Biometrika Trust

On a Correction to be Made to the Correlation Ratio η

Author(s): Karl Pearson

Source: *Biometrika*, Vol. 8, No. 1/2 (Jul., 1911), pp. 254-256

Published by: [Biometrika Trust](#)

Stable URL: <http://www.jstor.org/stable/2331454>

Accessed: 10/06/2014 07:13

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Biometrika Trust is collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*.

<http://www.jstor.org>

Hence

$$\chi^2 = 705 \times 590 \times .000,0273 = 11.36.$$

This gives us $P = .0265$, or the odds are about 37 to 1 against the boys and girls being random samples of the same material. These are nothing like the odds obtained in the scarlet fever and measles differentiation of pigmentation; but they are sufficiently considerable to render it desirable in future investigations to keep the sexes separate.

The above illustrations were simply selected because I happen to have Dr Macdonald's results before me, and it seemed desirable to ascertain whether there was a selective disease incidence by a method which would not appeal to statistics of the general population made for another purpose. The main point brought out is that the excess incidence of measles over scarlet fever in the persons recorded as fair-haired is much beyond the limits of any random sampling differences.

Many other problems to which the method can be applied will occur to the biometrician, who is "in active practice."

IX. On a Correction to be made to the Correlation Ratio η .

By KARL PEARSON, F.R.S.

It is well known that as the square of the correlation ratio always involves the sum of the squares of differences of means, it must always take a positive value. Hence its mean value even, when its value for an indefinitely large population would be zero, must for finite samples be positive. Thus such values as .05 to .10 of η may not denote small but significant values of the correlation; they may denote solely the measure of η 's mean value for actual zero correlation. The observed value of η ought to be compared with

$$\bar{\eta} \pm E_{\eta},$$

where $\bar{\eta}$ is the mean value of η for zero correlation and E_{η} the probable error of η . E_{η} has been determined by me for any value of η and $= .67449/\sqrt{N}$, when there is no correlation*.

Let n_{x_p} be the total frequency of the population falling in the subrange, centred at x_p , and let $\bar{y}_{x_p}$ be the mean value of all the y -characters associated with this subgroup of x 's. Let $\bar{y}$ be the mean, and σ_y the standard-deviation of all the y -characters. Then by definition:

$$\begin{aligned} \eta^2 &= \frac{S \{n_{x_p} (\bar{y}_{x_p} - \bar{y})^2\}}{N \sigma_y^2} = \frac{\sigma_M^2}{\sigma_y^2}, \\ \delta(\eta^2) &= \frac{\sigma_M^2 + \delta\sigma_M^2}{\sigma_y^2 + \delta\sigma_y^2} - \frac{\sigma_M^2}{\sigma_y^2} \\ &= \frac{\delta\sigma_M^2 \sigma_y^2 - \sigma_M^2 \delta\sigma_y^2}{\sigma_y^2 (\sigma_y^2 + \delta\sigma_y^2)} \end{aligned}$$

or, neglecting products of small quantities,

$$\begin{aligned} &= \frac{\delta\sigma_M^2}{\sigma_y^2} - \frac{\sigma_M^2}{\sigma_y^2} \frac{\delta\sigma_y^2}{\sigma_y^2} \\ &= \frac{\delta(\sigma_M^2)}{\sigma_y^2} - \eta^2 \frac{\delta\sigma_y^2}{\sigma_y^2}. \end{aligned}$$

Now the variation of σ_y^2 can either be positive or negative; hence if we take the mean values of both sides of the above equation

$$\text{Mean } \delta(\eta^2) = \frac{\text{Mean } \delta(\sigma_M^2)}{\sigma_y^2}.$$

Now if σ_M^2 be sufficiently large, η fairly large, $\delta(\sigma_M^2)$ may be either positive or negative with random sampling, but if η be either zero or unity, the variation of σ_M^2 must always be wholly

* "On the General Theory of Skew Correlation and Non-linear Regression." *Drapers' Company Research Memoirs*, 1905, Dulau and Co.

positive or negative, i.e. the mean value of $\delta\sigma_M^2$ must be a positive quantity or a negative quantity and cannot be zero.

Let $\bar{y}_{x_p}$, n_{x_p} and $\bar{y}$ represent the values in an indefinitely large population. Let $n_{x_p} + \delta n_{x_p}$, $\bar{y}_{x_p} + \delta \bar{y}_{x_p}$, $\bar{y} + \delta \bar{y}$ represent the values in any random sample, then

$$\begin{aligned} N\delta\sigma_M^2 &= S \{ (n_{x_p} + \delta n_{x_p}) (\bar{y}_{x_p} + \delta \bar{y}_{x_p} - \bar{y} - \delta \bar{y})^2 \} - S \{ n_{x_p} (\bar{y}_{x_p} - \bar{y})^2 \} \\ &= S \{ n_{x_p} (\delta \bar{y}_{x_p} - \delta \bar{y})^2 \} + 2S \{ n_{x_p} (\bar{y}_{x_p} - \bar{y}) (\delta \bar{y}_{x_p} - \delta \bar{y}) \} \\ &\quad + S \{ \delta n_{x_p} (\bar{y}_{x_p} - \bar{y})^2 \} + 2S \{ \delta n_{x_p} (\delta \bar{y}_{x_p} - \delta \bar{y}) (\bar{y}_{x_p} - \bar{y}) \}. \end{aligned}$$

