VOLUME I AUGUST, 1902 No. 4

THE MOST SUITABLE PROPORTION BETWEEN THE VALUES OF FIRST AND SECOND PRIZES.

BY FRANCIS GALTON, F.R.S.

A CERTAIN sum, say £100, is available for two prizes to be awarded at a forthcoming competition; the larger one for the first of the competitors, the smaller one for the second. How should the £100 be most suitably divided between the two? What ratio should a first prize bear to that of a second one? Does it depend on the number of the competitors, and if so, in what way? Similar questions may be asked, but will not be answered here, when the number of prizes exceeds two. What should be the division of the £100 when three prizes are to be given, or four, or any larger number?

The interest of this memoir does not depend solely upon the answer to the above questions, but more especially on its bringing to evidence a new property of the law of frequency of error, upon which I stumbled while engaged upon a side branch of the inquiry. The problem then before me (of which the results are still unpublished) was the probability that the winner of a first or of a second prize in a given year, would succeed in winning first or second prizes in subsequent years. The data assumed the following form:—100 winners of a first place supplied $m(1)$ winners of a first place, and $n(1)$ winners of a second place in subsequent years, while 100 winners of a second place supply $m(2)$ winners of a first place and $n(2)$ winners of a second place. What are the future prize-winning capacities of winners of first and second places respectively? Let the most appropriate values of first and second prizes be called α and β, then

$$\frac{\alpha}{\beta} = \frac{\alpha . m(1) + \beta . n(1)}{\alpha . m(2) + \beta . n(2)},$$

whence $\frac{\alpha}{\beta}$ can be determined.

Having found its value for the cases with which I was dealing, I sought to compare it with another obtained through the ordinary law of frequency of error, on the following bases:

(1) I concluded that when only two prizes α and β are given, their values should not be proportioned to the absolute merits of the two competitors, but to their respective *excesses* of merit above the third competitor, who receives no prize at all. Let $[A]$, $[B]$, and $[C]$ be the first, second, and third competitors, and a, b, c the marks allotted to them, then I conceive the most suitable relation of α to β is as $(a-c)$ to $(b-c)$, and not as a to b.

(2) If there be n competitors, considered as random samples from a large body among whom merit is normally distributed, the most reasonable presumption is that they will tend to occupy n equally probable positions. In the ordinary table of the Probability Integral the argument is $\pm hx$, whose values range from 0 to $\pm$ infinity, and the tabular values are those of $\Theta(hx)$, ranging from 0 to ± 1. For the present purposes $\Theta(hx)$ must be taken as the argument, running from -1, through 0, to $+1$, and hx becomes the tabular value. If there be n competitors the most equable, and therefore the most probable distribution of them along the scale of $\pm\Theta(hx)$, is that one competitor should fall into each of the n equidistant stalls ($\frac{1}{2}n$ stalls lying on either side of 0), the septa that enclose those stalls being situated at $0, +2, +4, \ldots +n$ on the positive side and at $0, -2, -4, \ldots -n$ on the negative side. I assume that each competitor fills his stall, and that his position is expressed with needful precision by the middle of the stall. Consequently the places of the several competitors will be taken to be at $+1, +3, +5, \ldots +(n-1)$ on the positive side and at $-1, -3, -5, \ldots -(n-1)$ on the negative side. Their position is purely a question of evenly distributed probabilities, entirely unconnected with the law by which the values of hx to which they refer are established. At the same time I am aware that others may hold that this method fails in accuracy, by treating the curve of distribution as a polygon, but I shall not stop to argue the point further because the difference of result is too small to weigh in the present argument. Following a nomenclature already adopted, in which the words 'centile' and 'decile' occur, I will call the n values in any array corresponding to those of $\Theta(hx) = \pm 1, \pm 3, \pm 5, \ldots \pm(n-1)$, by the name of "equi-postiles," and those of the septa between which they stand by that of "equi-partiles."

(3) Thus far it has been implied that the value of n is known, but, as a matter of fact, it seems usually impossible to arrive at even a grossly approximate idea of the number of *virtual* competitors; which far exceeds their *actual* number in all important competitions. The number of runners in the Derby are few, but they include the best horses out of a multitude of thoroughbreds, who are all qualified for entry but whose owners keep them back because their chance of winning was found by trial performances to be *nil*. The same happens in University scholarships, in the principal athletic sports, and in all competitions that arouse a widely felt and keen desire for distinction.

Therefore being ignorant of n, I selected a few widely different values of it for trial and worked out the $\Theta(hx)$ values of $[A]$, $[B]$, and $[C]$ by the formula

$\frac{1}{n}((n-1), (n-3), (n-5))$. Then I took from the Probability Integral Tables the corresponding values of hx.

As an example of the complete process let $n=10$, then the most probable values of $\Theta(hx)$ for the ten competitors will, according to my assumption, be

$$-0{\cdot}9,\ -0{\cdot}7,\ -0{\cdot}5,\ -0{\cdot}3,\ -0{\cdot}1,\ +0{\cdot}1,\ +0{\cdot}3,\ +0{\cdot}5,\ +0{\cdot}7,\ +0{\cdot}9.$$

They are separated by equal distances from one another and by the half of those distances from the septa, including the terminals, that enclose them.

Confining ourselves to the first three terms on the positive side, that is to $+0{\cdot}9$, $+0{\cdot}7$ and $+0{\cdot}5$, we find from the Probability Integral Tables that the corresponding values of hx are $+1{\cdot}1631$, $+0{\cdot}7329$, $+0{\cdot}4770$.

