X. On a Coefficient of Class Heterogeneity or Divergence.

By KARL PEARSON, F.R.S.

(1) In considering the sub-groups of a population—especially in dealing with local races in man, animals or plants—a problem of the following character has not infrequently arisen: It is found that a sub-class, for example a local sample, differs considerably from the general population. This divergence may have any magnitude upwards from the probable limits of random sampling. We require some coefficient which will express by a single number the relative divergence from the general population of each sub-class or local group. For example, we take the frequency of alternative characteristics of the local population and find these are represented by certain percentages in the general population; we know also the percentages in the sub-group. We can, of course, take the difference of each individual percentage and of the general population percentage and find the probable error of this difference, but this gives us a series of numbers, and not a single measure of the general heterogeneity of the group. These numbers may also belong to correlated characters, and when one number marks a great excess in percentage we may expect a great defect in a second percentage for this very reason. But this makes the weight to be given to such a complex system of numbers extremely difficult to estimate.

The necessity for some general coefficient of class heterogeneity was impressed upon me, while discussing with Mr J. F. Tocher his reduction of the Anthropometrical Surveys recently made of the inmates of Scottish Asylums and of the children in Scottish Schools. It was needful to find a single number, which would measure local heterogeneity, or the divergence from a random sample of the general population in a series of characters of the local population. The number chosen must be such (i) that allowance is made for the size of the sample, (ii) that the numbers for different sub-groups or localities are strictly comparable, and (iii) that we have some idea as to the size of its probable error. Following up a suggestion of Mr Tocher I have reached what I think is a workable *coefficient of divergence*, which may be useful in dealing with local race problems.

Suppose a contingency table formed in which the columns are marked by the alternative characters under consideration and each row is peculiar to a sub-group or district. Thus let the characters be $\alpha, \beta, \gamma, \delta \dots$ and the sub-groups $a, b, c, d, e \dots$. We have the table:

TABLE I.

	α	β	γ	δ	ϵ	η	—	ω	Totals
a	$n_{a\alpha}$	$n_{a\beta}$	$n_{a\gamma}$	$n_{a\delta}$	$n_{a\epsilon}$	$n_{a\eta}$	—	$n_{a\omega}$	n_a
b	$n_{b\alpha}$	$n_{b\beta}$	$n_{b\gamma}$	$n_{b\delta}$	$n_{b\epsilon}$	$n_{b\eta}$	—	$n_{b\omega}$	n_b
c	$n_{c\alpha}$	$n_{c\beta}$	$n_{c\gamma}$	$n_{c\delta}$	$n_{c\epsilon}$	$n_{c\eta}$	—	$n_{c\omega}$	n_c
d	$n_{d\alpha}$	$n_{d\beta}$	$n_{d\gamma}$	$n_{d\delta}$	$n_{d\epsilon}$	$n_{d\eta}$	—	$n_{d\omega}$	n_d
e	$n_{e\alpha}$	$n_{e\beta}$	$n_{e\gamma}$	$n_{e\delta}$	$n_{e\epsilon}$	$n_{e\eta}$	—	$n_{e\omega}$	n_e
—	—	—	—	—	—	—	—	—	—
s	$n_{s\alpha}$	$n_{s\beta}$	$n_{s\gamma}$	$n_{s\delta}$	$n_{s\epsilon}$	$n_{s\eta}$	—	$n_{s\omega}$	n_s
Totals	n_α	n_β	n_γ	n_δ	n_ϵ	n_η	—	n_ω	N

Here the first column gives all the districts or sub-groups which form the total population. The distribution of the alternative characters in the total population is given in the last row, while the last column gives the total frequency of each sub-group. Any number such as $n_{i,j}$

is the frequency of the alternative character γ in the sub-group or district i . For example $\alpha, \beta, \gamma \dots$ may be the alternative hair colours in a country of which the different districts are $a, b, c, d \dots z$. Such tables arise over and over again in anthropometric surveys. If now each sub-group or district were a random sample of the general population, then the coefficient of contingency of this table—say the coefficient of mean square contingency—should within the limits of probable error be zero. We have thus a table of the contingency between geographical sub-districts and the alternative characteristics. And the greater this contingency the more markedly are the sub-groups divergent from random samples of the general population. In other words the population is geographically* heterogeneous. Accordingly if we take the same or nearly the same set of characters for two populations and about the same number of sub-groups or districts, such tables as the above give by their coefficients of contingency a reasonable measure of inter-racial comparison. The population or race with the highest coefficient of contingency is clearly the most heterogeneous. The relative heterogeneity of Prussians, Swedes, Italians, Scottish and, perhaps, English could, I think, be now determined in this manner from published data for at least hair and eye colour.

