ON CERTAIN PROBABLE ERRORS AND CORRELATION COEFFICIENTS OF MULTIPLE FREQUENCY DISTRIBUTIONS WITH SKEW REGRESSION.

By L. ISSERLIS, D.Sc.

(1) In the systematic investigation of the statistical constants of multiple correlation and of their probable errors, it is important to have to hand the probable errors and the mutual correlations of the more fundamental constants—the means, the standard deviations and the correlation coefficients. For the case in which the frequency distribution follows the normal law this need is supplied in the memoir by Pearson and Filon entitled "On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation*."

As regards the more general case in which the regression is skew, the probable error of a correlation coefficient was first given by Sheppard (*Phil. Trans.* Vol. 192, A, p. 128).

The probable error of a mean and the correlation between deviations in the value of the mean and that of a standard deviation, or of a correlation coefficient, and the correlation between two standard deviations, are given by Pearson (*Biometrika*, Vol. IX. 1913, pp. 1–10).

For reference we give here the results for the case of normal distributions obtained by Pearson and Filon in the memoir referred to above:

$$\Sigma_{\sigma_1} = \sigma_1/\sqrt{2n} \qquad (1),$$

$$\Sigma_{r_{12}} = (1 - r^2_{12})/\sqrt{n} \qquad (2),$$

$$R_{\sigma_1\sigma_2} = r^2_{12} \qquad (3),$$

$$R_{\sigma_1 r_{12}} = r_{12}/\sqrt{2} \qquad (4),$$

$$R_{\sigma_1 r_{23}} = \frac{r_{12}(r_{13} - r_{12}r_{23}) + r_{13}(r_{12} - r_{13}r_{23})}{\sqrt{2}\,.\,(1 - r^2_{23})} \qquad (5),$$

$$R_{r_{12}r_{13}} = r_{23} - \frac{r_{12}r_{13}(1 - r^2_{23} - r^2_{13} - r^2_{12} + 2r_{12}r_{13}r_{23})}{2(1 - r^2_{12})(1 - r^2_{13})} \qquad (6),$$

$$R_{r_{12}r_{34}} = \frac{\left\{\begin{array}{l} \quad (r_{13} - r_{12}r_{23})(r_{24} - r_{23}r_{34}) + (r_{14} - r_{34}r_{13})(r_{23} - r_{12}r_{13}) \\ + (r_{13} - r_{14}r_{34})(r_{24} - r_{12}r_{14}) + (r_{14} - r_{12}r_{24})(r_{23} - r_{24}r_{34}) \end{array}\right\}}{2(1 - r^2_{12})(1 - r^2_{34})} \qquad (7)\dagger.$$

* *Phil. Trans.* Vol. 191, A (1898), pp. 229–311.

† *Phil. Trans.* Vol. 191, A, Equations (xv)–(xviii), (xxxvi), (xxxvii) and (xl).

In the present paper the corresponding results are obtained for the case of skew regression. The method employed is different and by supposing the regression to be linear and the distribution to be normal, a confirmation is obtained of the above results which in Pearson and Filon's memoir depend on very complicated analysis.

(2) We may begin by discussing the correlation that exists between deviations from their means in the case of two correlation coefficients r_{xy} and r_{zt}.

We have
$$r_{xy} = p_{xy}/\sqrt{(p_{x^2}p_{y^2})} \quad\text{..................................(8)},$$
$$r_{zt} = p_{zt}/\sqrt{(p_{z^2}p_{t^2})} \quad\text{..................................(9)},$$
where $p_{x^l y^m z^n t^k}$ is employed to denote the mixed moment of orders l, m, n, k in the variables, taken about the means so that
$$\frac{dr_{xy}}{r_{xy}} = \frac{dp_{xy}}{p_{xy}} - \frac{1}{2}\frac{dp_{x^2}}{p_{x^2}} - \frac{1}{2}\frac{dp_{y^2}}{p_{y^2}} \quad\text{........................(10)},$$
and
$$\frac{dr_{zt}}{r_{zt}} = \frac{dp_{zt}}{p_{zt}} - \frac{1}{2}\frac{dp_{z^2}}{p_{z^2}} - \frac{1}{2}\frac{dp_{t^2}}{p_{t^2}} \quad\text{........................(11)}.$$

It is clear that we shall require the correlations between any one of p_{xy}, p_{x^2}, p_{y^2} and any one of p_{zt}, p_{z^2} and p_{t^2}. It will suffice to find the correlation between p_{xy} and p_{zt}.