The percentage values of $(a-c)$ and $(b-c)$ (as described above in (2)), are quickly derived from these. We will call them X and Y, and their sum S.

$ha = 1{\cdot}1631$	$h(a-c) = 0{\cdot}6861$
$hb = 0{\cdot}7329$	$h(b-c) = 0{\cdot}2559$
$hc = 0{\cdot}4770$	$hS \quad = 0{\cdot}9420$

$$X : 100 :: h(a-c) : hS; \quad Y : 100 :: h(b-c) : hS.$$

Whence $X = 72{\cdot}8$, $Y = 27{\cdot}2$.

Thus h disappears from the result while m, the Mean, does not come under consideration. If it had been taken into consideration by writing $m+a$ for a, $m+b$ for b, and $m+c$ for c, it would have been eliminated by the subtractions, as h was by the divisions.

Similarly if n be taken $=1000$, the values of $\Theta(hx)$ for $[A]$, $[B]$, and $[C]$ would be $+0{\cdot}9990$, $+0{\cdot}9970$, and $+0{\cdot}9950$ which give $ha = +2{\cdot}3268$, $hb = +2{\cdot}0985$, and $hc = +1{\cdot}9849$.

Proceeding in this way for many widely different values of n, I found to my astonishment that the resultant X and Y values for those of $n=10$ and above, came out curiously alike, as is shown in Table I.

TABLE I.

n	X	Y	$X+Y$
3	66·7	33·3	100·0
5	71·0	29·0	,,
10	72·8	27·2	,,
20	73·8	26·2	,,
50	74·3	25·7	,,
100	74·5	25·5	,,
1,000	75·1	24·9	,,
10,000	75·3	24·7	,,
100,000	75·4	24·6	,,

The values of X between those corresponding to $n = 50$ and $n = 100{,}000$ range within a difference of 1·1. The smallest possible class in which c is not negative, consists of five individuals, and even here the proportion of X to Y is as 71·0 to 29·0, which does not differ grossly from that in a class of 100,000 where it is 75·4 to 24·6. Nay, even taking the smallest possible class which is of three individuals, in which the values of ha, hb, hc are respectively equal to $-c$, 0, and $+c$, the value of $h(a-c) = 2hc$ and that of $h(b-c) = hc$. Consequently $S = 3hc$, therefore $X = 100 \times \frac{2}{3}$, and $Y = 100 \times \frac{1}{3}$, $= 66{\cdot}7$ and $33{\cdot}3$ as in the Table.

The rationale of the approximate uniformity of the value of X and Y seems well worthy of a more searching mathematical investigation than I am competent to make. It seems difficult to doubt that this curious property of the terminal equi-postiles is associated with others whose character cannot now be foreseen.

Comparison with facts. Many serious objections present themselves *à priori* to the useful application of this theory, among which is the partial non-conformity of examination marks with the law of frequency, especially at either end of the series, one of which is precisely the part here in question. I therefore put the theory to test by procuring through the kindness of friends a large number of sets of marks in various Civil Service examinations. I took them just as they came and found the X and Y values for each case, as in the following example.

No. 268.

$a = 1801$	$a - c = 130$
$b = 1712$	$b - c = 41$
$c = 1671$	$S = 171$

$$X : 100 :: 130 : 171; \quad Y : 100 :: 41 : 171$$

$$X = 78{\cdot}0; \quad Y = 22{\cdot}0; \text{ Total } 100.$$

I grouped these values into fives, each page of my MS. book containing that number, then into twenty-fives, and so on. Individually their values ran very irregularly, but the groups of 25 began to give hopeful indications which were fully confirmed by larger groupings, as is shown in Table II. where the X values alone are entered. Those of Y are of course complementary to them.

Thus far the evidence that the calculation was correct in principle seemed conclusive, owing to its being so remarkably well confirmed by observation. In fact, I lived for a few days in a fool's paradise, thinking that such was the case, until with the desire of probing the matter more thoroughly, I made a Table of the distribution of the individual observations. The result is shown in Table III., which shattered my sanguine hopes. If the principle upon which the calculation is based had a contributory effect to any noticeable degree, in producing the mean value of 73·4 as shown in Table II., there would have been a concentration of values about that point in Table III. But there is nothing of the kind. The values are pretty equably distributed between 50 and 100, with a slight but

distinct tendency in the smaller values to be the more numerous. This seems due to the fact that the curve of distribution (*see Natural Inheritance*) is always convex towards its axis; consequently $b-c$ is on the average less than $\frac{1}{2}(a-c)$.

TABLE II.

Values of X derived from 300 *Lists of Marks in various Civil Service Examinations.*

Mean values of successive groups of		
25 cases	50 cases	100 cases
74·4	74·5	73·6
74·6		
70·6	72·6	
74·7		
75·7	73·3	72·9
70·9		
70·4	69·6	
68·8		
73·5	73·5	73·7
73·5		
74·7	73·9	
73·1		

Mean of all 300 cases, 73·4.

Subject to this qualification, the Mean is no more than the average of random values between certain limits. Those limits are created by the conditions (1) that b cannot exceed a though it may be equal to a, in which case one of the limits is $100(a-c)$ divided by $2(a-c)$, or 50; (2) that b cannot exceed c though it may be equal to c, in which case the other limit is $100(a-c)$ divided by $(a-c)+0$, or 100.

TABLE III.