But we require not only an inter-racial coefficient of heterogeneity, but an intra-racial coefficient, which will measure the relative heterogeneity of the various groups. To reach this, pick out any district or sub-class b and oppose it to the rest of the population in a table of the following kind :

TABLE II.

	α	β	γ	δ	ϵ	η	—	ω	Totals
b	n_{ba}	$n_{b\beta}$	$n_{b\gamma}$	$n_{b\delta}$	$n_{b\epsilon}$	$n_{b\eta}$	—	$n_{b\omega}$	n_b
Rest	$n_a - n_{ba}$	$n_\beta - n_{b\beta}$	$n_\gamma - n_{b\gamma}$	$n_\delta - n_{b\delta}$	$n_\epsilon - n_{b\epsilon}$	$n_\eta - n_{b\eta}$	—	$n_\omega - n_{b\omega}$	$N - n_b$
Totals	n_a	n_β	n_γ	n_δ	n_ϵ	n_η	—	n_ω	N

This is also a contingency table, of a very contracted character it is true, but none the less absolutely valid, if it be only used for relative purposes. Let the coefficient of mean square contingency of this table be found and be C_b , then the relative values of C_a, C_b, C_d , etc. will be measures of the class and local differences, or what we may call intra-racial differences. I suggest these C 's as the coefficients we are seeking. We will now investigate the nature of C_b . Let χ_b^2 be the mean square contingency, then :

$$\begin{aligned} \chi_b^2 &= \frac{1}{N} \sum_a \left\{ \frac{\left(n_{ba} - \frac{n_a n_b}{N} \right)^2}{\frac{n_a n_b}{N}} + \frac{\left(n_a - n_{ba} - \frac{n_a (N - n_b)}{N} \right)^2}{\frac{n_a (N - n_b)}{N}} \right\} \\ &= \frac{1}{N} \sum_a \left\{ \frac{\left(n_{ba} - \frac{n_a n_b}{N} \right)^2}{\frac{n_a n_b}{N}} \left(1 + \frac{n_b}{N - n_b} \right) \right\} \\ &= \frac{1}{N - n_b} \sum_a \left\{ \frac{\left(n_{ba} - \frac{n_a n_b}{N} \right)^2}{\frac{n_a n_b}{N}} \right\} \dots \dots \dots (i). \end{aligned}$$

* The sub-groups need not of course be 'geographical'; they might be economically, socially, or otherwise differentiated.

Now if ϕ^2 be the mean square contingency of Table I. and if ϕ_b^2 be the contribution to it of the b row we have :

$$\phi^2 = \phi_a^2 + \phi_b^2 + \phi_c^2 + \dots + \phi_s^2$$

where

$$\phi_b^2 = \frac{1}{N} \bar{S}_a \left\{ \frac{\left(n_{ba} - \frac{n_a n_b}{N} \right)^2}{\frac{n_a n_b}{N}} \right\} \dots \dots \dots (ii).$$

Or, we have :

$$\chi_b^2 = \frac{N}{N - n_b} \phi_b^2.$$

But:

$$C_b = \sqrt{\frac{\chi_b^2}{1 + \phi_b^2}} = \sqrt{\frac{\phi_b^2}{1 - \frac{n_b}{N} + \phi_b^2}} \dots \dots \dots (iii).$$

Thus we have the following rule:

Start with Table I. and determine the contributions $\phi_a^2, \phi_b^2, \phi_c^2 \dots \phi_s^2$ of each sub-group or locality to the total mean square contingency of this table. Then $C_a, C_b, \dots C_s$ determined as above are the "coefficients of divergence" of the respective sub-groups or classes or localities from the general population, and their relative magnitudes measure the relative divergency of such groups or localities.