Now
$$Np_{xy} = \underset{x}{S}\,\underset{y}{S}\{n_{xy}(x-\bar{x})(y-\bar{y})\} \quad\text{........................(12)},$$
$$\therefore\quad Ndp_{xy} = \underset{x}{S}\,\underset{y}{S}\{dn_{xy}(x-\bar{x})(y-\bar{y})\}$$
$$+\underset{x}{S}\,\underset{y}{S}\{-n_{xy}(y-\bar{y})\,d\bar{x}\} + \underset{x}{S}\,\underset{y}{S}\{-n_{xy}(x-\bar{x})\,d\bar{y}\} \quad\text{......(13)},$$
or
$$Ndp_{xy} = \underset{x}{S}\,\underset{y}{S}\{dn_{xy}XY\} \quad\text{..............................(14)},$$
if we denote the total population by N, $x-\bar{x}$ by X, $y-\bar{y}$ by Y and remember that
$$\underset{x}{S}\{(x-\bar{x})\,n_{xy}\} = \underset{y}{S}\{(y-\bar{y})\,n_{xy}\} = 0.$$
Similarly
$$Ndp_{zt} = \underset{z}{S}\,\underset{t}{S}\{dn_{zt}ZT\} \quad\text{..............................(15)}.$$

The mean value of $dn_{xy}dn_{zt}$ in many samples is the mean value of
$$\underset{h}{S}\,\underset{h'}{S}\{dn_{x_s y_{s'} z_h t_{h'}}\} \times \underset{k}{S}\,\underset{k'}{S}\{dn_{x_k y_{k'} z_l t_{l'}}\}$$
$$= -\frac{1}{N}\underset{h}{S}\,\underset{h'}{S}\,\underset{k}{S}\,\underset{k'}{S}\{dn_{x_s y_{s'} z_h t_{h'}}\,dn_{x_k y_{k'} z_l t_{l'}}\} + n_{x_s y_{s'} z_l t_{l'}}(1 - n_{x_s y_{s'} z_l t_{l'}}/N) \quad\text{...(16)},$$
where in the fourfold summation the term
$$(dn_{x_s y_{s'} z_l t_{l'}})^2$$
is omitted.

But clearly the right-hand member of (16) reduces to
$$n_{x_s y_{s'} z_l t_{l'}} - n_{x_s y_{s'}} n_{z_l t_{l'}}/N,$$
hence the mean value of
$$N^2 dp_{xy}dp_{zt} \quad\text{is}\quad N(p_{xyzt} - p_{xy}p_{zt}) \quad\text{....................(17)}.$$

Putting $t = z$ we deduce from (17)

$$\text{Mean value of } dp_{xy}\,dp_{z^2} = (p_{xyz^2} - p_{xy}p_{z^2})/N \quad \text{..............(18)},$$

and putting $y = x$ in this result,

$$\text{Mean value of } dp_{x^2}\,dp_{z^2} = (p_{x^2z^2} - p_{x^2}p_{z^2})/N \quad \text{..............(19)}.$$