Distribution of 300 *Observed Values of X.*

50—	*55—*	*60—*	*65—*	*70—*	*75—*	*80—*	*85—*	*90—*	*95—*	Total
40	36	27	31	32	23	34	30	26	21	300
	76		58		55		64		47	300
			166				134			300

Therefore it appears to be merely a coincidence that calculation and observation lead to much the same conclusion. The principle on which the former is based

is practically neutral in its effect on the observed results, neither contributing to nor conflicting with them in a sensible degree. The curious property of the foremost equi-postiles that it discloses, must rest its claims to interest upon its own merits and not upon any effective aid that it might be supposed to afford to solving the question of the most suitable proportion between the values of first and second prizes.

What I profess to have shown is

(1) that in the three topmost equi-postiles of a normal series, whose measures are a, b, and c, the value of $(a-c)$ is roughly three times as great as that of $(b-c)$, almost independently of the number of individuals in the series and quite independently of its Mean and of its Modulus of Variability.

(2) that observation leads to practically the same result as calculation, but almost wholly for a different reason.

(3) that when only two prizes are given in any competition, the first prize ought to be closely three times the value of the second.

I now commend the subject to mathematicians in the belief that those who are capable, which I am not, of treating it more thoroughly, may find that further investigations will repay trouble in unexpected directions.

Note on Francis Galton's Problem.

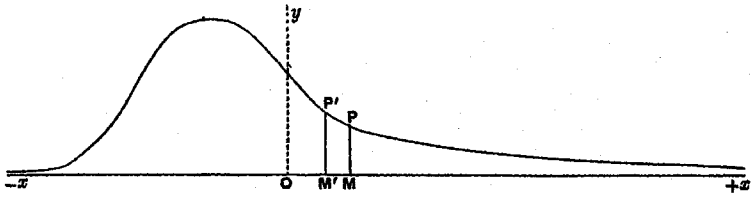
(1) THE problem proposed by Mr Galton is one of very great interest and, somewhat generalised, probably of wide application to a number of important biometrical investigations. In its generalised form it seems to open up possibilities of deducing statistical constants from comparatively small samples, for it provides us for the first time, I believe, with the most probable relationships between the individuals forming a random sample. I would state the problem as follows:

A random sample of n individuals is taken from a population of N members which when N is very large may be taken to obey any law of frequency expressed by the curve $y=N\phi(x)$, $y\delta x$ being the total frequency of individuals with characters or organs lying between x and $x+\delta x$. It is required to find an expression for the average difference in character between the p^{th} and the $(p+1)^{th}$ individuals when the sample is arranged in order of magnitude of the character.*

I propose to call this general problem: *Francis Galton's Individual Difference Problem in Statistics*, or more briefly *Galton's Difference Problem.* It will be seen at once to carry us from the consideration of the means and standard deviations of mass aggregates and arrays to the average interval between individuals of those aggregates. We may still deal with *averages*, but we fix our attention no longer on the whole population, but on definite individuals in its ordered array. This I believe to be a real advance in statistical theory.

* Clearly a knowledge of the average difference in character of adjacent individuals involves also a knowledge of the average difference in character between any two individuals.

(2) Let the figure represent any frequency distribution given by $y=N\phi(x)$, where we may suppose the limits, if finite, to be extended, if necessary, from $+\infty$ to $-\infty$ by zero ordinates. We make no hypothesis as to the nature of the distribution, or the position of the origin; as a



corollary we will deal with the case of normal distribution. Let N be the number of individuals or the area of the curve*, A the area to the left of any ordinate $PM=y$, at a character-value $OM=x$. Thus the area to the right is $N-A$. Then, if $a=A/N$, we shall have:

$$a=\int_{-\infty}^{+x}\phi(x)\,dx \qquad \text{.....(i),}$$

an integral which may be supposed known when the distribution of the general population is known.

We first note that the chance of any random individual having a character less than $x=A/N=a$, and having a character greater than $x=(N-A)/N=1-a$. Now let $OM=x_p$ correspond to the p^{th} individual's character reckoned downwards and $OM'=x_{p+1}$, to the next or $(p+1)^{\text{th}}$ individual's character. Then we require first to find the mean value of $M'M=x_p-x_{p+1}$, there being $p-1$ individuals to right of PM and $n-p-1$ individuals to left of $P'M'$ in the sample of n individuals we are selecting out of the population. The chance of an individual falling at M is given by $y_p\delta x_p/N$, and of one at M' by $y_{p+1}\delta x_{p+1}/N$; the chance of an individual to left of $P'M'=A_{p+1}/N$ and to right of $PM=(N-A_p)/N$. The total chance therefore of an individual at M, another at M' and $n-p-1$ to left of $P'M'$ and $p-1$ to right of PM

$$=\frac{y_p\delta x_p}{N}\times\frac{y_{p+1}\delta x_{p+1}}{N}\times\left(\frac{A_{p+1}}{N}\right)^{n-p-1}\times\left(\frac{N-A_p}{N}\right)^{p-1}.$$

But clearly we could permute the two individuals as well as those to right and left of PM and $P'M'$ and must introduce the factor $\lfloor n\,/\,(\lfloor n-p-1\rfloor p-1)$.† To get the average we must multiply the chance thus obtained by the corresponding x_p-x_{p+1} and first integrate from $x_{p+1}=-\infty$ to x_p and then for x_p from $-\infty$ to $+\infty$. For, the p^{th} and $(p+1)^{\text{th}}$ individuals may be anywhere in the range provided (i) there are no individuals between them, (ii) the $(p+1)^{\text{th}}$ is anywhere below the p^{th}, (iii) $p-1$ individuals fall above the latter, and (iv) $n-p-1$ individuals below the former. Hence if we write x' for x_{p+1}, x for x_p, a' for A_{p+1}/N, a for A_p/N, y_0' for y_{p+1}/N, y_0 for y_p/N, we have for χ_p the average interval between the p^{th} and $(p+1)^{\text{th}}$ individuals:

$$\chi_p=\frac{\lfloor n}{\lfloor n-p-1\rfloor p-1}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{+x}dx'\,y_0/y_0'\,a'^{n-p-1}(1-a)^{p-1}(x-x') \qquad \text{.....(ii),}$$

where by (i)

$$\frac{da'}{dx'}=y_0',\quad \frac{da}{dx}=y_0 \qquad \text{.....(iii).}$$

* Since $\int_{-\infty}^{+\infty}y\,dx=N$, it follows that $\int_{-\infty}^{+\infty}\phi(x)\,dx=1$.