(2) If the Class b were, for example, merely a random sample of the general population, we should have $\phi_b^2 = 0$ and $C_b = 0$. It becomes accordingly of importance to determine the probable error of C_b on the assumption that Class b is a random sample. If C_b differs from zero by several times its probable error, the divergence of the Class is almost certainly significant. The general expression for the probable error of a coefficient of mean square contingency has been dealt with in another paper*. In the notation of that paper

$$\phi^2 \sigma_{\phi^2} = S_{pq} \left(\phi^2 \frac{n_{pq}}{n_p n_q} \right) + \frac{1}{2} S_{pq} \left(\phi_p^2 \phi_q^2 \frac{n_{pq}}{n_p n_q} \right) - \frac{1}{2} S_p \left(\frac{\phi_p^4}{n_p} \right) - \frac{1}{2} S_q \left(\frac{\phi_q^4}{n_q} \right) \dots \dots \dots (iv).$$

ϕ^2 will now be χ_b^2 and we have to perform the summations for the two rowed table, Table II, above. The q summation will be from a to ω and the p summation for the two rows of our table. I take the terms in order.

(i) $S_{pq} \left(\phi^2 \frac{n_{pq}}{n_p n_q} \right)$. This in our present notation stands for

$$\bar{S}_a \left(\phi_{ba}^2 \frac{n_{ba}}{n_b n_a} \right) + \frac{n_b}{N - n_b} \bar{S}_a \left(\phi_{ba}^2 \frac{n_a - n_{ba}}{n_a (N - n_b)} \right),$$

since ϕ_{pq}^2 for any constituent of the second row is by the line above Equation (i) $= \frac{n_b}{N - n_b} \phi_{ba}^2$, where ϕ_{ba}^2 is the contribution to the mean square contingency from the first row constituent immediately above. Let us write $v_b = n_b / (N - n_b)$. Then we have, if we write

$$\tau_b^2 = \bar{S}_a \{ \phi_{ba}^2 n_{ba} / (n_b n_a) \} \dots \dots \dots (v),$$

$$S_{pq} \left(\phi^2 \frac{n_{pq}}{n_p n_q} \right) = \tau_b^2 (1 - v_b^2) + v_b^2 \phi_b^2 / n_b \dots \dots \dots (vi).$$

(ii) $\frac{1}{2} S_{pq} \left(\phi_p^2 \phi_q^2 \frac{n_{pq}}{n_p n_q} \right)$ For the first line $\phi_q^2 = \phi_b^2$ and for the second line $= v_b \phi_b^2$. Hence

* "On the Probable Error of Mean Square Contingency," see Equation (xxxvi), *Biometrika*, Vol. v. p. 195.

the value of this, if we remember that $\phi_p^2 = \phi_{ba}^2 + r_b \phi_{ba}^2$ for the a column, is given by

$$\frac{1}{2} \phi_b^2 \sum_a \left\{ (1+r_b) \phi_{ba}^2 \frac{n_{ba}}{n_b n_a} + r_b (1+r_b) \phi_{ba}^2 \frac{n_a - n_{ba}}{n_a (N - n_b)} \right\} \\ = \frac{1}{2} (1+r_b) (1-r_b^2) \phi_b^2 r_b^2 + \frac{1}{2} (1+r_b) r_b^2 \phi_b^2 / n_b \dots\dots\dots (vii).$$

$$(iii) \quad \frac{1}{2} S_p \left(\frac{\phi_p^4}{n_p} \right) = \frac{1}{2} (1+r_b)^2 \sum_a \left(\frac{\phi_{ba}^4}{n_a} \right) = \frac{1}{2} (1+r_b)^2 \omega_b^4 \dots\dots\dots (viii),$$

$$\text{if} \quad \omega_b^4 = \sum_a \left(\frac{\phi_{ba}^4}{n_a} \right) \dots\dots\dots (ix).$$

$$(iv) \quad \frac{1}{2} S_q \left(\frac{\phi_q^4}{n_q} \right). \quad \text{This since the summation is for only two rows is given by}$$

$$\frac{1}{2} \left(\frac{\phi_b^4}{n_b} + r_b^2 \frac{\phi_b^4}{N - n_b} \right)$$

$$\text{or} \quad \frac{1}{2} S_q \left(\frac{\phi_q^4}{n_q} \right) = \frac{1}{2} (1+r_b^2) \phi_b^4 / n_b \dots\dots\dots (x).$$