Sum for all values and divide by number m of such values, then remembering :

- (i) That the sum of all linear variations is zero, i.e. $\Sigma (\delta n_{x_p})/m = \Sigma (\delta \bar{y}_{x_p})/m = \Sigma (\delta \bar{y})/m = 0$, and,
- (ii) That there is no correlation between the number in any array and the mean of that array or between the mean of all the y 's and the number in any x -array.

We have :
$$N \times \text{mean value of } \delta(\sigma_M^2) = S \left\{ n_{x_p} \frac{\Sigma (\delta \bar{y}_{x_p} - \delta \bar{y})^2}{m} \right\}.$$

Now :

$$\begin{aligned} \frac{\Sigma (\delta \bar{y}_{x_p})^2}{m} &= \frac{\sigma_{x_p}^2}{n_{x_p}}, \\ \frac{\Sigma (\delta \bar{y})^2}{m} &= \frac{\sigma_y^2}{N}, \\ \frac{\Sigma (\delta \bar{y}_{x_p} \delta \bar{y})}{m} &= \frac{\sigma_{x_p y}^2}{N}. \end{aligned}$$

The first two simply represent the square of the standard deviations of the mean of an array of y 's at x_p containing n_{x_p} individuals, and of the y mean of the whole population containing N individuals. The last result is proved in the footnote below*. Hence substituting these values we have :

$$N \times \text{mean value of } \delta(\sigma_M^2) = S (\sigma_{x_p}^2) + S \left(n_{x_p} \frac{\sigma_y^2}{N} \right) - 2S \frac{(n_{x_p} \sigma_{x_p y}^2)}{N}.$$

* We have :

$$\begin{aligned} n_{x_p} \bar{y}_{x_p} &= n_{x_p} y_1 + n_{x_p} y_2 + \dots + n_{x_p} y_s + \dots, \\ N \bar{y} &= n_{y_1} y_1 + n_{y_2} y_2 + \dots + n_{y_s} y_s + \dots \end{aligned}$$

Hence :

$$\begin{aligned} n_{x_p} \delta \bar{y}_{x_p} &= \delta n_{x_p} y_1 + \delta n_{x_p} y_2 + \dots + \delta n_{x_p} y_s + \dots - \bar{y}_{x_p} \delta n_{x_p}, \\ N \delta \bar{y} &= \delta n_{y_1} y_1 + \delta n_{y_2} y_2 + \dots + \delta n_{y_s} y_s + \dots \end{aligned}$$

Of if m be the number of samples :

$$\begin{aligned} n_{x_p} N \frac{\Sigma (\delta \bar{y}_{x_p} \delta \bar{y})}{m} &= - \bar{y}_{x_p} S_s \left\{ y_s \frac{\Sigma (\delta n_{y_s} \delta n_{x_p})}{m} \right\} \\ &\quad + S_s \left\{ y_s^2 \frac{\Sigma (\delta n_{y_s} \delta n_{x_p} y_s)}{m} \right\} \\ &\quad + S_{ss'} \left\{ \frac{y_s y_{s'} \Sigma (\delta n_{y_s} \delta n_{x_p} y_{s'} + \delta n_{y_{s'}} \delta n_{x_p} y_s)}{m} \right\}. \end{aligned}$$

But † :

$$\begin{aligned} \Sigma \left(\frac{\delta n_{y_s} \delta n_{x_p}}{m} \right) &= n_{x_p} y_s - \frac{n_{y_s} n_{x_p}}{N}, \\ \Sigma \left(\frac{\delta n_{y_s} \delta n_{x_p} y_s}{m} \right) &= n_{x_p} y_s \left(1 - \frac{n_{y_s}}{N} \right), \\ \Sigma \left(\frac{\delta n_{y_s} \delta n_{x_p} y_{s'}}{m} \right) &= - \frac{n_{x_p} y_s n_{y_{s'}}}{N}. \end{aligned}$$

† See *Drapers' Research Memoirs*, "Skew Correlation," p. 12, and *Biometrika*, Vol. v. pp. 191, 192.

But :

$$\begin{aligned} S(n_{x\rho} \sigma_y^2)/N &= \sigma_y^2 \times S(n_{x\rho})/N = \sigma_y^2, \\ S(n_{x\rho} \sigma_{y_{x\rho}}^2)/N &= (1-\eta^2) \sigma_y^2 \\ &= \sigma_y^2, \text{ if } \eta \text{ be zero,} \\ S(\sigma_{y_{x\rho}}^2) &= \sigma_{y_{x_1}}^2 + \sigma_{y_{x_2}}^2 + \sigma_{y_{x_3}}^2 + \dots + \sigma_{y_{x_\kappa}}^2, \text{ if there be } \kappa \text{ arrays,} \\ &= \kappa \sigma_y^2, \text{ if } \eta \text{ be zero,} \end{aligned}$$

for all the arrays will be samples of the general population.