† We have to find the permutations of n things which may be distributed into four groups which contain respectively $p-1$, $n-p-1$, 1, and 1 individuals. This is the same as the number of ways in which out of n factors $(x+y+z+u)$ we can pick out $(p-1)$ x's, $(n-p-1)$ y's, one z and one u, i.e. the coefficient of $x^{p-1}y^{n-p-1}zu$ in $(x+y+z+u)^n$. But this coefficient is $\lfloor n\,/\,(\lfloor p-1\rfloor n-p-1\rfloor 1\rfloor 1)$. I owe this method of looking at the factor to Dr L. N. G. Filon.

Consider first the x' integral, i.e.

$$I=\int_{-\infty}^{+x} dx'\, y_0'\, a'^{n-p-1}(x-x')=\int_{-\infty}^{+x} da'\, a'^{n-p-1}(x-x'),$$

and integrate it by parts. It equals:

$$\left[\frac{a'^{n-p}}{n-p}(x-x')\right]_{-\infty}^{+x}+\int_{-\infty}^{+x}\frac{a'^{n-p}}{n-p}dx',$$

or between limits:

$$=\frac{1}{n-p}\int_{-\infty}^{+x} a'^{n-p}dx'=\frac{1}{n-p}\,U, \text{ say.}$$

Thus:

$$\chi_p=\frac{\lfloor n}{\lfloor n-p \lfloor p-1}\int_{-\infty}^{+\infty} y_0\, U(1-a)^{p-1}dx,$$

$$=\frac{\lfloor n}{\lfloor n-p \lfloor p-1}\int_{-\infty}^{+\infty} U(1-a)^{p-1}da, \text{ by (iii)},$$

$$=\frac{\lfloor n}{\lfloor n-p \lfloor p}\left\{\left[-U(1-a)^p\right]_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{dU}{dx}(1-a)^p dx\right\},$$

or, taking the value between limits and substituting $\frac{dU}{dx}$, we have

$$\chi_p=\frac{\lfloor n}{\lfloor n-p \lfloor p}\int_{-\infty}^{+\infty} a^{n-p}(1-a)^p dx \quad \text{.....(iv).}$$

This is the complete solution of Galton's difference problem*.

An interesting theorem which results from this has been given me by Mr W. F. Sheppard; namely: the average differences between successive individuals are the successive terms in

$$\int_{-\infty}^{+\infty}\{a+(1-a)\}^n dx$$

when the subject of integration is expanded by the binomial theorem.

Given any law of frequency $y_0=\phi(x)$, we must first find a from (i), and then when tables of a have been made, calculate χ_p by quadratures from (iv). This will be fairly easy, if the distribution be assumed to be normal, for then tables of a, or tables which readily give a, already exist, and quadratures may be used on (iv) to any degree of accuracy required. This has been done by Mr Sheppard in the cases cited below for comparison with Mr Galton's results.

It will be seen that the fundamental difference between the above theory and Mr Galton's lies in the assumption of the latter, that the individual results of a special examination give a sensibly normal distribution. The above theory only assumes that the competitors are a perfectly random sample from material which if it were indefinitely large would obey the law of frequency $y_0=\phi(x)$. Of course, if we want to compare with Mr Galton's results, we must assume this law to be the normal law, but we still have the great generalisation that the actual competitors are only a random sample from a great bulk of material following this law. In any individual examination, it may be quite possible—especially if the competitors are few—that the first man stands anywhere, even below mediocrity, and the chance of this is allowed for in this the full mathematical theory.

* This result is due independently to Mr W. F. Sheppard and myself. I had stated Mr Galton's problem to him, and said that I had reduced it to a determination of $\int_{-\infty}^{+\infty} A^r dx$. He sent me, practically by return of post, the answer in the above notation, suggesting quadratures as the best practical solution, and pointing out the theorem referred to in the text.

(3) Another method of reducing the integral in (iv) without quadratures is, perhaps, of interest. I have found it convenient in other cases, where the integral limits are, or can be safely extended to, $\pm\infty$. Suppose we require to find:

$$I=\int_{-\infty}^{+\infty} U\,dx.$$

Let m be the value of x for which U reaches the maximum value U_m and let $u=\log U$; thus $(du/dx)_m=0$, unless $U_m=\infty$. Then we find:

$$\begin{aligned}U=U_m e^{\frac{1}{2}\left(\frac{d^2u}{dx^2}\right)_m\xi^2}\Big\{&1+\tfrac{1}{6}\left(\frac{d^3u}{dx^3}\right)_m\xi^3+\tfrac{1}{24}\left(\frac{d^4u}{dx^4}\right)_m\xi^4\\&+\tfrac{1}{120}\left(\frac{d^5u}{dx^5}\right)_m\xi^5+\tfrac{1}{720}\left[\left(\frac{d^6u}{dx^6}\right)_m+10\left(\frac{d^3u}{dx^3}\right)_m^2\right]\xi^6\\&+\tfrac{1}{5040}\left[\left(\frac{d^7u}{dx^7}\right)_m+35\left(\frac{d^3u}{dx^3}\right)_m\left(\frac{d^4u}{dx^4}\right)_m\right]\xi^7\\&+\tfrac{1}{40320}\left[\left(\frac{d^8u}{dx^8}\right)_m+56\left(\frac{d^3u}{dx^3}\right)_m\left(\frac{d^5u}{dx^5}\right)_m+35\left(\frac{d^4u}{dx^4}\right)_m^2\right]\xi^8\\&+\text{terms in }\xi^9\text{ and higher powers}\Big\}\quad\text{.....(v).}\end{aligned}$$

