[Notes on a series of lectures given at the Glasgow University, January to March 1913].

Examples of Distributions.

(*a*) CONTINUOUS VARIATION.

Stature in Inches.	Symmetrical Distribution. Stature in Man.* Nos. in each Class.		Slightly Asymmetrical. Weight in Man. † Weight in Lbs.	Nos. in each Class.	Markedly Asymmetrical. Age Distribution. ‡ Cases of Scarlet Fever. Age.	No. of Cases.	Rental of Houses. § Rent under	No. of Houses in Thousands.
51-58	2		90-100	2	0-1	246	£10	3175
58-59	4	22	100-110	34	1-2	773	£10-20	1451
59-60	14		110-120	152	2-3	1399		
60-61	41		120-130	390	3-4	1874	£20-30	442
61-62	83	294	130-140	867	4-5	2009		
62-63	169		140-150	1623	5-6	1931	£30-40	260
63-64	394		150-160	1559	6-7	1704	£40-50	151
64-65	669	2053	160-170	1326	7-8	1533	£50-60	90
65-66	990		170-180	787	8-9	1236		
66-67	1223		180-190	476	9-10	1014	£60-80	104
67-68	1329	3782	190-200	263	10-15	2921		
68-69	1230		200-210	107	15-20	921	£80-100	47
69-70	1063		210-220	85	20-25	417		
70-71	646	2101	220-230	41	25-35	327		
71-72	392		230-240	16	35-45	85		
72-73	202		240-250	11	45-	32	Above £100	110
73-74	79	313	250-260	8				
74-75	32		260-270	1				
75-76	16		270-280	...				
76-77	5	23	280-290	1				
77-78	2							
Totals,	8388			7749				

* *Report B.A.* 1883, p. 256. † *Report B.A.* 1883. ‡ *Manchester Health Reports.*
§ Goschen, quoted by Pearson, *Phil. Trans. Roy. Soc.* vol. 186 A, pp. 343-414.

(*b*) DISCRETE VARIATION.

Number of Sepals in Flowers of *Anemone nemorosa.**			Number of Petals in *Ranunculus Bulbosum.* †	
No. of Sepals.	No. of Instances. Example (*a*).	No. of Instances. Example (*b*).	No. of Petals.	No. of Instances.
4	...	3	5	133
5	7	31	6	55
6	515	657	7	23
7	419	271	8	7
8	49	35	9	2
9	13	2	10	2
10	1	1	11	...
11	1	...		
12	...	...		

* Yule, *Biometrika*, vol. i. p. 308. † De Vries, quoted by Pearson, *loc. cit.*

Constants of Distributions.

(1) *Mean.* If there be a number of quantities of definite measurement, then the term *mean* is used to denote the sum of these measurements divided by the total number of the quantities, or if $X_1, X_2, \ldots X_n$ be the measurements, n in number, and M the mean,

$$M = \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n).$$

The term *mean* as used in statistics is equivalent to the term *arithmetical mean* in algebra.

(2) *Median.* The median is the central value of the group when the measurements are arranged in order of magnitude, so that the number of instances above the median is equal to that below the median. If the groups are at all numerous, it is most easily calculated by simple proportion. Thus, taking the weights of British adults, we find 7749 instances. Of these 3068 are under 150 pounds, a defect of 806·5 below the median, and 3122 above 160 pounds, so that the median will be very approximately given by

$$\text{Median} = 150 + \frac{806 \cdot 5}{1559} \times 10 = 155 \cdot 2 \text{ lbs.}$$

The first number is the weight at which the group begins: the multiplier 10 is the value of the group difference, and the fraction the proportional number of the group 1559 to be expected.

(3) *Mode.* This is the most frequent group in asymmetrical distributions, and in symmetrical distributions coincides with the mean and the median. The group in which the mode is situated can usually be easily seen, and if the middle point of this be denoted by zero, the distance of the mode from this can be calculated approximately by the formula

$$\text{Mode} = \frac{m_1 - m_3}{2(m_1 - 2m_2 + m_3)},$$

where m_1, m_2, m_3 are the numbers included in the successive groups, m_2 being that of the group in which the mode is expected.

More accurately, the median in general lies between the mode and the mean, so that its distance from the former is twice that from the latter, that is

$$2(\text{Mean} - \text{Median}) = \text{Median} - \text{Mode},$$

$$\text{or} \quad \text{Mode} = 3 \times \text{Median} - 2 \times \text{Mean}$$

$$= 150 \cdot 8 \text{ lbs. in the case previously considered.}$$

By the formula given, $\text{Mode} = 145 + \frac{-792}{-2 \times 822} \times 10 = 149 \cdot 2 \text{ lbs.},$

or one per cent. of difference.

(4) *Standard deviation.* This expresses the degree of scatter in the distribution: thus, take two distributions (*a*) and (*b*), as shown in the table below, of equal numbers, having the same mean. Let the measurements be 1, 2, 3, 4, 5.

Size.	Number.		Deviations.	No. × sqr. of Deviations.	
	a	*b*		(*a*)	(*b*)
1	...	1	−2	...	4
2	4	4	−1	4	4
3	8	6	0	0	0
4	4	4	1	4	4
5	...	1	2	...	4
Totals,	16	16		8	16

Both these groups have the same mean measuring 3, but the variation in the latter is much greater than in the former. This difference is measured by the standard deviation, which is defined as the square root of the mean of the squares of all the deviations, the latter being measured from the mean value of the quantity. In the above example the mean is at 3, therefore the deviations are as given in the fourth column. In the fifth and sixth columns the square of the deviations are multiplied by the number of each group occurring. The sums are 8 and 16, so that the respective standard deviations are

$$\sqrt{\frac{8}{16}} \text{ and } \sqrt{\frac{16}{16}}, \text{ or } \frac{\sqrt{2}}{2} \text{ and } 1.$$

The standard deviation is usually denoted by σ.

(5) *Skewness.* This measures the degree of asymmetry of a distribution, and is defined as the ratio of the distance between the mode and the mean divided by the standard deviation,

$$i.e. \quad \text{Sk.} = \frac{\text{Mode} - \text{Mean}}{\text{Standard Deviation}} = \frac{d}{\sigma} \text{ in the usual notation.}$$

(6) *Coefficient of Variation.* This coefficient is defined as 100 times the ratio of the standard deviation to the mean and is denoted by v, so that

$$v = 100\frac{\sigma}{M}.$$

This coefficient allows comparisons to be made; for a variation of 2 inches above or below the mean is very much greater when the mean is equal to 10 inches than when it is equal to 40 inches.

Method of Calculating the Mean and Standard Deviation of a Series of Observations.

EXAMPLE. Theoretical number of aces where three cards, one of which is an ace, are dealt in groups of seven.

No. of Aces.	No. of instances x.	Deviation from chosen zero f.	$x \times f$	$x \times f^2$	$x \times f^3$	$x \times f^4$
0	128	−2	−256	512	−1024	2048
1	448	−1	−448	448	− 448	448
			−704	960	−1472	2496
2	672	0				
3	560	1	560	560	560	560
4	280	2	560	1120	2240	4480
5	84	3	252	756	2268	6804
6	14	4	56	224	896	3584
7	1	5	5	25	125	625
Total, -	2187		1433	2685	6089	16053
			−704	+960	−1472	+2496
			729	3645	4617	18549

First choose by inspection an origin at a point as near as possible to the mean. Then ν_1, ν_2, ν_3, ν_4 denote the first, second, third, and fourth moments round this chosen origin, so that

$$\nu_1 = \frac{\Sigma x \times f}{N}, \quad \nu_2 = \frac{\Sigma x \times f^2}{N}, \quad \text{etc.,}$$

where N denotes the total number of instances.

In this case,

$$\nu_1 = \frac{729}{2187} = \frac{1}{3}, \quad \nu_2 = \frac{3645}{2187} = \frac{5}{3}, \quad \nu_3 = \frac{4617}{2187} = \frac{19}{9}, \quad \nu_4 = \frac{18549}{2187} = \frac{229}{27}.$$

Here $\nu_1 = \frac{1}{3}$, *i.e.* the Mean is at a distance $+\frac{1}{3}$ from the chosen origin, or at $2 + \dot{3}$ or $2\cdot\dot{3}$ units from the real zero.

The moments about the mean are the most important, and are denoted by μ_1, μ_2, μ_3, μ_4; μ_1 being equal to zero by definition. These are obtained from the preceding moments by the formulae:

$$\mu_2 = \nu_2 - \nu_1^2, \quad \text{(very important)}$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3,$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4.$$

In this case,

$$\mu_2 = \frac{5}{3} - \left(\frac{1}{3}\right)^2 = \frac{14}{19},$$

$$\mu_3 = \frac{19}{9} - 3 \cdot \frac{1}{3} \cdot \frac{5}{3} + 2\left(\frac{1}{3}\right)^3 = \frac{14}{27},$$

$$\mu_4 = \frac{229}{27} - 4 \cdot \frac{1}{3} \cdot \frac{19}{9} + 6\left(\frac{1}{3}\right)^2 \cdot \frac{5}{3} - 3 \cdot \left(\frac{1}{3}\right)^4 = \frac{182}{27}.$$

The standard deviation σ is equal to the $\sqrt{\mu_2}$.

I. Correlation.

EXAMPLE. Typical Correlation Table showing the relationship of the stature of Fathers and Sons in inches.*

		STATURE OF FATHER.						
		58·5–61·5	61·5–64·5	64·5–67·5	67·5–70·5	70·5–73·5	73·5–76·5	Totals
STATURE OF SON.	58·5–61·5	-6 — ·05	-3 **1·5** ·36	0 **2** 1·20	3 — 1·32	6 — ·50	9 — ·03	**3·5**
	61·5–64·5	-4 **3·5** ·84	-2 **19** 6·50	0 **33** 21·75	2 **5·5** 23·87	4 **1·5** 9·02	6 — ·55	**62·5**
	64·5–67·5	-2 **8·5** 4·02	-1 **53·75** 31·07	0 **148** 104·03	1 **80·5** 114·15	2 **8·25** 43·14	3 — 2·64	**299**
	67·5–70·5	0 **2·5** 6·07	0 **33·25** 46·86	0 **149·25** 156·90	0 **202·25** 172·17	0 **60·25** 65·06	0 **3·5** 3·97	**451**
	70·5–73·5	2 — 2·87	1 **3·5** 22·13	0 **39·75** 74·10	1 **104·25** 81·31	2 **62·0** 30·73	3 **3·5** 1·88	**213**
	73·5–76·5	4 — ·56	2 **1** 4·31	0 **3** 14·44	2 **14·5** 15·84	4 **20·5** 5·99	6 **2·5** ·37	**41·5**
	76·5–79·5	6 — ·10	3 — ·77	0 — 2·59	3 **4·5** 2·84	6 **3·0** 1·07	9 — ·07	**7·5**
	Totals	**14·5**	**112**	**375**	**411·5**	**155·5**	**9·5**	**1078**

NOTE. The numbers above the figures in the middle of each square are the products of the deviations from the chosen origin.

Coefficient of Correlation is defined by $r=\frac{\Sigma xy}{\sigma_1\sigma_2}$, where Σxy is equal to the sum of all the observations multiplied by the product of their deviations from the mean in the vertical and horizontal directions. It is termed the *product moment*.

If $h_1 h_2$ be the distances of the mean from the chosen origin, and $\Sigma x'y'$ the product moment round that origin, $\Sigma xy = \Sigma x'y' - Nh_1h_2$.

* Pearson, *Biometrika*, vol. ii. p. 415.

