating r^2/ϵ, we find:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{3\beta_1 - 2\beta_2 + 6}.$$

This gives r, then :

$$\epsilon = \frac{r^2}{4 + \frac{1}{4}\beta_1 (r+2)^2/(r+1)},$$

$$b^2 = \frac{\mu_2 r^2 (r+1)}{\epsilon} = \mu_2 \frac{\beta_1 (r+2)^2 + 16 (r+1)}{4},$$

or

$$b = \frac{\sqrt{\mu_2}\{\beta_1 (r+2)^2 + 16 (r+1)\}^{\frac{1}{2}}}{2}.$$

Since

$$r = m'_1 + m'_2, \quad \epsilon = m'_1 m'_2,$$

m'_1 and m'_2 are roots of

$$m'^2 - rm' + \epsilon = 0.$$

Thus $m_1 = m'_1 - 1$ and $m_2 = m'_2 - 1$ are determined.

Further, $a_1 + a_2 = b$, $a_1/a_2 = m_1/m_2$, and $\nu = m_1/a_1$ are all determined.

Lastly :

$$y_0 = \eta\, m_1^{m_1} m_2^{m_2}/(m_1 + m_2)^{m_1 + m_2},$$

and

$$\alpha = b\eta\, \Gamma(m_1 + 1)\, \Gamma(m_2 + 1)/\Gamma(m_1 + m_2 + 2),$$

give :

$$y_0 = \frac{\alpha}{b} \frac{m_1^{m_1} m_2^{m_2}}{(m_1 + m_2)^{m_1 + m_2}} \frac{\Gamma(m_1 + m_2 + 2)}{\Gamma(m_1 + 1)\, \Gamma(m_2 + 1)},$$

which completes the solution,* if a Table of Γ functions is to hand.

Remarks.—It is clear that the solution is *unique*.

It is necessary in order that the solution may be real, that m'_1 and m'_2 should be real or $r^2 > 4\epsilon$. Hence, if ϵ be negative, there is certainly a solution, because r is always real. The solution forms, however, one of the sub-types referred to in our Art. 13, (ii) and (iii).

If ϵ be positive, we must have $r^2/\epsilon - 4$ positive, or

$$\frac{\beta_1 (3 + \beta_2)^2}{(6 + 3\beta_1 - 2\beta_2)(4\beta_2 - 3\beta_1)} > 0.$$

Now it is easy to prove that for any curve $4\beta_2 - 3\beta_1$ or $4\mu_4\mu_2 - 3\mu_3^2$ is positive, for $\mu_4\mu_2$ is always greater than μ_3^2.

Thus, we must have

$$6 + 3\beta_1 - 2\beta_2 > 0,$$

or

$$2\mu_2 (3\mu_2^2 - \mu_4) + 3\mu_3^2 > 0.$$

* Very often with sufficient accuracy we may take :

$$y_0 = \frac{\alpha}{b} \frac{(m_1 + m_2 + 1)\sqrt{(m_1 + m_2)}}{\sqrt{(2\pi m_1 m_2)}}\, e^{\frac{1}{12}\left\{\frac{1}{m_1 + m_2} - \frac{1}{m_1} - \frac{1}{m_2}\right\}}.$$

Now it is theoretically impossible to fit a normal curve ($\mu_4 = 3\mu_2^2$) to a frequency distribution for which $\mu_4 > 3\mu_2^2$. It is, however, possible to fit this generalised curve of Type I., although μ^4 be $> 3\mu_2^2$, provided there is sufficient skewness to render

$$3\mu_3^2 > 2\mu_2(\mu_4 - 3\mu_2^2).$$

Hence the first stage in determining the type of curve suitable for a given set of observations is to ascertain the value of

$$2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2.$$

If this expression be positive, we see that a limited range of variation is a possibility.

Passing from range to *skewness* we remark that the distance d between the centroid vertical and the maximum ordinate

$$\begin{aligned} &= a_1 - \mu'_1 = a_1 - bm'_1/(m'_1 + m'_2), \\ &= \frac{a_1 m'_2 - a_2 m'_1}{m'_1 + m'_2} \\ &= \frac{b(m_1 - m_2)}{(m_1 + m_2)(m_1 + m_2 + 2)}. \end{aligned}$$

Now it might seem that d/b would form a good measure of skewness, and it would be so if all curves had a limited range. But, as they have not, it seems to me better to take as the measure of skewness the ratio of the distance between the maximum ordinate and the centroid to the length of the swing radius of the curve about the centroid vertical, *i.e.*, the quantity $d/\sqrt{\mu_2}$.

In our case we have accordingly,

$$\begin{aligned} \text{skewness} &= \frac{m_1 - m_2}{m_1 + m_2}\sqrt{\left(\frac{m_1 + m_2 + 3}{(m_1 + 1)(m_2 + 1)}\right)}, \\ &= \tfrac{1}{2}\sqrt{\beta_1}\,\frac{r + 2}{r - 2} \end{aligned}$$

in our previous notation.*

Thus range and skewness are determined in Type I.

(15.) A very considerable simplification of the above analysis arises when the range is given by the conditions of the problem itself, *e.g.*, guessing between two given tints. In this we only require the moments μ'_1 and μ'_2 about one end of the range, and the solution becomes as easy as in the case of fitting a normal curve.

Since b, μ'_1 and μ'_2 are known, let

$$\gamma_1 = \mu'_1/b \quad \text{and} \quad \gamma_2 = \mu'_2/(\mu'_1 b).$$

* The points of inflexion of the curve are at distances $\pm\sqrt{a_1 a_2/(m_1 + m_2 - 1)}$ on either side of the maximum ordinate.

Then

$$\gamma_1 = m'_1/(m'_1 + m'_2), \qquad \gamma_2 = \frac{m'_1 + 1}{m'_1 + m'_2 + 1},$$

and we have at once

$$m'_1 = \frac{\gamma_1(\gamma_2 - 1)}{\gamma_1 - \gamma_2}, \qquad m'_2 = \frac{(\gamma_2 - 1)(1 - \gamma_1)}{\gamma_1 - \gamma_2}.$$

Then $a_1/a_2 = (m'_1 - 1)/(m'_2 - 1)$, and $a_1 + a_2 = b$ give a_1 and a_2. Finally y_0 is given as before by

$$y_0 = \frac{\alpha}{b} \frac{m_1^{m_1} m_2^{m_2}}{(m_1 + m_2)^{m_1 + m_2}} \frac{\Gamma(m_1 + m_2 + 2)}{\Gamma(m_1 + 1)\,\Gamma(m_2 + 1)}.$$

(16.) A perhaps still more interesting and usual case arises when one end of the range is given, *i.e.*, when μ'_1, but not b, is known. For example, a curve of distribution of disease with age, the liability to the disease starting with birth. Here we require to calculate from the observations α_1 μ'_1, μ'_2 and μ'_3. The solution is as follows:

Let

$$\mu'_2/\mu'^2_1 = \chi_2, \qquad \mu'_3/(\mu'_2\mu'_1) = \chi_3;$$

then

$$\chi_2 = \frac{(m'_1 + 1)(m'_1 + m'_2)}{m'_1(m'_1 + m'_2 + 1)} = \frac{1 + v}{1 + u},$$

$$\chi_3 = \frac{(m'_1 + 2)(m'_1 + m'_2)}{m'_1(m'_1 + m'_2 + 2)} = \frac{1 + 2v}{1 + 2u},$$

if $v = 1/m'_1$ and $u = 1/(m'_1 + m'_2)$.

Solving

$$u = \frac{1 + \chi_3 - 2\chi_2}{2(\chi_2 - \chi_3)}, \qquad v = \frac{2\chi_3 - \chi_2 - \chi_2\chi_3}{2(\chi_2 - \chi_3)}.$$

Thus,

$$m'_1 = \frac{2(\chi_2 - \chi_3)}{2\chi_3 - \chi_2 - \chi_2\chi_3}, \quad m'_2 = \frac{2(\chi_2 - \chi_3)(\chi_3 - 1)(1 - \chi_2)}{(1 + \chi_3 - 2\chi_2)(2\chi_3 - \chi_2 - \chi_2\chi_3)}.$$

$$b = \mu'_1(m'_1 + m'_2)/m'_1 = \mu'_1 v/u$$

$$= \mu'_1 \frac{2\chi_3 - \chi_2 - \chi_2\chi_3}{2(\chi_2 - \chi_3)},$$

determines the range.

Hence, since

$$\alpha_1 + a_2 = b, \quad \text{and} \quad a_1/a_2 = \frac{m'_1 - 1}{m'_2 - 1},$$

we have, with the aid of the previous expression for y_0, the complete solution of the problem.

(17.) *Generalised probability curve of Type II. Limited Range and Symmetry.*

$$y = y_0 (1 - x^2/a^2)^m.$$

The solution in this case follows very easily from (14) by putting $\beta_1 = 0$, we have at once

$$2(m+1) = r = \frac{6(\beta_2 - 1)}{6 - 2\beta_2},$$

or

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)} = \frac{5\mu_4 - 9\mu_2^2}{2(3\mu_2^2 - \mu_4)}.$$

$$\text{Since } \mu_2 = \frac{b^2\epsilon}{2(r+1)}, \text{ and clearly } \epsilon = r^2/4,$$

we have $$b = 2a = 2\sqrt{\{\mu_2 (r+1)\}},$$

or

$$a = \frac{\sqrt{(2\mu_2\beta_2)}}{\sqrt{(3 - \beta_2)}} = \frac{\sqrt{(2\mu_2\mu_4)}}{\sqrt{(3\mu_2^2 - \mu_4)}}.$$

Finally

$$y_0 = \frac{\alpha}{b} \frac{m^{2m}}{(2m)^{2m}} \frac{\Gamma(2m+2)}{\{\Gamma(m+1)\}^2},$$

$$= \frac{\alpha\sqrt{(3-\beta_2)}}{2\sqrt{(2\mu_2\beta_2)}} \frac{\Gamma(2m+2)}{2^{2m}\{\Gamma(m+1)\}^2},$$

$$= \alpha \sqrt{\frac{3\mu_2^2 - \mu_4}{2\mu_2\mu_4}} \frac{\Gamma(2m+2)}{2^{2m+1}\{\Gamma(m+1)\}^2},$$

$$= \alpha \sqrt{\frac{3\mu_2^2 - \mu_4}{2\mu_2\mu_4}} \frac{\Gamma(m+1{\cdot}5)}{\sqrt{\pi}\,\Gamma(m+1)}.$$

For the normal frequency curve $\mu_4 = 3\mu_2^2$, for a symmetrical point-polygon $3\mu_2^2 > \mu_4$. Hence, whenever a symmetrical frequency curve differs from the normal curve on the side of the point-binomial, we can better the normal solution by taking a symmetrical frequency curve of limited range.

Since

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^{\frac{m}{a^2} \cdot a^2},$$

and

$$\frac{m}{a^2} = \frac{5\beta_2 - 9}{4\mu_2\beta_2} = \frac{1}{2\mu_2},$$

if $\beta_2 = 3$, we easily trace the transition from the limited symmetrical curve to the normal curve with infinite range.

Quite apart from the extremely interesting problem of finding the range, it is clear that better fits will be obtained for symmetrical distributions by the aid of this limited range curve for all cases in which $3\mu_2^2 > \mu_4$.

(18.) *Generalised Probability Curve of the Type III. Range limited in one direction only.*

$$y = y_0 (1 + x/a)^{\gamma a} e^{-\gamma x}.$$

In this case we have no need to determine the value of μ_4, and the analysis is much simplified by the replacement of the B function by a single Γ function.

Take $z = \gamma (a + x)$ and write $\gamma a = p$, we have

$$y = \frac{y_0 e^p}{p^p} z^p e^{-z}.$$

Further, $x = -a, z = 0$, $x = \infty , z = \infty$. Thus we find

$$\alpha \mu_n = \int_{-a}^{\infty} y (x + a)^n \, dx = \frac{y_0 e^p}{p^p \gamma^{n+1}} \int_0^{\infty} z^{p+n} e^{-z} \, dz.$$

Hence

$$\alpha = \frac{y_0 a e^p}{p^{p+1}} \Gamma (p + 1), \quad \alpha \mu'_n = \frac{\alpha \Gamma (p + n + 1)}{\gamma^n \Gamma (p + 1)},$$

whence

$$\mu'_1 = \frac{p + 1}{\gamma}, \quad \mu_2' = \frac{(p + 1)(p + 2)}{\gamma^2},$$

$$\mu_3' = \frac{(p + 1)(p + 2)(p + 3)}{\gamma^3}, \quad \mu_4' = \frac{(p + 1)(p + 2)(p + 3)(p + 4)}{\gamma^4}.$$

Or, transposing to the centroid-vertical, we have

$$\mu_2 = \frac{p + 1}{\gamma^2}, \quad \mu_3 = \frac{2 (p + 1)}{\gamma^3}, \quad \mu_4 = \frac{3(p + 1)(p + 3)}{\gamma^4}.$$

The first two results give us at once

$$\gamma = 2\mu_2/\mu_3, \qquad p = 4\mu_2^3/\mu_3^2 - 1,$$

whence

$$a = \frac{p}{\gamma} = \frac{2\mu_2^2}{\mu_3} - \frac{\mu_3}{2\mu_2}, \quad \text{and} \quad y_0 = \frac{\alpha}{a} \frac{p^{p+1}}{e^p \Gamma (p + 1)}.$$

This completes the solution of the problem, which is seen to require only the determination of μ_2 and μ_3.

Remarks.—The distance d of the centroid-vertical from the axis of y or maximum ordinate d, is given by

$$d = \mu_1 - a = \tfrac{1}{2}\mu_3/\mu_2.$$

Thus

$$\text{skewness} = d/\sqrt{\mu_2} = \tfrac{1}{2}\mu_3/\mu_2^{3/2}.$$

If we transfer the origin to the centroid-vertical we have

$$y = y_1 \left(1 + \frac{x}{2\mu_2^2/\mu_3}\right)^{4\mu_2^3/\mu_3^2 - 1} e^{-2\mu_2 x/\mu_3}$$

where

$$y_1 = \frac{\alpha}{\sqrt{(2\pi\mu_2)}} \frac{\sqrt{\{2\pi(p+1)\}}\, e^{-(p+1)}\, (p+1)^p}{\Gamma(p+1)}.$$

It is interesting to note how this skew curve passes into the normal curve when μ_3 is made vanishingly small, or $p = \infty$.

By WALLIS'S theorem the limit to $y_1 = \alpha/\sqrt{2\pi\mu_2}$.

It remains to find the limit of

$$\left(1 + \frac{x}{2\mu_2^2/\mu_3}\right)^{4\mu_2^3/\mu_3^2 - 1} e^{-2\mu_2 x/\mu_3} = \left(1 + \frac{1}{\sqrt{(p+1)}} \frac{x}{\sqrt{(\mu_2)}}\right)^p e^{-\sqrt{p+1}\, x/\sqrt{\mu_2}}$$

$$= \left[\{(1+u)\, e^{-u}\}^{\frac{1}{u^2}}\right]_{u=0}^{x^2/\mu_2}.$$

Now the limit of $\{(1+u)\, e^{-u}\}^{1/u^2}$ for $u = 0$ is easily found to be $e^{-\frac{1}{2}}$, hence

$$y = \alpha e^{-x^2/2\mu_2}/\sqrt{(2\pi\mu_2)},$$

the normal form.

Returning to the value we have found for μ_4 and eliminating p and γ between μ_2, μ_3, and μ_4 we find

$$2\mu_2 (3\mu_2^2 - \mu_4) + 3\mu_3^2 = 0.$$

This is the expression (see p. 398) which must be positive in the case of limited range. It is zero also for the normal curve, because both $3\mu_2^2 - \mu_4$ and μ_3 vanish. Hence the more nearly the quantity $2\mu_2 (3\mu_2^2 - \mu_4) + 3\mu_3^2$ approaches to zero, the more nearly are we able to fit our statistics with a skew frequency-curve having a range limited in one direction only.

(18 *bis*).—The skew frequency-curve of Type III. deserves especial notice. It is intermediate between those of Type I. and Type IV., and they differ very little from it in appearance. Hence, if the reader has once studied the various forms which Type III. can take as we alter its constants, he will grasp at once the forms taken by Types I. and IV., by simply considering the range doubly limited or doubly unlimited. To assist the process of realising Type III., Plate 9, fig. 5, has been constructed; it contains seven sub-types of this species, varying from fig. I., in which the curve is asymptotic to the maximum frequency-ordinate to fig. VII., which is practically identical with the normal curve. Taking $y = y_0 (1 + x/a)^p e^{-\gamma' x/a}$ for the equation to the curve, we have the following values for the constants p_1 and γ':—

I.	$p = -\cdot 67$	$\gamma' = \cdot 3$
II.	$p = \cdot 001$	$\gamma' = \cdot 2505$
III.	$p = \cdot 265$	$\gamma' = \cdot 363$
IV.	$p = 1\cdot 021$	$\gamma' = \cdot 7676$
V.	$p = 1$	$\gamma' = \cdot 5$
VI.	$p = 6\cdot 5625$	$\gamma' = 4\cdot 3125$
VII.	$p = 1890$	$\gamma' = 1700$

In the diagrams vertical and horizontal scales (y_0 and α) have been chosen so as to illustrate best the changes of shape in the curve. The general correspondence of this series with actual types of frequency curve, as indicated in Plate 7, fig. 1, will at once strike the reader.

The mean, the median, and the mode or maximum-ordinate are marked by *bb*, *cc*, and *aa*, respectively, and as soon as the curves were drawn, a remarkable relation manifested itself between the position of these three quantities: the median, so long as p was positive, was seen to be about one-third from the mean towards the maximum. For p negative and between 0 and -1, this relation was not true. The distance between the maximum-ordinate and the mean is, if the equation to the curve be

$$y = y_0 x^p e^{-\gamma x},$$

equal to $1/\gamma$. Now the maximum cannot be accurately determined from observation, but a fair approximation can be made to the median. Hence the constant γ could, if the above graphical relation were shown to be always true, be determined approximately as the *inverse of thrice the distance between median and mean.*

Now distance of mean from origin $= (p+1)/\gamma$,
and ,, maximum ,, $= p/\gamma$.

Hence, supposing distance of median $= (p+c)/\gamma$, we should expect to find $c = 2/3$ about.

Equating the integral which gives the area up to the median to half the total area, we have

$$y_0 \int_{\frac{p+c}{\gamma}}^{\infty} x^p e^{-\gamma x}\, dx = \tfrac{1}{2} y_0 \int_0^{\infty} x^p e^{-\gamma x}\, dx,$$

or,

$$\int_{p+c}^{\infty} z^p e^{-z}\, dz = \tfrac{1}{2} \int_0^{\infty} z^p e^{-z}\, dz.$$

This is the equation for c. Unable to solve it generally I gave p a series of integer values and found in all cases c nearly $\cdot 67$. Its value, however, decreased as p

increased. I, therefore, assumed c to be really of the form $c = c_1 + c_2/p$, and determining c_1 and c_2 by the method of least squares, found

$$c = \cdot 6691 + \cdot 0094/p.$$

Probably this is only the beginning of a rapidly converging series in inverse powers of p, but it would appear to suffice for most practical purposes. It is only true for $p > 1$ and does not explain why, when p is positive and fractional, c is still apparently near $\frac{2}{3}$; thus its value for $p = 0$ has only risen to $\cdot 6931$. We have then the following fairly simple means of determining *roughly* the constants of a skew curve of this type:

(1.) Find the mean and the median; these gives γ, approximately.

(2.) Find μ_2 for the mean; this gives p, since $\mu_2 = (p + 1)/\gamma^2$.

(3.) Knowing p, correct the value of γ by using the above value for c, and so obtain a corrected p.