Now since U is a maximum, d^2U/dx^2 and generally d^2u/dx^2 will be negative. The limits of ξ where $x=m+\xi$ will also be $\pm\infty$, and the integral of U can thus be expressed in terms of the well-known area and moments of the probability curve. In the first place, if $1/\sigma^2=-d^2u/dx^2$

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\xi^2/\sigma^2}\xi^{2i+1}d\xi=0$$

if i be an integer.

Further:

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\xi^2/\sigma^2}\xi^{2i}\,d\xi=(2i-1)(2i-3).....3.1\sqrt{2\pi}\sigma^{2i+1}.$$

Hence writing $a_q=\left(\frac{d^qu}{dx^q}\right)_m$, we find:

$$I=\int_{-\infty}^{+\infty}Udx=U_m\sqrt{2\pi}\frac{1}{\sqrt{-a_2}}\left\{1+\frac{a_4}{8a_2^2}-\frac{a_6+10a_3^2}{48a_2^3}+\frac{a_8+56a_5a_3+35a_4^2}{384a_2^4}-\text{etc.}\right\}\text{.....(vi).}$$

The successive terms often converge with such rapidity that two or three of them are quite sufficient for practical purposes.

To apply this to our special case, we note

$$U=a^{n-p}(1-a)^p,$$

$$u=\log U=(n-p)\log a+p\log(1-a),$$

$$\frac{du}{dx}=\frac{1}{U}\frac{dU}{dx}=\left(\frac{n-p}{a}-\frac{p}{1-a}\right)\frac{da}{dx}.$$

Hence if U be a maximum, we have $du/dx=0$, and

$$a=(n-p)/n,\qquad 1-a=p/n\text{.....(vii).}$$

Thus m is to be found from

$$\frac{n-p}{n}=\int_{-\infty}^{m}y_0\,dx,$$

or, since

$$1=\int_{-\infty}^{+\infty} y_0\,dx,$$

$$\frac{p}{n}=\int_{m}^{\infty} y_0\,dx \qquad \text{(viii)}.$$

We find at once:

$$U_m=\frac{(n-p)^{n-p}\,p^p}{n^n} \qquad \text{(ix)}.$$

It remains to find the successive differentials of u for $x=m$. Let us write the value of y_0 at $x=m$, simply y_m, and we shall then have

$$(da/dx)_m=y_m,\quad d^2a/dx^2=y'_m,\quad d^3a/dx^3=y''_m,\ \text{etc.}$$

We find:

$$a_2=-\frac{n^3}{(n-p)\,p}y^2{}_m$$

$$a_3=-\frac{2n^4(n-2p)}{(n-p)^2\,p^2}y^3{}_m-\frac{3n^3}{(n-p)\,p}y_m y'_m$$

$$a_4=-6n^4\left(\frac{1}{(n-p)^3}+\frac{1}{p^3}\right)y^4{}_m-\frac{12n^4(n-2p)}{(n-p)^2\,p^2}y^2{}_m y'_m-\frac{n^3}{(n-p)\,p}(3y'^2{}_m+4y_m y''_m)$$

$$a_5=-24n^5\left(\frac{1}{p^4}-\frac{1}{(n-p)^4}\right)y^5{}_m-60n^4\left(\frac{1}{(n-p)^3}+\frac{1}{p^3}\right)y^3{}_m y'_m$$
$$-10n^3\left(\frac{1}{p^2}-\frac{1}{(n-p)^2}\right)(3y_m y'^2{}_m+2y^2{}_m y''_m)$$
$$-5n^2\left(\frac{1}{p}+\frac{1}{n-p}\right)(2y'_m y''_m+y_m y'''_m)$$

$$a_6=-120n^6\left(\frac{1}{(n-p)^5}+\frac{1}{p^5}\right)y^6{}_m-360n^5\left(\frac{1}{p^4}-\frac{1}{(n-p)^4}\right)y^4{}_m y'_m$$
$$-30n^4\left(\frac{1}{(n-p)^3}+\frac{1}{p^3}\right)(9y^2{}_m y'^2{}_m+4y^3{}_m y''_m)$$
$$-30n^3\left(\frac{1}{p^2}-\frac{1}{(n-p)^2}\right)(y'^3{}_m+4y_m y'_m y''_m+y^2{}_m y'''_m)$$
$$-n^2\left(\frac{1}{n-p}+\frac{1}{p}\right)(10y''^2{}_m+15y'_m y'''_m+6y_m y^{iv}{}_m)$$
$$\text{etc. etc.} \qquad \text{(x)}.$$

These quantities may be calculated fairly easily when y is known as a function of x, the coefficients of the y terms in n and p repeating themselves in each a.