If we multiply (10) and (11), sum for all samples and divide by the number of samples we deduce

$$\begin{aligned} N\sigma_{r_{xy}}\sigma_{r_{zt}}R_{r_{xy},r_{zt}}/r_{xy}r_{zt} &= \frac{p_{xyzt} - p_{xy}p_{zt}}{p_{xy}p_{zt}} - \frac{1}{2}\frac{(p_{xyz^2} - p_{xy}p_{z^2})}{p_{xy}p_{z^2}} - \frac{1}{2}\frac{(p_{xyt^2} - p_{xy}p_{t^2})}{p_{xy}p_{t^2}} \\ &- \frac{1}{2}\frac{(p_{x^2zt} - p_{x^2}p_{zt})}{p_{x^2}p_{zt}} - \frac{1}{2}\frac{(p_{y^2zt} - p_{y^2}p_{zt})}{p_{y^2}p_{zt}} + \frac{1}{4}\frac{(p_{x^2z^2} - p_{x^2}p_{z^2})}{p_{x^2}p_{z^2}} \\ &+ \frac{1}{4}\frac{(p_{x^2t^2} - p_{x^2}p_{t^2})}{p_{x^2}p_{t^2}} + \frac{1}{4}\frac{(p_{y^2z^2} - p_{y^2}p_{z^2})}{p_{y^2}p_{z^2}} + \frac{1}{4}\frac{(p_{y^2t^2} - p_{y^2}p_{t^2})}{p_{y^2}p_{t^2}} \quad \text{......(20)}. \end{aligned}$$

This result like Sheppard's formula for $\sigma^2_{r_{xy}}$ is much simpler when expressed in *reduced moments*. Let us write

$$\frac{p_{x^ly^mz^nt^k}}{\sigma_x{}^l\sigma_y{}^m\sigma_z{}^n\sigma_t{}^k} = q_{x^ly^mz^nt^k},$$

so that q_{x^2} is unity and $q_{xy} = r_{xy}$. The numerical term in (20) is

$$-1 + \tfrac{1}{2}(4) + \tfrac{1}{4}(-4)$$

or zero, hence

$$\begin{aligned} N\sigma_{r_{xy}}\sigma_{r_{zt}}R_{r_{xy},r_{zt}}/r_{xy}r_{zt} &= \frac{q_{xyzt}}{r_{xy}r_{zt}} - \frac{1}{2}\left(\frac{q_{xyz^2} + q_{xyt^2}}{r_{xy}} + \frac{q_{x^2zt} + q_{y^2zt}}{r_{zt}}\right) + \frac{1}{4}(q_{x^2z^2} + q_{x^2t^2} + q_{y^2z^2} + q_{y^2t^2}) \\ &\quad \text{.........(21)}. \end{aligned}$$

In the same notation Sheppard's formula becomes

$$\frac{\sigma^2_{r_{xy}}}{r^2_{xy}} = \frac{1}{N}\left\{\frac{q_{x^2y^2}}{r^2_{xy}} + \tfrac{1}{4}(\beta_2 + \beta'_2) + q_{x^2y^2} - \frac{q_{x^3y} + q_{xy^3}}{r_{xy}}\right\} \quad \text{.........(22)*}.$$

To find the correlation between r_{xy} and r_{xz} we have only to replace t by x in (21), thus

$$\begin{aligned} N\sigma_{r_{xy}}\sigma_{r_{xz}}R_{r_{xy},r_{xz}}/r_{xy}r_{xz} &= \frac{q_{x^2yz}}{r_{xy}r_{xz}} - \frac{1}{2}\left(\frac{q_{xyz^2} + q_{x^3y}}{r_{xy}} + \frac{q_{x^3z} + q_{xy^2z}}{r_{xz}}\right) + \frac{1}{4}(q_{x^2z^2} + q_{x^4} + q_{x^2y^2} + q_{y^2z^2}) \\ &\quad \text{.........(23)}. \end{aligned}$$

(3) These correlation coefficients will simplify if the regression be linear and simplify to a considerable extent if at the same time the distribution be normal. For with *linear regression*

$$\begin{aligned} Np_{x^2yz} &= \underset{x}{S}\,\underset{y}{S}\,\underset{z}{S}\,(n_{xyz}x^2yz)\dagger \\ &= \underset{x}{S}\,\underset{y}{S}\,(n_{xy}x^2y \times \bar{z}_{xy}), \end{aligned}$$

where $\bar{z}_{xy}$ is the mean value of z for given values of x and y.

* For the denominator of left-hand side, cf. *Biometrika*, Vol. IX. p. 4.

† The origin being taken at the mean.