(4.) Determine y_0 from the area.

This method is not very laborious and may be of service in some cases.* It will, of course, fail for any curves in which p is negative, and must only be applied when the curve is known to be of Type III. If the beginning of the range is definitely known, we may save stage (2) above and find p from the distance of the mean from the start of the range.

(19.) *Generalised Probability Curve of Type IV. Range unlimited, but form skew.*

$$y = \frac{y_0}{\{1 + (x/a)^2\}^m} e^{-\nu \tan^{-1}(x/a)}.$$

Put $x = a \tan \theta$, hence

$$y = y_0 \cos^{2m} \theta \, e^{-\nu\theta}.$$

$$\alpha\mu'_n = \int_{-\infty}^{+\infty} y x^n \, dx = y_0 a^{n+1} \int_{-\pi/2}^{\pi/2} \cos^{2m-n-2}\theta \sin^n\theta \, e^{-\nu\theta} \, d\theta,$$

$$= y_0 a^{n+1} \int_{-\pi/2}^{\pi/2} \cos^{r-n}\theta \sin^n \theta \, e^{-\nu\theta} \, d\theta, \text{ if } r = 2m - 2,$$

$$= \frac{y_0 a^{n+1}}{r - n + 1} \left\{ (n-1) \int_{-\pi/2}^{\pi/2} \cos^{r-n+2}\theta \sin^{n-2}\theta \, e^{-\nu\theta} \, d\theta - \nu \int_{-\pi/2}^{\pi/2} \cos^{r-n+1}\theta \sin^{n-1}\theta \, e^{-\nu\theta} \, d\theta \right\}$$

$$= \frac{a}{r - n + 1} \left\{ (n-1) a\mu'_{n-2} - \nu\mu'_{n-1} \right\} \alpha,$$

provided $r > n - 1$.

* The points of inflexion may also occasionally be found from the observations; they are at distances $\pm \sqrt{p}/\gamma$ on either side of the maximum ordinate.

Thus, if we know α and μ'_1, we can find the successive μ"s. Now

$$\alpha = y_0 a \int_{-\pi/2}^{\pi/2} \cos^r\theta\, e^{-\nu\theta}\, d\theta,$$
$$= y_0 a e^{-\frac{1}{2}\nu\pi} \int_0^{\pi} \sin^r\theta\, e^{+\nu\theta}\, d\theta,$$

and depends on the integral $\int_0^{\pi} \sin^r\theta\, e^{+\nu\theta} d\theta$, which I propose to write G (r, ν). The result above for μ'_n shows us that the more general integral $\int_0^{\pi} \cos^p\theta \sin^q\theta\, e^{+\nu\theta}\, d\theta$ can always be expressed in terms of G-functions. Further:

$$a\mu'_1 = y_0 a^2 \int_{-\pi/2}^{\pi/2} \cos^{r-1}\theta \sin\theta\, e^{-\nu\theta}\, d\theta,$$
$$= -\frac{y_0 a^2 \nu}{r} \int_{-\pi/2}^{\pi/2} \cos^r\theta\, e^{-\nu\theta}\, d\theta = -\frac{a\nu}{r}\alpha.$$

Thus we find by the formula of reduction above:

$$\mu'_2 = \frac{a^2}{r(r-1)}(r+\nu^2), \quad \mu'_3 = -\frac{a^3\nu}{r(r-1)(r-2)}(3r-2+\nu^2),$$
$$\mu'_4 = \frac{a^4}{r(r-1)(r-2)(r-3)}\{3r(r-2)+\nu^2(6r-8)+\nu^4\}.$$

Referring to centroid vertical, we have:

$$\mu_2 = \frac{a^2}{r^2(r-1)}(r^2+\nu^2), \quad \mu_3 = -\frac{4a^3\nu(r^2+\nu^2)}{r^3(r-1)(r-2)},$$
$$\mu_4 = \frac{3a^4(r^2+\nu^2)\{(r+6)(r^2+\nu^2)-8r^2\}}{r^4(r-1)(r-2)(r-3)}.$$

These may be rewritten, if $z = r^2 + \nu^2$,

$$\mu_2 = \frac{a^2 z}{r^2(r-1)}, \quad \mu_3 = -\frac{4a^3 z\sqrt{(z-r^2)}}{r^3(r-1)(r-2)}$$
$$\mu_4 = \frac{3a^4 z\{(r+6)z-8r^2\}}{r^4(r-1)(r-2)(r-3)}.$$

As before, putting $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$, we have

$$-\frac{\beta_1(r-2)^2}{2(r-1)} = 8\frac{r^2}{z} - 8,$$
$$\frac{\beta_2(r-2)(r-3)}{3(r-1)} = r + 6 - 8\frac{r}{z}.$$

Adding and dividing out by $r-2$, we have

$$r = \frac{6\,(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6},$$

hence

$$m = \tfrac{1}{2}\,(r+2)$$

is known. Further

$$z = \frac{\gamma^2}{1 - \dfrac{\beta_1}{16}\dfrac{(r-2)^2}{r-1}}$$

is known, whence

$$\nu = \sqrt{(z - \gamma^2)}$$

is given.* Finally

$$a = r\sqrt{\left(\frac{\mu_2\,(r-1)}{z}\right)}$$

and

$$y_0 = \frac{\alpha e^{\frac{1}{2}\nu\pi}}{a\int_0^{\pi} \sin^r\theta\, e^{\nu\theta}\, d\theta}$$

completely determine the problem.

Remarks. The solution is clearly *unique*.

(i.) To determine the skewness we must find the position of the ordinate for which $dy/dx = 0$; this is $x_0 = -\nu a/(2m) = -\nu a/(r+2)$.

But

$$d = -\mu'_1 + x_0 = \frac{\nu a}{r} - \frac{\nu a}{r+2} = \frac{2\nu a}{r\,(r+2)}.$$

Hence

$$\text{skewness} = d/\sqrt{\mu_2}$$
$$= \frac{2\nu}{r+2}\sqrt{\left(\frac{r-1}{r^2+\nu^2}\right)} = \tfrac{1}{2}\sqrt{\beta_1}\,\frac{r-2}{r+2} \quad (cf.\ \text{p. } 370).$$

(ii.) We further notice that

$$r - 1 = \frac{4\beta_2 - 3\beta_1}{2\beta_2 - 3\beta_1 - 6}.$$

Hence, since $4\beta_2$ is always $> 3\beta_1$ (see p. 369), it follows, since $r > 1$, that we must have

$$2\beta_2 - 3\beta_1 - 6 > 0,$$

or

$$2\mu_2\,(3\mu_2^2 - \mu_4) + 3\mu_3^2 < 0.$$

* Whether we give ν the $-$ or $+$ sign will depend upon the sign of μ_3 in the actual statistics.

Thus this expression is again critical for the class of curve with which we are dealing. We may say that a skew frequency curve will have limited range, range limited in one direction only, or unlimited range according as

$$2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2$$

is greater than, equal to or less than zero. Thus the calculation of this expression is the first step towards the classification of a frequency curve given by observation.

(iii.) It is noteworthy that the values we have obtained for r, z, a, ν and y_0 will be real and possible if $r > 1$. On the other hand we have required in our work that r should be > 3. I propose now to return to this point. So long as $r > 1$ the values of both μ'_1 and μ_2 will be finite, but the values of μ'_3 and μ'_4 and consequently of μ_3 and μ_4 will be infinite if r be < 3. That is to say, the third and fourth moments of the curve about the centroid vertical become infinite. This is quite conceivable from the geometrical standpoint, and various interesting questions, of purely theoretical value however, arise according as $r > 1$ and < 2, *i.e.*, μ_4 and μ_3 are both infinite, or $r > 2$ and < 3, *i.e.*, μ_4 alone is infinite. The solution we have given fails in these cases. We should obtain, however, finite relations between the four constants of the equation to the curve by taking the first and second moments $\alpha\mu''_1$ and $\alpha\mu''_2$ round the axis of x; we find in this case

$$\alpha\mu''_1 = \tfrac{1}{2}y_0^2\, a \int_{-\pi/2}^{\pi/2} \cos^{2r+2}\theta\, e^{-2\nu\theta}\, d\theta,$$

$$\alpha\mu''_2 = \tfrac{1}{3}y_0^3\, a \int_{-\pi/2}^{\pi/2} \cos^{3r+4}\theta\, e^{-3\nu\theta}\, d\theta,$$

or,

$$\mu''_1 = \tfrac{1}{2}y_0\, e^{-\frac{1}{2}\nu\pi}\, \mathrm{G}(2r+2,\, 2\nu)/\mathrm{G}(r, \nu),$$

$$\mu''_2 = \tfrac{1}{3}y_0^2\, e^{-\nu\pi}\, \mathrm{G}(3r+4,\, 3\nu)/\mathrm{G}(r, \nu).$$

These results together with

$$\mu_2 = \frac{a^2(r^2+\nu^2)}{r^2(r-1)}, \quad \alpha = y_0 a e^{-\frac{1}{2}\nu\pi}\, \mathrm{G}(r, \nu),$$

are *theoretically* sufficient to determine the four constants r, ν, y_0 and a. *Practically* they would hardly be of service without very elaborate tables of the G functions.

As a matter of fact, we are very unlikely in dealing with actual statistics to meet with cases in which μ_3 and μ_4 become infinite, because neither the range of observations, nor the size of the groups observed at great distances from the origin can be infinite. With finite values of μ_3 and μ_4, it is, however, easy to see that we always obtain from our solution on page 377 a value of $r > 3$, so that the solution is self-consistent.

(iv.) It remains to say a few words about the integral

$$\mathrm{G}(r, \nu) = \int_0^\pi \sin^r\theta \, e^{\nu\theta} \, d\theta.$$

Provided $r > 1$, we find a formula of reduction

$$\mathrm{G}(r, \nu) = \frac{r(r-1)}{r^2 + \nu^2} \mathrm{G}(r-2, \nu).$$

Thus the value of the integral from $r = 0$ to $r = 2$ only will be required for diverse values of ν. The integral does not yet appear to have been studied at length or tabulated. Dr. A. R. FORSYTH* has kindly answered my inquiry for a fairly easy method of reducing G (r, ν) for purposes of calculation, by sending me the formula

$$\mathrm{G}(r, \nu) = \frac{2^{-r}\pi e^{\frac{1}{2}\pi\nu}\,\Pi(r)}{\Pi(\frac{1}{2}r - \frac{1}{2}\nu i)\,\Pi(\frac{1}{2}r + \frac{1}{2}\nu i)},$$

where Π is GAUSS'S function such that

$$\Pi(n) = \Gamma(n+1).$$

Taking as definition of Π that

$$\Pi(z) = \text{limit of } \frac{1.2\ldots n}{(z+1)(z+2)\ldots(z+n)} n^z$$

when n is infinite, we can reduce the above expression to the form

$$\mathrm{G}(r, \nu) = \frac{2^{-\nu}\pi e^{\frac{1}{2}\pi\nu}\,\Gamma(r+1)}{\text{Product}_{n=1}^{n=\infty}\left(1 + \frac{\nu^2 + r^2}{4n^2(1 + r/n)}\right)}.$$

Here, since r can always be supposed to lie between 0 and 2, when ν is small a few terms of the product will generally suffice for the calculation of G (r, ν) to the degree of accuracy required in statistical practice.

On the other hand when r is large, *i.e.*, generally in cases of slight skewness, I find if $\tan\phi = \nu/r$

$$\Pi(\tfrac{1}{2}r - \tfrac{1}{2}\nu i)\,\Pi(\tfrac{1}{2}r + \tfrac{1}{2}\nu i) = \frac{\pi r}{\cos\phi} e^{-r}\left(\frac{\nu}{2\cos\phi}\right)^r e^{\frac{\cos^2\phi}{3r} - \phi r\tan\phi}$$

very nearly.

Hence

$$y_0 = \frac{\alpha}{a}\sqrt{\frac{r}{2\pi}}\,\frac{e^{\frac{\cos^2\phi}{3r} - \frac{1}{12r} - \phi r\tan\phi}}{(\cos\phi)^{r+1}}$$

very nearly.

* "Evaluation of two Definite Integrals," 'Quarterly Journal of Mathematics,' January, 1895.

(20.) We have now considered methods for fully investigating whether a given system of measurements has a limited range, and for ascertaining the degree of skewness of the system.

Analytically, our work may be expressed as follows:—

The slope of the normal curve is given by a relation of the form

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x}{c_1}.$$

The slope of the curve correlated to the skew binomial as the normal curve to the symmetrical binomial is given by a relation of the form

$$\frac{1}{y}\frac{dy}{dx} = \frac{-x}{c_1 + c_2 x}.$$

Finally, the slope of the curve correlated to the hypergeometrical series (which expresses a probability distribution in which the "contributory causes" are not independent, and not equally likely to give equal deviations in excess and defect) as the above curves to their respective binomials is given by a relation of the form

$$\frac{1}{y}\frac{dy}{dx} = \frac{-x}{c_1 + c_2 x + c_3 x^2}.$$

This latter curve comprises the other two as special cases, and so far as my investigations have yet gone practically covers all *homogeneous* statistics that I have had to deal with. Something still more general may be conceivable, but I have hitherto found no necessity for it.

To demonstrate its fitness and the importance of these generalised frequency distributions for various problems in physics, economics, and biology, I have devoted the remainder of this paper to the consideration of special cases of actual statistics.

Part II.—Statistical Examples.

(21.) Quetelet, who often foreshadowed statistical advances without perceiving the method by which they might be scientifically dealt with, has treated of the subject of limits in Lettre XXII of his "Lettres sur la Théorie des Probabilités" (1846). He seems to have been conscious that certain variations in excess or defect might biologically or physically be impossible, and he accordingly introduces the terms *Limites extraordinaires en plus et en moins* to mark the range of possible variation. He makes no attempt to show how this range may be found from a given set of statistics.

"Lorsqu'on suppose le nombre des observations infini, ou peut porter les écarts à des

distances également infinies de la moyenne, et trouver toujours des probabilités qui y correspondent. Cette conception mathématique ne peut évidemment s'accorder avec ce qui est dans la nature. . . . Les limites extraordinaires au delà desquelles se trouvent les *monstruosités*, me semblent plus difficiles à fixer."

Indeed QUETELET's attempt to fix these limits in the case of the height of human beings at 2·801 and ·433 metres is purely empirical, and scientifically worthless.

I propose in this the first section of the practical part of this paper to consider how far the theory we have developed in the first part, enables us to find the range in various groups of physical and biological phenomena.

Example I. The Range of the Barometer.—The following results for the curve of barometric heights are given on p. 352.

$$\alpha = 171{\cdot}6 \qquad \mu_2 = 10{\cdot}14$$
$$\mu_0 = 15{\cdot}95 \qquad \mu_4 = 326{\cdot}34.$$

We have accordingly :

$$2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2 = 400{\cdot}581,$$

that is, this expression is positive, and we have a limited range.

We have further : $\beta_1 = {\cdot}24401$, $\beta_2 = 3{\cdot}17391$.

Hence, determining the constants in the manner described in §14, we have :

$$r = 30{\cdot}1382 \qquad \epsilon = 150{\cdot}7954$$
$$b = 43{\cdot}61016,$$
$$m_1 = 5{\cdot}3352 \qquad a_1 = 8{\cdot}2688$$
$$m_2 = 22{\cdot}8030 \qquad a_2 = 35{\cdot}3414.$$

Next to find d, giving the distances of the centroid from the origin, or the distance on barometer between mean and maximum, we have by p. 370

$$d = -{\cdot}8983.$$

Thus

Range of barometer above mean = 9·1671
" " below " = 34·4431.

Now, in the scale upon which our curve is drawn in Plate 10, fig. 6, each centimetre equals $\frac{1}{10}$ inch, and the mean barometer in Dr. VENN's results equals about 29″·931. Thus the *maximum possible* = 30″·85 and the *minimum possible* = 26″·49 ; the range of the barometer being about 4″·36. Now, the highest barometer in Dr. VENN's record = 30″·7, and the lowest 28″·7 ; it is clear, therefore, that we reach much nearer in

practice to the upper than to the lower limit of the barometric range.* The result here obtained for the barometric range is of course only tentative and approximate. Far larger statistics must be dealt with, and for a greater variety of places, we shall then be better able to judge how far the range, as ascertained from Dr. VENN'S statistics, is local, or if general, what modification or correction may be required.

Calculating the value of y_0, we find for the curve of barometric heights:

$$y = 21{\cdot}642\,(1 + x/8{\cdot}2688)^{5{\cdot}3352}\,(1 - x/35{\cdot}3414)^{22{\cdot}8030}.$$

This curve is traced on Plate 10, fig. 6. It will be seen to be extremely close to the observations.

Although the expression $2\mu_2\,(3\mu_2^2 - \mu_4) + 3\mu_3^2$ is not zero, it is interesting to see with what closeness the skew curve which is the limit to a point binomial can be fitted to the barometric observations. This is the curve of Type III. Calculating its constants by aid of § 18, we find

$$y = 22\,(1 + x/12{\cdot}1063)^{15{\cdot}393}\,e^{-1{\cdot}2715x},$$

while d, the distance between the maximum ordinate and the centroid-vertical, $= {\cdot}7864$. This gives a maximum possible height of the barometer of $31''{\cdot}22$ instead of $30''{\cdot}85$, there being of course no lower limit. The curve is shown in Plate 10, fig. 6, and will be seen to give a very close correspondence with the observations. The "skewness" of barometric results as given by the curve with limited range $= {\cdot}8983/3{\cdot}184 = {\cdot}2821$, and as given by the curve of Type III. $= {\cdot}7864/3{\cdot}184 = {\cdot}2470$,—no very great difference.

The areal deviations of the two curves are almost exactly the same, being about 7·1 sq. centims. or percentage error of 4·1. The normal curve is also drawn on the same plate. It diverges widely from the observations, the areal deviation $=26$ sq. centims. or the percentage error 15·1,—about 3·7 times as great as in the case of either skew probability curve.

Till a wider range of barometric observations have been analysed, it may be wiser not to draw too definite conclusions from the above results, contenting ourselves with the remark that the new skew curve gives far better results than the old normal curve of errors.

* I am unaware if Dr. VENN'S results are reduced to sea-level. The lowest recorded barometric height for the British Isles reduced to sea-level is $27''{\cdot}333$ (at Ochertyre, Perthshire, January 26, 1884) and the highest (at Roche's Point, Cork, February 20, 1882) is $30''{\cdot}93$. A statement that the barometer stood at $31''{\cdot}046$ at Gordon Castle, in January, 1820, has hardly sufficient evidence. Supposing Dr. VENN'S statistics to be unreduced Cambridge statistics, the expression theoretically found for the barometric range seems to be on the whole satisfactory. I have at present in hand other series of barometric heights.

Example II. Professor WELDON'S *Crab Measurements No.* 4. The details of these are given in 'Phil. Trans.,' vol. 185, p. 96.