(4) Let us apply these results to the special case when the distribution from which the material is drawn is supposed to obey the normal law. In this case, if s be the standard deviation of the material from which the sample is made:

$$y=\frac{1}{\sqrt{2\pi}\,s}\,e^{-\frac{1}{2}x^2/s^2},$$

$$a=\int_{-\infty}^{+x} y\,dx,$$

$$\chi_p=c\times\int_{-\infty}^{+\infty}a^{n-p}(1-a)^p\,dx,\ \text{if}\ c=\frac{\lfloor n}{\lfloor n-p\ \lfloor p}.$$

Write $x=sx'$, then, if $ys=y'$

$$y'=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x'^2},$$

$$a=\int_{-\infty}^{+x'}y'\,dx',$$

$$\chi_p=cs\int_{-\infty}^{+\infty}a^{n-p}(1-a)^p\,dx' \quad \text{.....(xi).}$$

Hence dropping dashes we have:

$$\frac{n-p}{n}=\int_{-\infty}^{+m}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx,$$

or:

$$\frac{n-2p}{n}=\sqrt{\frac{2}{\pi}}\int_0^m e^{-\frac{1}{2}x^2}\,dx \quad \text{.....(xii).}$$

Thus as soon as n and p are known m can be found from tables of the probability integral. Then we may find y_m from

$$y_m=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}m^2} \quad \text{.....(xiii),}$$

or tables of the ordinates of the normal curve.

We easily find by differentiating (xiii) that:

$$y'_m=-my_m,\quad y''_m=(m^2-1)y_m,\quad y'''_m=m(3-m^2)y_m$$
$$y^{iv}{}_m=(3-6m^2+m^4)y_m,\quad y^{v}{}_m=m(10m^2-15-m^4)y_m \quad \text{.....(xiv).}$$

Substituting in (x) we find:

$$a_2=-n^2\left\{\frac{1}{n-p}+\frac{1}{p}\right\}y^2{}_m \quad \text{.....(xv).}$$

$$a_3=-2n^3\left\{\frac{1}{p^2}-\frac{1}{(n-p)^2}\right\}y^3{}_m+3n^2\left\{\frac{1}{n-p}+\frac{1}{p}\right\}my^2{}_m \quad \text{.....(xvi).}$$

$$a_4=-6n^4\left\{\frac{1}{p^3}+\frac{1}{(n-p)^3}\right\}y^4{}_m+12n^3\left\{\frac{1}{p^2}-\frac{1}{(n-p)^2}\right\}my^3{}_m$$
$$-n^2\left\{\frac{1}{n-p}+\frac{1}{p}\right\}(7m^2-4)y^2{}_m \quad \text{.....(xvii).}$$

$$a_5=-24n^5\left\{\frac{1}{p^4}-\frac{1}{(n-p)^4}\right\}y^5{}_m+60n^4\left\{\frac{1}{p^3}+\frac{1}{(n-p)^3}\right\}my^4{}_m$$
$$-10n^3\left\{\frac{1}{p^2}-\frac{1}{(n-p)^2}\right\}(5m^2-2)y^3{}_m+5n^2\left\{\frac{1}{p}+\frac{1}{n-p}\right\}m(3m^2-5)y^2{}_m \quad \text{...(xviii).}$$

$$a_6=-120n^6\left\{\frac{1}{p^5}+\frac{1}{(n-p)^5}\right\}y^6{}_m+360n^5\left\{\frac{1}{p^4}-\frac{1}{(n-p)^4}\right\}my^5{}_m$$
$$-30n^4\left\{\frac{1}{p^3}+\frac{1}{(n-p)^3}\right\}(13m^2-4)y^4{}_m+30n^3\left\{\frac{1}{p^2}-\frac{1}{(n-p)^2}\right\}m(6m^2-7)y^3{}_m$$
$$-n^2\left(\frac{1}{p}+\frac{1}{n-p}\right)(31m^4-101m^2+28)y^2{}_m \quad \text{.....(xix).}$$

If the a's be found from these equations, then by (xi) and (vi):

$$\chi_p=s\frac{\lfloor n}{\lfloor n-p\ \lfloor p}\frac{(n-p)^{n-p}p^p}{n^n}\sqrt{2\pi}\sqrt{\frac{(n-p)p}{n^3}}\frac{1}{y_m}$$
$$\times\left\{1+\frac{a_4}{8a_2{}^2}-\frac{a_6+10a_3{}^2}{48a_2{}^3}+\text{etc.}\right\} \quad \text{.....(xx).}$$

Here the term in a_3^2 is generally the largest, a_4 the next and a_6 the least.

We can write the terms in the curled brackets:

$$c_1=-\frac{10a_3^2}{48a_2^3}=\tfrac{5}{6}\frac{(n-2p)^2}{n(n-p)p}-\tfrac{5}{2}\frac{n-2p}{n^2}\frac{m}{y_m}+1\tfrac{5}{8}\frac{(n-p)p}{n^3}\frac{m^2}{y^2_m} \quad\text{.....(xxii).}$$

$$c_2=\frac{a_4}{8a_2^2}=-\tfrac{3}{4}\frac{(n-p)^3+p^3}{n^2(n-p)p}+\tfrac{3}{2}\frac{n-2p}{n^2}\frac{m}{y_m}-\tfrac{1}{8}\frac{(n-p)p}{n^3}\frac{7m^2-4}{y^2_m} \quad\text{.....(xxiii).}$$

$$\begin{aligned}c_3=-\frac{a_6}{48a_2^3}&=-\tfrac{5}{2}\frac{(n-p)^5+p^5}{n^3(n-p)^2p^2}+1\tfrac{5}{2}\frac{(n-p)^4-p^4}{n^4(n-p)p}\frac{m}{y_m}\\&-\tfrac{5}{8}\frac{(n-p)^3+p^3}{n^5}\frac{13m^2-4}{y^2_m}+\tfrac{5}{8}\frac{(n-2p)(n-p)p}{n^5}\frac{m(6m^2-7)}{y^3_m}\\&-\tfrac{1}{48}\frac{(n-p)^2p^2}{n^6}\frac{31m^4-101m^2+28}{y^4_m}\end{aligned} \quad\text{.....(xxiv).}$$