II. Contingency.

In the above table, beneath each figure is printed in smaller characters a figure which shows the number of cases to be expected if there were no relationship between the stature of fathers and sons. These are obtained by taking the total of each horizontal series and dividing it in the same proportions as are given in the horizontal series showing the totals in each column.

To obtain the coefficient of contingency, take the difference of each theoretical number from the corresponding actual number, square this difference, and divide by the theoretical number.

Thus, in the first row and third column, the theoretical number is 1·20 and the actual is 2, whence we have $\frac{(-\cdot 8)^2}{1\cdot 2} = \cdot 53$.

All the numbers found in this way are summed. The total is denoted by χ^2. This total divided by N, the total number of observations, is further denoted by ϕ^2, whence we have the coefficient of contingency

$$r = \sqrt{\frac{\phi^2}{1+\phi^2}}.$$

III. Fourfold Division Method.

EXAMPLE. Smallpox and Vaccination. Sheffield, 1887–88.*

	Recoveries.	Deaths.	Totals.
Vaccinated -	3951	200	4151
Unvaccinated -	278	274	552
Totals -	4229	474	4703

If the same method as is shown in paragraph 1 above is used, and if the fourfold division be

a	b
c	d

$$r = \frac{ad - bc}{\{(a+b)(b+c)(c+d)(d+a)\}^{\frac{1}{2}}}.$$

No method, however, applied to this fourfold division is satisfactory.

Pearson's fourfold division method gives

$$r = \cdot 77.$$

*Macdonell, *Biometrika*, vol. i. p. 376.

Examples of Correlation.

II. AGES AT MARRIAGE OF BACHELORS AND SPINSTERS, ENGLAND AND WALES, 1901.

AGES OF BACHELORS.	AGES OF SPINSTERS.														
	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80	Totals	Mean
15-20	2,606	1,356	75	4	2	—	—	—	—	—	—	—	—	4,043	19·39
20-25	14,821	73,430	12,989	1,110	123	20	4	—	—	—	—	—	—	102,497	22·54
25-30	2,785	37,317	33,229	5,249	648	76	9	2	—	—	—	—	—	79,315	25·23
30-35	482	6,657	10,184	5,908	1,217	182	24	4	—	—	1	—	—	24,659	27·78
35-40	103	1,317	2,545	2,246	1,432	326	57	3	2	1	—	—	—	8,032	30·51
40-45	15	322	594	726	631	427	96	17	2	—	—	—	—	2,830	33·50
45-50	3	67	158	221	250	206	112	19	6	1	—	—	—	1,043	36·28
50-55	3	3	35	73	86	87	67	39	11	2	—	—	—	406	40·31
55-60	—	—	9	22	32	24	30	17	14	5	—	—	—	153	43·25
60-65	—	1	6	8	6	12	11	5	9	9	—	—	—	67	45·50
65-70	—	—	2	2	3	1	5	3	4	2	1	—	—	23	47·50
70-75	—	—	—	—	1	1	—	—	1	1	1	2	1	8	61·35
75-80	—	—	1	—	—	1	—	1	—	—	—	1	1	5	54·50
Totals	20,818	120,470	59,827	15,569	4,431	1,363	415	110	49	21	3	3	2	223,081	24·56
Mean	22·87	24·78	27·91	32·36	35·88	40·74	45·59	50·36	55·76	59·4	57·1	67·3	75·0	26·38	

II. CORRELATION OF NUMBER OF MÜLLERIAN GLANDS ON THE RIGHT AND LEFT LEGS OF 2000 SWINE (DAVENPORT).

NUMBER OF GLANDS ON LEFT LEG.	NUMBER OF GLANDS ON RIGHT LEG.											
	0	1	2	3	4	5	6	7	8	9	10	Totals
0	8	4	2	—	—	—	—	—	—	—	—	14
1	5	151	65	14	5	1	—	—	—	—	—	241
2	2	58	154	88	27	7	—	—	—	—	—	336
3	—	9	96	173	119	24	8	1	—	—	—	430
4	—	3	28	128	153	92	16	8	1	—	—	429
5	—	—	7	28	77	101	58	20	3	1	—	295
6	—	—	1	6	26	52	48	18	5	3	—	159
7	—	—	—	—	3	11	16	17	3	3	—	53
8	—	—	—	—	1	9	7	9	2	2	—	30
9	—	—	—	—	—	—	—	5	2	2	1	10
10	—	—	—	—	—	—	2	—	—	1	—	3
Totals	15	225	353	437	411	297	155	78	16	12	1	2000
Mean	·60	1·36	2·31	3·20	3·89	4·78	5·51	6·14	6·50	7·33	9·00	

Mean No. of Glands : right leg = 3·55 ; left leg = 3·54. Standard Deviation : right leg = 1·72 ; left leg = 1·73.
Correlation : $r = 0{\cdot}792$.

Partial Correlation Coefficients.

IF there be three variables which are all correlated with one another, and if $r_{1.2}$, $r_{2.3}$, $r_{1.3}$ denote the correlations of each pair, the *partial* correlation coefficients are given by

$$r_{1.2:3} = \frac{r_{1.2} - r_{1.3} \times r_{2.3}}{\sqrt{1 - r_{1.3}^2} \times \sqrt{1 - r_{2.3}^2}}, \text{ etc.};$$

whence $r_{1.2:3}$ denotes the correlation between the first and second variables when the third is constant.

EXAMPLE. Correlation between the amount of Summer Diarrhœa, the Mean Temperature in July and the Rainfall in the same month.

TABLE I.—Death rate from Diarrhœa per million per year in London, and the Mean Temperature of July (Greenwich).

DEATH RATE FROM DIARRHŒA.	TEMPERATURE—JULY.					
	57°–59°	59°–61°	61°–63°	63°–65°	65°–67°	Totals.
300–500	–	2	2	–	–	4
500–700	2	6	3	1	–	12
700–900	–	1	1	2	4	8
900–1100	–	–	2	5	4	11
Totals, -	2	9	8	8	8	35

$r_{T.D} = \cdot 650.$

TABLE II.—Correlation between Temperature and Rainfall.

TEMPERATURE.	RAINFALL IN INCHES.							
	0–1	1–2	2–3	3–4	4–5	5–6	6–7	Totals.
57°–59°	–	–	–	1	–	–	1	2
59°–61°	1	2	2	3	1	–	–	9
61°–63°	–	1	2	4	–	1	–	8
63°–65°	3	2	1	1	–	1	–	8
65°–67°	2	3	2	–	–	–	1	8
Totals, -	6	8	7	9	1	2	2	35

$$r_{T.R} = -\cdot 395.$$

TABLE III.—Correlation Diarrhœa and Rainfall.

DEATH RATE FROM DIARRHŒA.	RAINFALL IN INCHES.							
	0–1	1–2	2–3	3–4	4–5	5–6	6–7	Totals.
300–500	1	–	–	3	–	–	–	4
500–700	1	2	3	3	1	1	1	12
700–900	1	4	1	2	–	–	–	8
900–1100	3	2	3	1	–	1	1	11
	6	8	7	9	1	2	2	35

$$r_{D.R} = -\cdot 295.$$

PARTIAL CORRELATIONS.

$$r_{D.R:T} = \frac{-\cdot 295 - (\cdot 650) \times (-\cdot 395)}{\sqrt{1-\cdot 650^2}\sqrt{1-\cdot 395^2}}$$

$$= -\cdot 070.$$

$$r_{D.T:R} = \frac{\cdot 650 - (-\cdot 295)(-\cdot 395)}{\sqrt{1-\cdot 650^2}\sqrt{1-\cdot 395^2}}$$

$$= \cdot 643.$$

Probable Error.

Example of the variation in groups of small numbers. Number of deaths in parallel series of fifties among the admissions of patients suffering from scarlet fever to Belvidere Hospital, 1900-1908.

1, –	2, 1	3, 2	2, 1	1, 4	2, 1
2, –	5, 4	3, 3	1, 2	–, 3	3, 1
1, 3	1, 3	1, 4	3, 2	3, 2	4, 5
1, 3	3, 2	3, 3	3, –	2, 3	7, 4
3, –	1, 4	1, 3	2, 6	–, 1	2, 1
3, –	4, 1	2, 1	2, 2	2, 2	2, 1
3, 1	3, 2	3, 2	–, 1	–, 1	1, 2
1, –	2, 7	3, 4	2, 4	1, 1	

Table showing the distributions of these figures when classified.

No. of Deaths.	No. of Times observed.	Theoretical Number.	Difference.	Difference squared divided by the Theoretical Number.
0	10	10·4	·4	·00
1	25	23·4	1·6	·11
2	23	25·8	2·8	·30
3	22	18·7	3·3	·59
4	9	9·8	·8	·07
5	2	4·0	2·0	1·00
6 7	1 2 } 3	1·3 ·4 } 1·7	1·3	1·00
Total -	94	93·8	—	$3{\cdot}07 = \chi^2$

$$M = 2{\cdot}16 \text{ deaths}, \quad \sigma = 1{\cdot}437, \quad \frac{\sigma}{\sqrt{N}} = {\cdot}148.$$

If a quantity be measured, then the *probable error* of that quantity is defined by the limits within which it is equally likely that the quantity would in a long series of measurements be found to lie inside and outside.

If σ be the standard deviation of the quantity, the limits above defined are given by $\pm {\cdot}674\sigma$.

It is better to take, however, σ itself and call it the "standard error."

If $\pm 2\sigma$ are taken as the limits of error, then the odds are 21 to 1 that the real value of the quantity lies between these limits.

Standard Errors.

(1) Of the mean h of a number N of quantities $= \dfrac{\sigma}{\sqrt{N}}$.

(2) Of a correlative coefficient r when N observations has been made $= \dfrac{1 - r^2}{\sqrt{N}}$.

Test of Goodness of Fit between Theory and Observation.

Take each group, subtract the actual from the theoretical value, square, divide by the theoretical value, and sum. The sum is denoted by χ^2.

Find P or the probability that in a certain number of trials more difference between theory and observation would be found. A short table of this function giving the values of χ^2 and the number of groups compared, N is printed below.

EXAMPLE.

Table showing Days of Sickening in 907 Cases of Scarlet Fever.

	No. of Cases.	Theoretical Value.	Difference.	(Difference)^2.	(Difference)^2. / Theoretical Value.
Sunday - - -	124	129·6	−5·6	31·36	·24
Monday - -	143	129·6	13·4	179·56	1·38
Tuesday - -	117	129·6	−12·6	158·76	1·22
Wednesday - -	134	126·6	4·4	19·36	·15
Thursday - -	120	129·6	−9·6	92·16	·71
Friday - - -	143	129·6	13·4	179·56	1·38
Saturday - -	126	129·6	−3·6	12·96	·10
Total - -	907	907		673·72	5·18

Thus $\chi^2 = 5{\cdot}18$ or $P = {\cdot}522$ or in half the trials made as much divergence would be found.

Table showing the Values of χ^2 for certain Values of P and N.

N	Values of P.								
	·9	·8	·7	·6	·5	·4	·3	·2	·1
3	—	—	·7	1·0	1·2	1·8	2·4	3·4	4·6
4	—	1·00	1·4	1·9	2·4	2·9	3·6	4·6	6·1
5	1·0	1·7	2·2	2·8	3·4	4·1	4·9	5·9	7·7
6	1·6	2·3	3·0	3·6	4·3	5·1	6·0	7·3	9·2
7	2·2	3·0	3·8	4·6	5·3	6·2	7·2	8·6	10·5
8	2·9	3·8	4·7	5·4	6·3	7·3	8·4	9·7	12·0
9	3·5	4·6	5·5	6·4	7·4	8·4	9·5	11·0	13·2
10	4·2	5·4	6·4	7·4	8·4	9·4	10·6	12·2	14·6
12	5·6	7·0	8·2	9·3	10·4	11·5	12·8	14·5	17·2
14	7·0	8·7	9·9	11·1	12·3	13·6	15·1	16·9	19·6
16	8·6	10·3	11·8	13·0	14·4	15·7	16·3	19·3	22·3

THE THEORY OF CHANCE DISTRIBUTION AS APPLIED TO BIOLOGY.

by

John Brownlee, M.D.