We have

$$\alpha = 999, \qquad \mu_2 = 7{\cdot}6759, \qquad \mu_3 = 3{\cdot}4751,$$
$$\mu_4 = 184{\cdot}3039, \qquad \beta_1 = {\cdot}0267022, \qquad \beta_2 = 3{\cdot}12807.$$

In this case

$$2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2 = \mu_2^3(6 + 3\beta_1 - 2\beta_2) = -\mu_2^3 \times {\cdot}1760334,$$

and is accordingly *negative*. In *Example I.* of the barometric heights we had

$$2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2 = \mu_2^3 \times {\cdot}38421.$$

Since, in the latter case, this value was sufficiently small to give a good curve of Type III., we may expect the like result in this case. There is, indeed, a slight but sensible skewness even in this the most symmetrical of all Professor WELDON'S crab measurements, and the skew curve of Type III. is really a better fit than the normal curve. But clearly since the critical function is *negative*, we are dealing properly with a case of a curve of Type IV. The ratio of the organs dealt with in No. 4 series of measurements does not give a "limited range" of variation. Proceeding by the method indicated in § 19, we find for the constants

$$r = 71{\cdot}624, \qquad m = 36{\cdot}812, \qquad \nu = 25{\cdot}7616,$$
$$a = 21{\cdot}909, \qquad \mu'_1 = -\ 7{\cdot}8802,$$
$$d = {\cdot}21407, \qquad \text{Skewness} = {\cdot}077267, \qquad y_0 = 1{\cdot}75509.*$$

Thus the equation to the curve is :

$$y = 1{\cdot}75509\, \frac{e^{-25{\cdot}7616 \tan^{-1}(x/21{\cdot}909)}}{[1 + x^2/(21{\cdot}909)^2]^{36{\cdot}812}}$$

To trace the curve, take :

$$x = 21{\cdot}909 \tan\theta,$$
$$y = 1{\cdot}75509 \cos^{73{\cdot}624}\theta\, e^{-25{\cdot}7616\theta}.$$

If we take a skew curve of Type III., we find for its equation :

$$y = 144{\cdot}22\,(1 + x/33{\cdot}683)^{148{\cdot}8}\, e^{-4{\cdot}41766x},$$

where, for the centroid

$$d = {\cdot}226364,$$

and the skewness

$$= {\cdot}081704.$$

For the normal curve we have :

$$y = 143{\cdot}85\, e^{-x^2/(2{\cdot}77054)^2}.$$

* y_0 was calculated by aid of the approximate formula on p. 380.

All three curves are drawn in fig. 4 of Plate 8. It will be seen that they are all very close to the observations. So far as skewness is concerned, curves of Types III. and IV. give practically the same result (·082 and ·077); in both cases the skewness is small. The areal deviations are in the three cases respectively: 4·4 sq. centims., 5·9 sq. centims., and 6·7 sq. centims., or we have mean percentage errors in frequency of 4·4, 5·9, and 6·7 nearly; the percentage error for the closest point binomial is 10·5. We thus conclude that even in a case which has been selected as the most typically symmetrical series of measurements out of a very considerable set of careful statistics, the generalised probability curve is one-third as good again as the normal curve, while the special case of that generalised probability curve—which is not the most appropriate to our observations—is itself distinctly better than the normal curve. This result has been confirmed by a considerable application of these generalised curves; in good cases of normal curve fitting, the generalised curves are always sensibly better; in cases where the normal curve is almost useless, as in the case of barometric observations, the new curve, *if of the appropriate type*, will represent with a 4 to 5 per cent. mean accuracy many observations not yet reduced to statistical theory. It is, perhaps, unnecessary to repeat that this mean percentage is much less than the average of what has been allowed to pass muster hitherto in both physical and biological measurements. Professor EDGEWORTH'S view* thus seems untenable; a curve with a comparatively easy theory of its constants has been found which excels the accuracy of the hitherto adopted normal curve. And this for the simple reason that it would pass into the normal curve, if that curve were itself the best fit.

23. *Example III.*—The following statistics of *height* for 25,878 recruits in the United States Army, are given by J. H. BAXTER, 'Medical Statistics of the Provost-Marshal-General's Bureau,' vol. 1, Plate 80, 1875.

78″–77″	2	64″–63″	1947
77–76	6	63–62	1237
76–75	9	62–61	526
75–74	42	61–60	50
74–73	118	60–59	15
73–72	343	59–58	10
72–71	680	58–57	6
71–70	1485	57–56	7
70–69	2075	56–55	3
69–68	3133	55–54	1
68–67	3631	54–53	2
67–66	4054	53–52	1
66–65	3475	52–51	1
65–64	3019		

* 'Phil. Mag.,' vol. 24, p. 334, 1887.

I find :

Mean height = 67″·2989.
Standard deviation = 2″·5848.
Maximum ordinate, 3994·04.

This gives a very close-fitting normal curve.
The data for a generalised curve are

$$\mu_2 = 6{\cdot}68122 \qquad \beta_1 = {\cdot}005769$$
$$\mu_3 = -1{\cdot}31168 \qquad \beta_2 = 3{\cdot}024801.$$
$$\mu_4 = 135{\cdot}02324$$

Thus,

$$2\beta_2 - 3\beta_1 - 6 = {\cdot}032295,$$

and being positive, we see the curve belongs to Type IV. There is, thus, exactly as in the previous examples of crab measurements, no range of a limited character for these statistics of height.* For a true normal curve, β_1, β_2 ought to be 0 and 3 respectively; we have therefore a still closer approach (3·025) than in the case of the crabs (3·128) to normality. In this case r is about 400, and on any reasonable scale, there is no sensible difference between the normal and the generalised curves. The skewness is very slight, = ·038 about, or about half its value in the case of the crabs.

24. *Example IV.—Height of* 2192 *St. Louis School Girls, aged* 8.—The following statistics are given by W. T. Porter, "The Growth of St. Louis Children," 'Trans. of Acad. of Sci. of St. Louis,' vol. 6, p. 279, 1894.

Heights at intervals of 2 centims.	Number.	Heights at intervals of 2 centims.	Number.
centims.		centims.	
141 and 142	1	119 and 120	342
139 ,, 140	0	117 ,, 118	321
137 ,, 138	1	115 ,, 116	297
135 ,, 136	5	113 ,, 114	222
133 ,, 134	10	111 ,, 112	137
131 ,, 132	21	109 ,, 110	84
129 ,, 130	28	107 ,, 108	42
127 ,, 128	79	105 ,, 106	27
125 ,, 126	138	103 ,, 104	8
123 ,, 124	183	101 ,, 102	2
121 ,, 122	243	99 ,, 100	1

The following are the calculated values of the constants† :—

* If, notwithstanding, we take a curve of Type III., we find the range limited on the 'dwarf' side at about ·7645″.

† The unit of all these constants = 2 centims., except in the case of the mean height. The standard deviation = 5·55244 centims., which gives a probable deviation of 3·745 centims. The mean

$\mu_2 =$ 7·70739, Mean height = 118·271 centims.,
$\mu_3 = -$ 2·38064, Standard deviation = 2·77622,
$\mu_4 =$ 192·17419, y_0 for normal curve = 314·99,
$\beta_1 =$ ·0123784, $\beta_2 = 3{\cdot}235045$.

Thus $2\beta_2 - 3\beta_1 - 6$ is positive, and the curve is again of Type IV.

We have

$d =$ ·135606, Skewness = ·04885,
$r = 30{\cdot}8023$, $m = 16{\cdot}4011$,
$\nu =$ 4·56967, $a = 14{\cdot}9917$,
$y_0 = 235{\cdot}323$,

or, for the equation to the curve :—

$$x = 14{\cdot}9917 \tan\theta,$$
$$y = 235{\cdot}323 \cos^{32{\cdot}8023}\theta \, e^{-4{\cdot}56967\theta},$$

the axis of x being positive towards dwarfs and the origin 2·2241 on the positive side of the centroid-vertical.

The maximum ordinate = 324·18 and occurs at $x = -$ 2·0884.

The curve of Type IV., together with the normal curve, is drawn (Plate 10, fig. 7).

If we attempt to fit a curve of Type III., we find p about 322·14, and the range limited on the dwarf side at about 99·812 centims. from the mean, or at a height of about 18·5 centims. The largeness of p causes this curve to coincide with the normal curve to the scale of our diagram. The areal deviations are for the curve of Type IV. and for the normal curve 6·1 and 8·3 centims., giving percentage mean errors of 5·56 and 7·66 in the ordinates respectively. The advantage is again on the side of the generalised curve. It will be seen at once that the normal curve by no means well represents the number of girls of giant height. The theoretical probability that these giants should occur is small, and their actual redundancy over the numbers indicated by the normal curve suggests some peculiarity in this direction ; it is fully met by the curve of Type IV. The asymmetry of the curves given by anthropometrical measurements on children has been noted both by Bowditch* and Porter,† but in their published papers, to which I have had access, they do not give their raw material, only the ogive curve arising from Galton's method of percentiles. Unfortunately, theoretical evaluation of the skewness of anthropometric statistics can only be applied or verified when we have raw material, and not integral frequency

height and probable deviation, as given by Mr. Porter, are 118·36 and 3·698. The latter is obtained from the mean deviation, but I do not know how the former is to be accounted for.

* 'Growth of Children, studied by Galton's Method of Percentiles.' Boston, 1891, p. 496.

† 'Growth of St. Louis Children.' St. Louis, 1894, p. 299.

curves, the integral of the frequency in all suggested forms of the frequency curve being not expressible in terms of undetermined constants. Valuable as is the method of percentiles for representing popularly the numerical facts of anthropometry, it is to be regretted that percentile statistics are replacing the raw material in so many publications. The raw material of Professor WELDON'S crab-measurements and BOWDITCH and PORTER'S child-measurements ought to be preserved and circulated in print, as a means of developing and testing statistical theory.

(25.) *Example V. Length-Breadth Index of* 900 *Bavarian Skulls.*—The following statistics are taken from Tables I.–VI., VIII.–X., inclusive, of J. RANKE'S 'Beiträge zur physischen Anthropologie der Baiern, München, 1883.' They include all the material, which may be treated as typically "Alt-Baierisch," both male and female skulls.

Index.	Frequency.	Index.	Frequency.	Index.	Frequency.
70	1	80	71·5	90	10
71	1	81	82	91	8
72	0	82	116	92	3
73	2·5*	83	98	93	1·5
74	1·5	84	107	94	2
75	3·5	85	82	95	1·5
76	12·5	86	74	96	0
77	17	87	58	97	0
78	37	88	34·5	98	1
79	55	89	19	99	0

We find, as before,

Position of centroid-vertical, 83·07111,

$\sigma =$ 3·468, $y_0 =$ 103·532 (for normal curve),
$\mu_2 =$ 12·027166, $\beta_1 =$ ·0078995,
$\mu_3 =$ 3·707179, $\beta_2 =$ 3·649553,
$\mu_4 =$ 527·91696, $r =$ 12·42734,
$d =$ ·111388, Skewness = ·0321186,
$m =$ 7·21367, $\nu =$ ·853,771, $a =$ 11·69583, $y_0 =$ 107·4706.

Thus we see that the curve is again of Type IV. This result seems of considerable significance, but it requires, of course, wider examination of cases than I have yet been able to make. But, so far as I have gone, in both anthropometric and biological statistics, whether relative or absolute measurements of organs, the frequency curves all deviate from the normal curve—however slight the deviation—in the direction of Type IV. That is to say, the distribution of chances upon which the frequency of variation of an organ depends, appears to resemble the drawing of a

* Indices such as 73·5 have been divided between 73 and 74 groups.

limited amount from a limited mixture. So far as this goes, it is evidence against the usual hypothesis that in biological matters the chances of deviations on either side of the mean are equal, and the "contributory causes" independent and indefinitely great in number. Thus we appear in biological statistics to be dealing with a chance system corresponding, not to a binomial, but to a hypergeometrical series, such as that discussed in § 11.

If it be remarked that Type IV. dismisses at once the problem of *range* from biological investigations, we must notice that, while this is theoretically correct so long as we are dealing with the *continuous curve* by which we replace the hypergeometrical series, it is not true the moment we fall back from the curve on the point series (see p. 361). If the r of that page (or the qn) be an integer, the series is limited in range. It seems very possible that discreteness, rather than continuity, is characteristic of the ultimate elements of variation; in other words, if we replaced the curve by a discrete series of points, we should find a limited range. It is the analytical transition from this series to a closely fitting curve which replaces the limited by an unlimited range. Exactly the same transition occurs when we pass from the symmetrical point binomial to the normal curve. Thus, while Type I. marks an absolutely limited range, the occurrence of Type IV. does not necessarily mean that the range is actually unlimited.*

For the equation to the curve we have

$$x = 11{\cdot}69583 \tan \theta,$$
$$y = 107{\cdot}4706 \cos^{14{\cdot}42734}\theta \, e^{-{\cdot}8537710},$$

the origin being at a distance ·803515 on the *positive* side of the centroid vertical.

The normal curve as well as the curve of Type IV. are shown (Plate 11, fig. 8). The result in both cases is quite good for this type of statistics—*i.e.*, the skulls came from eight different districts and include 100 female skulls. With the planimeter the areal deviation *in both cases* = 6·8 square centims., giving in either case an average percentage error of 7·56. That the generalised curve does not in this case give a decidedly better result than the normal curve I attribute to the heterogeneity of the material. It clearly accounts better for the extreme dolichocephalic and brachycephalic skulls than the normal curve. The same 900 skulls have been fitted with a normal curve by STIEDA,† but neither the constants of his normal distribution nor

* I reserve for the present the fitting of hypergeometrical point series to statistical results. The discussion is related to curves of Type IV., as the fitting of point binomials to curves of Type III. It will, I think, throw considerable light on the nature of chance in the field of biological variation, especially with regard to limitation of the material to be drawn upon, to which I referred above, and which, I believe, finds confirmation in skull statistics.

† "Ueber die Anwendung der Wahrscheinlichkeitsrechnung in der anthropologischen Statistik," 'Archiv für Anthropologie,' Bd. 14. Braunschweig, 1882.

his plotting of RANKE's observations agree with mine. He has added together under 83, for example, all indices from 83 to 83·9. Thus, for the indices 81, 82, 83, 84 he gives the frequencies 106, 92, 111, 99, while I find 82, 116, 98, 107, a very sensible difference.* STIEDA's method can introduce very sensible errors. In this particular case it transfers the maximum frequency of observation from 82 to 84.

The last four examples have dealt with cases where the statistician has hitherto been content to assume symmetry. They have been given to indicate (i.) an apparently uniform trend in biological statistics of variation, and (ii.) the improved fitting of theory to practice which arises from using the generalised curve. I now pass to cases of obvious skewness, where the statistician has hitherto had no satisfactory theory.

(26.) *Example VI. Distribution of* 8689 *Cases of Enteric Fever Received into the Metropolitan Asylums Board Fever Hospitals*, 1871–93.

Age.	Number of cases.	Age.	Number of cases.
Under 5	266	35–40	299
5–10	1143	40–45	163
10–15	2019	45–50	98
15–20	1955	50–55	40
20–25	1319	55–60	14
25–30	857	Above 60	13
30–35	503		

I considered that the 13 cases "above 60" might be distributed as follows: 60–65, 8; 65–70, 4; 70–75, 1.

Taking five years as the unit I found

$$\mu_2 = 4{\cdot}070554, \qquad \mu_3 = 7{\cdot}598196, \qquad \mu_4 = 69{\cdot}379605.$$

The centroid-vertical is at 18·9691 years, *i.e.*, ·29382 unit from 15–20.

Thus $2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2 = 13{\cdot}05102$, or the curve is of Type I. Since, however, $3\beta_1 - 2\beta_2 + 6 = {\cdot}1935$ is small, a curve of Type III. will also be a good fit.

We have for the other constants

$$r = 72{\cdot}28642, \qquad d = {\cdot}98643,$$
$$\epsilon = 259{\cdot}78912, \qquad \text{Skewness} = {\cdot}488922,$$
$$b = 77{\cdot}28312,$$
$$m_1 = 2{\cdot}79291, \qquad m_2 = 67{\cdot}49351,$$
$$a_1 = 3{\cdot}07801, \qquad a_2 = 74{\cdot}20511,$$
$$y_0 = 1890{\cdot}83.$$

* I class as 83 all from 82·6 to 83·4, dividing 82·5 between 82 and 83 evenly, and 83·5 between 83 and 84 evenly. Thus in the Table above certain frequencies will be found with such values as 12·5 or 71·5 skulls.

Thus we have for the curve of Type I.

$$y = 1890{\cdot}83\left(1 + \frac{x}{3{\cdot}07801}\right)^{2{\cdot}79291}\left(1 - \frac{x}{74{\cdot}20511}\right)^{67{\cdot}49351},$$

where the centroid is ·98643 unit from axis of y.

The curve of Type III. is

$$y = 1894{\cdot}57\left(1 + \frac{x}{3{\cdot}428094}\right)^{3{\cdot}673042} e^{-1{\cdot}071453x}.$$

The centroid is in this case ·933313 unit on the positive side of the origin and the skewness = ·462594.

It will be noticed that the curve of Type I. extends ·2706 unit or 1·353 years, and the curve of Type III. ·5676 unit or 2·838 years before birth. In both cases the chances of an "antenatal" death from enteric fever are very, very small. Curve of Type I. is in this respect better than the curve of Type III. The latter curve gives no maximum limit, the former a limit of about 77 units or 385 years. In both cases, however, the chances of a case of enteric fever with the subject over 100 years are vanishingly small. These statistics of enteric fever thus set a maximum limit to the duration of life, but it is a limit so high as to have little suggestiveness.

In order to see what is the nature of the difference made, when we suppose the liability to enteric fever to commence *with birth*, I will treat these statistics as a case falling under § 16.

If then μ'_1, μ'_2, and μ'_3 be the first three moments about the vertical through 0 years we have

$\mu'_1 =$ 3·79382,	$\mu'_2 =$ 18·46362,
$\mu'_3 =$ 108·53175,	
$\chi_2 =$ 1·282813,	$\chi_3 =$ 1·549399,
$u =$ ·030435,	$v =$ ·321856,
$m_1 =$ 2·14296,	$m_2 =$ 28·71414,
$b =$ 40·1206,	$y_0 =$ 1873·39,
$a_1 =$ 2·78629,	$a_2 =$ 37·33431.

whence we have for the curve

$$y = 1873{\cdot}39\left(1 + \frac{x}{2{\cdot}78629}\right)^{2{\cdot}14296}\left(1 - \frac{x}{37{\cdot}33431}\right)^{28{\cdot}71414}.$$

Here the duration of life is 200 years about, and the maximum incidence of the disease is at 13·93 years.

Lastly for the normal curve, we have the constants $\sigma = 2{\cdot}01756$ units = 10·0878 years and $y_0 = 1718{\cdot}12$.

All the above four curves are drawn, Plate 12, fig. 9.

We see at once that the normal curve is perfectly incapable of expressing statistical results like these. It gives an average error in the ordinate of 25·8 per cent. and no less than 260 antenatal deaths!

For the remaining three curves we have the following results:—

	Percentage error in ordinate.	Antenatal cases.
Curve of Type I. (closest fit) . . .	5·75	3
,, ,, (starting at birth) .	7·3	0
Curve of Type III.	5·98	9

The percentage errors here are well within those usually passed by statisticians. If they are slightly larger than what we have found in previous cases the source of the error is not far to seek. We have combined both male and female cases, but the distributions of enteric fever for both sexes are not the same. The fever curves for either sex differ in some cases markedly, although less for enteric fever than for diphtheria, for example. We have thus, in reality, a compound curve. I have found for about 700 male cases only a percentage error of about 5.*

Another point needing notice is the question of antenatal cases, which may at first strike the reader as absurd. The closest fitting curve of Type I. runs, as we have seen, 1·35 years about before birth, and gives three antenatal cases. Three antenatal cases (or, indeed, 9 in the case of the curve of Type III.) is a very small percentage of 8689 cases, and not of importance from the statistician's standpoint. But the fact that a curve starting *before birth* gives a better fit than one starting at birth, is significant, and there is every probability that a curve starting from about — ·75 year would give a still less percentage error than one from 1·35 year or from birth.†

In dealing with mortality curves for infancy I have found it impossible to get good fitting theoretical curves, without carrying these curves backward to a limit of something less than a year. The "theoretical" statistics thus obtained of antenatal deaths seem to be fairly well in accordance with the actual statistics of maternity charities. In vital statistics therefore we must be prepared in most diseases for small percentages of antenatal cases and antenatal deaths, and it is just possible that theory in this matter will be able to indicate lines of profitable inquiry to the medical statistician.