And thus:

$$\chi_p=s\frac{\lfloor n}{\lfloor n-p\rfloor p}\frac{(n-p)^{n-p}p^p}{n^n}\sqrt{2\pi}\sqrt{\frac{(n-p)p}{n^3}}\frac{1}{y_m}\{1+c_1+c_2+c_3+\ldots\} \quad\text{.....(xxv).}$$

The solution of the problem is now purely arithmetical, although of course laborious.

(5) We may note some special cases.

Corollary (i). *Suppose both n and p large and not nearly equal.*

Since if q be large

$$\lfloor q=\sqrt{2\pi q}\,q^{-q}e^{-q},$$

we have

$$\chi_p=s\frac{1}{n\,y_m}\{1+c_1+c_2+c_3+\ldots\} \quad\text{.....(xxvi),}$$

a much simpler form.

Corollary (ii). *Suppose n large and p small.*

$$\chi_p=s\frac{\sqrt{2\pi p}\,p^p\,e^{-p}}{\lfloor p}\frac{1}{ny_m}\{1+c_1+c_2+c_3+\ldots\} \quad\text{.....(xxvii).}$$

Corollary (iii). Suppose n large, and that we consider Mr Galton's special problem of the ratio of the distance between the first and second to the distance between the second and third in a graduated array. Then

$$\frac{\chi_1}{\chi_2}=\frac{e}{2\sqrt{2}}\frac{y'_m}{y_m}\frac{1+c_1+c_2+c_3+\ldots}{1+c_1'+c_2'+c_3'+\ldots} \quad\text{.....(xxviii),}$$

where undashed letters refer to quantities for $p=1$ and dashed letters to the same quantities when $p=2$.

(6) As a first series of illustrations, let us apply these results to Mr Galton's consideration of the proportion of money to be given in prizes, supposing only two prizes, for the cases $n=3$, 10, 50, 100, 1000.

The following table contains the chief values*. We write:

$$\chi_p=s\times\phi(p)(1+c_1+c_2+c_3+\ldots) \quad\text{.....(xxix).}$$

Then, if $d_{rr'}$ be the difference measured in variability units between the r^{th} and r'^{th} individuals,

$$d_{rr'}=\{\chi_r+\chi_{r+1}+\chi_{r+2}+\ldots\ldots+\chi_{r'-1}\}/s,$$

* I owe to Dr Alice Lee, not only a careful revision of my numbers, but an extension of this table.

and Mr Galton takes as a reasonable measure of the prizes $100d_{13}/(d_{13}+d_{23})$ and $100d_{23}/(d_{13}+d_{23})$ per cent. of the prize money. These are obtained from the last two rows of the table.

Table of Data for Two-Prize Ratios.

$n=$	3	10	50	100	1000
m	·43074	1·28155	2·05375	2·32635	3·09040
$\log y_m$	$\bar{1}$·560,6213	$\bar{1}$·244,2739	$\bar{2}$·685,0071	$\bar{2}$·425,7300	$\bar{3}$·527,0311
m'	−·43074	·84162	1·75069	2·05375	2·87830
$\log y_{m'}$	$\bar{1}$·560,6213	$\bar{1}$·447,0995	$\bar{2}$·935,3726	$\bar{2}$·685,0071	$\bar{3}$·801,9239
$\phi(1)$	·833,910 *	·524,952 *	·380,906 †	·345,992 †	·274,009 †
$\phi(2)$	·833,910 *	·342,013 *	·222,691 †	·198,170 †	·151,399 †
c_1	+·004,736	+·031,971	+·070,072	+·084,161	+·119,233
c_2	+·011,633	−·005,875	−·032,216	−·042,268	−·066,830
c_3	−·002,055	+·000,204	+·002,068	+·001,656	−·001,876
$1+c_1+c_2+c_3$	1·0143	1·0263	1·0399	1·0435	1·0505
c_1'	+·004,736	+·007,686	+·027,170	+·035,035	+·055,246
c_2'	+·011,633	+·002,355	−·010,553	−·016,108	−·030,375
c_3'	−·002,055	−·000,327	+·000,443	+·000,517	−·000,170
$1+c_1'+c_2'+c_3'$	1·0143	1·0097	1·0171	1·0194	1·0247
$\chi_1/s=d_{12}$	·8458	·5388	·3969	·3611 ‡	·2879
$\chi_2/s=d_{23}$	·8458	·3453	·2265	·2020 ‡	·1551
$d_{13}/(d_{13}+d_{23})$	·667	·719	·733	·736	·741
$d_{23}/(d_{13}+d_{23})$	·333	·281	·267	·264	·259

The results are in fairly close agreement with those obtained from Mr Galton's investigation, which puts the first and second individuals in the places they would hold if the sample of the competitive population were actually arranged according to the normal law. His proposition that if there be two prizes they should embrace 75 and 25 per cent. respectively of the prize money is seen to be a sound rule for practical purposes when n is at all large, and might well be impressed upon the powers that rule such distributions not only in the educational world, but in rifle, athletic, sporting and agricultural competitions.