THE THEORY OF CHANCE DISTRIBUTION
AS APPLIED TO BIOLOGY.

In recent years a large amount of work has been done regarding the forms of distribution which occur in biological measurements, but most of this work has been inductive and most of the methods proposed, though permitting elegant mathematical developments, throw little light on the causation of different types of distribution. The points which are specially important from the biological side have been much neglected. Few attempts have been made to enquire into the reasons which determine the distributions observed, to estimate how far they represent some real vital factor or to ascertain the extent to which they are "artefacto", that is, the result of the application of some particular standard of measurement.

I propose, then, in this paper to discuss frequency distributions with a view to ascertain how types of distribution arise and, when such types have arisen, to see how far they conform to the results of biological observation and to examine how far the reverse problem of reasoning from a curve to a biological process can be justified.

Chance distributions have, for a couple of centuries, been the subject of much discussion. The theory was first put on a scientific basis by Laplace and Gauss. Both of these reached by somewhat different processes the curve now

known as the normal curve of error, a curve which has been found to give an adequate description to many of the measurements made in the organic and inorganic world. Further developments to allow for asymmetry have been made by McAllister, Edgeworth and Pearson in this country, and by Kapteyn, Fechner and Thiele abroad. These in different ways arrive at forms of curves which in many cases very closely describe groups of statistics which have asymmetrical distributions. In spite of this fact, I feel that the initial reasoning shows far too little consideration to the requirements of biology. Though the mathematics is fascinating, it cannot but be felt, that no clear idea can be formed as to th the meaning the initial hypotheses have in the world of life.

In seeking a foundation hypothesis I think that one can d do no better than choose the simple binomial $(p+q)^n$, the form selected by Laplace as the starting point of his investigation. Its meaning is always clear. If p = q this form is indistinguishable from the normal curve of error even though _n_ be comparatively small. If _p_ be greater than _q_ the normal curve also results is _n_ be large, but if _n_ be small the resulting curve taked the form known as Pearson's Type III If p = q positive and negative errors are equally probable. In all cases variation is independent of the other, although a small amount of correlation as shown by Edgeworth does

not effect the ultimate form of the curve to any great extent. In biology, if $p = q$, such a condition as is seen in the inheritance of stature, where the mean stature of the offspring is determined by the means of the elements derived from the parents and the distribution of both is equally normal, is represented. On the other hand, if $p \neq q$, sucha condition as dominance is described. As an example of this in the same range it may be stated that if certain tall varieties of pea be mated with certain dwarf varieties all the offspring in the first generaltion are tall and in the stable population there are ultimately three tall plants to one dwarf. This is represented by the form $(p \quad q)'$ where $p = \frac{3}{4}$ and $q = \frac{1}{4}$. If tallness depend upon a number of factors the variation of the mixed race would in the simplest case correspond to $(\frac{3}{4} + \frac{1}{4})^n$, where n denotes the number of pairs of qualities on which stature depends, assuming that all the pairs determining tallness and shortness come originally from the same plants. Allowing for such a condition as partial dominance where the offspring takes more markedly after one parent than after the other and also for coupling where definite pairs of elements seem to have some special affinity, it can be shown that all the different distributions which have been biologically found are adequately expressed as direct derivatives of one or other of the two expressions given above.

Before proceeding further the basis of the theories of asymmetrical distribution as developed by Profs. Edgeworth and Pearson, which may be taken as representative of all, will be briefly outlined. The former calls his first method, "the method of translation" of the normal curve. (1) In this the normal curve is taken as the general law and the frequency of some quality assumed to vary in this manner. The frequency corresponding to a particular element of abscissa in this curve, is:-

$$y_0 e^{-\frac{x^2}{2\sigma^2}} dx$$

If now $x = f(\xi)$ so that ξ varies as $f^{-1}(x)$ the corresponding frequency of ξ is equal to

$$y_0 f'(\xi) e^{-\frac{(f(\xi))^2}{2\sigma^2}} d\xi$$

or

$$y = y_0 f'(\xi) e^{-\frac{(f\xi)^2}{2\sigma^2}}$$

is the equation of the new distribution. This is evidently quite sound reasoning, provided a justification for the application of the process can be found. The formula, however,

will cover almost any distribution. In general for practical purposes $f(\xi)$ is assumed to converge rapidly. Later Prof. Edgeworth arrives at a more general formula derived from the interaction of many factors. (2) (3) "The rationale of this method consists not merely, and not principally, in its exactly representing those cases in which each member of the frequency group under consideration is a definite function of some member of a group distributed according to the normal law of error. For instance, if the velocities of winds, or of any other objects, are distributed normally, then the corresponding energies will be distributed according to a frequency curve which is an exact translation of a normal curve. So if the diameters of oranges or other spherical bodies vary normally, the solid contents are proportionate to the cubes of the diameters". He says further, "The cases in which translation in the formula are doubtless not uncommon (compare "Journal of teh Statistical Society", Vol. LXI, 1898, p. 678) but the reason of the method lies deeper. It consists in the affinity of the formula to that universal law which is ever and everywhere approximately fulfilled through the whole realm of Statistics - Statistics proper as distinguished from arithmetic by sporadic or fortuitous dispersion".

Prof. Pearson on the other hand takes as his basis the differential equation:- (6) (7)

$$-\frac{1}{y}\frac{dy}{dx} = \frac{x}{a + bx + cx^2}$$

This he obtains from the differential equation of the normal curve $-\frac{1}{y}\frac{dy}{dx} = \frac{x}{c}$ by substituting a function of x for the constant in the denominator. This function is considered sufficiently given by the first three terms of its expansion by Maclaurin's theorem. He criticises Prof. Edgeworth's method of translation by asking what is the character that obeys the normal law, and atates that this had no real existence as a biological entity. To this Prof. Edgeworth makes the following reply, "This objection might be applicable if the proposed form was advocated as a pa particular real curve related to some real attribute distributed normally, say, as enery is to velocity. But the objection is not equally applicable to the position now taken. The best defence of this position is that it is the same as Prof. Karl Pearson's. For his Types, as here interpreted, are but particular representative curves formed by a judicious divergence from the normal law of error (a divergence well indicated by himself)".

My position in this paper is quite distinct from either of these. It is much more close to Prof. Edgeworth's first method of translation. I do not object at all to the use of a graduation formula for special purposes. But graduation formulae such as those of Prof. Pearson and Edgeworth tell us nothing about the biological processes which determine the variations and it is these specially which we wish to

investigate. Further, as will be seen, neither of these methods completely accounts for certain curves which arise directly in biological measurements.

The modes in which the normal curve $y = y_0 e^{-\frac{x^2}{2\sigma^2}}$ arises are of special interest. A full treatment is impossible here and from the point of biology has yet to be written. The curve is usually deduced by taking the limit of $(\frac{1}{2} + \frac{1}{2})^n$ when n is great, but as shown by Prof. Pearson, (5), even when n is comparatively small the approximation is very close. The same proof holds when $(p + q)^n$ is taken as the basis of the development, if p is nearly equal to q and if n is large. The reasoning on which the theorem is based can be easily extended to three dimensional space when a "normal" surface results.

The normal curve or surface, however, arises in many different ways. For instance, the solution of the "random walk" problem by Prof. Pearson is an example of how even a considerable assumption leads to the normal surface.

If again, two races differing is a specific quality mix, so long as the mean of the two elements determining a quality in the parents represents the average value of that quality in the offspring, then the curve of distribution of the hybrid will be much nearer the normal curve than that of either race, and if quality depend on several elements may be almost indistinguishable from it. This is easily seen from a proof given in a former paper (7) where it was shown

that if the moments of the two frequencies round their respective centres of gravity were known, those of a frequency which was the distributed product of these round its own centre of gravity could be at once written down. As the problem here considered from the assumption made is equivalent to the distributive multiplication of two curves, that proof applies. Denoting the moments respectively of the original distribution by $\zeta_2, \zeta_3, \zeta_4$ and $\zeta_2', \zeta_3', \zeta_4'$ and those of the product by μ_2, μ_3, μ_4 we have

$$\mu_2 = \zeta_2 + \zeta_2'$$
$$\mu_3 = \zeta_3 + \zeta_3'$$
$$\mu_4 = \zeta_4 + \zeta_4' + 6\zeta_2\zeta_2'$$

This enables us at once to see how the normal curve arises; letting for simplicity $\zeta_2 = \zeta_2'$ etc.

$$\mu_2 = 2\zeta_2$$
$$\mu_3 = 2\zeta_3$$
$$\mu_4 = 2\zeta_4 + 6\zeta_2^2$$

so that if β_1, β_2 be Pearson's constants for the first distributions and β_1', β_2' for the derived distribution respectively

$$\beta_1' = \frac{4\zeta_3^2}{8\zeta_2^3} = \frac{1}{2}\frac{\zeta_3^2}{\zeta_2^3} = \frac{1}{2}\beta_1$$

$$\beta_2' = \frac{2\zeta_4^2 + 6\zeta_2^2}{4\zeta_2^2} = \frac{1}{2}\beta_2 + 1.5$$

or the curve of distribution of the hybrid is much nearer the

normal than that of the parent, since for the normal curve
$\beta_1 = 0.\ \beta_2 = 3$

Further when free mating between the original races themselves and (of) the hybrid thakes place the normal surface results if the quality depends on more than one element. To show how rapidly this takes place a series of diagrams of a stable mixture of two races of different mean quality, when the quality depends on one, two and three elements respectively, is given for one of the simplest cases.

Diagram here.

This derivation of the normal curve is expressible perhaps more easily in terms of Mendelism when the average quality in the offspring is assumed to be equal to the average quality of their parents. If two races with a quality depending on two elements mix, we may denote the two pure races by

$$\left|\begin{matrix} A & A \\ B & B \end{matrix}\right| \qquad \left|\begin{matrix} a & a \\ b & b \end{matrix}\right|$$

These will form a stable race with the proportions

$$1\left|\begin{matrix} A & A \\ B & B \end{matrix}\right| \quad 2\left|\begin{matrix} A & a \\ B & B \end{matrix}\right| \quad 4\left|\begin{matrix} A & a \\ B & b \end{matrix}\right| \quad 2\left|\begin{matrix} A & a \\ b & b \end{matrix}\right| \quad 1\left|\begin{matrix} a & a \\ b & b \end{matrix}\right|$$

$$2\left|\begin{matrix} A & A \\ B & b \end{matrix}\right| \quad \begin{matrix} 1\left|\begin{matrix} a & a \\ B & B \end{matrix}\right| \\ 1\left|\begin{matrix} A & A \\ b & b \end{matrix}\right| \end{matrix} \quad 2\left|\begin{matrix} a & a \\ B & b \end{matrix}\right|$$

Hence by our assumption that A a is of mean stature between A A and a a etc., the resulting proportions are:-

1 4 6 4 1

or $(1+1)^4$ which is the expression which most easily gives rise to the normal curve of error. The grouping with any number of pairs of elements can easily be deduced from the above. Even when coupling is a marked feature this distribution may be fairly maintained if the number of elements determining the quality be large.