(27.) *Example VII.*—As an example of the method of Section 15, I take the following statistics of guessing a tint. Nine mixtures of black and white were taken,

* I propose on another occasion to deal with the age distribution of fever cases. My object at present is only to give typical illustrations of the method of calculating skew curves.

† In fact the case of a pregnant woman with enteric fever is to be considered as a case also of antenatal enteric fever.

so as to get a series of tints in arithmetical progression 1, 2, 3, 4, 5, 6, 7, 8, and 9. These tints were then placed in non-consecutive order, and 231 persons asked to guess a tint by affixed letters lying between 1 and 9. The results were as follows :—

Tint.	Frequency of guess.	Tint.	Frequency of guess.
1	0	6	54
2	8	7	94
3	7	8	40
4	6	9	0
5	22		

Now, obviously, the number of tints and the number of persons guessing were far too limited to draw any definite conclusions as to the distribution of tint guesses.* I propose here merely to use these statistics to illustrate the calculation of a skew frequency curve with a given limited range. I do not wish to propound any theory of tint guessing, nor to assert that these guesses actually distribute themselves according to the curves dealt with in this paper.

Calculating the moments about the centroid in the usual manner, we have

$$\left.\begin{aligned} \mu_2 &= \quad 2{\cdot}1417 \\ \mu_3 &= -\ 3{\cdot}70067 \\ \mu_4 &= \quad 19{\cdot}6255 \end{aligned}\right\}$$ Centroid lies at a distance of 5·376624 units from the tint 1.

We easily find $2\mu_2(3\mu_2^2 - \mu_4) + 3\mu_3^2 = 15{\cdot}96335$, or the observations fall into a curve of Type I., that is to say, have a *limited range*.

We obtain

$$\beta_1 = 1{\cdot}39407, \qquad \beta_2 = 4{\cdot}27862,$$
$$r = 6{\cdot}95847, \qquad \epsilon = 6{\cdot}443186.$$

hence the range

$$b = 11{\cdot}31768.$$

Further

$$m_1 = 4{\cdot}858705, \qquad m_2 = {\cdot}099765,$$
$$a_1 = 11{\cdot}08997, \qquad a_2 = {\cdot}22769,$$
$$d = 1{\cdot}561012, \qquad \text{Skewness} = 1{\cdot}06666.$$

Thus the range of the theoretical curve runs from a point 4·15233 units before tint 1, and concludes at a point ·734674 unit before tint 9. The curve is, however,

* I hope later to deal with the subject of tint guesses falling within a limited range, as my material increases in bulk. I would only note here, that the geometrical mean frequency curve does not seem to give results according well with experiment.

practically insensible before tint 1. Considering the roughness of the experimental method, the obtaining an actual range of about 11 instead of 8, and its covering very nearly the range of 8 must be held to be fairly encouraging for the method. I shall accordingly calculate the constants of the curve on the assumption that the range lies between Tints 1 and 9, using the method of § 15.

We find

$$\mu'_1 = 2{\cdot}623376, \quad \mu'_2 = 9{\cdot}023803,$$
$$\gamma_1 = \ {\cdot}327922, \quad \gamma_2 = \ {\cdot}429971.$$

Whence

$$m_2 = 2{\cdot}75412, \quad m_1 = \ {\cdot}83172,$$
$$a_2 = 6{\cdot}144435, \quad a_1 = 1{\cdot}855565,$$

and

$$y_0 = 59{\cdot}5996.$$

Thus we may take for the curve

$$y = 59{\cdot}6\left(1 + \frac{x}{6{\cdot}144435}\right)^{2{\cdot}75412}\left(1 - \frac{x}{1{\cdot}855565}\right)^{{\cdot}83172}.$$

The curve is figured, Plate 11, fig. 10, with the first "smooth" of the observations. It will be seen to give the general character of the distribution, but much more elaborate experiments would be required before any statement could be made as to whether frequency of tint guesses really does follow a curve with limited range of Type I. On the same plate the frequency of 128 guesses distributed over 18 tints is given, the approximation to a curve of Type I. is fairly close considering the paucity of guesses.

(28.) *Example VIII.*—The question may be raised, how are we to discriminate between a true curve of skew type and a compound curve, supposing we have no reason to suspect our statistics *à priori* of mixture. I have at present been unable to find any general condition among the moments, which would be impossible for a skew curve and possible for a compound, and so indicate compoundedness. I do not, however, despair of one being found. It is a fact, possibly of some significance, that the best fitting skew curve to several compound curves that I have tested is a curve of Type I., and not that of Type IV. which appears to be the more usual type in biological statistics. Taking, as an example, the statistics for the "foreheads" of Naples crabs due to Professor WELDON, and resolved into their components in my memoir, 'Phil. Trans.' A, vol. 185, p. 85, *et seq.*, I find for the best fitting skew curve the equation

$$y = 83{\cdot}2526\left(1 + \frac{x}{40{\cdot}9296}\right)^{14{\cdot}77264}\left(1 - \frac{x}{11{\cdot}2125}\right)^{4{\cdot}0469},$$

where the origin is at 1·4274 horizontal units from the centroid-vertical in the

positive sense of the horizontal scale. If, now, we place this skew curve and the compound curve of Plate 1, 'Phil. Trans.,' vol. 185, on top of the observations (see Plate 13, fig. 11), we see at once how much better is the fit of the compound curve. The skew curve gives a mean percentage error in the ordinates of 10·4, the compound curve of only 7·4. The determination of the best skew curve, when the compound curve is known, is easy, for all its details are already practically calculated.

A criterion of whether a compound or skew curve is to be sought for *ab initio*, would be, however, of great value.

(29.) *Example IX.*—A more markedly skew curve than any we have yet dealt with is that giving the frequency of divorce with duration of marriage. I take my statistics from a paper by Dr. W. F. Willcox, entitled "The Divorce Problem, a Study in Statistics" ('Studies in History, Economics, and Public Law,' Columbia College, vol. 1, p. 25). They are as follows:—

Duration of marriage in years.	Divorces (1882–6).	Duration of marriage in years.	Divorces (1882–6).
1	5314	12	4089
2	7483	13	3563
3	9426	14	3144
4	9671	15	2931
5	9014	16	2721
6	8274	17	2217
7	7021	18	1877
8	6093	19	1577
9	5305	20	1459
10	5002	21 and over	9401
11	4384		

Total number of divorces granted, 109,966.

Now these statistics suffer from a defect common to many of the class—the want of careful enumeration of the frequencies near the beginning and end of the series. It cannot be too often insisted upon that careful details of the frequencies in the start and finish of the distribution are requisite if we are to fit skew distributions with their appropriate skew curves. How, in this case for example, are we to distribute the 9401 divorces which occur after 21 years of married life? How, on the other hand, does the curve start? It is impossible to place 5314 divorces at the mean—6 months—of the one year duration. It is obvious that the applications for divorce will be far more numerous in the last half-year than the first half-year of matrimony. The very time required to institute legal proceedings and get a divorce granted must ensure this if nothing else did. Yet these two tails of 5314 and 9401, of which the accurate distributions are not given, are between $\frac{1}{7}$ and $\frac{1}{8}$ of the total number of divorces, and until we know how they are exactly distributed, we cannot hope for the very exact fitting of a theoretical curve.

In order to make the best of the "tails" under the circumstances, their moments were calculated on two hypotheses, (i.) that they were triangles, (ii.) that they were logarithmic curves, and the mean of these extreme results taken.

I found

$$\mu_2 = 60{\cdot}7376, \qquad \mu_3 = 809{\cdot}15,$$
$$\gamma = {\cdot}150127, \qquad p = {\cdot}36891.$$

Distance of centroid from start of curve = 9·1183,
,, maximum ,, ,, = 2·4373,

y_0 = maximum frequency = 8882·45.

Here the curve is assumed, owing to the obviously long tail to the right and the abrupt start to the left, to be of Type III. Its equation is accordingly

$$y = 8882{\cdot}45\left(1 + \frac{x}{2{\cdot}4573}\right)^{{\cdot}36891} e^{-{\cdot}150127x}, \qquad \text{Skewness} = {\cdot}8547.$$

The curve is figured, Plate 11, fig. 12, and will be seen to rise abruptly at about ·47 of a year's duration. It may be doubted whether legal proceedings even in America are so rapid that a divorce suit can be complete within six months of marriage. The curve gives fairly well the general form of the frequency statistics. Could the moments have been determined with greater accuracy, most probably a better fit would have resulted. As it is the mean percentage error is above 6.

(30.) *Example X.*—A still more extreme case may be selected from the field of economics. I take the following numbers from the 1887 Presidential Address of Mr. GOSCHEN to the Royal Statistical Society ('Journal,' vol. 50, Appendix II. pp. 610–2). I have grouped together both houses and shops, because the details of the two are not in Mr. GOSCHEN's returns separated for values under £20.

VALUATION of House Property, England and Wales, years 1885 to 1886.

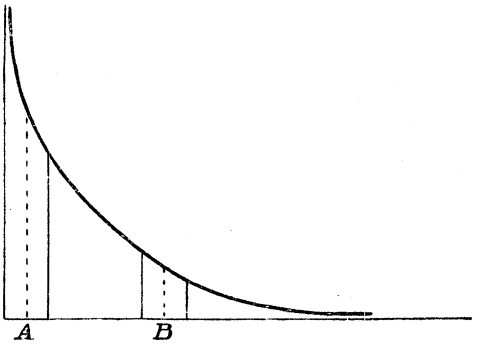
	Number of houses.		Number of houses.
Under £10	3,174,806	£80 to £100	47,326
£10 to £20	1,450,781	100 ,, 150	58,871
20 ,, 30	441,595	150 ,, 300	37,988
30 ,, 40	259,756	300 ,, 500	8,781
40 ,, 50	150,968	500 ,, 1,000	3,002
50 ,, 60	90,432	1,000 ,, 1,500	1,036
60 ,, 80	104,128		

Here clearly the curve *starts* with the maximum frequency, and further to any

scale to which the curve can be drawn, it tails away indefinitely to the right. This justifies us in the assumption that the curve will be fairly approximated to by a form of type

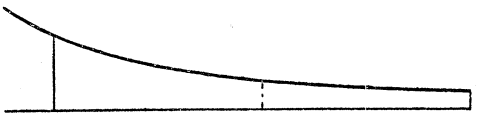
$$y = y_0 x^p e^{-\gamma x},$$

where p would turn out to be a negative quantity lying between 0 and 1. But the details given us of the start and finish of the curve are far too scanty to allow us to proceed by moments. In the first place, to measure an element of area of the frequency curve by an element of value into its mid-ordinate is perfectly legitimate



at such a point as B; it fails entirely, however, at such a point as A, which includes the part of the curve which is asymptotic to the ordinate of maximum frequency. The area at such a point is much greater than the element into the mid-ordinate, and the calculation of moments on the assumption that 3,174,806 houses may be concentrated at £5, is purely idle. The ordinate obtained from the area in this manner may often differ 30 per cent. from the true ordinate, and yet about three-fifths of the total number of houses fall into this first group.

Further treating the area as ordinate into element of value is also true only if the element of value be *small*. For "elements" such as £150, £200, or even £500, which are all that are given in the tail of these statistics, it is perfectly idle to concentrate the area at the mid-ordinate. The centroid of a piece of tail such as the accompanying figure suggests lies far to the left of the mid-ordinate In other words, to attack the



problem by the method of moments, we require to have the "tail" as carefully recorded as the body of statistics. Unfortunately the practical collectors of statistics often neglect this first need of theoretical investigation, and proceed by a method of "lumping together" at the extremes of their statistical series.

Still three further points in regard to the present series of statistics. First, they are

very unlikely to be homogeneous. Houses with an annual valuation of over £300 hardly fall under the same series of causes as the bulk of houses in the kingdom which fall under £100. Secondly, when we are told that 3,174,806 houses are valued *under* £10, it can hardly mean that any houses are valued at 0, certainly not the maximum number. Hence our frequency curve in theory must not be expected to rise from zero, but from some point between 0 and £10, which corresponds to the customary minimum at which a cottage can be rented.

Lastly, there is one special cause at work tending to upset, about the value of £20, the general distribution due to a great variety of small causes. This is the value at which taxation commences, and we should expect a larger proportion of houses to be built just under the taxable value than is given by a chance distribution.

Notwithstanding the many disadvantages of these results, I determined to obtain if possible a skew curve approximating to the main portion of the distribution. I took £10 as my unit of value and 1000 houses as my unit of frequency. I started with the ordinary method of moments, concentrating each area at its centroid as given by the total valuation of the group, also recorded by Mr. GOSCHEN, and found a curve of the type

$$y = y_0 x^p e^{-\gamma x},$$

with

$$p = -\cdot 65448, \qquad \gamma = \cdot 2003.$$

This was so far satisfactory that it showed even by this rough method that p was negative, and between 0 and 1. Thus the theoretical curve gave an infinite ordinate, but *finite area* at its start.

A laborious method of trial and error was then adopted, and by varying p and γ slightly, as well as y_0 and the origin of the curve, I sought to improve the fit given by the rough method (in this case) of moments. The fundamental consideration was to keep the total areas under £100 value as nearly as possible the same in the theoretical curve and the statistics. This portion of the curve I treated as practically referring to homogeneous material. Ultimately I found the following curve:

$$y = 1388{\cdot}32\, x^{-\cdot 690077}\, e^{-\cdot 3057256x},$$

with the origin as ·45 unit from zero. Thus the minimum annual valuation was £4 10*s.*, or, to a weekly valuation, of 1*s.* 7½*d.* This would connote probably a weekly rental of 1*s.* 8*d.* to 2*s.* The total area of this theoretical curve was 5795 in thousands of houses; of these 5729 had a valuation under £100 and 66 over £100; the corresponding numbers for the statistics themselves are 5720 and 110. The additional 44 over £100 I assume to be due to the heterogeneity of the statistics—high values corresponding to blocks of chambers, large hotels and other buildings hardly falling into the same category as the small house under £100 in value. Unfortunately the "tail" of the statistics is so defectively recorded that there is no hope of reaching a separate distribution for this high class property.

Returning now to the curve and statistics, we have the following comparative results :—

Value.	Number of 1000's of houses.	
	Theory.	Statistics.
£		
Under 10	3580 } 4625	3175 } 4626
10–20	1045 }	1451 }
20–30	452	442
30–40	253	260
40–50	153	151
50–60	97	90
60–80	102	104
80–100	46	47
Above 100	66	110

The general accordance here is very marked, the chief divergences being accounted for by the special causes to which we have referred above, *i.e.* (i) the crowding of houses just below the limit of taxation, and (ii) the divergent character of the causes at work determining the frequency of low and high class house property.

The results are depicted, Plate 14, fig. 13.

It will be observed that so far as the observations can be plotted to the theoretical curve, it leaves little to be desired. The histogram* shows, however, the amount of deviation at the extremes of the curve.

(31.) *Example XI.*—Frequency curves of the type considered in Example X. are so common that it is needful to make a few further remarks with regard to them, and illustrate them by further examples. Such curves occur in many economical instances (income tax, house valuation, probate duty), in vital statistics (infantile mortality), and not uncommonly in botanical statistics of the frequency of variations in the petals or other characteristics of flowers.

As we have noted, the method of moments developed in this memoir cannot be directly applied, or only applied to obtain a first approximation to the constants required. This first approximation, however, will often assist us to obtain with quite sufficient accuracy the value of the moments of portions of the area, especially if the position of the initial or asymptotic ordinate is known.

For example, consider the curve of limited range :

$$y = y_0 x^{-p} (b - x)^n$$

where p lies between 0 and 1. Then if α be its area, $\alpha\mu''_s$ = the s^{th} moment about the asymptotic ordinate of the area up to x :

* Introduced by the writer in his lectures on statistics as a term for a common form of graphical representation, *i.e.*, by columns marking as areas the frequency corresponding to the range of their base.

$$\alpha\mu''_s = \int_0^x y_0 x^{s-p}(b-x)^n\,dx$$

$$= y_0 b^n \left\{\frac{1}{1+s-p} - \frac{nx}{b(2+s-p)} + \frac{n.\overline{n-1}}{1.2\,(3+s-p)}\left(\frac{x}{b}\right)^2 - \&c.\right\} x^{1+s-p}.$$

Hence, if the range b be large and x be small, this series converges very rapidly, and we may often take with sufficient approximation even only its first term. Thus

$$\left.\begin{aligned} \mu''_1 &= x\,\frac{1-p}{2-p} \\ \mu''_2 &= x^2\,\frac{1-p}{3-p} \\ \mu''_3 &= x^3\,\frac{1-p}{4-p} \\ \alpha &= \frac{y_0 b^n}{1-p}\,x^{1-p} \end{aligned}\right\} \text{nearly.}$$

Now α is given by the statistics, and we note that if p has been determined to a first approximation by the method of moments, we can now improve the values of the moments of the areas near the asymptotic ordinate by the use of the above expressions.

For example, if $p = \cdot 5$ as a first approximation, we have

$$\mu''_1 = \tfrac{1}{3}x, \quad \mu''_2 = \tfrac{1}{5}x^2, \quad \mu''_3 = \tfrac{1}{7}x^3.$$

Concentration along the mid-ordinate in the usual manner would have given us

$$\mu''_1 = \tfrac{1}{2}x, \quad \mu''_2 = \tfrac{1}{4}x^2, \quad \mu''_3 = \tfrac{1}{8}x^3,$$

and as the area up to a short distance from the asymptotic ordinate is generally a considerable proportion of the total area, the above values very considerably modify the calculated moments.

In the case of the curve

$$y = y_0 x^{-p} e^{-\gamma x},$$

we have the result

$$\alpha\mu''_s = y_0 x^{s+1-p}\left\{\frac{1}{1+s-p} - \frac{\gamma x}{2+s-p} + \frac{\gamma^2 x^2}{1.2.(3+s-p)} - \right\}.$$

Hence, as before, if γ and x be small,

$$\mu''_s = \frac{1-p}{1+s-p}\,x^s, \text{ approximately.}^*$$

Results such as the above enable us to approximate fairly rapidly to the constants of a frequency curve.

As a special example, I take the following. In 1887, Herr H. DE VRIES transferred several plants of *Ranunculus bulbosus* to his flower garden, and counted the petals of 222 of their flowers in the following year. He found ('Berichte der deutschen botanischen Gesellschaft,' Jahrg. 12, pp. 203–4, 1894):

Petals	5	6	7	8	9	10
Frequency . . .	133	55	23	7	2	2

Now the series here proceeds by discrete units, and corresponds probably to a hypergeometrical series, but remembering how closely the results of tossing ten coins can be represented by a normal frequency curve, I was not without hope that the areas of a skew frequency curve would give results close to these numbers. The buttercups start with 5 petals and run to 10, I therefore took my origin at 4·5 and determined the constants to a second approximation in the manner above indicated. There resulted,

$$y = \cdot 211225x^{-\cdot 322}(7\cdot 3253 - x)^{3\cdot 142},$$

a curve of Type I., with limited range, the asymptotic ordinate being at 4·5 petals, or practically a distribution ranging from 5 to 11 petals.