Thus, for a system without correlation

$$\begin{aligned} N \times \text{mean value of } \delta(\sigma_M^2) &= \kappa \sigma_y^2 + \sigma_y^2 - 2\sigma_y^2 \\ &= (\kappa - 1) \sigma_y^2. \end{aligned}$$

$$\text{Mean value of } \eta^2 = \frac{\kappa - 1}{N}, \text{ for true } \eta = 0.$$

$$\text{Or, mean value of } \eta = \sqrt{\frac{\kappa - 1}{N}}.$$

Hence no value of η is really significant if it does not differ sensibly from

$$\sqrt{\frac{\kappa - 1}{N}} \pm 67449 \frac{1}{\sqrt{N}}.$$

If we do not suppose η zero, the mean addition to the value of η^2 will be

$$N \times \delta(\eta^2) = \kappa \frac{S(\sigma_{y_{x\rho}}^2)}{\kappa} \frac{1}{\sigma_y^2} + 1 - 2(1 - \eta^2).$$

The question arises as to what value should be given to the mean of the *unweighted* standard deviations squared of the arrays. We shall not err much in this small order expression, if we give it the value it would take if the arrays were weighted, or if the distribution were homoscedastic, i.e. $(1 - \eta^2) \sigma_y^2$. Accordingly we have :

$$\delta(\eta^2) = \{(\kappa - 2)(1 - \eta^2) + 1\}/N.$$

$$\text{Or : Mean observed value of } \eta^2 = \text{true value of } \eta^2 + \frac{(\kappa - 2)(1 - \eta^2) + 1}{N}.$$

$$\text{Or : } \eta^2 = \bar{\eta}^2 - \frac{(\kappa - 2)(1 - \eta^2) + 1}{N};$$

we see therefore that the correction for $\bar{\eta}^2$ rapidly diminishes as η approaches unity. Of course we have, if we prefer, instead of writing η^2 on the left, the observed value

$$\eta^2 = \{\bar{\eta}^2 - (\kappa - 1)/N\} / \{1 - (\kappa - 2)/N\},$$

where we shall put for $\bar{\eta}^2$, the mean value, the observed value as most probably coincident with that mean value.

Hence, substituting we find :

$$\begin{aligned} n_{x\rho} N \frac{\sum (\delta \bar{y}_{x\rho} \delta \bar{y})}{m} &= -\bar{y}_{x\rho}^2 n_{x\rho} + \frac{n_{x\rho}}{N} \bar{y}_{x\rho} N \bar{y} \\ &\quad + S_s(y_s^2 n_{x\rho} v_s) - S_s \frac{y_s n_{x\rho} v_s S_{s'}(n_{y_s' v_{s'}})}{N} \\ &= n_{x\rho} (-\bar{y}_{x\rho}^2 + \bar{y}_{x\rho} \bar{y} - \bar{y}_{x\rho} \bar{y}) + S_s(y_s^2 n_{x\rho} v_s) \\ &= S_s \{n_{x\rho} v_s (y_s^2 - \bar{y}_{x\rho}^2)\} = S_s \{n_{x\rho} v_s (y_s - \bar{y}_{x\rho})^2\} \\ &= n_{x\rho} \sigma_{y_{x\rho}}^2, \\ \therefore \frac{\sum (\delta \bar{y}_{x\rho} \delta \bar{y})}{m} &= \frac{\sigma_{y_{x\rho}}^2}{N}, \end{aligned}$$

which had to be proved.

C. Same cranium with L. pars condyloidea and pars basilaris removed.

B. Left pars condyloidea and pars basilaris removed.

A. Complete synchondrosis on left, and on right to within 8 mm.
Cranium E 167.

B. Bilateral incomplete synchondrosis. Cranium E 576.

A. Unilateral incomplete synchondrosis on right. Cranium E 553.

B. Bilateral incomplete synchondrosis. Cranium E 454.

A. Bilateral slight synchondrosis. Cranium E 749.

B. Synchondrosis on left side with ossicle at masto-occipital end. Cranium E 160.

A. Bilateral and complete synchondrosis from masto-occipital suture to foramen magnum. Child, Dr Derry, No. 14.

B. Trace of synchondrosis on both sides; well-marked traces of *basitotique* of Albrecht and pressure facets on hinder margin of foramen magnum. Cranium E 801.

A. Aged male skull, with slight trace of synchondrosis on right and opisthial notch. Cranium E 822.

Development of the Human Occipital (Anatomical Museum, University College, London).

- a.* Foetus of about five months.
- b.* Infant at birth.
- c.* Child of one year.

- d.* Child of about three years.
- e.* Child of about five years.