But from the usual regression equation

$$\bar{z}_{xy} = x\left(\frac{r_{xz} - r_{xy}r_{yz}}{1 - r^2_{xy}}\right)\frac{\sigma_z}{\sigma_x} + y\left(\frac{r_{yz} - r_{xz}r_{xy}}{1 - r^2_{xy}}\right)\frac{\sigma_z}{\sigma_y},$$

so that for linear regression

$$p_{x^2yz} = \left(\frac{r_{xz} - r_{xy}r_{yz}}{1 - r^2_{xy}}\right)\frac{\sigma_z}{\sigma_x}\, p_{x^2y} + \left(\frac{r_{yz} - r_{xz}r_{xy}}{1 - r^2_{xy}}\right)\frac{\sigma_z}{\sigma_y}\, p_{x^2y^2} \;\ldots\ldots\ldots\ldots(24).$$

Further

$$p_{x^2y} = \frac{1}{N}\underset{x}{S}\,\underset{y}{S}\,(n_{xy}x^2y)$$

$$= \frac{1}{N}\underset{x}{S}\,(n_x x^2\bar{y}_x)$$

$$= \frac{1}{N}\underset{x}{S}\left(n_x x^4\frac{\sigma_y}{\sigma_x}r_{xy}\right),$$

or

$$p_{x^2y} = p_x\, r_{xy}\frac{\sigma_y}{\sigma_x} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(25),$$

while

$$q_{x^2y^2} = 1 + r^2_{xy}\sqrt{(\beta_2 - 1)(\beta'_2 - 1)} \text{ approximately } \ldots\ldots\ldots(26)^*,$$

so that (23) can be evaluated approximately by the use of simple moment coefficients and correlation coefficients only. *If in addition the distribution be normal*, we know that

$$p_{x^3y} = 3p_{xy}p_{x^2} \text{ and } p_{x^2y^2} = (1 + 2r^2_{xy})/p_{x^2}p_{y^2},$$

so that for normal distributions

$$\frac{p_{x^2yz}}{p_{xy}p_{xz}} = \left(\frac{r_{xz} - r_{xy}r_{yz}}{1 - r^2_{xy}}\right)\frac{3}{r_{xz}} + \frac{r_{yz} - r_{xz}r_{xy}}{1 - r^2_{xy}}\,\frac{1 + 2r^2_{xy}}{r_{xy}r_{xz}},$$

or

$$\frac{q_{x^2yz}}{r_{xy}r_{xz}} = 2 + r_{yz}/r_{xy}r_{xz} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(27).$$

Similarly

$$\frac{q_{xyz^2}}{r_{xy}} = 1 + \frac{2r_{yz}r_{xz}}{r_{xy}} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(28),$$

and

$$\frac{q_{xy^2z}}{r_{xz}} = 1 + 2r_{yz}r_{xy}/r_{xz} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(29),$$

Substituting these values in (23) we obtain, after some reduction and using

$$\sigma_{r_{xy}} = (1 - r^2_{xy})/\sqrt{N},$$

$$(1 - r^2_{xy})(1 - r^2_{xz})\,R_{r_{xy}r_{xz}}$$
$$= r_{yz}(1 - r^2_{xy})(1 - r^2_{xz}) - \tfrac{1}{2}r_{xy}r_{xz}(1 - r^2_{xy} - r^2_{yz} - r^2_{xz} + 2r_{xy}r_{yz}r_{xz}) \;\ldots\ldots\ldots(30),$$

agreeing with (6) the value obtained by Pearson and Filon for normal distributions.