Calculating the areas, there results,

Petals		5	6	7	8	9	10	11
Frequency	Theory . . .	136·9	48·5	22·6	9·6	3·4	·8	·2
	Observation .	133	55	23	7	2	2	0

The agreement here is very satisfactory considering the comparative paucity of the observations.† The results are exhibited by curve and histogram, Plate 15, fig. 14; the two points on the "observation curve" corresponding to five and six petals are deduced from the areas given by the statistics by the same percentage reduction as

* Another very serviceable formula is due to SCHLÖMILCH. It gives the area of the "tail" of $y = y_0 x^{-p} e^{-\gamma x}$ from $x = x$ to $x = \infty$ in a rapidly converging series, *i.e.*,

$$\text{area} = \frac{y_0 x^{-p} e^{-\gamma x}}{\gamma}\left\{1 - \frac{p}{\gamma x+1} + \frac{p^2}{(\gamma x+1)(\gamma x+2)} - \frac{p(p^2+1)}{(\gamma x+1)(\gamma x+2)(\gamma x+3)} + \&c.\right\}.$$

† 2048 tosses of 10 shillings at a time gave a mean 3 per cent. deviation between theory and experiment, 100 tosses gave about 9 per cent. The above series corresponds to about 7·2 per cent., and thus is quite within the range of accuracy of coin-tossing experiments.

converts the theoretical areas into the ordinates of the theoretical curve. For other petals, ordinates and areas practically coincide in value.

(32.) *Example XII.*—Another example of a similar kind may be taken from Herr DE VRIES' memoir (*loc. cit.*, p. 202). He cultivated under the name of *perum-bellatum* a race of *Trifolium repens*, in which the axis is very frequently prolonged beyond the head of the flower, and bears one to ten blossoms. In the summer of 1892 he had a bed of such clover, produce of a single plant, and in July counted the extent of this variation on 630 flowers. In 325 cases the axis, according to DE VRIES, had not grown through the head of the flower, in 83 cases it had grown through and bore one blossom, in 66 cases two blossoms, and so on. The complete statistics are as follows :—

High blossoms	0	1	2	3	4	5	6	7	8	9	10
Frequency . .	325	83	66	51	36	36	18	7	6	1	1

Taking moments in the manner of the earlier part of this memoir, I found as a first approximation to the frequency curve :

$$y = 4{\cdot}52842\, x^{-{\cdot}442817}\,(10{\cdot}69114 - x)^{1{\cdot}525944},$$

with the origin at ·47813 to left of maximum ordinate. This first approximation seemed to justify three things : (i.) starting at ·5 to the left of the maximum ordinate; (ii.) assuming a range, 11, which just covered the whole series of observations, *i.e.*, from ·5 to 10·5 ; and (iii.) that the moments of the areas might be found from a value of p not far from ·5.

A second approximation was then made, and taking moments round the asymptotic ordinate, I found :

$$\mu'_1 = 1{\cdot}8680, \qquad \mu'_2 = 7{\cdot}77028,$$

whence, in the manner of §16, we have :

$$\chi_1 = {\cdot}1698182, \qquad \chi_2 = {\cdot}3781526,$$

and ultimately :

$$m_1 = -{\cdot}493118, \qquad m_2 = 1{\cdot}47797,$$

and

$$y_0 = 4{\cdot}65148.$$

The equation to the frequency curve is therefore :

$$y = 4{\cdot}65148\, x^{-{\cdot}493118}\,(11 - x)^{1{\cdot}47797}.$$

The value found for p, *i.e.*, ·493, justifies our calculation of the moments on the assumption that it was ·5.

Placing statistics and theory side by side, we have :

High blossoms	0	1	2	3	4	5	6	7	8	9	10
Statistics	325	83	66	51	36	36	18	7	6	1	1
Theory	303·22	106·12	69·99	49·27	35·23	24·93	17·07	10·96	6·27	2·79	·52

The agreement between theory and observation is here all that could be desired, except in the case of 0 and 1 high blossoms. Here 22 blossoms have in actual counting been transferred from the theoretical group of 1 to the theoretical group of zero high blossom. I consider it highly probable that the theory here gives better results than the actual statistics ; and this, for the simple reason that it must be very difficult to distinguish between any one of the low blossoms and a very *slightly extended axis bearing only one blossom*, that is to say, the extension of the axis passes insensibly into one of the low blossoms, or *vice versâ*, and in a certain proportion of cases it must be difficult to distinguish between the categories 0 and 1. The comparison between theory and observation is represented by curve and histogram, Plate 15, fig. 15.

Examples X. to XII. will suffice to illustrate the application of our theory to extreme cases of skew distribution.

(33.) *Example XIII.*—It must not be supposed that in every case of variation by units (as in the buttercup and clover examples), the curve will be found to be of Types I. or III. It is impossible to illustrate, in anything short of a treatise on statistics, the infinite variety of statistical distributions, but the occurrence of Type IV. in zoological, as distinguished from botanical measurements, is so persistent that it seems well to illustrate this for the special case of discontinuous variation. Professor WELDON has kindly given me the following statistics of dorsal teeth on the rostrum of 915 ♂ and ♀ specimens of *Palæmonetes varians* from Saltram Park, Plymouth.

Teeth.	Cases.
1	2
2	18
3	123
4	372
5	349
6	50
7	1

The centroid-vertical here lies ·313661 of a tooth beyond 4, *i.e.*, at 4·313661 teeth. The following are the moments about centroid-vertical :—

$$\left.\begin{aligned}\mu_2 &= \cdot 910906\\ \mu_3 &= \cdot 233908\\ \mu_4 &= 2\cdot 625896\end{aligned}\right\}\text{where the unit} = 1\text{ tooth.}$$

For the normal curve these give

$$\begin{aligned}\text{Standard deviation} &= \cdot 9544,\\ \text{Maximum ordinate} &= 382\cdot 5.\end{aligned}$$

For the skew curve we have

$$\beta_1 = \cdot 072222, \qquad \beta_2 = 3\cdot 164684.$$

Hence

$$2\beta_2 - 3\beta_1 - 6 = \cdot 122702,$$

or, we have a curve of Type IV. The values of β_1 and β_2, however, show that it will not differ very widely from the normal type.

Proceeding to determine the other constants we find

$$\begin{aligned}r &= 111\cdot 398,\\ \nu &= -109\cdot 047\ (\nu \text{ is negative since } \mu_3 \text{ is positive}),\\ a &= 7\cdot 16613, \qquad\qquad m = 56\cdot 699.\end{aligned}$$

Distance of origin from centroid-vertical $= 7\cdot 0149$,

Thus

$$\log y_0 = \overline{18}\cdot 4431056.$$

$$y = y_0 \cos^{113\cdot 398}\theta\, e^{109\cdot 047\theta},$$
$$x = 7\cdot 16613 \tan\theta$$

give the form of the curve. This curve, the normal curve, and the observations are drawn, Plate 13, fig. 16. A comparison of the observations and the normal curve shows an amount of skewness in the tails of the former, which would be very improbable if the normal curve really expresses the distribution. The skew curve really accounts for this divergence and is a sensibly better fit. The mean percentage errors in the ordinates are for the two cases 8·67 and 3·88. The skew curve is thus an excellent fit.

The discontinuity in these teeth probably corresponds to a hypergeometrical polygon, of which the skew curve is a limiting form.

(34.) *Example XIV.*—Another extremely interesting illustration of skew variation will be found in the statistics of pauperism for England and Wales, to which my attention was drawn by Mr. G. U. Yule, who had plotted the statistics from the raw material provided in Appendix I. of Mr. Charles Booth's 'Aged Poor; Condition'

In Plate 14, fig. 17, we have 632 unions distributed over a range of pauperism varying from 100 to 850 per 10,000 of the population for the year 1891. The observations

are at once seen to give a markedly skew distribution. Taking 50 paupers as unit of variation, we find

$$\mu_2 = 6{\cdot}31889, \qquad \beta_2 = 3{\cdot}060017,$$
$$\mu_3 = 6{\cdot}62465, \qquad \beta_1 = {\cdot}173942.$$
$$\mu_4 = 122{\cdot}1815,$$

Hence

$$3\beta_1 - 2\beta_2 + 6 = {\cdot}401791,$$

or the curve is of Type I.

The other constants were found to be

$$r = 28{\cdot}165013,$$
$$\epsilon = 148{\cdot}0886,$$
$$m_1 = 20{\cdot}169714, \qquad a_1 = 24{\cdot}2203$$
$$m_2 = 5{\cdot}995305, \qquad a_2 = 7{\cdot}199312$$
$$y_0 = 99{\cdot}9065.$$

Range = 31·4196.

Maximum = ·60434 to left of centroid vertical.

Skewness = ·24.

The equation to the curve is thus

$$y = 99{\cdot}9065\left(1 + \frac{x}{7{\cdot}1993}\right)^{5{\cdot}9953}\left(1 - \frac{x}{24{\cdot}2203}\right)^{20{\cdot}1697}.$$

For the normal curve,

Standard deviation = 2·514,

Maximum ordinate = 100·301.

Both skew curve and normal curve are drawn on Plate 14, fig. 13. The former is at once seen to be an excellent fit. We might fairly have simplified our work by taking zero paupers as the commencement of our range, but preference was given to the more general results in order to demonstrate that they give no appreciable amount of "negative pauperism." The range determines a limit of about 15 per cent. as the greatest possible amount of pauperism. The normal curve is seen to diverge very widely from the statistics besides giving an appreciable amount (3 to 4 unions) with "negative pauperism." The point-binomial for these statistics is also figured on the plate. Its constants are $p = {\cdot}833$, $q = {\cdot}167$, $n = 14{\cdot}4834$, $c = 1{\cdot}70306$, the start of the binomial being 5·81503 to the left of the centroid-vertical: see § 5. The fit is a very close one, the mean error of ordinate = 5·37, and the suggestiveness of such results for social problems needs no emphasising.

The case is of peculiar interest, because the statistics of pauperism are known to give a definite trend to the distribution, *i.e.*, if the statistical curve of pauperism for 1881 be compared with that of 1891, for example, the maximum frequency of the

earlier will be found at a much higher percentage. The whole frequency curve is sliding across from right to left. Now it is of interest to notice that in this, as in other cases where the trend of the variation is known *à priori*, the skew curve is shifted away from the normal curve in the direction in which variation is taking place with lapse of time. It is not safe at present to extend this to all biological instances, but the result *suggests*, for example, that there is a secular progression towards brachycephaly in Bavarian skulls (fig. 8), towards reduced antero-lateral margin in crabs (fig. 4), towards increased height in St. Louis school-girls (fig. 7), and towards long-sightedness in Marlborough School boys.* I believe most suggestive and important results might be obtained for the theory of evolution, if we only had the series of skew curves for a biological case of progressive variation in the same manner as we have for pauper percentages.

(35.) *Example XV.* The theoretical resolution of heterogeneous material into *two* components, each having skew variation, is not so hard a problem as might at first appear, and I propose to deal at length with the subject later. If there be more than two components, the equations become unmanageable. In this case however, if the components have rather divergent means, a tentative process will often lead to practically useful results. To illustrate this I propose to conclude this paper by an example of a mortality curve resolved into its chief components. By a mortality curve I understand one in which frequency of death (for 1,000, 10,000, or 100,000 born in the same year) is plotted up to age. I have worked out the resolution for English males, and for French of both sexes. The generally close accordance of the results for both cases has given me confidence in their approximate accuracy. The method adopted was the following: An attempt was made to fit a generalised frequency curve to the old age portion of the whole mortality curve, the constants of this curve being determined from the data for four or five selected ages by the method of least squares; the frequency curve so determined was subtracted from the total curve, and a frequency curve fitted by the same method to the tail of the remainder. This second component was again subtracted and the process repeated, until the remainder left could itself be expressed by a single frequency curve. The components thus obtained were added together, and a tentative process adopted of slightly modifying their constants and position, so that the total areas of the components and of the whole mortality curve coincided. It was soon obvious that no very great change either in the constants or position was permissible, if the sum of the components was to give the known resultant curve, hence I feel very confident that whatever be the combination of causes which result in the mortality curve, that curve is very approximately to be considered as the compound of five types of mortality centering about five different ages. The allied character of the results obtained for both French and English statistics confirms this view.

* Dr. ROBERTS' statistics, which I have reduced to skew curves, but have not reproduced in this memoir.

Professor LEXIS has already suggested that the old age distribution of mortality is given by a normal curve.* Now, although the rougher French statistics give a fair approximation to a normal curve, this is not true for English males. The curve for old age is of Type I., but for all practical purposes it may be treated as one of Type III. Whatever be the chief causes of old age mortality, they extend very sensibly through middle life, and less sensibly through youth, only becoming inappreciable in childhood. Hence, if we speak of our first component as the "mortality of old age," the name is to be understood as referring to a group of causes especially active in old age mortality, but not excluded from other portions of life. The second and third components I found to be skew curves, but so nearly normal that to my degree of approximation no stress could be laid on the skewness obtained. The fourth component was a markedly skew curve, also closely given by a curve of Type III., and corresponding in general shape to the mortality curves of fevers peculiarly dangerous in childhood (*e.g.*, diphtheria, scarlet fever, enteric fever, &c.). These three components I have termed respectively the mortality of middle life, of youth, and of childhood. I found it impossible to fit the remainder of the original mortality curve with any type of generalised curve, so long as I supposed the mortality frequency to commence with birth. I was therefore compelled to suppose the set of causes giving rise to "infantile mortality" extended into the period of gestation, and I obtained a satisfactory fit for the infantile mortality frequency, when the range of the curve started about ·75 of a year before birth. The form taken by the curve is the extreme type in which the curve is asymptotic to the ordinate of maximum frequency (*cf.* Examples X.–XII.). The five fundamental components of the mortality curve for English males are the following, the numbers referring to 1000 contemporaries, or persons born in same year:—

(A.) *Old Age Mortality.*

Total frequency = 484·1.
Centroid-vertical at 67 years.
Maximum mortality = 15·2 at 71·5 years.

The equation is†

$$y = 15{\cdot}2\left(1 - \frac{x}{35}\right)^{7{\cdot}7525} e^{{\cdot}2215x},$$

the axis of y being the maximum ordinate and the positive direction of x towards age. The skewness of the curve = ·345, and its range concludes at 106·5 years.

The corresponding French component = 411, but the maximum mortality (16·4) occurs at 72·5 years.

* 'Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft,' § 46. Freiburg, 1877.

† Unit of x = 1 year, unit of y = 1 death per year.

(B.) *Mortality of Middle Life.*

Total frequency = 173·2.
Centroid-vertical at 41·5 years.
Maximum mortality = 5·4.

The curve is very approximately normal, and has a standard deviation of 12·8 years. The corresponding French component = 180 deaths, standard deviation 12 years, with a maximum of 6 at 45 years.

(C.) *Mortality of Youth.*

Total frequency = 50·8.
Centroid-vertical at 22·5 years.
Maximum mortality = 2·6.

The curve is very approximately normal, with a standard deviation of 7·8 years.* The corresponding French component gives a total mortality of 78, standard deviation of 6 years, and a maximum of 5·2 at 22·5 years.

The greater and more concentrated French mortality of youth is noteworthy.

(D.) *Mortality of Childhood.*

Total frequency = 46·4.
Centroid-vertical at 6·06 years.
Maximum mortality = 9 at 3 years.

The equation to the curve, the axis of y being maximum ordinate, is

$$y = 9\ (1 + x)^{\cdot 3271}\ e^{-\cdot 3271x}.$$

Thus the skewness of the curve = ·87, and the range commences at 2 years.

The French component appears to be shifted further towards youth. It gives a total of 47 deaths, centroid at 8·75 years, and a maximum of 5·8 at 5·75 years, skewness = ·71. Childish mortality is therefore, if these results be correct, more concentrated, and at an earlier age in England than in France.

(E.) *Infantile Mortality.*

Total frequency *after birth* = 245·7.
Maximum frequency *after birth* occurs in first year and equals 156·2.

The equation to the frequency curve is

$$y = 236\cdot 8\ (x + \cdot 75)^{-\cdot 5}\ e^{-\cdot 75x},$$

the origin being at birth, the skewness ·707, and the centroid at ·083 year, = 1 month nearly, before birth. Taking the corresponding French component, we have a total frequency after birth of 284, with 186 deaths in the first year of life. Infantile mortality is therefore considerably greater in France.

* The mortality of youth would be better expressed by a curve of type $y = y_0\left(1 - \frac{x^2}{a^2}\right)^m$: see our § 13 (v.).

If we investigate the areas of our infantile mortality curve, we have the following deaths :—

	Theory.	Statistics.
1st year of life	156·2	158·5
2nd year of life . . .	53·5	51·2

After this the mortality of childhood begins to sensibly increase the infantile mortality. Turning to the "antenatal" portion of the curve, we have the following results, of course not verifiable from ordinary mortality statistics :—

(i.) The total "antenatal" deaths for the 9 months preceding birth are 605 for every 1000 actually born and registered.

(ii.) "Antenatal" deaths for the 6 months immediately preceding birth are 214 for every 1000 born.

(iii.) "Antenatal" deaths for the 3 months immediately preceding birth are 83 for every 1000 born at the proper period.

The 391 "deaths" of the first three months of pregnancy would not be recorded, and in many cases possibly pass without notice. The 214 deaths of the remaining six months would be considered as miscarriages or still-births. The proportion of 1 in 6 of such accidents to births of the normal kind does not appear excessive. On the average Dr. GALAPIN says such an occurrence is "the experience of every woman who has borne children and reached the limit of the child-bearing age." So far then there appears nothing to contradict our theoretical results in what is known of the first six months of antenatal life.

For the last three months we have more definite data. According to our curve we have 83 deaths (per 1000 born) in the last three months before birth, or 83 in 1083 pregnancies = about 7·7 per cent. Now this percentage must consist of two factors—still-born children and children who, born before their time, die shortly after birth, and who would not be recorded in any proper proportions in statistics based on census returns, nor as a rule in the returns of maternity charities.

For statistics of still-births, I find :

	per cent.
Dublin Rotunda Hospital (1847–54)	6·9
,, ,, ,, (1871–75)	6·1
Dr. J. H. DAVIS for 14,000 births for a large maternity charity in St. Pancras	4
Guy's Hospital Lying-in Charity, 25,777 births, 1,127 born dead or died within a few hours, 1000 corresponding to births in the last three months of pregnancy	3·84
NEWSHOLME's "Vital Statistics" (no authority cited) . . .	4

It would thus appear that there are 4 to 5 per cent. of still-births, thus leaving 2·7 to 3·7 per cent. of deaths to be accounted for—if there is any validity in our analysis—by deaths of children born before their proper time and dying before their proper birthdays. Such deaths would not appear in the category of still-born children in the returns of the maternity charities, nor in any true proportion in the census returns.

Thus, while it is impossible to assert any validity for the antenatal part of our curve of infantile mortality, while, indeed, the constants of that curve, and consequently the percentages of antenatal deaths, might be considerably modified had we surer data of the actual deaths in the first year of life; still there appears to be nothing wildly impossible in the results obtained, and they may at any rate be suggestive, if only as to the nature of those statistics of "antenatal" deaths, which it would be of the greatest interest to procure.

The absolute necessity of skew curves in all questions of vital statistics is sufficiently evidenced in this resolution of the general mortality curve. A complete picture of the resolution into components of the mortality curve is given (Plate 16, fig. 18), with a separate figure on an enlarged scale of infantile mortality.

(36.) In conclusion, there are several points on which it seems worth while to insist. The normal curve of errors connotes three equally important principles:

(i.) An indefinitely great number of "contributory" causes.

(ii.) Each contributory cause is in itself equally likely to give rise to a deviation of the same magnitude in excess and defect.

(iii.) The contributory causes are independent.

The frequency of each possible number of heads in repeatedly throwing several hundred coins in a group together, practically fulfils all the above three conditions.

Condition (ii.) is not, however, fulfilled if a number of dice be thrown or a number of teetotums of the same kind be spun together. Condition (iii.) is still fulfilled.