Lastly the normal curve may arise in time through the fact that some quality varies in time according to the inverse exponential. This law which holds with regard to a large number of precesses in physico-biological chemistry. Thus I find diseases which attack about the mean age of life have often an approximately normal distribution (cf. setc. IX) and many epidemics seem to run a similar course. cf. (5) In this case as I have before shown (<u>16</u>) but repeat here for the sakeof completeness, if p be the value of the infectivity at the beginning of the period and q the fraction determining the rate per unit time at which the infectivity is lost the resulting curve will be of the form $y = p^x q^{\frac{1}{2}(x-1)(x-2)}$ which as q is less than unity is the normal curve of error.

B. Frequencies derivable from $(p+q)^n$

The different methods by which asymmetrical distribution

and also symmetrical but not normal distributions may arise will now be considered seriatim, by the selection of a number of typical cases in which the causes of the departure from the form $(p + q)^m$ can be easily seen. The The asymmetrical as much the most important will be discussed in the first place.

(a) Asymmetrical or Skew Distributions.

I. Skew distributions arise simply from the method of measurement. Two methods of measurement may often be equally probable a priori; for example, notes in music may be measured either by their position on the scale or by the number of their vibrations. It is obvious that if the distribution be symmetrical under either of these conditions it cannot be symmetrical under the other.

II. They arise when areas or masses are taken as units of measurement if the variation of linear dimensions is symmetrical.

III. They arise when ratios are used in place of direct measurements, if the direct measurements are symmetrically distributed. The typical variation of ratios is skew, marked degrees of skewness being readily obtainable. Symmetry is really accidental.

IV. They arise again when the quality which is being measured has some inverse ratio to the quality which actually varies.

V. They arise when the variation of the quantity measured is due indirectly to symmetrical variation. This is very common.

VI. They arise when unequal numbers of races mix freely

though the original race may have varied in a symmetrical manner.

VII. They arise from the fact that many qualities in life vary so as to be easily graduated by the terms of a geometrical progression.

VIII. They arise when frequency distributions have time as the independent variable.

IX. They arise in such distributions as represent the frequency of disease at different ages.

I. As an example of the changes made by the choice of a scale of measurement, the number of notes of definite pitch occurring in the soprano songs of Schubert has been chosen. As in any song of definite key, certain notes, the tonic, the median etc., tend to occur with much greater frequency than others; one soprano song in each of the twelve keys has been taken and the first fifty notes of that song apportioned to their proper pitch. In this case we have two obvious modes of measurement, that by octaves, which is the method adopted in musical notation, and that by vibration which is the method in science. In the first case, the number of octaves above and below the mean pitch is theoretically infinite in number, though but a few are audible as music. In the second case, the actual number of

vibrations extends from zero to infinity. It is obvious that the logarithm of the independent variable of the latter scale is the independent variable of the former. The relationship between the frequencies on these two scales is therefore immediately determined. If the normal curve fits the distribution of notes as measured by octaves, the distribution measured by the scale of vibrations will be the Galton-Macalister curve. For if we take $y = y_0 e^{-\frac{X^2}{2\sigma^2}}$ to represent the distribution on the first supposition with X as the independent variable, the complete change is obtained by making X equal to Log. x. As the amount present on each element of abscissa in the first case $y_0 e^{-\frac{X^2}{2\sigma^2}} dX$ in the transformed curve, the amount resting on the new element is $\frac{dx}{x} e^{-\frac{(\log x)^2}{2\sigma^2}}$

This is the Galton-Macalister curve which is thus seen to be a simple case of "translation" of the normal curve. If on the other hand the musician's ear really estimates the number of vibrations, only using octaves for convenience of notation, a normal curve might be reasonable expected to represent the frequency on the scale of vibrations. The curve of frequency on the scale of octaves will be obtained by substituting e^x for X

It is obviously

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

The example chosen suffers under various defects. It is exceedingly difficult to get twelve characteristic songs upon twelve different keys. This was found especially with regard to the songs on the keys of C sharp and G flat. These songs have an undue number of repetitions of the same notes and when the notes are grouped for calculation it is found that out of one hundred notes taken from these keys, thirty-nine fallin one group at a low point of the scale. As, however, it seems impossible to select sufficient songs on different keys from any other composer I have not tried to repeat the experiment. The frequencies have been fitted to the normal curve on both hypotheses. The observed values and the values obtained on the two different hypotheses are tabulated in parallel columns with the values of χ^2, P, β_1 and β_2 added for comparison. In neither case is the fit a good one, but in the case where the vibrations are taken as the independent variable there is very little divergence between the facts and theory, except at the point already referred to, which accounts for two-thirds of the value of χ^2. Where the octaves are the independent variable the fit is not nearly so good on the whole.

As far as this one example goes it may be taken to show that the musician's ear is attuned more to vibrations

than to octaves. Exactly the same kind of remark must apply to such problems as guessing at tints. We do not know *a priori* what mixture of influences are at play. The biological question is to find the scale which most nearly measures our psychical processes.

Table I showing the concordance of fact and theory when the notes of Schubert's songs are fitted to the Normal Curve when the pitch is measured by octaves and by vibrations.

Notes.	Actual Numbers.	Theoretical Numbers Scale of Octaves.	Theoretical Numbers Scale of Vibrations
840 A. B♭	1	3.74	.54
G. A♭	7	12.60	6.10
F. F♯	28	32.32	31.36
E♭ E.	90	64.58	79.53
C♯ D.	105	102.51	119.81
B. C.	127	120.66	128.31
420 A. B♭	96	113.97	98.89
G. A♭	65	79.83	64.28
F. F♯	56	44.86	36.90
D♯ E.	19	18.69	19.11
C♯ D.	7	6.27	9.04
B. C.	2	1.57	4.5
χ^2		24.03	19.38
P		.013	.131
β_1		.075	.062
β_2		2.920	2.746

II and III. Frequencies of Products and Indices.

Asymmetrical curves arise when ratios or products are chosen as independent variables. The area of a leaf or the ratio of two lengths will not be normal in distribution, if the linear measurements on which they are based are normal. For an examination of the principles it will be clearer to begin with a first approximation by taking the frequency distribution of the primary qualities, as that of Type III.

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

The curve approaches very closely in form to the normal curve when n is large and is chosen on account of the ease with which it can be dealt with mathematically. The error of the results is small and easily allowed for. Let the measurements σ, ρ have frequencies

$$y_0 \rho^{n_1} e^{-\gamma_1 \rho} \quad , \quad y_0' \sigma^{n_2} e^{-\gamma_2 \sigma}$$

For products and ratios new independent variables m and p are introduced so that $\rho\sigma = m$ and $\frac{\rho}{\sigma} = p$.

A. The first case considered is that of the frequency distribution of m where $\rho\sigma = m$ and where ρ and σ are assumed quite uncorrelated. The surface which represents the frequency of all products is

$$z = y_0 y_0' \rho^{n_1} \sigma^{n_2} e^{-\gamma_1 \rho} e^{-\gamma_2 \sigma}$$

The integral desired is obviously the volume resting on the strip between $\varsigma\sigma = m$ and $\varsigma\sigma = m + dm$ contained between the plane of ς and σ and the surface given above. This frequency is most easily obtained by summing all values up to the curve $\varsigma\sigma = m$ and then taking the differential of that sum. The expression for this is easily seen to be

$$u = y_0 y_0' \int_0^\infty d\varsigma\, \varsigma^{n_1} e^{-\gamma_1\varsigma} \int_0^{\frac{m}{\varsigma}} \sigma^{n_2} e^{-\gamma_2\sigma} d\sigma$$

from which volume resting on strip of area required namely $\frac{du}{dm}dm$ is found.

i.e. $$\frac{du}{dm}dm = y_0 y_1' m^{n_2} dm \int_0^\infty \varsigma^{n_1 - n_2 - 1} e^{-\gamma_1\varsigma - \frac{\gamma_2 m}{\varsigma}} d\varsigma$$

so that the frequency of each value of m is given by

$$y = y_0 y_0' m^{n_1} \int_0^\infty \varsigma^{n_1 - n_2 - 1} e^{-\gamma_1\varsigma - \frac{\gamma_2 m}{\varsigma}} d\varsigma$$

an integral which is a solution of Bessel's equation. To illustrate the variation on this curve we may take

$n_1 = n_2$ and $\gamma_1 = \gamma_2$

Denoting the frequency of each value of m by ξ

$$\xi = y_0 y_0' m^n \int_0^\infty \frac{1}{\varsigma} e^{-\gamma\varsigma - \frac{m\gamma}{\varsigma}} d\varsigma$$

so that

$$\int_0^\infty \xi\, dm = y_0 y_0' \left(\frac{\Gamma n+1}{\gamma^{n+1}}\right)^2$$

and likewise for the integrals

$$\int_0^\infty m\xi\, d\xi, \quad \int_0^\infty m^2 \xi\, d\xi \qquad \text{etc.}$$

Whence the moments round the origin are

$$\mu_1' = \frac{(n+1)^2}{\gamma^2}$$
$$\mu_2' = \frac{(n+1)^2(n+2)^2}{\gamma^4}$$

etc.

giving

$$\mu_2 = \frac{(n+2)^2(2n+3)}{\gamma^4}$$
$$\mu_3 = \frac{(n+1)^2(10n^2+32n+26)}{\gamma^6}$$
$$\mu_4 = \frac{3(n+1)^2(4n^4+48n^3+179n^2+274n+151)}{\gamma^8}$$

whence β_1 and β_2 are easily obtained. In the case of volume where a third variable τ comes in and $\varsigma\sigma\tau = \phi$ the corresponding curve is

$$y_0 y_0' y_0'' \phi^n \int_0^\infty \int_0^\infty \frac{1}{m\varsigma} e^{-\gamma\varsigma - \frac{m^4}{\varsigma}} d\varsigma\, dm$$

B. The case when ς and σ are completely correlated is more simple.[x] There is little less generality if we assume them always directly proportional so that $\varsigma^2 = m$

[x] cf. (1) page 670, note V.

becomes the limiting equation. Here the frequency of ς^2 is the same as that of ς, namely, $y_0 \varsigma^n e^{-\gamma\varsigma} d\varsigma$

But this element is transferred to a different distance on the axis and placed on a different element of abscissa. The relation is by assumption

$$2\varsigma d\varsigma = dm$$
$$\text{a}\ d\varsigma = \frac{1}{2}\frac{dm}{\sqrt{m}}$$

so that the curve of frequency of each value m is

$$y = \frac{a}{2} m^{\frac{n}{2}-1} e^{-\gamma\sqrt{m}}$$

The moments are easily obtained and are

$$\mu_2 = \frac{(n+1)(n+2)(4n+10)}{\gamma^4}$$
$$\mu_3 = \frac{(n+1)(n+2)(40n^2+216n+296)}{\gamma^6}$$
$$\mu_4 = \frac{12(n+1)(n+2)(4n^4+88n^3+577n^2+1535n+1462)}{\gamma^8}$$

C. The case of the ratio $\frac{\varsigma}{\sigma}$ can be solved in the same way but more easily otherwise ✻

Let $\varsigma = r\cos\theta \qquad \sigma = r\sin\theta$

✻ cf. (8) & (9) where either is treated quite differently.

