for the more serious business of the Session. To name all who assisted in both respects would be impossible, but they will be rewarded by the grateful appreciation of all who participated in the memorable fourteenth session of the Institute.

The "banquet d'adieu" was on this occasion, although a farewell to our hosts at Vienna, not the signal for the complete break-up of the gathering. Many members were able to accept the cordial invitation of the Municipality of Prague and the Czech "syndicats d'initiatives" to visit their city, and never were guests made to feel more welcome. Leaving Vienna on Sunday, the members on arrival in Prague, in the evening, attended an informal reception in the beautifully decorated rooms of the Public Hall. On Monday morning there was a formal reception at the old Hôtel de Ville, and afterwards the guests were taken in tramcars through some of the interesting parts of the city and up to the Castle. After visiting the Castle and Cathedral, lunch was served on the Belvedere, with its glorious view over river and city. Another tour on the tramcars followed, and the guests reassembled for a banquet in the concert hall of the Public Hall at 5.30, from which they were conducted to the Opera House for a performance of Dvořák's "Rusalka." The following morning those who could do so visited the Statistical Bureau and the Ethnological Museum, where is preserved a collection illustrating the arts and industries of the Bohemian peasants. The visitors carried away the most delightful recollections of their welcome to the beautiful city of Prague, and to Dr. Malý in particular the thanks of the English members are due.

On the Criterion of Goodness of Fit of the Regression Lines and on the Best Method of Fitting them to the Data.

By E. Slutsky, *Lecturer in Mathematical Statistics, The Commercial Institute, Kiev (Russia).*

I.

Suppose we have an uncorrelated system of variables with the deviations from their means $x_1, x_2, \ldots \ldots x_n$ and with standard deviations $\sigma_1, \sigma_2, \ldots \ldots \sigma_n$. On the hypothesis of the normal distribution the equation to the frequency surface will be—

$$(1) \quad \ldots \ldots \ldots \quad Z = Ce^{-\frac{1}{2}S\left(\frac{x_i^2}{\sigma_i^2}\right)}$$

and the equation to the generalised ellipsoid, giving the system of equally probable values of $x_1, x_2, \ldots \ldots x_n$:—

$$(2) \quad \ldots \ldots \ldots \quad S\left(\frac{x_1^2}{\sigma_1^2}\right) = \chi^2.$$

Now the equations (1) and (2) are only particular forms of the more general expressions dealt with by Prof. Pearson in his memoir "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling."* We conclude, therefore, that the probability of an uncorrelated system of n errors occurring with a frequency as great as or less than that of the observed system will be given by the same expressions which have been found by Prof. Pearson in the paper cited.

Thus we shall have—

$$P = \sqrt{\frac{2}{\pi}}\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2} d\chi + \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}\chi^2}\left(\frac{\chi}{1} + \frac{\chi^3}{1.3} + \frac{\chi^5}{1.3.5} + \ldots + \frac{\chi^{n-2}}{1.3.5.\ldots.\overline{n-2}}\right)$$

if n be odd, and

$$P = e^{-\frac{1}{2}\chi^2}\left(1 + \frac{\chi^2}{2} + \frac{\chi^4}{2.4} + \frac{\chi^6}{2.4.6} + \ldots + \frac{\chi^{n-2}}{2.4.6\ldots.\overline{n-2}}\right)$$

if n be even.

The values of P have been tabulated by Palin Elderton,† so that to find our P we must only enter the tables with the arguments χ^2 given above by (2), and $n' = n' + 1$.

These results we will apply to the problem of testing the goodness of fit of the theoretical regression line.

Let $y_{x_1}, y_{x_2}, \ldots . y_{x_n}$ be the means of the x-arrays and $Y_1, Y_2, \ldots . Y_n$ the ordinates of the regression line with the equation—

$$y = f(x, a_1, a_2 \ldots . a_p).$$

Now it is known that there is no correlation between the deviations in the mean of an x-array and in the mean of a second x-array.‡ These deviations being—

$$e_1 = Y_1 - y_{x_1}, e_2 = Y_2 - y_{x_2}, \ldots . e_n = Y_n - y_{x_n}$$

their standard deviations can easily be found if we know the standard deviations of y (σ_{n_x}) and the frequencies (n_x) in each x-array.