Condition (iii.) is not fulfilled if p cards be drawn out of a pack of nr cards containing r equal suits, supposing the p cards to be drawn at one time. Now, it appears to me that we cannot say *à priori* whether the example of tossing, of teetotum-spinning, or of card-drawing is more likely to fit the proceedings of nature. There is, I think, now sufficient evidence to show that the conditions (i.) to (iii.) are not fulfilled, or not exactly fulfilled, in many cases—in economic, in physical, in zoometric, and botanical statistics. We are, therefore, justified in seeing what results we shall obtain by supposing one or more of the above conditions which lead to the normal curve to be suspended. The analogy of teetotums and cards leads us to a system of skew frequency curves which in this paper have been shown to give a very close approximation to observed frequency in a wide number of cases—an approximation quite as close as the writer has himself obtained between theory and experiment in very wide experiments in tossing, card-drawing, ball-drawing, and

lotteries. But the introduction of these skew curves leads us to two important conclusions:—

(i.) If a material be heterogeneous we have no right to suppose it must be made up of groups of homogeneous material each obeying the *normal* law of distribution. Each homogeneous group may follow its own skew distribution.

(ii.) If material obeys a law of skew distribution, the theory of correlation as developed by GALTON and DICKSON requires very considerable modification.

We may note two points bearing on these two conclusions, which do not seem without interest for the general problem of evolution. Fever mortality curves are skew curves. The general mortality curve—frequency of death at different ages—is a compound of many diseases, but with sufficient approximation, it can be resolved into five components; three of these components are markedly skew, the other two less so. Selection, according to age, is thus distributed with different degrees of skewness about five stages in life; this at least *suggests* that selection according to the size or weight of an organ may be compound, if we take a considerable range of size, and that the components may have varying degrees of skewness.

The correlation of the ages of husband and wife at marriage is a subject with regard to which we have a very fair amount of material. For a given age of the husband, the frequency of marriage with the age of the wife fits very closely a curve of Type IV., and with sufficient exactness very often a curve of Type III.* The sections of the surface of frequency are oval curves differing entirely from the ellipses of the GALTON-DICKSON theory, but resembling in general the "oval" polygons obtained by taking horizontal sections of the frequency polyhedron for the correlation of cards of the same suit in two players' hands at whist. Plate 9, fig. 19, shows how widely these differ from ellipses. There seems therefore to be considerable danger in assuming in vital statistics, whether in man or the lower animals, that the "contributory" causes are independent. All the statistics for sizes of organs in animals, which I have yet analysed, if they are not compound, seem to agree in following a curve of Type IV., and suggest this kind of inter-dependence of the "contributory" causes. Their correlation surfaces of frequency will thus have for lines of level skew ovals—what for want of a better name may be termed "whist ovals" as distinguished from the ellipses which flow from the normal frequency surface. The remarks from quite a different standpoint of RANKE on skull measurements seem to lead to the same conclusion. I propose on another occasion to illustrate the resolution of compound curves into skew components, and further to deal with the main features of correlation in cases of a skew frequency distribution.

* I have fitted some of PEROZZO's marriage statistics with skew curves, but reserve their discussion for the present, as they belong properly to the theory of skew correlation.

NOTE.

Added May 24, 1895.

[Since writing the above memoir my attention has been drawn to a note in Dr. WESTERGAARD'S "Theorie der Statistik," referring to Professor T. N. THIELE'S treatment of skew frequency curves. I have procured and read his book, 'Forelaesninger over Almindelig Iagttagelseslaere,' Kjøbenhavn, 1889. It seems to me a very valuable work, and is, I think, suggestive of several lines for new advance. It does not cover any of the essential parts of the present memoir. Dr. THIELE does indeed suggest the formation of certain "half-invariants," which are functions of the higher-moments of the observation—quantities corresponding to the $\mu_4 - 3\mu_2^2$, $\mu_5 - 10\mu_2\mu_3$, &c., of the above memoir. He further states (pp. 21–2) that a study of these half-invariants for any series of observations would provide us with information as to the nature of the frequency distribution. They are not used, however, to discriminate between various types of generalised curves, nor to calculate the constants of such types. A method is given of expressing any frequency distribution by a series of differences of inverse factorials with arbitrary constants. Thus if

$$\beta_n(x) = \frac{\lfloor n}{\lfloor x \, \lfloor n - x}$$

and

$$\Delta\beta_n(x) = \beta_n(x + \tfrac{1}{2}) - \beta_n(x - \tfrac{1}{2})$$

we can express any law of frequency $y = f(x)$ by

$$f(x) = b_0\beta_n(x) + b_1\Delta\beta_{n-1}(x) + \ldots + b_n\Delta^n\beta_0(x),$$

where the constants b_0, $b_1 \ldots b_n$ can be determined numerically when the frequency of $n + 1$ chosen derivation-elements is known.

I see a possibility of more than one theoretical development of interest, especially in relation to compound material, from this development of Dr. THIELE'S, but I doubt whether it can be of practical statistical service even as an empirical expression for frequency. Instead of having the 3 to 5 constants of our generalised curves, the full value of Dr. THIELE'S expression requires as many constants as there are recorded frequencies, and then expresses the result in functions like $\Delta^r\beta_s(x)$, by no means easily realised or likely to appeal to the practical statistician. It is true the complete series gives absolutely accurately the frequency of all the points used in the calculation, but it does not, like the generalised curves, indicate the purely accidental variations of the frequency. If, on the other hand, we take, as Dr. THIELE suggests, some half-dozen terms only of the series—which give the really essential character of the

frequency—we obtain results which, although more complex in form, are not as satisfactory as those given by the generalised curve.

For example, Dr. THIELE gives the following series (p. 12):—

Values	7	8	9	10	11	12	13	14	15	16	17	18	19
Frequency	3	7	35	101	89	94	70	46	30	15	4	5	1

His "Faktiske Fejllove" gives

$$\begin{aligned} y = {} & \cdot1221\beta_{12}(x) + \cdot278\,\Delta\beta_{11}(x) + \cdot600\,\Delta^2\beta_{10}(x) \\ & + \cdot216\,\Delta^3\beta_9(x) + \cdot278\,\Delta^4\beta_8(x) - \cdot318\,\Delta^5\beta_7(x) \\ & + \cdot574\,\Delta^6\beta_6(x) + \cdot596\,\Delta^7\beta_5(x) + \cdot499\,\Delta^8\beta_4(x) \\ & + \cdot259\,\Delta^9\beta_3(x) - \cdot0645\,\Delta^{10}\beta_2(x) - \cdot0303\,\Delta^{11}\beta_1(x) \\ & - \cdot0088\,\Delta^{12}\beta_0(x). \end{aligned}$$

He tells us that 6 terms practically suffice, the additional terms merely accounting for the individual irregularities of this particular 500 observations. Without specifying what the observations are, he tells us that the possible values run from 4 to 28, or that the range is really limited.

If we fit our generalised curve of Type I., we find for its equation:

$$y = 98\cdot801\left(1 + \frac{x}{4\cdot5191}\right)^{3\cdot89708}\left(1 - \frac{x}{20\cdot0296}\right)^{17\cdot27285},$$

the origin is at 11·191, or the range runs from 6·6715 to 31·1202, *i.e.*, is a range of 24·5487 instead of 25, but is shifted some 2 to 3 units. Considering the small number of observations, this is not a bad approximation to a marked feature of the distribution not indicated on the surface by the observations, nor discoverable from the "Faktiske Fejllove."

Comparing our curve (i.) with (ii.) the actual statistics—all 13 terms of the "Faktiske Fejllove" series, and with (iii.) the first 6 terms of the same series, we have the following results:—

Values .	7	8	9	10	11	12	13	14	15	16	17	18	19
(i.) .	1	10	42	80	99	92	70	48	29	15	6	3	1
(ii.) .	3	7	35	101	89	94	70	46	30	15	4	5	1
(iii.) .	1	11	40	82	103	92	70	48	26	13	8	4	1

The generalised curve here gives slightly the better results in addition to its more easily realised form, and its fewer constants (iv.).

On the other hand, there are, I think, some points of first-class theoretical importance in the mode adopted by Dr. THIELE for expressing frequency; it gives us a means of expanding all varieties of frequency curves in a series of factorial functions which may lead to important theorems in the analysis of heterogeneous material.]

PLATES.

The scale of the accompanying figures is not that of the original drawings, and the clearness and distinctness of the several curves of the same figure have been, in several instances, partially lost by the process of reproduction and reduction. In every case the square element of the figure corresponds to the square centimetre of the original diagram, and is spoken of both in the text of the memoir and on the figures themselves as a square centimetre. The scale of actual reduction is indicated by a fraction placed at the lower right-hand corner of the figure.

Fig. 2.

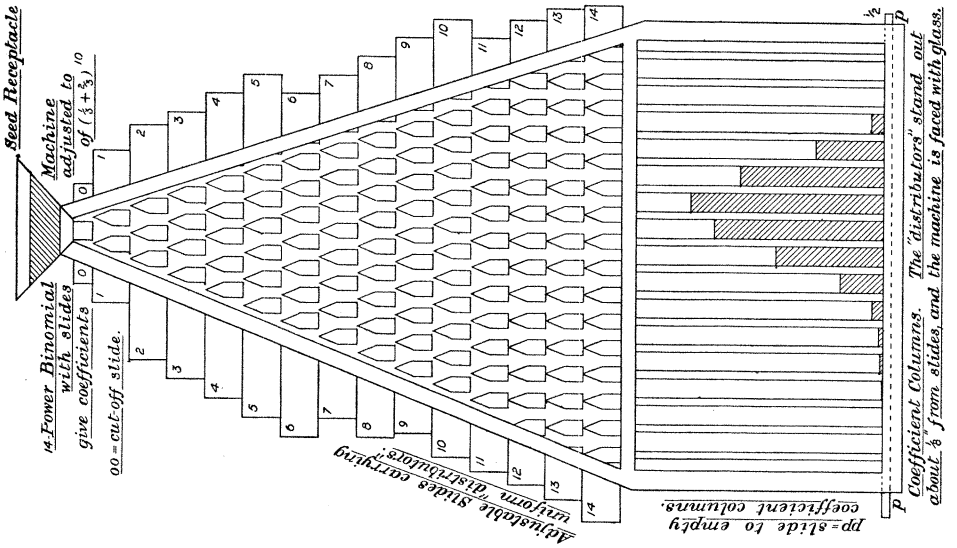


Fig. 1.

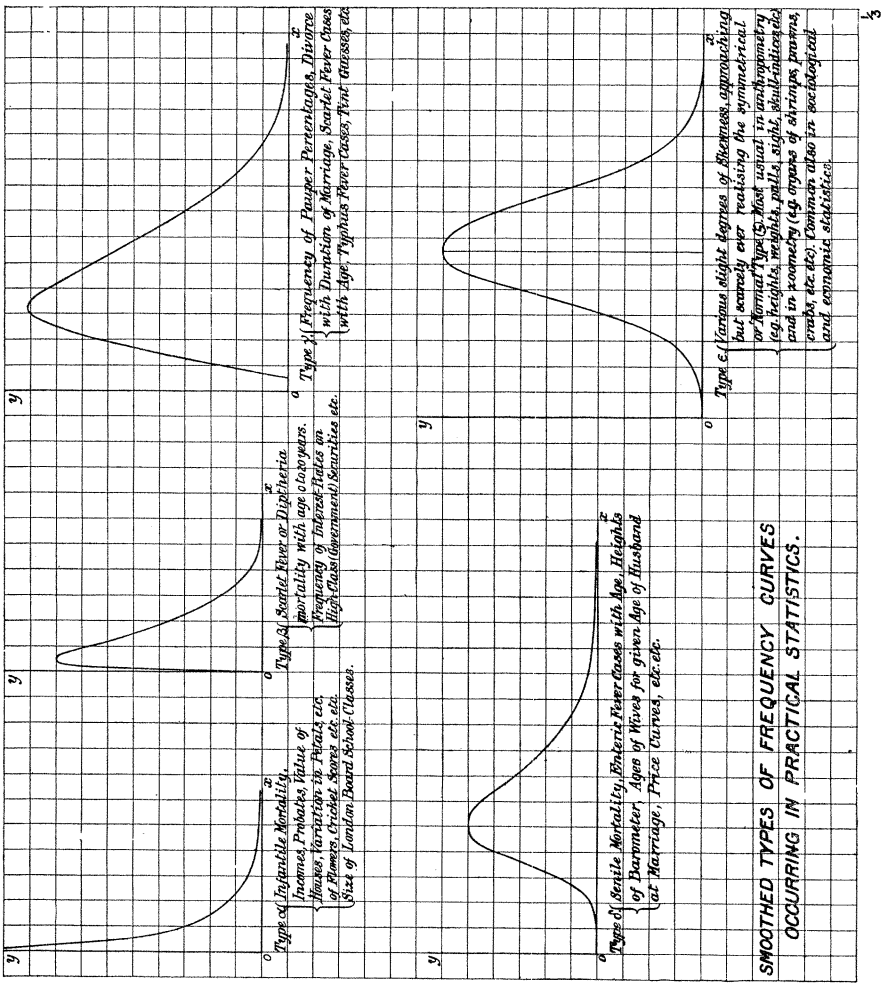


Fig. 3.

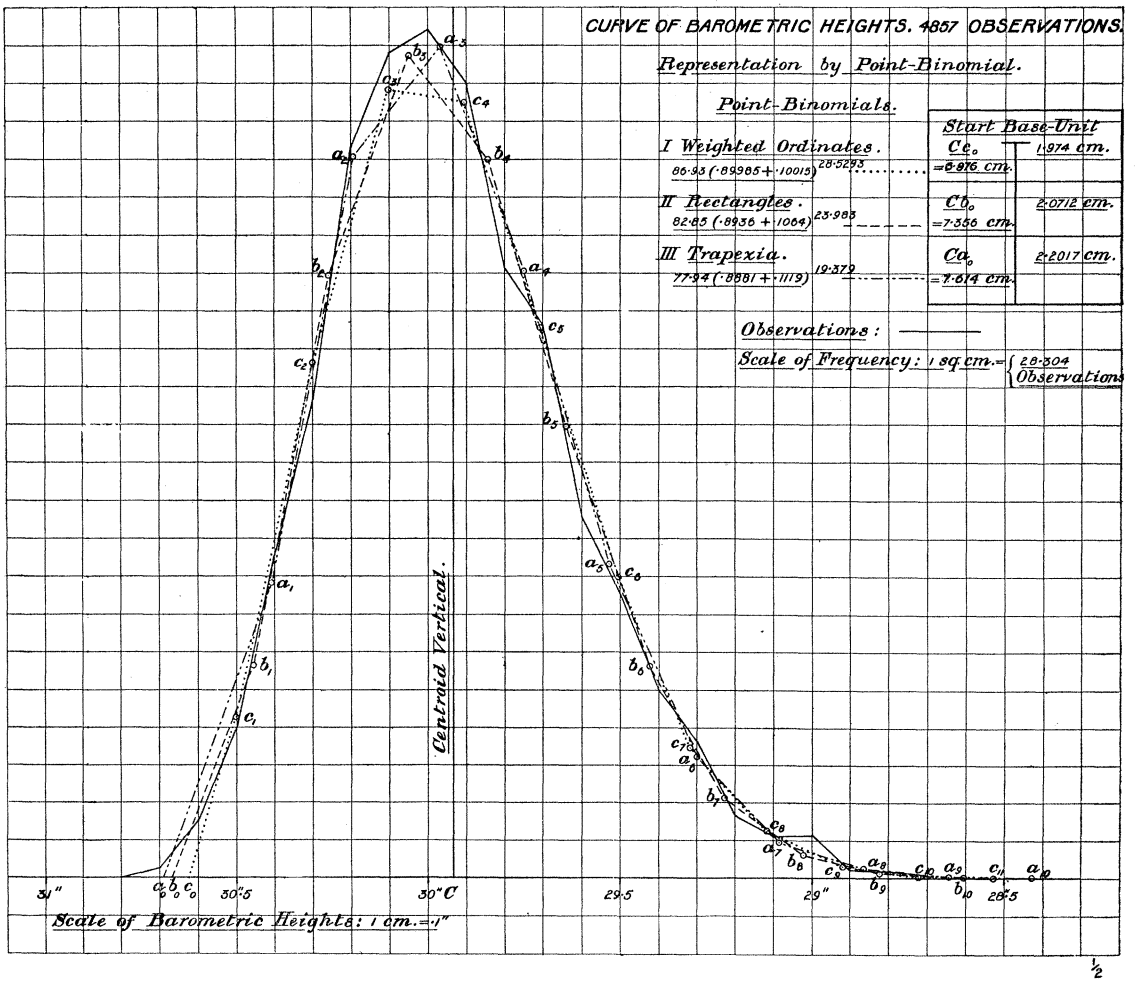


Fig. 4.

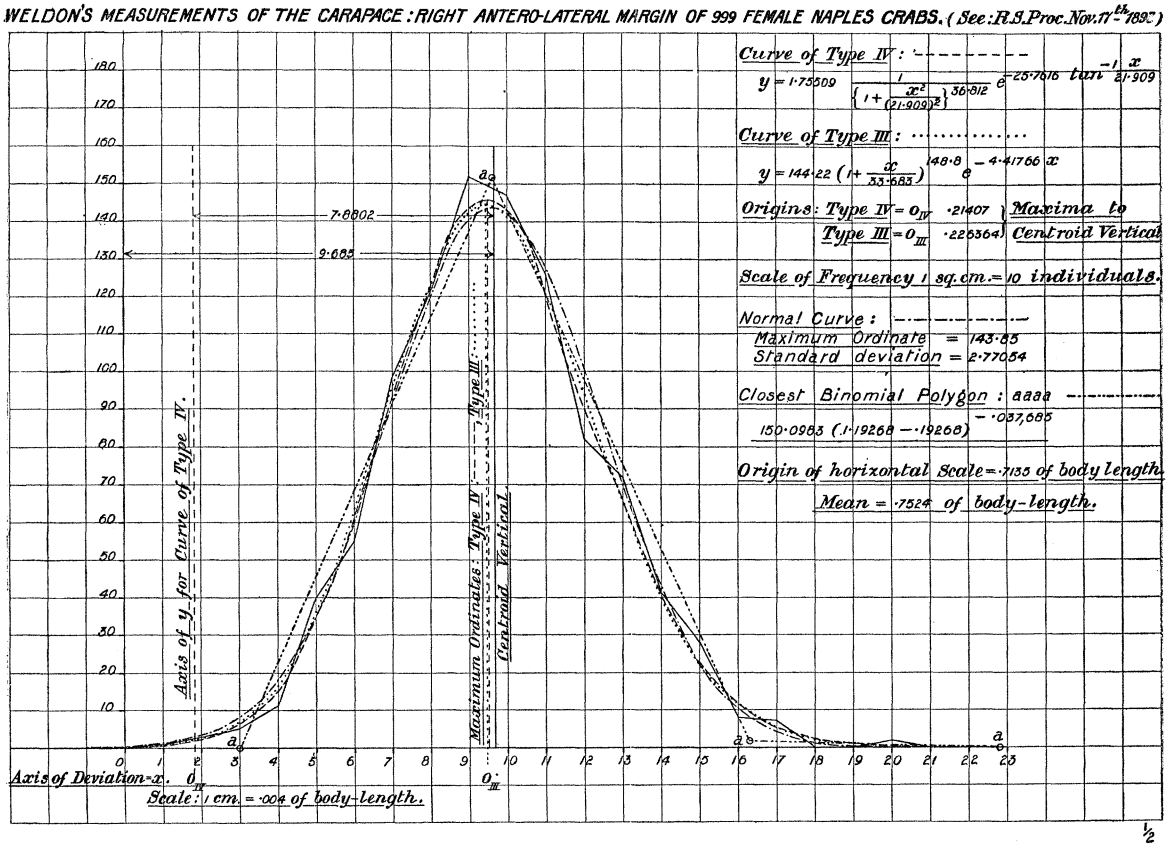


Fig. 5.