Writing (vi), (vii), (viii) and (x) in (iv) we find:

$$\chi_b^2 \sigma_{\chi_b}^2 = \{ r_b^2 (1-r_b^2) + r_b^2 \phi_b^2 / n_b \} \left\{ 1 + \frac{1}{2} (1+r_b) \phi_b^2 \right\} - \frac{1}{2} \left\{ (1+r_b^2) \frac{\phi_b^4}{n_b} + (1+r_b)^2 \omega_b^4 \right\}.$$

But by the line above equation (iii) $\chi_b^2 = (1+r_b) \phi_b^2$. Hence:

$$\sigma_{\chi_b}^2 = \frac{1}{n_b} \left\{ \frac{r_b^2}{1+r_b} + \frac{n_b r_b^2}{\phi_b^2} (1-r_b) \right\} \left\{ 1 + \frac{1}{2} (1+r_b) \phi_b^2 \right\} \\ - \frac{3}{4 n_b} \left\{ (1-r_b+r_b^2) \phi_b^2 + (1+r_b) \frac{n_b \omega_b^4}{\phi_b^2} \right\} \dots\dots\dots (xi).$$

This involves a knowledge of ϕ_b^2 , r_b^2 and ω_b^4 . The first will have been found in determining the contingency coefficient of the entire table; the second in determining its probable error, and the third only has to be specially calculated.

Finally we have*

$$\sigma_{C_b} = \{1 - C_b^2\}^{\frac{1}{2}} \sigma_{\chi_b} \dots\dots\dots (xii).$$

Or the

$$\text{Probable Error of } C_b = .67449 \{1 - C_b^2\}^{\frac{1}{2}} \sigma_{\chi_b}.$$

(3) I propose to illustrate this numerically on a table already largely worked out in the paper referred to above. It has been shown that handwriting is contingent in a certain degree on grade of intelligence. I propose to investigate which group of handwriters has a distribution of intelligence most markedly different from that of the general population, i.e. which is intellectually most heterogeneous. This is not in itself a problem of any importance but it will serve to illustrate the application of the above formulæ, and the numerical work needful for their evaluation. Turning to the table, p. 197, I extracted the results given in Table III. The only new quantities to be calculated are the values of

$$\omega_b^4 = \sum_a \left(\frac{\phi_{ba}^4}{n_a} \right) = \frac{1}{N^2} \sum_a \left(\frac{(N \phi_{ba}^2)^2}{n_a} \right).$$

Now $N \phi_{ba}^2$ is the fourth number in each constituent of the table on p. 197. The squares of these from Barlow's Tables are the first number in each constituent of Table IV.; n_a is given under the total at the foot and immediately above n_a , its reciprocal. These reciprocals placed successively on the calculator and multiplied by the first number in each column

* *loc. cit.* p. 194.

TABLE III.

Handwriting	n_b	n	$1+n_b^2$	ϕ_b^2	$N\phi_b^2$	$N^2\tau_b^2$	$N^2\omega_b^4$	σ_{χ_b}	C_b	P.E. of C_b
Very Good	128	0752	1.0004	0185	32.26	53.19	2.232	0321	1395	0210
Good ...	657.5	5750	1.1901	0210	37.73	31.39	1.749	0239	1787	0154
Moderate ...	712.5	6548	1.2805	0118	21.26	18.79	.631	0238	1384	0156
Poor ...	244	1567	1.0038	0161	29.06	43.54	2.237	0296	1354	0194
Bad ...	45	0256	1.0000	0081	14.56	31.43	.776	0367	0907	0244
Very Bad ...	16	0090	1.0000	0204	36.65	261.93	15.246	0714	1418	0467

TABLE IV.
Intelligence.

Handwriting.	Quick Intelligent	Intelligent	Slow Intelligent	Slow	Slow Dull	Very Dull	$N^2\omega_b^4$
Very Good	512.57 2.006	.04 .000	3.28 .006	42.38 .182	1.42 .014	.85 .024	2.232
Good ...	25.60 .100	90.63 .144	1.28 .002	105.47 .453	61.94 .632	15.05 .418	1.749
Moderate	132.02 .517	1.66 .003	12.53 .023	20.07 .086	.21 .002	.00 .000	.631
Poor ...	3.61 .014	41.22 .065	.00 .000	73.10 .314	20.16 .206	58.98 1.638	2.237
Bad ...	.00 .000	2.04 .003	2.59 .005	9.00 .039	71.40 .729	.00 .000	.776
Very Bad	.00 .000	5.62 .009	.03 .000	.32 .001	122.32 1.248	503.55 13.988	15.246
Reciprocals Totals ...	.003914 255.5	.001586 630.5	.001825 548	.004292 233	.010204 98	.027778 36	.555247/10 ³ 1801