As regards the more general case dealing with the correlation between deviations of r_{xy} and those of r_{zt} given by equation (22), we have *when the regression is linear*

$$p_{xyzt} = \underset{x}{S}\,\underset{y}{S}\,\underset{z}{S}\,\underset{t}{S}\,\{n_{xyzt}(xyzt)\}$$

$$= \underset{x}{S}\,\underset{y}{S}\,\underset{z}{S}\,\{n_{xyz}(xyz\bar{t}_{xyz})\},$$

* *Biometrika*, Vol. IX. p. 4.

where $\bar{t}_{xyz}$ is the mean value of t for given values of x, y and z, so that as is well known

$$\bar{t}_{xyz} = -x\frac{\sigma_t}{\sigma_x}\frac{\Delta_{xt}}{\Delta_{tt}} - y\frac{\sigma_t}{\sigma_y}\frac{\Delta_{yt}}{\Delta_{tt}} - z\frac{\sigma_t}{\sigma_z}\frac{\Delta_{zt}}{\Delta_{tt}} \quad \text{..................(31)},$$

where

$$\Delta = \begin{vmatrix} 1, & r_{xy}, & r_{xz}, & r_{xt} \\ r_{xy}, & 1, & r_{yz}, & r_{yt} \\ r_{xz}, & r_{yz}, & 1, & r_{zt} \\ r_{xt}, & r_{yt}, & r_{zt}, & 1 \end{vmatrix} \quad \text{..............................(32)},$$

and Δ_{pq} is the minor corresponding to r_{pq}. Thus

$$q_{xyzt} = -(\Delta_{xt}q_{x^2yz} + \Delta_{yt}q_{xy^2z} + \Delta_{zt}q_{xyz^2})/\Delta_{tt} \quad \text{...............(33)},$$

so that $R_{r_{xy}\cdot r_{zt}}$ can be evaluated approximately in the case of linear regression without employing any mixed moments beyond the simple product moment occurring in a correlation coefficient.

For *normal distributions* we may use (27), 28) and (29) giving

$$q_{xyzt} = -[\Delta_{xt}(2r_{xy}r_{xz} + r_{yz}) + \Delta_{yt}(2r_{xy}r_{yz} + r_{xz}) + \Delta_{zt}(2r_{xz}r_{yz} + r_{xy})]/\Delta_{tt} \text{ ...(34)}.$$

By well-known properties of first minors of a determinant we have from (32)

$$\Delta_{xt} + r_{xy}\Delta_{yt} + r_{xz}\Delta_{zt} + r_{xt}\Delta_{tt} = 0 \quad \text{........................(35)},$$

$$r_{xy}\Delta_{xt} + \Delta_{yt} + r_{yz}\Delta_{zt} + r_{yt}\Delta_{tt} = 0 \quad \text{........................(36)},$$

$$r_{xz}\Delta_{xt} + r_{yz}\Delta_{yt} + \Delta_{zt} + r_{zt}\Delta_{tt} = 0 \quad \text{........................(37)}.$$

Multiply these equations by r_{yz}, r_{xz}, r_{xy} respectively and add,

$$\therefore\ (r_{yz} + 2r_{xz}r_{xy})\Delta_{xt} + (r_{xz} + 2r_{yz}r_{xy})\Delta_{yt}$$
$$+ (r_{xy} + 2r_{yz}r_{xz})\Delta_{zt} + (r_{yz}r_{xt} + r_{xz}r_{yt} + r_{xy}r_{zt})\Delta_{tt} = 0 \text{......(38)}.$$

Combining this result with (33) we see that for normal distributions

$$q_{xyzt} = r_{xy}r_{zt} + r_{yz}r_{xt} + r_{xz}r_{yt} \quad \text{........................(39)*},$$

an interesting result likely to prove useful in other applications and probably capable of generalisation Particular cases of (39) are obtained by putting $t = x$ so that $q_{x^2yz} = r_{yz} + 2r_{xy}r_{xz}$ which is (27) and $t = x$, $z = y$ giving $q_{x^2y^2} = 1 + 2r^2_{xy}$ which is well known.