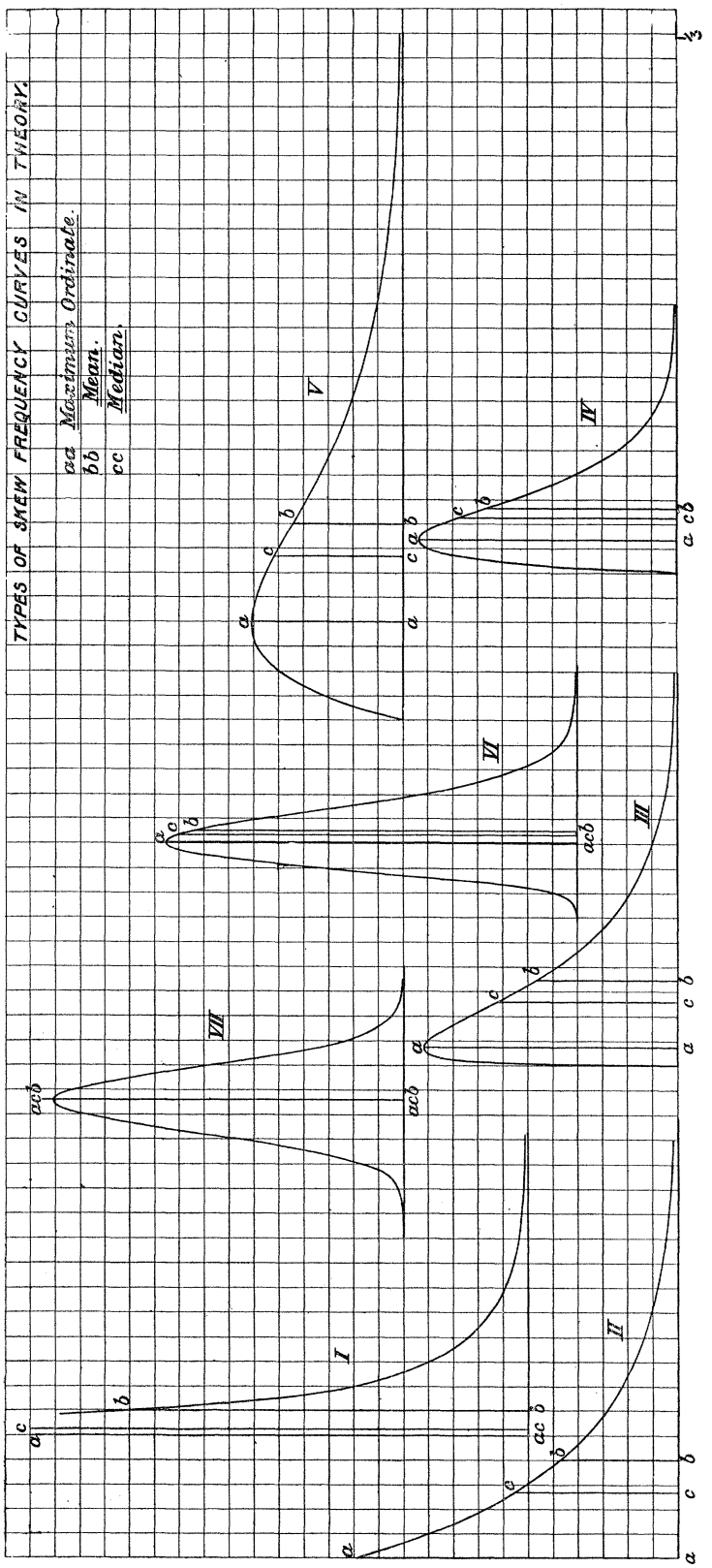


Fig. 19.

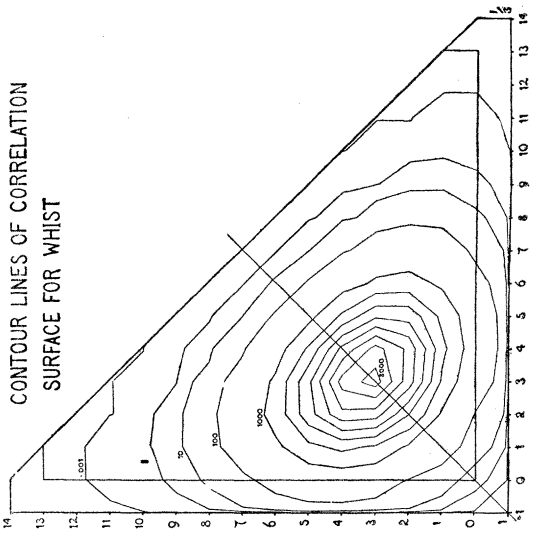


Fig. 6.

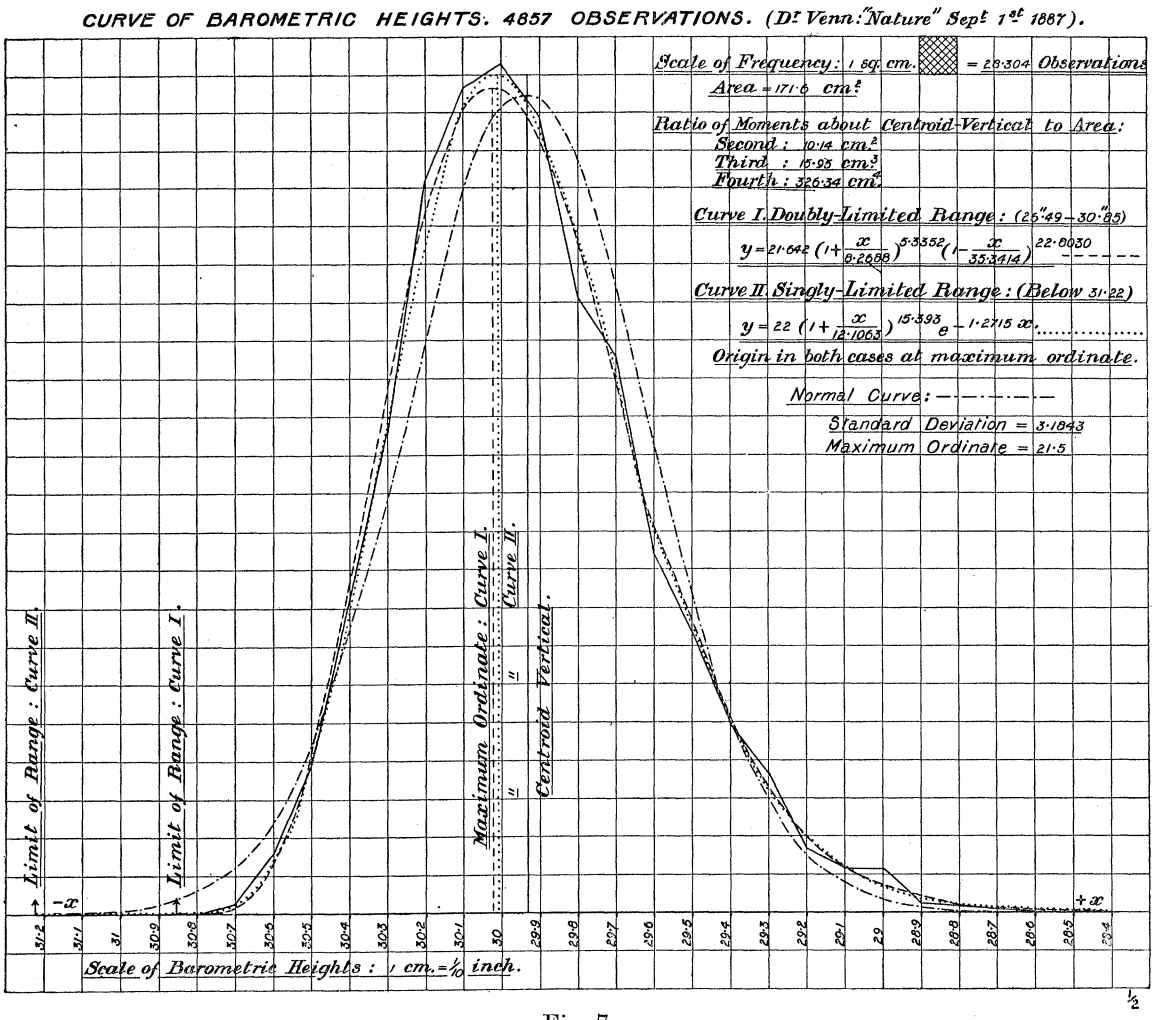


Fig. 7.

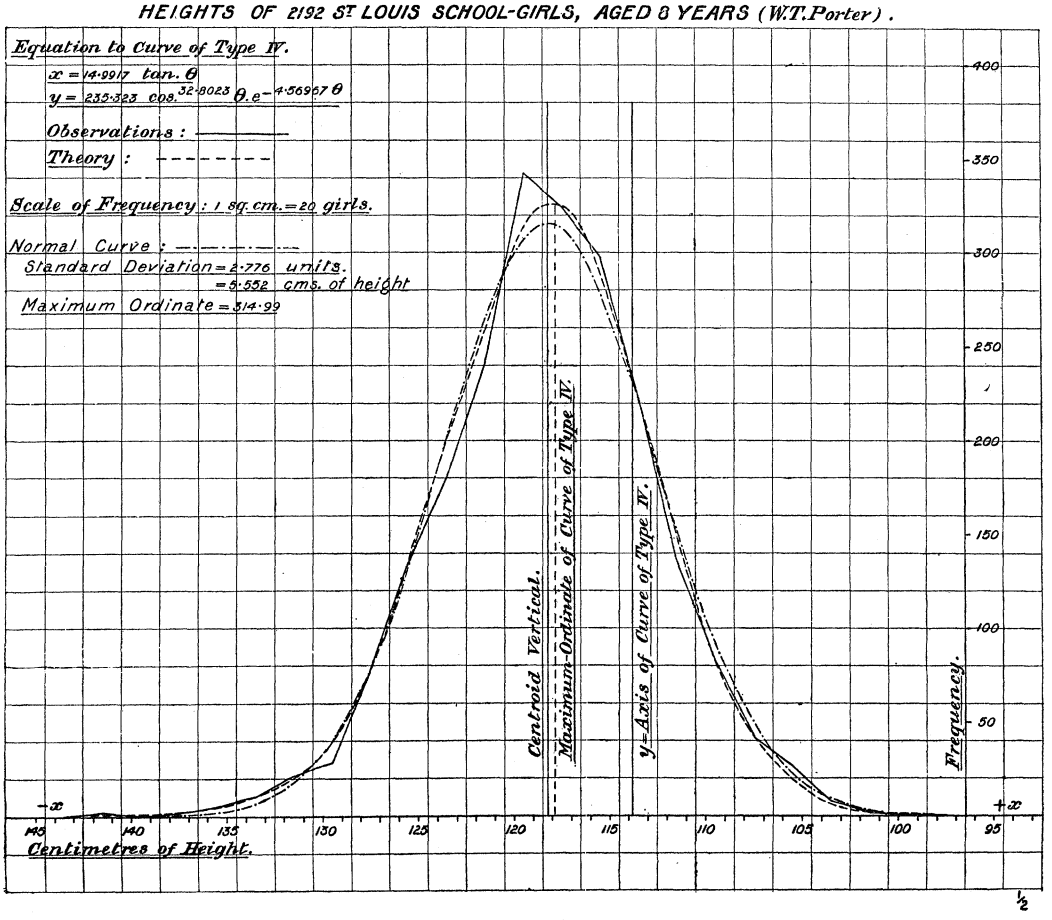


Fig. 8.

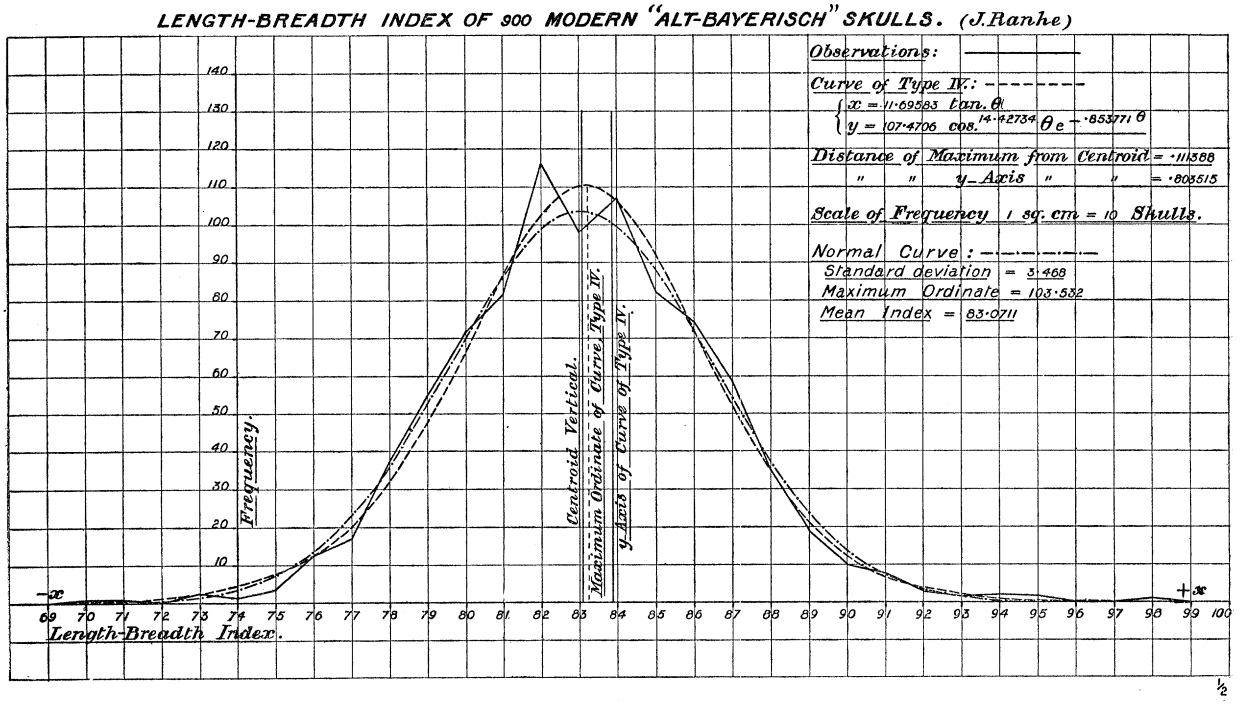


Fig. 10.

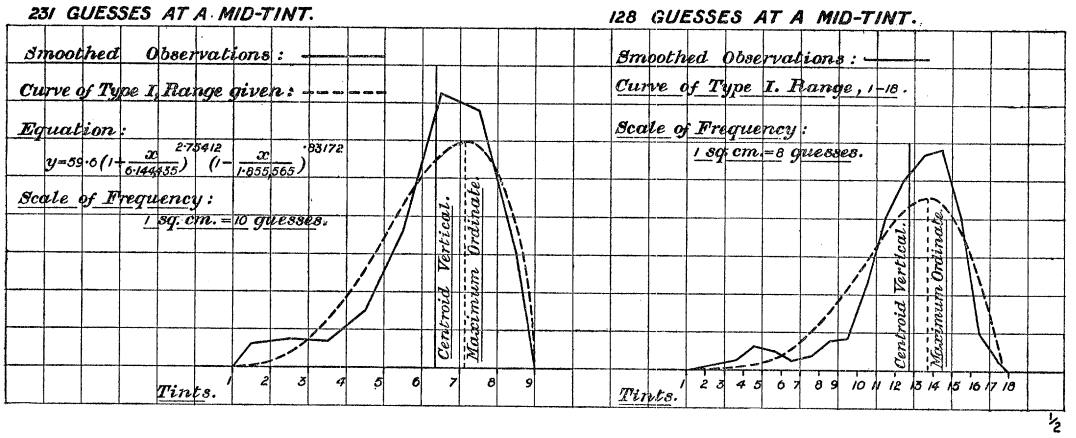


Fig. 12.

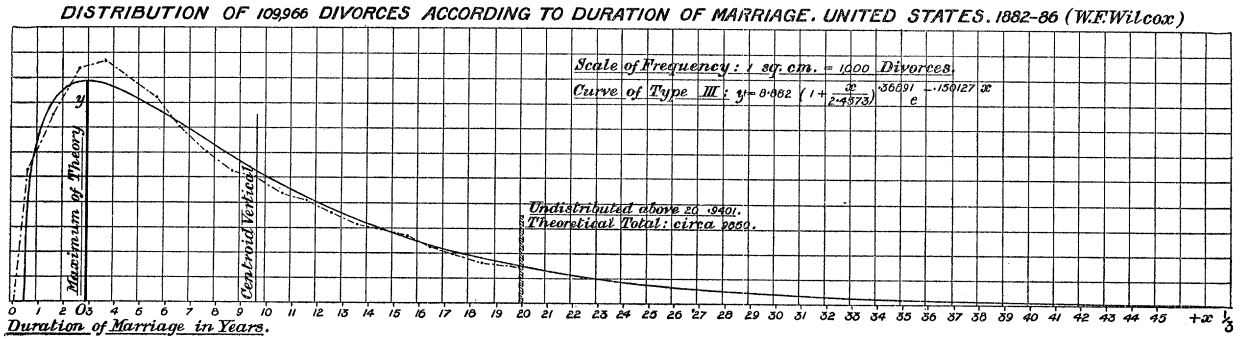


Fig. 9.

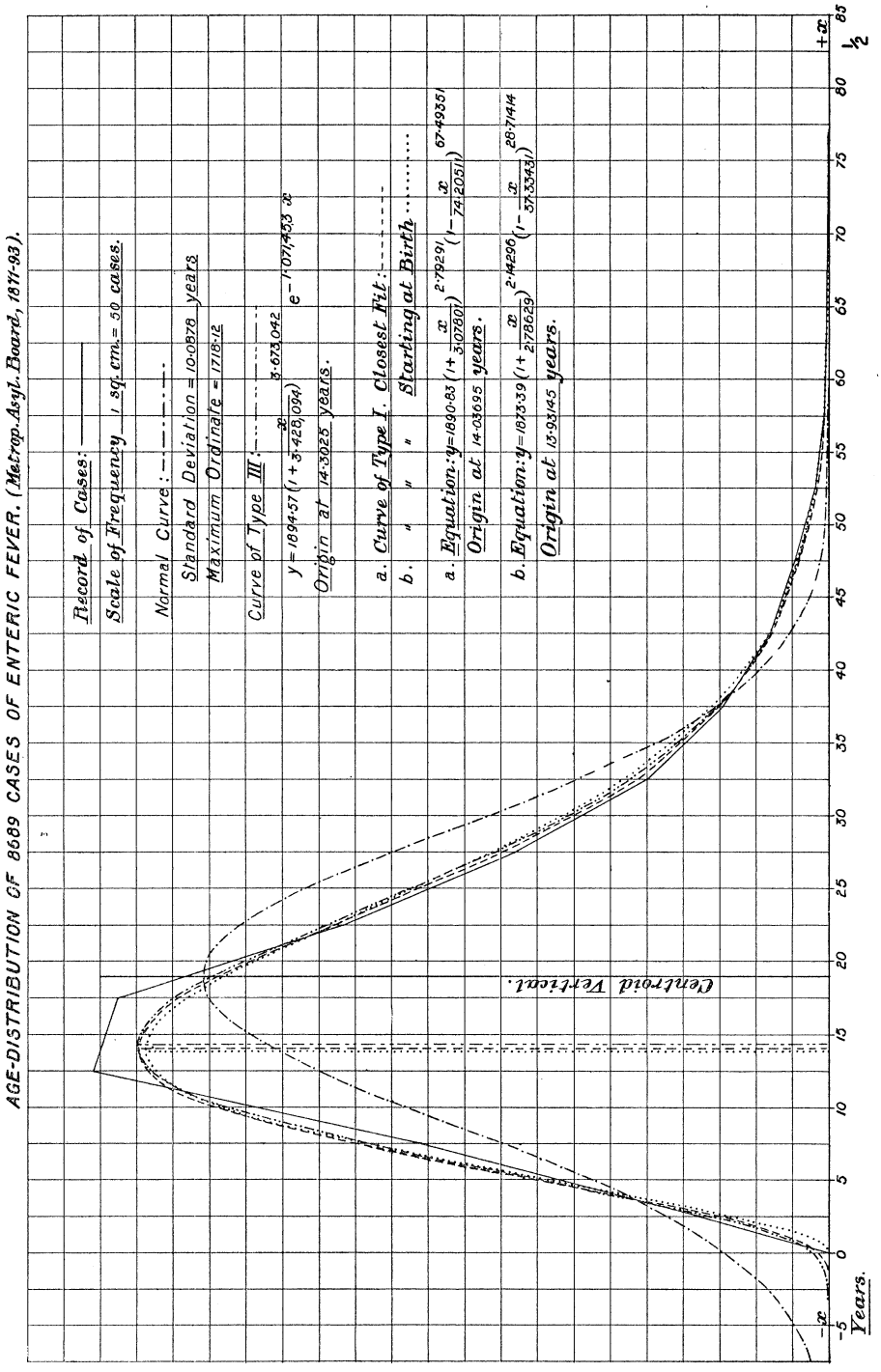


Fig. 13.