Then the element is

$$y_0 y_0' r^{n_1+n_2} \cos^{n_1}\theta \sin^{n_2}\theta \, e^{-(\gamma_1\cos\theta+\gamma_2\sin\theta)r} \, r\,dr\,d\theta$$

If θ be constant so that $\frac{\rho}{\sigma} = m = \tan\theta$ all that is necessary is to integrate from 0 to ∞ with reference to r. For each value of θ we have the frequency

$$y_0 y_0' \cos^{n_1}\theta \sin^{n_2}\theta \int_0^\infty r^{n_1+n_2+1} e^{-r(\gamma_1\cos\theta+\gamma_2\sin\theta)} \, dr$$

$$= \frac{y_0 y_0' \cos^{n_1}\theta \sin^{n_2}\theta \, \Gamma(n_1+n_2+2)}{(\gamma_1\cos\theta+\gamma_2\sin\theta)^{n_1+n_2+2}}$$

$$= \frac{y_0 y_0' m^{n_2} \Gamma(n_1+n_2+2)}{(\gamma+\gamma_2 m)^{n_1+n_2+2}} \, dm$$

or the curve in Pearson's Type VI. Let $\gamma_1 = \gamma_2$ $n_1 = n_2$ and the equation becomes

$$\xi = \frac{y_0 y_0' m^n \Gamma(2n+2)}{\gamma^{2n+2}(1+m)^{2n+2}}$$

The moments of this are known. The degree of asymmetry introduced by tabulating areas or indices can now be seen. The values of n are given in the annexed table for values of n from 0 to 1000 for the curve Type III. and for the three instances worked out above.

Table showing how asymmetry arises when areas or indices are tabulated.

	Type III.		Area: no Correlation		Area: perfect Correlation		Index: no Correlation.	
[illegible]	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	4	9	25.037	50.333	43.808	87.750	....	....
1	2	6	9.248	19.68	18.507	37.408	....	....
2	1.33	5	5.474	12.674	11.268	23.488	....	....
5	.67	4	2.403	7.394	4.978	11.766	7.784	20.485
10	.36	3.55	1.231	5.099	2.526	7.364	2.493	7.297
20	.19	3.29	.621	4.052	1.262	5.153	1.047	4.670
50	.08	3.12	.249	3.426	.503	3.850	.381	3.661
100	.04	3.06	.125	3.210	.261	3.423	.185	3.281
1000	.004	3.006	.00125	3.021	.025	3.042	.018	3.027

D. The general theorem of a ratio when the frequencies of the two variables are given by a normal surface is as simple as that given, but the expressions are more complex. The ordinary element of volume of a normal correlation surface is given by

$$e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}} dx\,dy.$$

Changing to polar co-ordinates and transferring the origin to distances $x = a$, $y = b$ which correspond to the means of the qualities measured, we have

$$r\,d\theta\,dr\; e^{-\frac{1}{2(1-\rho^2)}\left\{r^2\left(\frac{\cos^2\theta}{\sigma_1^2} - \frac{2\rho\sin\theta\cos\theta}{\sigma_1\sigma_2} + \frac{\sin^2\theta}{\sigma_2^2}\right) - 2r\left(\frac{a\cos\theta}{\sigma_1^2} + \frac{b\sin\theta}{\sigma_2^2} - \rho\frac{a\sin\theta\cos\theta}{\sigma_1\sigma_2}\right) + \left(\frac{a^2}{\sigma_1^2} - \frac{2\rho ab}{\sigma_1\sigma_2} + \frac{b^2}{\sigma_2^2}\right)\right\}}$$

As regards the integration, consider first

$$\int_0^\infty r\, e^{(r-a)^2}\, dr$$

which equals

$$\int_0^\infty r-a\; e^{-(r-a)^2}\, dr + a\int_0^\infty e^{-(r-a)^2}\, dr$$

$$= \int_{-a}^\infty r' e^{-r'^2} dr' + a\int_{-a}^\infty e^{-r'^2} dr' \quad \text{when } r' = r-a$$

$$= \frac{1}{2} e^{-a^2} + a\frac{\sqrt{\pi}}{2} + a\int_{-a}^0 e^{-r'^2} dr'$$

Now since a is never less than 10 in practice this last $= a\sqrt{\pi}$, so that for all practical purposes we may take t the limits of the first integral as $-\infty$ and ∞.

For then, $\frac{x}{y} = m = \tan\theta$ we may integrate between $-\infty$ and ∞ with regard to r so that for θ constant the amount on each element, when the origin is changed as above, is when the origin is changed as above

$$d\theta \times \frac{\frac{a\cos\theta}{\sigma_1^2} - \frac{\rho(a\sin\theta + b\cos\theta)}{\sigma_1\sigma_2} + \frac{b\sin\theta}{\sigma_2^2}}{\frac{\cos^2\theta}{\sigma_1^2} - \frac{2\rho\sin\theta\cos\theta}{\sigma_1\sigma_2} + \frac{\sin^2\theta}{\sigma_2^2}} \int_{-\infty}^{\infty} dr\, e^{\frac{-r^2}{2(1-\rho^2)}\left\{\frac{\cos^2\theta}{\sigma_1^2} - \frac{2\rho\sin\theta\cos\theta}{\sigma_1\sigma_2} + \frac{\sin^2\theta}{\sigma_2^2}\right\} + \frac{1}{2}\frac{a^2\sin^2\theta - 2ab\sin\theta\cos\theta + b^2\cos^2\theta}{\sigma_2^2\cos^2\theta - 2\rho\sigma_1\sigma_2\sin\theta\cos\theta + \sigma_1^2\sin^2\theta}}$$

$$= \sqrt{\pi}\, d\theta \frac{\frac{a\cos\theta}{\sigma_1^2} - \frac{\rho(a\sin\theta + b\cos\theta)}{\sigma_1\sigma_2} + \frac{b\sin\theta}{\sigma_2^2}}{\left\{\frac{\cos^2\theta}{\sigma_1^2} - \frac{2\rho\sin\theta\cos\theta}{\sigma_1\sigma_2} + \frac{\sin^2\theta}{\sigma_2^2}\right\}^{3/2}} e^{\frac{1}{2}\frac{a^2\sin^2\theta - 2ab\sin\theta\cos\theta + b^2\cos^2\theta}{\sigma_2^2\cos^2\theta - 2\rho\sigma_1\sigma_2\sin\theta\cos\theta + \sigma_1^2\sin^2\theta}}$$

$$= \sqrt{\pi}\, dm \frac{(b\sigma_1^2 - \rho a\sigma_1\sigma_2)m + (a\sigma_2^2 - \rho b\sigma_1\sigma_2)}{\{m^2\sigma_1^2 - 2\rho m\sigma_1\sigma_2 + \sigma_2^2\}^{3/2}} e^{+\frac{1}{2}\frac{(am-b)^2}{\sigma_1^2 m^2 - 2\rho\sigma_1\sigma_2 m + \sigma_2^2}}$$

which is seen at once to be a "translation form" of the normal curve

$$y = e^{-\frac{1}{2}X^2} \quad \text{if} \quad X^2 = \frac{(am-b)^2}{\sigma_1^2 m^2 - 2\rho\sigma_1\sigma_2 m + \sigma_2^2}$$

since the factor of the exponential is the differential of this latter expression with respect to m.

From this general equation some light can be thrown on the degree of skewness of ratio curves. Assuming $a = b$, $\sigma_1 = \sigma_2$ and $\rho = 0$ the curve becomes

The larger $\frac{a^2}{\sigma^2}$ where **a** is the mean value and σ the standard deviation of the variables the more symmetrical the curve becomes and in addition the more nearly normal it is. To compare with our former approximation in which Type III. was used we have since for that curve $\underline{a}$ is equal to the distance of the mode from the origin and $\frac{n+1}{\gamma^2}$ is equal to the second moment or square of the standard deviation the value corresponding to $\frac{a}{\sigma}$ is $\frac{\gamma a}{\sqrt{n+1}}$ or $\frac{n}{\sqrt{n+1}}$ so that if n is moderately large we can $n = \frac{a}{\sigma}$

In the table given below we see that for the cephalic index $\frac{a}{\sigma} = 30$, whence $n = 1000$, and for the sacral index $\frac{a}{\sigma} = 10$, or $n = 100$. From Table II. it is thus seen that the distribution curve of the former ratio would in the case first considered be practically normal and the last curve have a skewness lying between the values given in the table for no correlation and perfect correlation, namely, $\beta_1 = .125$ and $.261$ and $\beta_2 = 3.21$ and 3.423 respectively. Skewness, however, may be much increased if $a \neq b$ and $\sigma \neq \sigma_2$. The actual values found by observation are given in Table IV.

T A B L E IV.

Showing values of constants regarding the dimensions of the human skull and the human sacrum in the female.

Skulls Length	L	189.06	6.267	30.16	.001	2.958
"		180.36	6.218	29.01	.047	3.109
Breadth	B	140.67	5.279	26.64	.026	4.312
"		134.68	4.773	28.22	.002	2.683
Height	H	132.04	5.560	23.75	.092	2.802
"		124.56	4.933	25.25	.032	3.282
Index	$\frac{B}{L}$	74.34	6.520	11.40	.001	3.473
"	"	74.73	5.963	12.53	.004	2.609
Index	$\frac{H}{L}$	69.97	6.448	10.85	.006	3.815
"	"	69.13	5.668	12.19	.167	2.792
Sacrum Length		10.00	1.093	9.15	.0184	3.189
Breadth		11.50	.707	16.26	.0044	2.835
Index	$\frac{B}{L}$	116.11	13.679	8.49	.7251	4.489

$$dx = \frac{a}{1+m} - \frac{am}{(1+m)^2}$$

$$= \frac{a\,dm}{(1+m)^2}$$

Note on Ratios and Inverse Laws.

The distributions of ratios are themselves so interesting that a few remarks on the subject may not be out of place. The simplest case is that of the distribution of the ratios which the two parts of a straight line divided at random bear to one another. Here, if the one part be denoted by x and the other be a - x the ratio is

$$\frac{x}{a - x} = m \qquad \text{so that}$$

$$x = \frac{am}{1+m} \quad \text{and} \quad dx = \frac{a\,dm}{(1+m)^2}$$

The frequency of each value m of this ratio when x is the independent variable is equal, whence the frequency of each value of m when m is the unit of abscissa is given by $\frac{a\,dm}{(1+m)^2}$ or $y = \frac{a}{(1+m)^2}$ is the curve of frequency. This result in itself is neither new or specially interesting. But the process is interesting when it is noticed that the above equation is the simplest case of pearson's Type IV., and corresponds to the case regarding the ratio already calculated for Type III. $y = x^n e^{-\gamma x}$ when n = 0. It is one of the many instances of curves which may be called "Symmetrical" ratio-curves. The symmetry is at once apparent is the independent variable be changed to θ where $m = \tan\theta$ for instance, the above curve takes the form $y = \frac{a}{(\sin\theta + \cos\theta)^2}$

The group of such curves is large. The Galton-Macalister curve is one; Type VI. where m = 2n 2 is another. Any curve of the form

$$y = \frac{1}{x} \sum_n f\left(x^n + \frac{1}{x^n}\right)$$

obviously fulfils the conditions and for even functions

$$y = \frac{1}{x} \sum_n f\left(x^n - \frac{1}{x^n}\right)$$

The curve just discussed is

$$y = \frac{a}{m\left(\sqrt{m} + \frac{1}{\sqrt{m}}\right)^2}.$$

One such form occurred lately in connection with Mendelian coupling and is noted specially. In this case the the number of instances is too small to allow of much discussion. The development will be found in a previous paper. The main facts are that from a fourfold division a, b, c, d, the value of the ratio $\frac{a\ d}{b\ c}$ was specially required. This ratio had the following frequency

Value of ratio	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	18-19	23-24
No. of instances	2	2	2	5	5	5	4	1	1	1

The mean of these M = 9.14 48. The median lies somewhere nearer 9 than 8. When 81 is taken as the square of the constant of inversion, and the curve inverted, the

new mean has an abscissa of 8.96, the median lying slightly on the opposite side of 9; so as far than as the figures go, it may be taken as a symmetrical ratio curve, a result which might be expected. When the distribution is fitted to one of Peatson's curves the constants give Type I., but the particular variety which starts with an infinite ordinate; in this case at a distance of 1.7 units from zero. In view of the previous points it has therefore no biological significance. This example was hardly worth much additional statistical labour, but as it was an actual complex ratio it seemed better to verify whether the curve might not be a solution of

$$-\frac{1}{y}\frac{dy}{dx} = \frac{x}{a+bx+cx^2+dx^3}$$

A trial was made to obtain a better fit by this means, but the type of curve found was identical.