They are—

$$\Sigma_{y_{x_1}} = \frac{\sigma_{n_{x_1}}}{\sqrt{n_{x_1}}}, \Sigma_{y_{x_2}} = \frac{\sigma_{n_{x_2}}}{\sqrt{n_{x_2}}}, \ldots . \Sigma_{y_{x_n}} = \frac{\sigma_{n_{x_n}}}{\sqrt{n_{x_n}}}.§$$

* *Phil. Mag.*, 5th series, vol. 1, 1900, pp. 157–175.

† *Biometrika*, vol. i, pp. 155–163.

‡ Karl Pearson, "On the General Theory of Skew Correlation and Non-Linear Regression." *Drap. Comp. Research Memoirs, Biometric*, Series II, p. 13.

§ *Ibid.*, p. 14, Proposition VI.

Then we have only to form the value—

$$(3) \quad . \quad . \quad . \quad . \quad \chi^2 = S\left(\frac{e_i^2}{\sigma_{n_{x_i}}^2/n_{x_i}}\right) = S\left(\frac{n_{x_i} e_i^2}{\sigma_{n_{x_i}}^2}\right)$$

and the tables of Palin Elderton will give us (for $n' = n + 1$) the value of the probability in question.

Illustration A.

Let us investigate the closeness of fit of the cubical parabola found by Prof. Pearson for the correlation between age and head height in girls.* I take the cubic (*c*), which is considered by the author as the best, the equation to which is—

$$(4) \quad . \quad . \quad . \quad . \quad Y_{x_p} = 0{\cdot}280194 + 0{\cdot}722886\, X_p - 0{\cdot}029580\, X_p^2 - 0{\cdot}002223\, X_p^3.$$

The standard deviations, given by Prof. Pearson in 2-mm. units, I express in the same units as the heights, *i.e.*, in millimetres, and obtain the following table (see Table 1) :—

TABLE 1.—*Mean auricular height of girl's head at given age.*

Age.	Height. Observed.	Height. Calculated from cubic (c).	Errors.	Frequencies.	Standard deviation.	
x_p.	y_{x_p}.	Y_{x_p}.	$e_p = \|Y_{x_p} - y_{x_p}\|$	n_{x_p}.	$\sigma_{n_{x_p}}$.	$\frac{n_{x_p} e_p^2}{\sigma^2_{n_{x_p}}}$.
3·5	115·25	116·90	1·65	1	5·7*	0·084
4·5	116·96	117·66	0·70	7	5·7706	0·103
5·5	117·47	118·42	0·95	18	5·8552	0·474
6·5	119·10	119·24	0·14	40	5·9282	0·022
7·5	120·30	120·08	0·22	76	5·9764	0·103
8·5	121·63	120·93	0·70	125	5·2732	2·203
9·5	121·72	121·78	0·06	177	6·7754	0·014
10·5	122·82	122·62	0·20	235	5·9306	0·267
11·5	123·14	123·42	0·28	261	6·4178	0·497
12·5	123·89	124·18	0·29	309	6·4122	0·632
13·5	124·86	124·88	0·02	263	6·7178	0·002
14·5	125·71	125·52	0·19	198	7·1730	0·139
15·5	126·16	126·07	0·09	214	6·9326	0·036
16·5	126·53	126·52	0·01	162	7·7392	0·000
17·5	126·91	126·87	0·04	95	6·3358	0·004
18·5	127·02	127·09	0·07	61	6·2470	0·008
19·5	129·56	127·18	2·38	13	9·6812	0·787
20·5	123·82	127·11	3·29	7	5·0622	2·955
21·5	126·50	126·88	0·38	8	8·2828	0·017
22·5	125·25	126·48	1·23	2	1·9148	0·825
	....	....	....	2,272	....	$\chi^2 = 9{\cdot}17$

* The frequency in this group being *unity* the standard deviation equals zero (Pearson, *l.c.*, Table III). It is clear, however, that we have here an error in σ_{n_x} due to random sampling, and that it would be quite reasonable to omit this group. I prefer to maintain it, assuming for σ_{n_x} a value obtained by a rough extrapolation.

* *Ibid.*, pp. 34–38.

Thus we find $\chi^2 = 9{\cdot}17$, $n' = 20 + 1$, and $P = 0{\cdot}98$ (from Palin Elderton's tables), and conclude that the fit is an extremely good one; then if we assume the values in the general population distributed in accordance with the cubic, the deviations due to random sampling equally improbable or more improbable than the observed ones, would occur 98 times in 100 cases.

Illustration B.

The following table (Table 2) shows the correlation between the mean monthly price of rye in Samara (y) and the mean monthly price of rye in the same town a month before (x). The headings of rows and columns give the prices in copecks per pud:—

TABLE 2.—*Correlation between prices of rye at monthly intervals.*
Prices of rye in Samara a month before.