If we now substitute these values in equation (21) we find

$$N\sigma_{r_{xy}}\sigma_{r_{zt}}R_{r_{xy}r_{zt}}$$
$$= 2r_{yz}r_{xt} + 2r_{xz}r_{yt} - 2r_{xz}r_{yz}r_{zt} - 2r_{xt}r_{yt}r_{zt}$$
$$- 2r_{xz}r_{xt}r_{xy} - 2r_{yz}r_{yt}r_{xy} + r_{xy}r_{zt}(r^2_{xz} + r^2_{xt} + r^2_{yz} + r^2_{yt}).$$

The right-hand member can be put in the form

$$\tfrac{1}{2}\{(r_{xt} - r_{xz}r_{zt})(r_{yz} - r_{xy}r_{xz}) + (r_{xt} - r_{xy}r_{yt})(r_{yz} - r_{yt}r_{zt})$$
$$+ (r_{xz} - r_{xy}r_{yz})(r_{yt} - r_{yz}r_{zt}) + (r_{xz} - r_{xt}r_{zt})(r_{yt} - r_{xy}r_{xt})\},$$

* This result, which is accurate for normal distributions, is given as approximately true for such distributions by H. E. Soper, *Biometrika*, Vol. IX. p. 100

and if we remember that for normal distributions

$$\sigma_{r_{xy}} = (1 - r^2_{xy})/\sqrt{N}, \quad \sigma_{r_{zt}} = (1 - r^2_{zt})/\sqrt{N},$$

this result agrees with Pearson and Filon's value quoted above as equation (7).

(4) To find the probable error of a standard deviation.

$$\sigma^2_x = p_{x^2},$$

$$\therefore \frac{d\sigma_x}{\sigma_x} = \frac{dp_{x^2}}{2p_{x^2}} \quad \text{...(40).}$$

Hence

$$\frac{\Sigma^2_{\sigma_x}}{\sigma^2_x} = \frac{p_{x^4} - p^2_{x^2}}{4p^2_{x^2}N} \text{ by (17)}$$

$$= \frac{\beta_2 - 1}{4N},$$

$$\therefore \Sigma_{\sigma_x} = \frac{\sigma_x\sqrt{\beta_2 - 1}}{2\sqrt{N}} \quad \text{....................................(41).}$$

This result is well known, and for normal distributions, i.e. when $\beta_2 = 3$, becomes $\Sigma_{\sigma_x} = \frac{\sigma_x}{\sqrt{2N}}$, agreeing with (1).

To find the correlation between a standard deviation σ_x and a correlation coefficient r_{yz}, we multiply (40) by the equation

$$\frac{dr_{yz}}{r_{yz}} = \frac{dp_{yz}}{p_{yz}} - \frac{dp_{y^2}}{2p_{y^2}} - \frac{dp_{z^2}}{2p_{z^2}},$$

and sum for all samples and divide by their number in the usual way, obtaining

$$2N\sigma_{\sigma_x}\sigma_{r_{yz}}R_{\sigma_x, r_{yz}}/\sigma_x r_{yz}$$

$$= (p_{x^2yz} - p_{x^2}p_{yz})/p_{x^2}p_{yz} - \tfrac{1}{2}(p_{x^2y^2} - p_{x^2}p_{y^2})/p_{x^2}p_{y^2} - \tfrac{1}{2}(p_{x^2z^2} - p_{x^2}p_{z^2})/p_{x^2}p_{z^2}$$

$$= \frac{q_{x^2yz}}{r_{yz}} - \frac{q_{x^2y^2} + q_{x^2z^2}}{2} \quad \text{...(42),}$$

a result which as before can be approximated to in the case of linear regression, and which for normal distributions becomes*

$$R_{\sigma_x, r_{yz}} = \frac{\dfrac{r_{yz} + 2r_{xy}r_{xz}}{r_{yz}} - \tfrac{1}{2}(2 + 2r^2_{xy} + 2r^2_{xz})}{2N \cdot \dfrac{\sigma_x}{\sqrt{2N}} \cdot \dfrac{(1 - r^2_{yz})}{\sqrt{N}} \cdot \dfrac{1}{\sigma_x r_{yz}}}$$

$$= \frac{2r_{xy}r_{xz} - (r^2_{xy} + r^2_{xz})\,r_{yz}}{\sqrt{2}\,.\,(1 - r^2_{yz})} \quad \text{..............................(43),}$$

agreeing with equation (5).

* For the case $z = x$, i.e. $R_{\sigma_x, r_{xy}}$, cf. *Biometrika*, Vol. IX. p. 8.