IV. Asymmetry may arise when the inverse of the quantity is measured rather than its direct value. For instance, if in cases of myxoedema the weights of the patients were measured, a distribution would be obtained probably more or less in inverse relationship to the amount of active secretion of the thyroid gland. The cases of asymmetry, however, due to inversion which are of most inportance occur in indices. It is either a matter of chance or of accidental convenience whether the index or its inverse is chosen and the degree of skewness will almost certainly be different in the two cases. The mathematics of inverse curves is very simple. Take, for instance, Type VI which is found to represent approximately the index curve to distributions which are normal in character. The element of area corresponding to a definite abscissa is in this case

$$\frac{y_0 x^n dx}{(a+x)^m}$$

If $x = \frac{a^2}{x'}$ the element of the new curve corresponding to the abscissa x' is

$$\frac{y_0 x'^{m-n-2} dx'}{(a+x')^m}$$

These are equivalent if m = 2n + 2, a form which may be termed a symmetrical index curve. If $m \neq 2n+2$ then the index curves will have different degrees of skewness.

F

For instance, if m = 20 and n = 8, the curve has a skewness of .5352 and int inverse one of .5490, while that of the symmetrical form when m = 20 n = 9 is .5410. The skewness of the index curve can thus be taken as having little or no biological significance. It is interesting to note in this connection that Type V. is the inverse of Type III. For Type III. the element of area is $y_0 x^{n} e^{-\gamma x} dx$ which becomes $y_0 \frac{e^{-\frac{\gamma}{x'}}}{x'^{n-2}} dx'$ if $x = \frac{1}{x'}$. Type V. is a rare curve in biology and it is quite possible that on some of the occasions in which it appears it is due to the quantity measured being dependent upon defect of a quality rather than on its presence. As Type V. is deduced from Type III., so an inverse curve can be obtained from the normal curve. In this case, if the equation of the latter be taken

$$y = y_0 e^{-\frac{(x+a)^2}{2\sigma^2}}$$

the origin being the point of inversion, we obtain the equation of the inverse as

$$y = \frac{a y_0}{x^2} e^{-\frac{a^2(x-a)^2}{2\sigma^2 x^2}}$$

It is noted that that portion of the normal curve corresponding to a negative abscissa has no biological significance

Time symmetry is also probably of inverse origin, but from a reason which will be considered under the appropiate heading.

V. In biology a great variety of shapes of curves may occur directly as the result of normal variation. This ca can be perhaps most easily illustrated by discussing a specific instance. Such a structure as the human pelvis affords a good example for an approximate solution of some of these problems. The inlet of the pelvis is practically a circular opening bounded by a chord corresponding to the sacrum which occupies a large part of the circumference. From observation the breadth of the sacrum varies very closely according to the normal curve of error.

For the purposes of a first approximation the pelvis may be imagined as circular in all its parts, excepting the sacral chord denoted by CD in the diagram. Let the middle point of this chord be denoted by B, and let a diameter be drawn at right angles through this point. Let this meet the circle again in A; for the approximation the arcs A C and A D may be imagined as rigid and hinged together at A. This gives a rough mechanism by which the normal variation of the chord C D may be examined. As a further construction a second diameter is drawn through the center O at right angles to the first. This meets the circle in F and H, F H is thus the transverse diameter of the pelvis and A B the anteroposterior diameter. If now the chord C D vary according to the normal curve of error, there will be corresponding variations in the anteroposterior diameter

A B and in the transverse diameter F H. (From the assumed condition tha transverse diameter will not really be F H but the distane between the points at which the tangents patellel to A B touch the curve). Now if A C and A D are joined these are obviously constant lengths which may be denoted by p. We may denote A B by b and its variation by x. In the same way we may denote C B by a and its variation by r. We have then, since A B C is a right angled triangle

$$b^2 = (a+r)^2 + (b+x)^2$$

which reduces to

$$0 = 2bx + x^2 + 2ar + r^2$$

b being negative when a is positive and vice versa. This gives

$$x = -b + \sqrt{b^2 - 2ar - r^2}$$

The positive sign of the root being taken since when $x = 0$ $r = a$ This gives the value of x in terms of r. When the equation of the frequency of x is given with x as the independent variable

$$y = c\,\frac{b-x}{\sqrt{a^2 - 2bx + x^2}}\; e^{-\frac{1}{\sigma^2}\left(-a + \sqrt{a^2 - 2bx + x^2}\right)^2}$$

It is obvious at once that if the variation of r is normal that the variation of x ia skew, the skewness increasing

as the standard deviation of r increases. A complete solution of the variations of the transverse diameter will also be givne, but if we consider that the vertical distance from the diameter A B of every given point of the chord A C varies normally, it will be seen that the variation of the transverse diameter must be symmetrical and not deviate so much from the normal curve of the anteroposterior, this is in fact what is found. The actual figures give the constants of the variation of the pelvis as follows:-

Sacral Breadth	.00445	2.8351
Transverse Diameter	.05395	3.0386
Anteroposterior Diameter	.1452	3.5988

One further development can be made to show how a one-sided frequency descending from an infinite quantity zero is dertimined by the normal variation of the chord C D. If the tangential transverse diameter cuts A B at the point of P then then the frequency of each value A P varies with this extreme shewness. This investigation is directly allied to with that of the true variations of the transverse diameter according to our mechanical hypothesis.

The true transverse diameter is the distance between two tangents parellel to the diameter A B. As A is fixed each arc A C or A D will rotate according to the hypothesis an equal angle in opposite directions which we may term .

It is obvious then that whether the variation of the chord C D is positive or negative the centre of the centre of the new circle will lie on the side of O opposite to C D since A is fixed and A O is constant. If A O = a the new centre will be at the point $x = a\cos\theta$ $y = a\sin\theta$, if A is taken as the origin of the co-ordinates. The equation of this circle is thus obviously

$$(x - a\cos\theta)^2 + (y - a\sin\theta)^2 = a^2$$

For pusposes of calculation C D may be supposed to subtend an angle of 60° at the circumference of the circle. As <u>a</u> is the radius of the circle

$$\frac{1}{2} CD = \frac{\sqrt{3}}{2} a$$

O B C being a right angled triangle with the angle C O B = 60°. A C is thus equal to $\sqrt{3}a$ and A B to $1.5a$. The types of variation can now be evaluated. If P is the point on the axis where the line joining the tangents parallel to A B cuts the abscissa A P is obviously equal to $a\cos\theta$

$2a(1 + \cos\theta)$ The Angle ϕ is the angle between the mean position of A C and its variation <u>x</u>

$$\text{i.e. } \phi = \sin^{-1}\frac{BC + x}{AC} - \sin^{-1}\frac{BC}{AC}$$

$$= \sin^{-1}\frac{b + x}{p} - \frac{\pi}{6}$$

For the variation of the transverse diameter if η be the actual variation

$$\eta = a \sin\theta \quad \text{since} \quad y = a(1+\sin\theta)$$

so that

$$\sin^{-1}\theta = \frac{\eta}{a}$$

but

$$\sin^{-1}\left(\frac{1}{2} + \frac{x}{p}\right) - \frac{\pi}{6} = \theta$$

or

$$\sin^{-1}\left(\frac{1}{2} + \frac{x}{p}\right) = \sin^{-1}\frac{\eta}{a} + \frac{\pi}{6}$$

$$a \quad x = \frac{p}{2}\left\{\eta\sqrt{3} + \sqrt{1-\eta^2} - 1\right\}$$

similarly

$$\text{if } AP = \xi = a\cos\theta$$

so that

$$x = \frac{p}{2}\left\{\xi + \sqrt{3}\sqrt{1-\xi^2} - 1\right\}$$

If we take as found by observation the frequency of the variation of <u>x</u> to be normal σ

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

we have the frequency of variation for the new curves given by; (a) for the transveʙse diameter by the ordinary "translation formula"

$$y = a\left\{\sqrt{3} - \frac{\eta}{\sqrt{1-\eta^2}}\right\} e^{-\frac{p^2}{4}\left\{\eta\sqrt{3} - \sqrt{1-\eta^2} - 1\right\}^2 \times \frac{1}{\sigma^2}}$$

and for A P

$$y' = a\left\{\frac{\sqrt{3}\,\xi}{(1-\xi^2)^{\frac{1}{2}}} - 1\right\} e^{-\frac{p^2}{2}\left\{\xi + \sqrt{3}\sqrt{1-\xi^2} - 1\right\}^2 \frac{1}{2\sigma^2}}$$

By observation $\sigma = \frac{1}{10} CD = \cdot 066$ if $a = 1$

and by hypothesis $p = \sqrt{3} \times a = 1\cdot 732$ if $a = 1$

The first curve is nearly a normal curve tha second has a form given in the subjoined table

x	ξ	η
.	1.000	
.0662	.999	19.463
.0932	.998	10.219
.1126	.997	6.394
.1455	.995	2.755
.2029	.990	.410
.2458	.985	.111
.2812	.980	.027

It is not pretended that the above description holds rigidly, it is a mere mathematical first approximation resulting in a series of translated normal curves of which the origin can be at once shown. It might be further extended if it was worth while to include the fact that the variation in the circumference of the iliac bones also obeyed the normal law but the discussion given is quite sufficient for the present purpose.

VI. Under the category of asymmetry due to the mixture of races I have not been able to find any figures which will repay the trouble of analysis. Adequate statistice of the mixture of two races which are symmetrically disposed with regard to some quality in the offspring is the mean of the parents, do not seem to exist at present. Experiments will require to be made.

Such a case would however come under the ordinary formula for a stable race namely

$$m^2 (A,A) \qquad 2\,m\,n\,(A,a) \qquad n^2\,(a,a)$$

The solution is nearly identical with that Prof. Pearson gives in his first memoir on the mathematice of evolution. There Prof. Pearson analyses an asymmetrical distribution into the sum of two symmetrical distributions each of which obeys the normal law. His chief example however refers to a ratio distribution which as we have seen is essentially assymmetric and therefore not a suitable case for the application of this method. Further I think that it is exceedingly unlikely that two races of the same species can exist commingled in nature without breeding together. The distribution will not thus be expressed by $m f_1(x) + n f_2(x)$ but by $\{m f_1 x + n f_2 x\}^2$. This is practically

? meaning of

an application of distributed multiplication. If there fore ζ_2, ζ_3 etc be the moments of $n_1 f_1(x) + n_2 f_2(x)$ and μ_2, μ_3 etc. the moments of the distributed square of this, we have

$$\mu_2 = 2\zeta_2$$
$$\mu_3 = 2\zeta_3$$
$$\mu_4 = 2\zeta_4 + 6\zeta_2^2$$
$$\mu_5 = 2\zeta_5 + 2\zeta_2\zeta_3$$

From the actual observations $\mu_2 \mu_3$ etc., are determined and thence ζ_2, ζ_3 etc., The problem is thus reduced to that solved by Prof. Pearson.