Copecks per pud. (*Prices of rye in Samara.*)	25.	30.	35.	40.	45.	50.	55.	60.	65.	70.	75.	Totals.
75										1	1	2
70									1	4	1	6
65						1	1	3½	5	1		11½
60							2	5½	5			12½
55						3	2	2½	1½			9
50					2	19	4	1				26
45			1	2	10	2						15
40			1	2	1							4
35		3	3		2							8
30	6	13	2									21
25	3	5	1									9
Totals	9	21	8	4	15	25	9	12½	12½	6	2	124

There are one hundred and twenty-four months, covering a period of eleven years (1893–1904), eight months not being included because of a gap in the data.* We find—

Means.	Standard deviations.
$h_x = 47{\cdot}16$ copecks per pud;	$\sigma_x = 13{\cdot}93$ copecks per pud.
$h_y = 47{\cdot}04$,,	$\sigma_y = 13{\cdot}84$,,
Correlation coefficient	$r_{xy} = +0{\cdot}93292$

and the equation to the regression line—

$$(5) \quad . \; . \; . \; . \; . \; . \quad Y = 0{\cdot}92689\,X + 3{\cdot}33.$$

Let us investigate now the closeness of fit. The data are exhibited in Table 3.

* "Prices of Commodities on the principal Russian and Foreign Markets in 1893," and the same publication for the following years till 1904. (Published yearly, in Russian, by the Department of Trade and Manufactures, now by the Ministry of Trade and Industry.)

TABLE 3.—*Mean monthly price of rye in Samara.*

Mean price for a month before.	Monthly price.		Errors.	Frequencies.	Standard deviations.	
	Observed.	Calculated.				
x_p.	y_{x_p}.	Y_{x_p}.	$e_p = \|Y_{x_p} - y_{x_p}\|$.	n_{x_p}.	$\sigma_{n_{x_p}}$.	$\frac{n_{x_p} e^2_p}{\sigma^2 n_{x_p}}$.
25	28·33	26·50	1·83	9	2·36	5·4
30	29·52	31·14	1·62	21	3·05	5·9
35	34·37	35·77	1·40	8	5·83	0·5
40	42·50	40·41	2·09	4	2·50	2·8
45	44·00	45·04	1·04	15	4·16	0·9
50	50·80	49·67	1·13	25	3·66	2·4
55	55·00	54·31	0·69	9	5·27	0.2
60	59·60	58·94	0·66	12½	4·45	0·3
65	62·20	63·58	1·38	12½	4·02	1·5
70	70·00	68·21	1·79	6	2·89	2·3
75	72·50	72·85	0·35	2	2·50	0·0
Total	—	—	—	124	—	$\chi^2 = 22·2$

Thus we obtain $\chi^2 = 22·2$, $n' = 11 + 1$, and $P = 0·02$. It may be concluded, therefore, that the fit is not impossibly bad. If we assume the values in the general population distributed in accordance with the regression line (5), deviations due to random sampling equally or more improbable than the observed ones would occur twice in 100 cases.

It must not be forgotten, however, that the formula $\Sigma_{y_{x_p}} = \frac{\sigma_{n_{x_p}}}{\sqrt{n_{x_p}}}$ is only approximate, and that $\sigma_{n_{x_p}}$ involved therein is the standard deviation of y in the p-th array in the general population. It follows that when using the empirical values of $\sigma_{n_{x_p}}$ errors of random sampling are made which in some, if not most, cases, tend to increase the value of the criterion χ^2. These errors may be considerable when the frequencies n_{x_p} are as small as in the Illustration B, and the question arises whether our criterion can be used in such cases.

The general solution of this problem cannot, however, be given here. We may only assume that in cases where the frequencies $n_{x_1}, n_{x_2}, \ldots n_{x_p}$ are great the error in χ^2 cannot be so considerable as to make idle the conclusions to be drawn from it. Further, we may suggest that when the frequencies are small the empirical values of the standard deviations ($\sigma_{n_{x_p}}$) must be graduated, at first in any reasonable manner and the values obtained in such way used in evaluating the formula for χ^2.