(7) We may next consider how the divergencies between individual members of an array vary when we take the pair close to one end of the array, or nearer to the centre. Let us suppose the array to contain 100 individuals; we already know the differences between the 1st and 2nd, and the 2nd and 3rd individuals. We will now find the differences between the 25th and 26th and the 50th and 51st. In other words we will determine $\chi_{25}(100)$ and $\chi_{50}(100)$. We can easily find these expressions in the more general case for n fairly large§; we have:

$$\chi_{\frac{1}{2}n}(n)=s\times\frac{2\cdot506,628}{n}\left(1+\frac{\cdot035,398}{n}-\frac{\cdot012,327}{n^2}\right)\qquad\text{(xxx)}$$

and

$$\chi_{\frac{1}{4}n}(n)=s\times\frac{3\cdot146,865}{n}\left(1+\frac{\cdot072,942}{n}-\frac{\cdot026,989}{n^2}\right)\qquad\text{(xxxi).}$$

* Calculated from (xxv).

† Calculated from (xxvii).

‡ Mr W. F. Sheppard sends me as the values for these constants deduced by quadratures ·3594 and ·2018, which thus show that our method is sufficiently approximate.

§ i.e. using (xxvi).

These will give the corrective terms in the brackets close enough, even if n be as small as 10. The terms outside the brackets will need determining by (xxvii) instead of (xxvi) if n be less than 30, say. We see that (xxx) gives us the average difference between the mediocre individuals and (xxxi) the difference between two individuals at the quartile. Roughly the differences in the two cases are as 5 to 6. But if we compare the extreme individuals' difference for $n=100$, we have

$$\chi_1 = \cdot 3611 \times s, \quad \chi_2 = \cdot 2020 \times s, \quad \chi_{25} = \cdot 0315 \times s, \quad \chi_{50} = \cdot 0251 \times s.$$

Thus the interval between extreme individuals is more than ten times the interval between mediocre individuals.

Now, of course, the normal distribution in a general sort of way indicates that the differences between modal, or what the biologists term 'normal,' individuals are very small. But Mr Galton's difference problem enables us for the first time to quantitatively appreciate how much wider the differences are between the extreme (or biologists' 'abnormal' individuals) and modal (or normal) individuals. Now the range of a distribution being somewhat about $6s$, we see that extreme individuals may be separated by as much as $\frac{1}{7}$ of the range, while modal individuals have only a difference of $\frac{1}{250}$th of the range, and even individuals at the quartile only a difference of $\frac{1}{200}$th of the range.

It is not possible to pass over the general bearing of such results on human relations. If we define 'individuality' as difference in character between a man and his immediate compeers, we see how immensely individuality is emphasised as we pass from the average or modal individuals to the exceptional man. Differences in ability, in power to create, to discover, to rule men, do not go by uniform stages. We know this by experience, but we see it here as a direct consequence of statistical theory, flowing from a characteristic and familiar chance distribution. We ought not to be surprised, as we frequently are, at the results of competitive examination, where the difference in marks between the first men is so much greater than occurs between men towards the middle of the list. In the same way the individuality of imbeciles and criminals at the other end of the intellectual and moral scales receives its due statistical appreciation.

We stand in a better position to judge the pathological from the merely exceptional, mere isolation no longer leads us to doubt the position of an extreme outlying error, observation or individual*.

In short Galton's difference problem leads us to look upon samples of populations and even on populations themselves, no longer as arrays of continuously varying individuals, but as systems of discrete units. We see discontinuity in every sample and in every population. We obtain a new and most valuable conception of a normal or standard population. It is one in which each individual is separated from his immediate neighbours, when the whole is arranged according to any character, by definite calculable intervals. These intervals are, of course, the *average* intervals which would be found by taking the mean of many such samples or populations, but they are none the less of extreme suggestiveness. Just as the *continuous* representation by a frequency curve is only an ideal representation of the observed facts, so we now reach an ideal representation of the actual *discontinuity* in the given population. As in the case of many physical investigations, so we find in statistical theory both continuous and discontinuous representations of the phenomena equally important and equally valid within the legitimate limits of interpretation.

(8) As a last illustration I propose to investigate the value of χ when $n=2$, and $p=1$. We easily find:

$$m=0, \quad y_m = \frac{1}{\sqrt{2\pi}},$$

* I propose on another occasion to consider the application of Galton's problem to a new theory for the rejection of outlying individuals.

and from (xxii)...(xxiv):

$$c_1=0, \quad c_2=\cdot 017{,}699, \quad c_3=-\cdot 003{,}081,$$
$$\chi=s\times 1\cdot 127.$$

Since a_5 the next term vanishes, I believe this result is probably true to the last figure. Anyhow I think we may say that if the individuals be taken at random from a population, then the probable value of the standard deviation of that population is nearly $\frac{23}{16}$ of the difference between the two individuals. Thus by averaging the differences between pairs of individuals taken at random we can obtain fairly readily an appreciation of the standard deviation, i.e. the variability of the general population. Further, if we take individuals, not quite at random, but from correlated groups, e.g. pairs of brothers selected at random, the $\frac{23}{16}$th of the average difference of the pairs will be the standard deviation of the correlated groups, e.g. a group of brethren; hence the degree of relationship between such correlated individuals may be determined. This is only a suggestion of one of the many possible uses of Galton's difference problem. It opens up, indeed, many new methods of inquiry, the effectiveness of which, however, can only be tested by their application in actual statistical practice. It must suffice for the present to have indicated that this difference problem marks a new, and very probably a most important, departure in statistical theory.

KARL PEARSON.