VII. Apparently as a result of the process of physical chemistry the geometrical progression describes many phenomina of life. The well known example given by de Vries of the number of flowers of Ranunculus Bulbosus witj each number of petals will illustrate this. It is an almost perfect example of a geometrical progression. Prof. Pearson has fitted it to the curve $y = a x^{-n} e^{-\gamma x}$ but the latter formula does not give nearly so close a fit. As the geometrical progression is itself a special case of Type III. this example shows incidently the difficulty of applying the method of moments to cases discrete variates when the curve terminatesabruptly. A table of the numbers theoretical and actual with thecorresponding values of χ^2 and P is annexed. It is seen that the geometrical progression has a probability $P = .95$ whereas TypeIII. has a probability only $P = .67$.

Firther examples of the geometrical progression are seen in the decline of the death-rate with each year of life with children suffering from measles, scarlet fever, etc. This subject is only mentioned at present. Along with the subject matter of the next two sections it will be dealt with separately at a later date.

Table I showing number of Petals in flowers of Ranunculus Bulbosus (de Vries).

No. of Petals.		Prof. Pearson's fitting.		Fitted to Geometrical Progerssion	
5.	133	136.9		134.37	
6.	55	48.9		53.32	
7.	23	22.6		21.16	
8.	7	9.6		8.40	
9.	2	3.4		3.33	
10.	2	.8	1	1.32	1.84
11.	0	.2		.52	
Total	222	222.0		222.42	
χ^2		3.27		1.09	
P		.67		.95	

VIII. In this connection that asymmetry which arises when time is tha independent variable in the curve of distribution is considered. When observations are made in such a matter as the date of flowering of a large number of plants, it is uniformly found that a skew distribution results. This is what would be expected <u>a priori</u> from such limited knowledge of the physiology of the plants which we at persent possess.

Flowering is the result of a completion of a process to which a certain summation of changes is necessary. For a first approximation this sum may be taken as expressed by the simple; if y is the rate at which the ch change takes place per unit of time ae^{yt} willrepresent the amount of substances necessary to determine flowering present at any specified time. If then a limit K must be reached before flowering can proceed, we have the distribution really the inverse relationship between

$$ae^{yt} = K \quad \text{or} \quad yt = K'$$

If y vary normally or according to Type, III. we have the frequency of flowering on each date given respectively by

$$y = \frac{m}{x^2} e^{-\frac{a^2(x-a)^2}{\sigma^2 x^2}}$$

$$\text{and} \quad y = \frac{n}{x^n} e^{-\frac{p}{x}}$$

A large number of observations will be necessary to settle the exact form of the rate of increase, but whatever law is found the resulting distribution can hardly be anything but asymmetrical.

(b) Symmetric forms derivable from $(p + q)^n$

I. Symmetric forms are much less interesting than asymmetric but two forms arise which are of some importance. An example of one of these is given by Mr. Yule on page 163 of his "Theory of Statistics" (10). In this example the proportions of male to female births are tabulated for the registration districts of England both as regards the ratio and the size of the district. As is naturally to be expected in the smaller districts, the range of variation is very considerable, while in districts such as those containing three hundred thousand inhabitants ans upwards, the range is very small. Here the theory of chance permits of easy evaluation of the resulting form of curve. The chance of the birth of a boy or a girl (1.04 to 1) being practically equal the asymmetry which exists may be neglected. The curve formed for each registration district of definite size will be essentially

normal and the numerical standard deviation will be $\frac{\sqrt{n}}{2}$ where n is the number of births in the district. The standard deviation of the ratio will then be $\frac{\sqrt{n}}{2}/n$ or $\frac{1}{2\sqrt{n}}$. All that is therefore necessary is to know the distribution of the registration districts in size. This is found to conform closely to a curve of Type, III., where the power of n is negative but greater than -1 i.e. the curve starts from infinity at the distance say c from zero. We have thus since for each size of district the distribution is given by $e^{-\frac{x^2 n}{a}}$ when a is a constant and since the distri--bution of n is given by $n^{-p} e^{-\gamma n}$ the final distribution given by the equation

$$y = M \int_0^\infty n^{-p} e^{-\gamma n} e^{-\frac{x^2(n+c)}{a}} dn = M \frac{e^{-\frac{cx^2}{a}} \Gamma p+1}{\left(\gamma + \frac{x^2}{a}\right)^{p+1}}$$

This is a curve of considerable complexity which does not lend itself to an easy formula for calculating the moments. It is allied to Type IV. and it is interesting to note that Type IV proper arises if the distribution in place of being that Form of Type III. found, is the ordinary form when the power of n is positive. The integral then becomes a simple Gamma function and gives the symmetrical

form of Type IV. or

$$y = \frac{m\,\Gamma p+1}{\left(1+\frac{x^2}{a}\right)^{p+1}}$$

To compare theory with facts the distribution as given by Mr. Yule has values of the essential constants β_1=.0167 β_2=7.0126 which gives a curve of Type, IV. The asymmetry being slight has been neglected in fitting. On this hypothesis the actual and theoretical figures of the distri--bution are given in the annexed table.

Number of Registration Districts with different proportions of Male and Female Births.

	465-70	471-76	477-82	483-88	489-94	495-500
Actual No.	1	1	2	4	10	33
Theoretical		.57	.97	3.01	10.74	40.29

513-18	519-24	525-30	531-36	537-42	543-
135	37	5	4	1	1
140.61	40.29	10.74	3.01	.97	.87

This gives χ^2= 5.92 or P=.92 ·

This form is thus very close to that which might be reasonably expected from the foregoing argument. It may be objected that this is one of the cases where the result might be expected a priori from the method ofProf. Pearson's proof, since Type, IV. arises when, from a mixture of an equal finite number of red and white balls, half are with--drawn time after time and the distribution of the results are tabulated, but in the example given, the number of balls of the analogy is clearly infinite, and the normal curve should be very approximately obeyed.

II. One other form is considered. It is a continuation of a subject discussed in a previous paper. It concerns that grouping of stature which results if in inheritance coupling either internal or external occurs, or if dominance is a factor in the herèdity. It has been shown before that if the element determining a quality such as stature is dominant the stable race for one quality is $(\frac{3}{4} \quad \frac{1}{4})$ as is well known; if a second element take part, the stable race is either $(\frac{3}{4} \quad \frac{1}{4})$ or $(\frac{3}{4} \quad \frac{1}{4})$ $(\frac{1}{4} \quad \frac{3}{4})$. The latter is of the form (1 n 1) and if inheritance depend on such numerous such pairs, we have (1 n 1) as the resulting form.

Now this form is easily approximated to, if p be

moderately large. The writing down of the terms of a distributed multiplication may be done as follows, a method, I think, not hitherto pointed out. Take for example (a b c) and write the product as follows:-

1	a^5		$5a^4c$		$10a^3c^2$		$10a^2c^3$		$5ac^4$		c^6
5b		a^4		$4a^3c$		$6a^2c^2$		$4ac^3$		c^4	
10b			a^3		$3a^2c$		$3ac^2$		c^3		
10b				a^2		$2ac$		c^2			
5b					a		c				
b						1					

The distributed terms are then obtained by adding together the terms in the vertical columns each multiplied by the corresponding terms in the first column showing the expansion of (b 1). Thus the middle term is:-

$$6a^2c^2 \times 5b \quad 2ac \times 10b^3 + 1 \times b^5$$

or

$$30a^2bc^2 + 20ab^3c + b^5$$

If a=c each row is approximated to by the normal curve with a standard deviation proportional to the vertical distance from the point where the distribution is represented by unity i.e. where $\sigma = 0$. The form of resulting distribu-

-tion can thus be found to be

$$y = M\int_0^\infty (p+c)^r e^{-y(p+c)} e^{-\frac{x^2}{ap}}\,dp$$

For purposes of approximation c may be taken as zero, and we have

$$\mu_2 = \frac{a}{2}\,\frac{r+\frac{3}{2}}{\gamma}$$

$$\mu_4 = \frac{3a^2}{4}\,\frac{r+\frac{5}{2}}{y}$$

or the curve partakes of the moment relationship of the symmetrical Type IV.

This same form also arises when there is a coupling as here again distributive multiplication really comes in. The ordinary form for a stable population is as shown before

$$a\begin{vmatrix} A & A \\ B & B \end{vmatrix} \qquad 2ab\begin{vmatrix} A & A \\ B & b \end{vmatrix} \qquad b\begin{vmatrix} A & A \\ b & b \end{vmatrix}$$

$$2ab\begin{vmatrix} A & a \\ B & B \end{vmatrix} \qquad 2a\ \ 2b\begin{vmatrix} A & a \\ B & b \end{vmatrix} \qquad 2ab\begin{vmatrix} A & a \\ b & b \end{vmatrix}$$

$$b\begin{vmatrix} a & a \\ B & B \end{vmatrix} \qquad 2ab\begin{vmatrix} a & a \\ B & b \end{vmatrix} \qquad a\begin{vmatrix} a & a \\ b & b \end{vmatrix}$$

Where $\frac{a^2}{b^2} = R$ or the coupling ratio.

If we consider stature determined by the same hypothesis as before (p.) then the above expression must be summed diagonally, so that we have the distribution given by

	a^2	$4ab$	$2a^2 + 4b^2$	$4ab$	a^2
instead of	1	4	6	4	1

This is however the same form as for the dominance given before; for it can be written

a^2	1		2		1
$4ab$		1		1	
$4b^2$			1		
	a^2	$4ab$	$2a^2 + 4b^2$	$4ab$	a^2

A Type IV. curve thus results in general, whether heredity is dependent on coupling or dominance.

The method of analysis of a complex form such as of stature can now be indicated. It requires the help of the theory of correlation. If stature depends on a number of elements inherited independently and if the combi-nation of two of these determine a mean condition in the offspring, the the co-efficient of correlation between

parent and offspringwill be given by $r =$ .5 and the curve of stature will be normal. If the curve approches Type IV. rather than the normal, this is explained by the fact that some of the elements are dominant. If all the elements are dominant $r = .\dot{3}$. Taking all distributions of stature, it is found that though nearly normal, there is usually a defect of fit shown by the normal curve at the maximum. This defect means a deviation in the direction of Type IV. and suggest a certain measure of dominance. For correlation figures we find for instance, that the correlation between father and son is r .514, between mother and son $r =$ .494, and father and mother $r =$ .280 showing that the real correlation between father and son is given by

$$r = \frac{.514 - .494 \times .280}{\sqrt{1-.494^2}\sqrt{1-.280^2}} = .45$$

Thus the same kind of defect from the theoretical correlation .5 is seen as has been observed between the actual and theoretical distributions.

CONCLUSIONS.

The examples given in the previous pages will have explained sufficiently the point of view from which this paper has been written. I do not think that any advantage can be obtained from general systems of curve derivation. In the first place a very small alteration in the value of the constants may make a very great difference in the type of the theoretical curve, and the figures at our disposal are rarely of sufficient worth to allow of fine distinctions being drawn between curves obtained on these systems, especially when in addition, the differences in the type of the curve represent really nothing in the facts. But there is another drawback. Neither Prof. Edgeworth.'s nor Prof. Pearson's methods account for many curves which arise. Thus by the former, the independent variable x of the normal curve is replaced by $f(x)$ and the latter assumed to be quickly convergent. In practise this is not so. If the original curves are normal we find the curve of a ratio given in the simple symmetrical case by

$$y = \frac{m(1+x)}{(1+x^2)^{\frac{3}{2}}} e^{\frac{a^2}{2\sigma^2}\frac{(1-x)^2}{1+x^2}}$$

an immediate "translation" form. The power of the exponential is finite for all values of x but it is not convergent in a series. Further the fact that it is as exact "translation" form, tells us

nothing biologically; we are left stranded on the "sands of surmise". Derived as shown above (p) the forms is perfectly intelligible. The same results apply to the inverse of a normal curve.