Returning now to our Illustration B, we find that the probable errors of the $\sigma_{n_{x_p}}$ must be so considerable that the differences between them cannot be regarded as truly significant. Thus we come to the conclusion that the distribution can be regarded as homoscedastic, the standard deviations $\sigma_{n_{x_p}}$ being probably nearly equal in the general population. The common value may be assumed to be not very different from the mean value—

$$\bar{\sigma}_{n_x} = \frac{1}{\text{N}} \text{S}(n_{x_p} \sigma_{n_{x_p}}) = 3{\cdot}8022,$$

and if we substitute it in the formula for χ^2 we obtain $\chi^2 = 15{\cdot}1$, and $\text{P} = 0{\cdot}18$.

II.

The criterion of goodness of fit given above allows us to resolve the fundamental problem of the theory of fitting the regression lines to the data, *i.e.*, to find the most probable regression curve from the whole family of curves belonging to the given type. The reasoning is quite straightforward.

Given an equation to the line—

$$y = f(x, a_1, a_2, \quad . \quad . \quad . \quad . \quad a_p),$$

the most probable values of the cœfficients $a_1, a_2, \quad . \quad . \quad . \quad . \quad a_p$ will be those which bring our χ^2 to its minimum. Thus we have the condition—

$$(6) \quad . \quad . \quad . \quad . \quad \chi^2 = \text{S}\left\{ \frac{n_x}{\sigma_{n_x}} \Big(y_x - f(x, a_1, a_2 \quad . \quad . \quad . \quad . \quad a_p) \Big)^2 \right\} = \text{min.}$$

From the analytical standpoint the process consists in the application of the method of least squares, the weights to be given to the values being proportional to the frequencies of the arrays divided by the squares of the standard deviations. In the case of homoscedasticity ($\sigma_{n_{x_1}} = \sigma_{n_{x_2}} = \quad . \quad . \quad . \quad . \quad = \sigma_{n_{x_p}}$), the weights will be proportional to the frequencies alone, and the method adopted by Prof. Pearson, in the memoir cited, on skew correlation will give the best results. The only difference in this case between the two methods will consist in the use of moments which enables one to fit the curve, not to the disparate points, but to the continuum. If we realise, however, the fact that in most cases of non-linear regression the arrays are not homoscedastic, we shall come to the conclusion that we must expect to obtain better (in the sense of more probable) results with the method given above than with that of Prof. Pearson.

The type equations resolving the problem, in the case of parabolæ of the p-th order, are easily obtainable.

Let the equation to the parabola be—

$$y = a_0 + a_1 x + a_2 x^2 + \quad . \quad . \quad . \quad . \quad a_p x^p,$$

and let us determine the cœfficients $a_0, a_1, \ldots \ldots a_p$ so as to satisfy the condition—

$$(7) \ldots \quad S\left\{\frac{n_{x_i}}{\sigma^2_{n_{x_i}}}\left(y_{x_i} - a_0 - a_1 x - a_2 x^2 - \ldots - a_p x^p\right)^2\right\} = \text{min.}$$

Now let us write—

$$(8) \ldots \quad w_i = \frac{n_{x_i}}{\sigma^2_{n_{x_i}}}, \quad m_r = S\,(w_i x_i^r), \quad u_r = S\,(w_i y_{x_i} x_i^r).$$

Then we obtain at once the linear system—

$$(9) \ldots \quad \begin{cases} m_0 a_0 + m_1 a_1 + \ldots + m_p a_p = u_0 \\ m_1 a_0 + m_2 a_1 + \ldots + m_{p+1} a_p = u_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ m_p a_0 + m_{p+1} a_1 + \ldots + m_{2p} a_p = u_p \end{cases}$$

The solution of which gives the required values of the coefficients.

Infantile Mortality and the Proportion of the Sexes.

By B. L. Hutchins.

In a previous paper (*Journal*, June, 1909, pp. 210–212) I ventured to suggest that the efforts made by sanitary authorities and others to reduce mortality would have, as a secondary consequence, a reduction in the excess of women, which has been so marked a feature at recent Censuses. That excess arises, in great part at least, from the greater mortality amongst males as compared with females. If the mortality in each sex, up to any given age, be reduced in the same ratio, the relative proportion of females amongst those surviving to that age will be reduced, and the same thing will still be true even if the reduction of female mortality in some degree exceed the reduction of male mortality. The case is most clearly illustrated by a comparison of the Life Tables for England and Wales (E and F) with the Healthy Districts Life Tables (J and K) in the Decennial Supplement of the Registrar-General for 1891–1900.

Age.	England and Wales. Survivors.		Healthy districts. Survivors.	
	Males.	Females.	Males.	Females.
0	509	491	509	491
20	362	364	407	404
40	313	321	365	365
60	208	232	281	292