constituent give $(N\phi_b^2)^2/n_b$ which is recorded as the second number in each constituent. The sum of these for each row gives $N^2\omega_b^4$ recorded to the right and also in Table III., $\sigma_{\chi_b}^2$ can now be found from the form

$$\sigma_{\chi_b}^2 = \frac{1}{N} \left[\left(n_b + \frac{N^2\tau_b^2}{N\phi_b^2} (1 - n_b) \right) (1 + \frac{1}{2} (1 + n_b) \phi_b^2) - \frac{1}{2} \left(\frac{1 + n_b^2}{n_b} \phi_b^2 + (1 + n_b) \frac{N^2\omega_b^4}{N\phi_b^2} \right) \right].$$

And again

$$C_b = \sqrt{\frac{\phi_b^2}{1 - \frac{n_b}{N} + \phi_b^2}} = \sqrt{\frac{\phi_b^2}{1/(1 + n_b) + \phi_b^2}} \text{ from (iii).}$$

These values are also recorded in Table III. Then σ_{C_b} was found from (xii) and so the probable error of C_b .

The values of C_i show us that the class of "Good" handwriters is most and that of "Bad" handwriters least divergent from the general population. The other four classes have values of C_i sensibly equal and equal to .14. The "Good" handwriters have .18 and the "Bad" .09, and the question is whether these are significantly different from .14, or from each other. The probable error of the difference is about .03. It would therefore be reasonable to assume that "Good" and "Bad" handwriters do differ from each other, though it is less easy to assert marked difference from the community at large. On the whole it seems reasonable to suggest that in distribution of intelligence the "Good" handwriters are less like a random sample of the general girl population than "Bad" handwriters. In other words heterogeneity of intelligence is more marked in the class "Good" than in the class "Bad."

As I have said, the illustration is one of numerical method only and not of interest in itself. The special purpose of the present note is the suggestion of a coefficient which may be of value in the many cases in which we wish to compare the deviation of local samples of a population from the proportions exhibited by a general population.

XI. Inheritance in the Female Line of Size of Litter in Poland China Sows.

By G. M. ROMMEL, B.S.A., and E. F. PHILLIPS, Ph.D., United States Department of Agriculture, Washington, D.C.

From the data of the American Poland China Record, the authors determined the inheritance of the size of litters from mother to daughter, using 6145 litters farrowed in 1902. The methods were those commonly employed in statistical studies of heredity.

The tabulation of the sizes of litters from mothers and daughters and the determination of the coefficient of correlation (r) shows that there is an actual correlation between the size of litters of two successive generations, and the authors feel justified in concluding that size of litter is a character transmitted from mother to daughter. The coefficient of correlation for the five years is small (.06) but it is appreciable and consequently it would appear proved that by judicious selection of sows from large litters, the average for the breed may be increased.

Correlation in size of Litter of Poland China Sows between Mother (M) and Daughter (D). American Poland China Record.

Age of Daughters	Number of Cases	Mean M	Mean D	σ_M	σ_D	r	P. E. of r
1 Year ...	2010	7.908	6.6451	2.0764	1.7582	.1088	$\pm .0149$
2 Years ...	2047	7.6927	7.5598	1.9818	1.9415	.0885	$\pm .0148$
3 Years ...	1157	7.5809	7.8799	1.9615	2.0693	.0883	$\pm .0197$
4 Years ...	606	7.6304	8.2821	1.9856	2.0661	.0379	$\pm .0274$
5 Years ...	325	7.6738	8.4031	2.1001	2.1571	.0032	$\pm .0375$
1—5 Years	6145	7.7349	7.4391	2.0202	2.0312	.0601	$\pm .0086$

The decrease from .1088 to practically zero (.0032) from the first to the fifth year does not necessarily mean that the inheritance of fecundity is lost as a sow grows older, but probably indicates that inheritance from the dam gradually plays relatively less and less of a part in the determination, while other factors, notably nutrition, play more. The correlation tables are given over page. This work is being followed with an investigation of the inheritance of size of litter through the male line and from the ancestors in the female line.