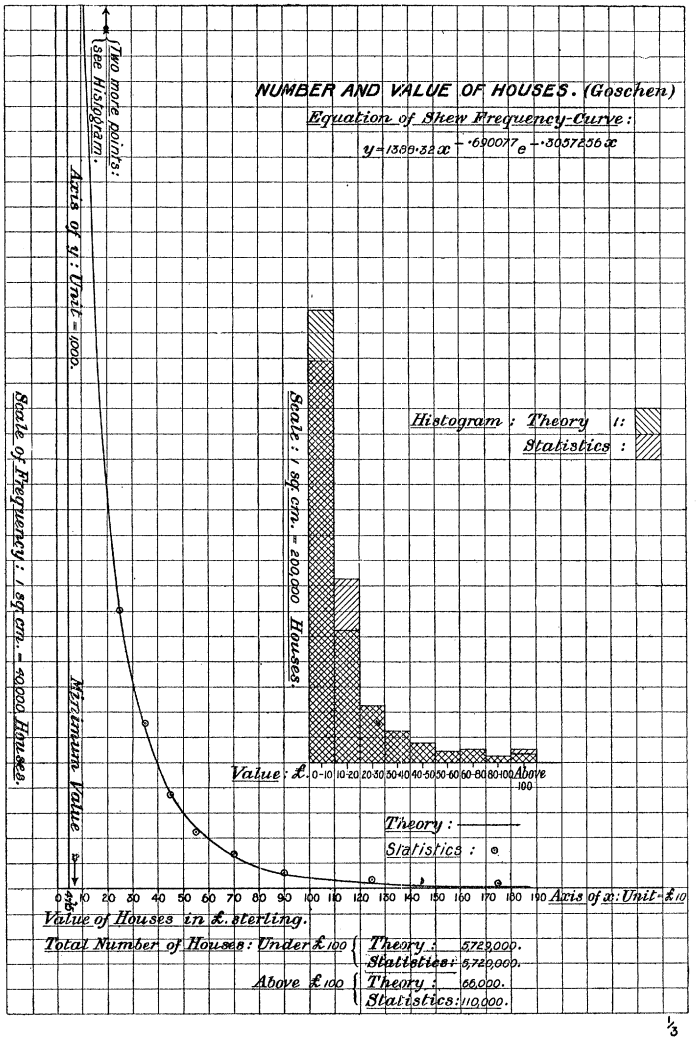


Fig. 17.

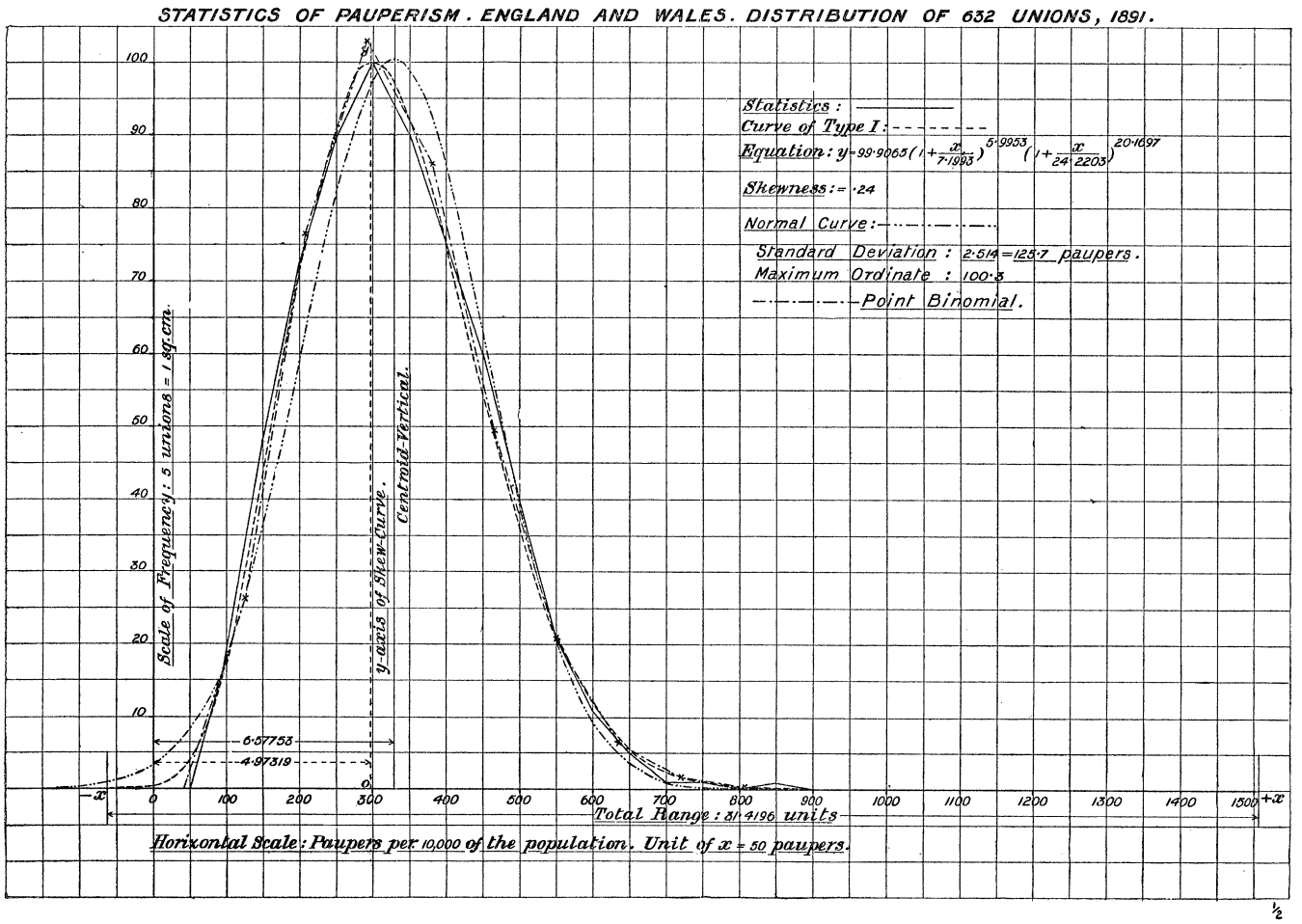


Fig 11..

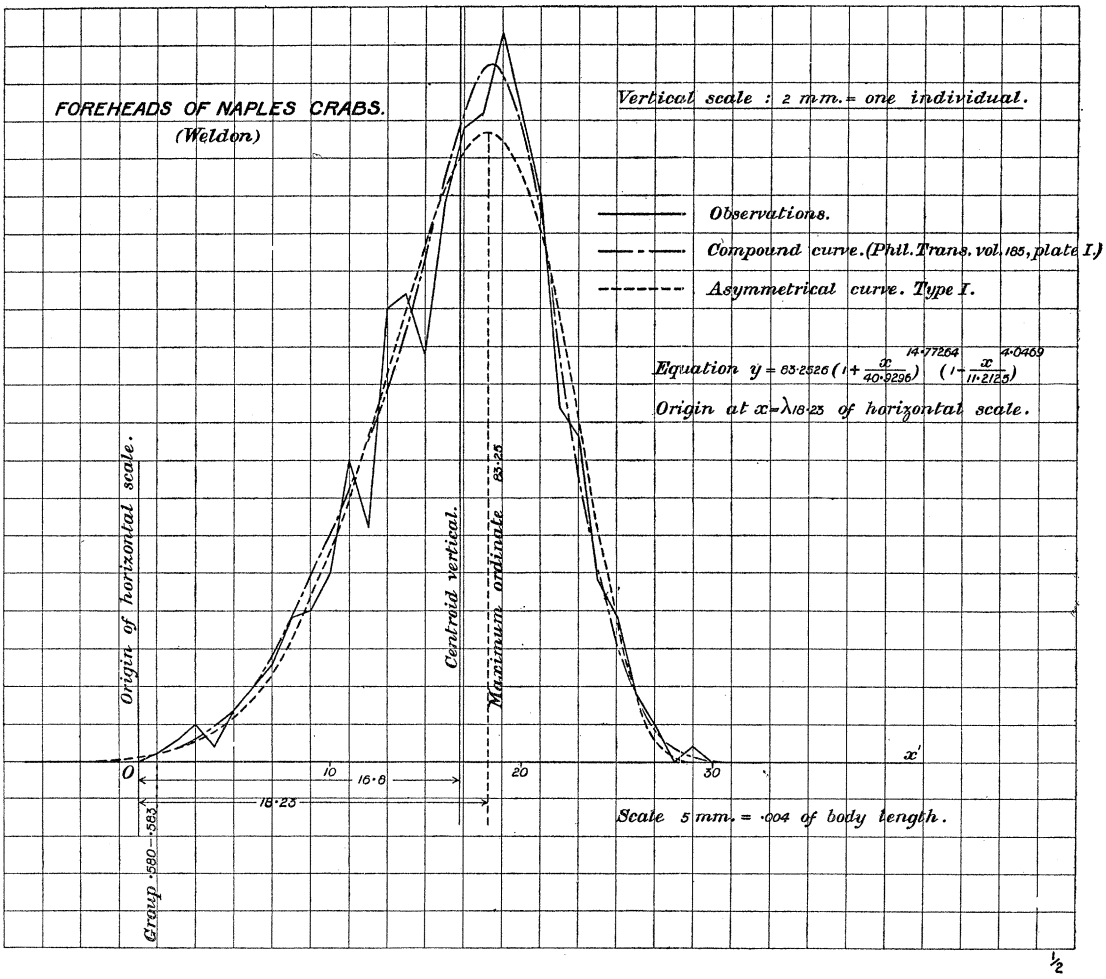


Fig. 16.

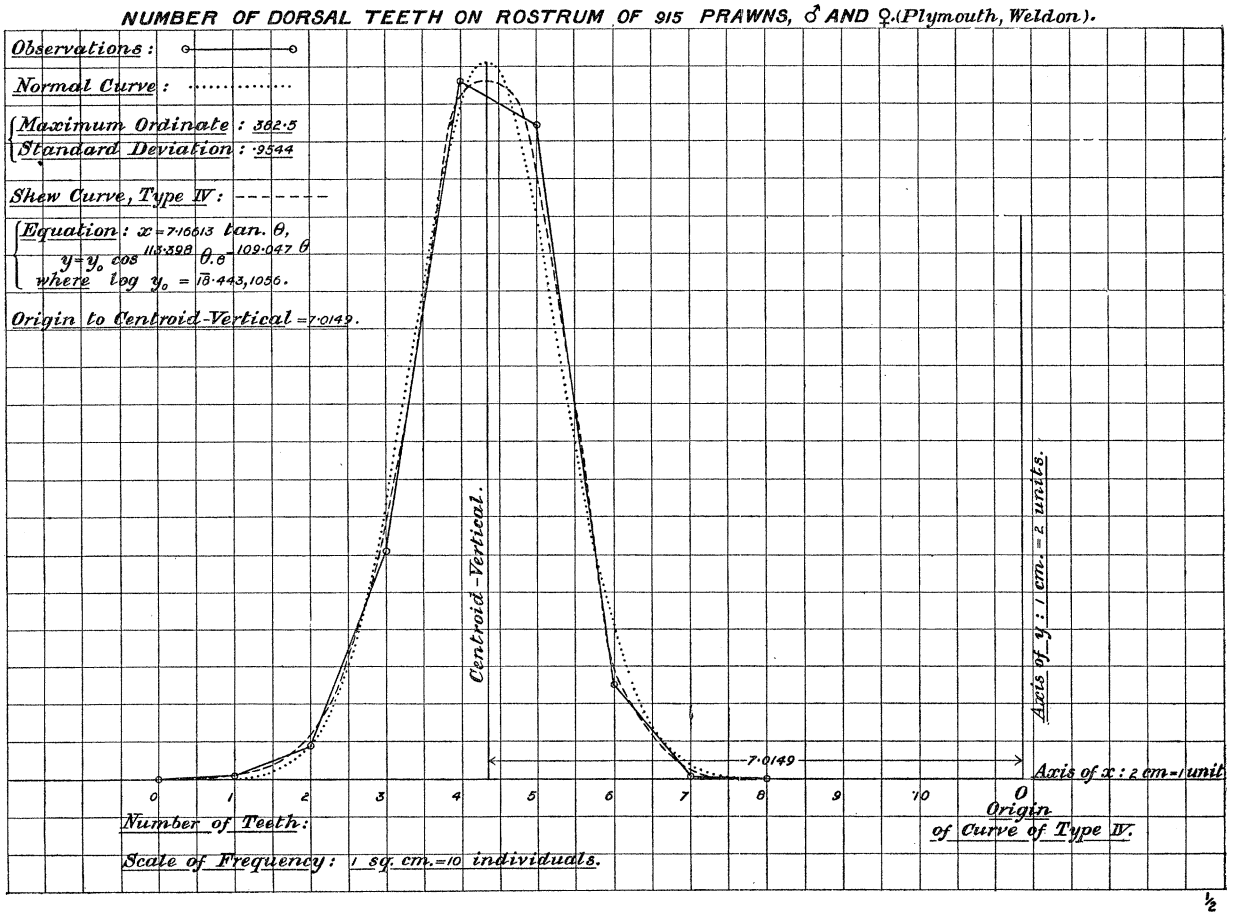


Fig. 15.

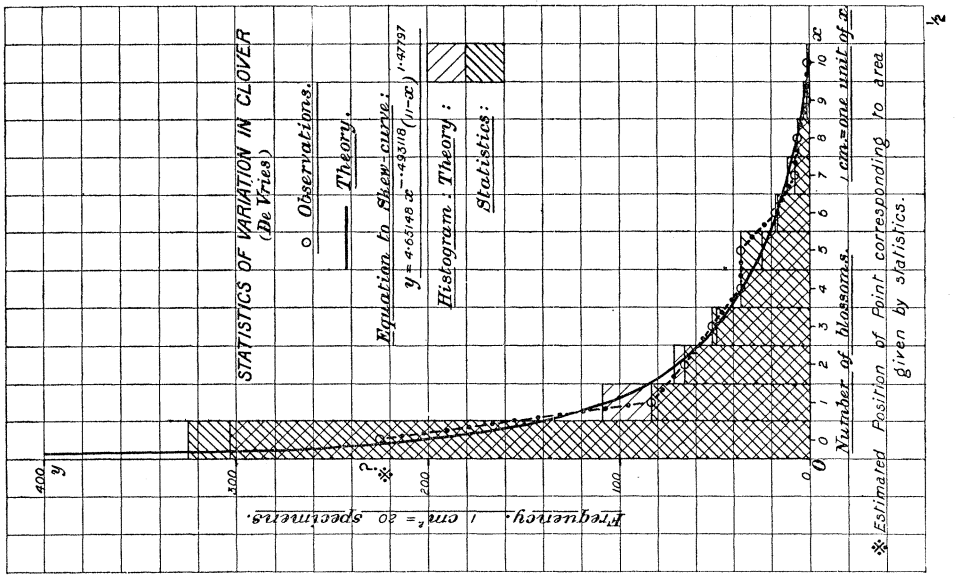


Fig. 14.

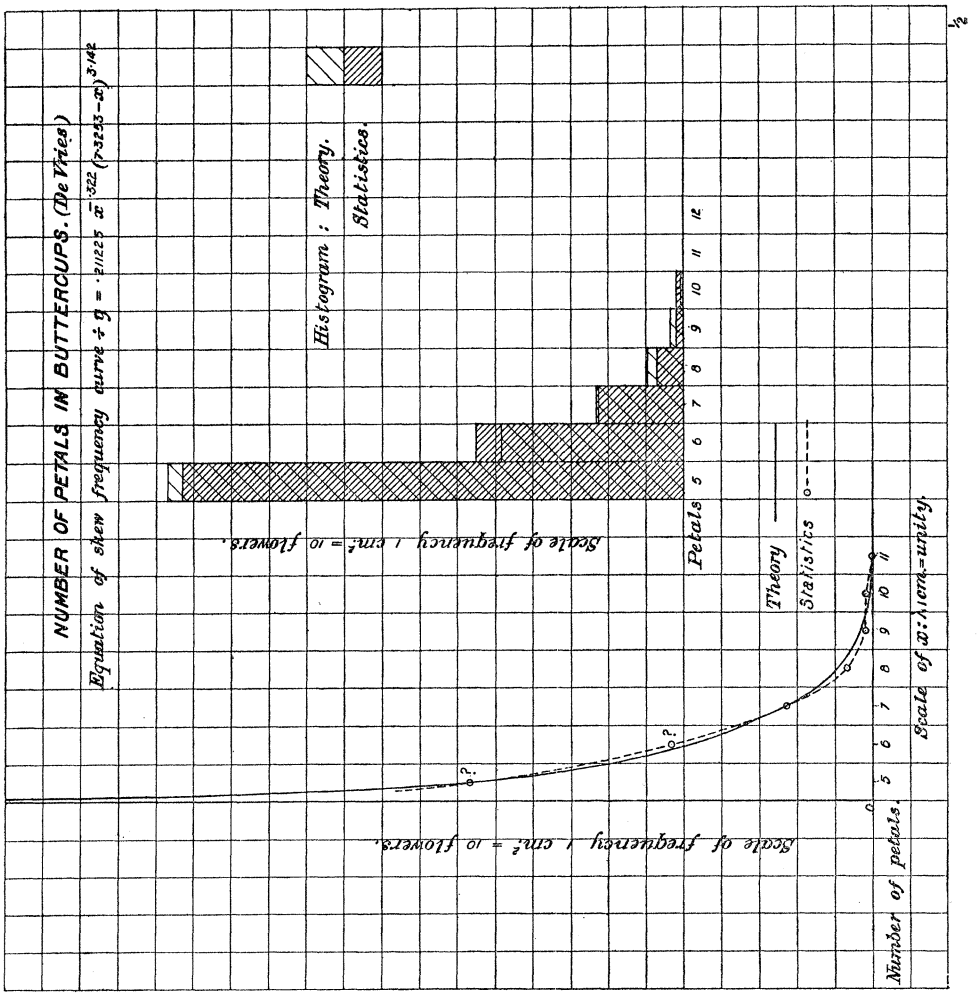


Fig. 18.