Such forms are also not comprehended by Prof. Pearson's formula. That just given:-

gives

$$-\frac{1}{y}\frac{dy}{dx} = -\frac{1}{1+x} + \frac{3x}{1+x^2} + \frac{a^2}{\sigma^2}\frac{(1-x)^2}{(1+x^2)^2}$$

while the form for a product

$$y = m^{n_2}\int_0^\infty r^{n_1-n_2+1} e^{-yr-\frac{ym}{r}} dr$$

is a derivative of a Bessel solution. In the same way others of thederived curves do not satisfy his equation but more probably an equation.

$$-\frac{1}{y}\frac{dy}{dx} = \frac{a+bx+cx^2+\dots fx^{n-1}}{\alpha+\beta x+\gamma x^2+\dots \delta x^n}$$

In conclusion it seems to me that in biology frequency distributions are a purely experimental science. Mathematical assistance is essential. The discoveries of many mathematicians have been most brilliant and illuminating, and but for their work the present biological statistical knowledge would not exist. But though this work has been brilliant, it is too much divorced from life. The real function of mathematics is to use observations and measurements to aid in finding real laws, and to

test such hypotheses as suggest themselves. Directly curves are thus of the first importance. As far as can be judged from the study of heredity, many of the elements are inherited independently or according to laws which afford easy mathematical expressions. Properly chosen experiments can easily be made to test these. But apart from this there is a great basis of physical chemistry underlying life with laws of its own, which determine the subsequent curves of distribution. Such curves can not be comprehended intelligently even by the most general of the "generalised" curves.

REFERENCES.

(1) Edgeworth, "On the Representation of Statistics by Mathmatical Formulae." Jour. Roy. Soc. vol. XLI. 1898, p. 670 Note, V. p. 695, vol XLII, 1899, pp 125, 373, 534.

(2) Edgeworth, "The Law of Error". Trans. Cam. Phil. Soc. vol. XX, 1908, pp 36, 113.

(3) Edgeworth, "The Method of Representing Statistics by Analytical Geometry". 8nter. Cong. Math. Cam, 1912.

(4) Pearson, "Contributions to the Mathematical Theory of Evolution".
"On the Dissection of Asymmetrical Frequency Curves". Phil. Trans. Roy. Soc. Vol. 185 A, part 1, 1894.

(5) "Skew Variation in Homogeneous Material". Phil. Trans. Roy. Soc. vol. 186 A, 1895.

(6) Pearson, "Das Fehlergesetz und seine Verellgemeinerungen durch Fechner und Pearson". A rejoinder. Biom. vol. 4, p. 169.

(8) Pearson "On the Constants of Index Distributions as Deduced from the Like Constants for the Components of the Ratio with Special Reference to the Opsonic Index". Biom. vol. VII. p. 531.

(9) Greenwood & White, "A Biometric Study of the Phagocytosis with Special Reference to the "Opsonic Inded". Biom. vol. VII. p. 531.

(10) Yule, "Theory of Statistics" 1912, p. 163.

(11) Galton, "The Geometrical Mean in Vital and Social Statistics". Proc. Roy. Soc. vol. XXIX. p. 365.

(12) Mc.Alister "The Law of the Geometrical Mean". Proc. Roy. Soc. vol. XXIX. p. 367.

(13) Leaks "Studies in Indian Cotton". Jour. of Gen. vol. I. 1911, p. 205. (Table XIV p.p. 254)

(14) Brownlee, "The Relation of the MOno-Molecular Reactinn to Life Processes and to Immunity". Proc. ROy. Soc. Ed. vol. XXXI.. 1911 p. 628.

(15) Brownlee, "On inheritance of Hair and Eye Colour". Proc. Roy. Soc. Ed. vol. XXXII. 1912, p. 458

(16) Brownlee, "The Law of the Epidemic". Proc. Roy. Soc. Ed. 1906. p. 502.

(17) Brownlee, "Point Binomials and Multinomials in Relation to Mendelian Distributions". Proc. Roy. Soc. Ed. vol. XXXII. p. 396. (p.398)

(18.) Rayer, "Suette Milaire en 1821". Paris 1822.

(19) Barrie, "Report on the Epidemic of Smallpox at Sheffield during 1887-88". 1889, p. 192.

(20) Kepteyn, "Skew Frequency Curves in Biology and Statistics". London. 1903.

(21) Macdonell, "A Study of the Variation and Correlation of the Human Skull with Special Reference to English Crania". Biom. vol. III. p. 191. (p. 228)

(22) Emmons, "A Study of the Variations in the Female Pelvis based on the Observation made on 217 Specimens of the American Indian Squaw Biom. vol. IX. p. 34.

Further Observations on the Treatment of Smallpox by the Serum of Vaccinated Heifers. (With Professor R.S. Thomson)

Lancet, 4th April, 1903. p. 947.

Age Incidence in Infectious Disease.

Proceedings of the Royal Philosophical Society of Glasgow, 1903-1904.

Obituary Note of J. B. Russell, Esq., M.D., L.L.D. (Written by request of the Council.)

Proceedings of the Royal Philosophical Society of Glasgow, 1904-1905.

Statistical Studies in Immunity.

(1) Smallpox and Vaccination.

Biometrika IV., November 1905.

(2) Natural Immunity and the Capacity for Acquiring Immunity in the Acute Infectious Diseases.

Journal of Hygiene, Vol. V., Oct. 1905. p. 514.

(3) The Theory of an Epidemic.

Proceedings of the Royal Society of Edinburgh, 18th June, 1906.

(4) The Incubation Period and the Crisis in Infectious Disease.

Proceedings of the Royal Philosophical Scoiety of Glasgow, 7th November, 1906.

(5) A Discussion of the Means of Estimating the Severity of Acute Disease.

Biometrika, Vol. V ., 1907.

An Account of Five Cases of Pyelitis in Enteric Fever. (With Dr. Chapman.)

Glasgow Medical Journal, December 1906.

The Problem of Fever Nurses.

Nursing Times, 6th July, 1907.

A Review of the Serum Treatment of Infectious Diseases in the Practice of the City of Glasgow Fever and Smallpox Hospitals, Belvidere.

Glasgow Medical Journal, September 1908.

Germinal Vitality: A Study of the Growth of Nations as an Instance of a hitherto undescribed Factor in Evolution.

Proceedings of the Royal Philosophical Society of Glasgow, 29th April, 1908.

Certain Considerations on the Causation and Course of Epidemics.

Proceedings of the Royal Society of Medicine, June 1909.

The Significance of the Correlation Coefficient when Applied to Mendelian Distributions.

Proceedings of the Royal Society of Edinburgh, 24th January, 1910.

The Mathematical Theory of Random Migration and Epidemic Distribution.

Proceedings of the Royal Society of Edinburgh, 16th May, 1910.

The Relation of Mono-molecular Reaction to Life Processes and to Immunity.

Proceedings of the Royal Society of Edinburgh, 9th January, 1911.

Note on the Calculation of the Probabilities of Life at High Ages. (With Dr. Morison.)

Journal of the Royal Statistical Society, Vol. LXXIV. January 1911.

On Inheritance of Hair and Eye Colour.

Proceedings of the Royal Scoiety of Edinburgh, Vol. XXXII. June 17th, 1912.

Notes on a Case in which certain Symptoms of Syringomyelia followed Enteric Fever.

Glasgow Medical Journal, November 1912.

Point Binomials and Multinomials in relation to Mendelian Distribution.

Royal Society Edinburgh, 16th Feb., 1912.

The Theory of Probable Error and its Application to Vital Statistics.

Journal of the Royal Sanitary Institute, Vol.XXXIV. 1913.

A Note on the Possibility of Analysing Race Mixture into their Original Elements by the Mendelian Formula.

Journal of the Royal Anthropological Institute of Great Britain & Ireland. Jan. to June 1911.

The Prodromal Rashes of Smallpox. (With Professor R. S. Thomson.)

Quarterly Journal of Medicine, Vol. II., No. 6. January 1909.

LIST OF PAPERS PUBLISHED BY JOHN BROWNLEE, M.D.

Notes on the Occurrence of a number of Cases of Epidemic Roseola or Rotheln in the City of Glasgow Fever Hospital, Kennedy Street.

Glasgow Medical Journal, March 1897.

The Administrative Treatment of Undefined Cases certified as Scarlet Fever. (With Dr. Dittmar.)

Lancet, 3rd April, 1897.

Observations on the Aerial Transmission of the Enteric Fever Poison, with a Record of an Outbreak presumably caused by that means of Infection.

Proceedings of the Philosophical Society of Glasgow, 1897-98; also reprinted in Public Health, 1898.

The Diagnosis and Hospital Treatment of Scarlet Fever.

Thesis for the Degree of Doctor of Medicine, 1897.

Observations on an Infectious Disease in Lascars having close Relations with Variola and Varicella. (With Professor R. S. Thomson.)

Lancet, 22nd October, 1898.

Soil and Vegetation.

Guernsey Health Lectures, 1899-1900.

Conditions Determining Disease.

Concluding Lecture, Guernsey Health Lectures, 1899-1900.

The Serotherapeutics of Plague.

Lancet, 17th August, 1901. p. 435

The Antitoxin Treatment of Diphtheria in the City of Glasgow Fever Hospital, Belvidere during Six and a Half Years.

Glasgow Medical Journal, April 1902.

Preliminary Note on the Parasites of Chickenpox and Smallpox. (With Professor R.S. Thomson.)

British Medical Journal, 1903.
January 31st p. 241.

VOLUME IV.

Contents.

Contents.

[illegible]ence Distribution in [illegible]and.

Australasian [illegible] Science [illegible]

The [illegible] of [illegible] and the [illegible] [illegible]
[illegible].

XVIIth International Congress of Medicine, London, 19[illegible]

Variations in the Breeding [illegible] [illegible] of [illegible]
Death Rates. I. The Relationship between "[illegible]"
Death Rates and Life Table Death Rates.

Journal of Hygiene, Vol. XIII, [illegible]

Notes on a Series of Lectures given at the [illegible]
University, January to [illegible]

(1) Examples of Distributions.

(2) Constants of Distributions.

(3) Method of Calculating the [illegible] [illegible] of [illegible]
of a Series of Observations.

(4) Correlation.

(5) Examples of Correlation.

(6) Partial Correlation Coefficients.

(7) Probable Error.

[illegible] in Infectious Diseases.

[illegible] [illegible] Society [illegible]

Investigations into the [illegible] of [illegible]
[illegible] application of a [illegible] [illegible]

Public Health, [illegible]

[illegible] Curve of the [illegible].

[illegible] [illegible]

Studies in the Meaning and Relationships of Birth and Death Rates. II. Density of Population and Death Rate (Farr's Law). III. The Constitution of a Death Rate. IV. On the Range of Instances in which Geometrical Progressions describe numerically precesses of life, i.e., those processes which might be explained by a monomolecular reaction. V. On the difficulty that in applying tha laws of physical chemistry to life precesses, indices occur which suggest the actions of fractions of a molecule.

Journal of Hygiene, 30th July, 1915.

Historical Note on Farr's Theory of the Epidemic.

British Medical Journal, 14th August, 1915.

The Theory of Chance Distribution as Applied to Biology.