E. E. Slutsky

Theory of Correlation
and Elements of the Doctrine of the Curves of Distribution
Manual for Studying Some Most Important Methods
# of Contemporary Statistics

Translated by Oscar Sheynin

Berlin, 2009

Е. Е. Слуцкий

Теория корреляции и элементы учения о кривых распределения
Пособие к изучению
некоторых важнейших методов современной статистики

Известия Киевского коммерческого института, кн. 16, 1912, 208с.

E. E. Slutsky

Théorie de la corrélation et Traité abregé de courbes de fréquence
Manuel pour servir à l'étude
de quelques méthodes principales de la statistique moderne

Annales de l'Institut Commercial de Kiew
vol. 16, 1912, 208pp.

Annotation

This is a translation of Slutsky's contribution of 1912 which was intended for Russian readers. He described the Pearson's theory of correlation drawing on the pertinent work of that founder of biometry and on many other British authors. At the time, Markov failed to appraise it properly although Chuprov had at once realized its value (and a few years later compiled a very positive reference for Slutsky), and even in 1948 Kolmogorov called it "important and interesting".

ISBN 3-938417-97-8

NG Verlag, Berlin

Contents

Foreword by Translator

1. Slutsky: life and work
2. The book on the theory of correlation
3. Foreword to Slutsky (1960)

1. Slutsky: Life and Work

1.1. General information. Evgeny Evgenievich Slutsky (1880 – 1948) was an economist, statistician and mathematician, in that chronological order. His life and work are described in Kolmogorov (1948), Smirnov (1948), Chetverikov (1959), Allen (1950), Sheynin (1999), Seneta (2001), with pertinent archival and newspaper sources quoted in Sheynin (1990). Slutsky himself (1938 and 1942, published 1999) compiled his biography. In two other unpublished pieces Wittich (2004; 2007) provides valuable data on Slutsky's life and a pertinent annotated bibliography. In another unpublished paper Rauscher & Wittich (2007) collected information about Slutsky the poet and connoisseur of literature, a side of his personality (as well as his being an artist) that remains unknown. Kolmogorov (1948/2002, p. 72) called Slutsky "a refined and witty conversationalist, a connoisseur of literature, a poet and an artist".

Slutsky's works include his student diploma (1910), the book of 1912 translated below, a paper (1914) which directly bears on a subject discussed in that book, and a most important economic contribution (1915), see also Chipman & Lenfant (2002) and Chipman (2004). His *Selected Works* (1960) contains his biography written by B. V. Gnedenko and an almost complete list of his works. In my § 3 below, I translate its Foreword.

In 1899, Slutsky enrolled in the mathematical department of Kiev university, was drafted into army with others for participating in the students' protest movement; released after nationwide shock; expelled in 1902 for similar activities and banned from entering any other academic institution. In 1902 – 1905 studied mechanical engineering at Munich Polytechnic School; obviously gained further knowledge in mathematics and physics, but remained disinclined to engineering. In 1905 was able to resume learning in Russia, graduated with a gold medal from the Law faculty of Kiev University (end of 1910). His book of 1912 ensured him a position at Kiev Commercial Institute. Became professor at a successor organisation of that institute but had to move to Moscow because of an official demand that teaching ought to be in the Ukrainian language.

Worked as consultant (a very high position) at the Conjuncture Institute and Central Statistical Directorate. Owing to the beginning of the Stalinist regime with horrible situation in statistics (Sheynin 1998), abandoned these occupations, turned to the applications of statistics in geophysics. Did not find suitable conditions for research, became engaged in mathematics. Worked at Moscow State University, received there the degree of Doctor of Physical and Mathematical Sciences *honoris causa* and (Slutsky 1942/2005, p. 145)

> *was entrusted with the chair of theory of probability and mathematical statistics.* […] *However, soon afterwards I convinced myself that that stage of life came to me too late, that I shall not experience the good fortune of having pupils. My transfer to the Steklov Mathematical Institute also created external conditions for my total concentration on research* […]

Until the end of his life Slutsky had been working at that Institute of the Academy of Sciences, became eminent as cofounder of the theory of stationary processes, died of lung cancer. Was happily married, but had no children. From 1912 to Chuprov's death in 1926 maintained most cordial relations with him.

A special remark is due to Allen (1950, pp. 213 – 214):

> *For a very long time before his death Slutsky remained almost inaccessible to economists and statisticians outside Russia.* […] *His assistance, or at least personal contacts with him would have been invaluable.*

Slutsky compiled his book in a very short time; in a letter to Markov of 1912 he (Sheynin 1990/1996, p. 45) explained that he had "experienced a direct impetus from Leontovich's book [1909 – 1911] […] as well as from information reaching me […]". So had he meant 1909 or 1911? He was more specific elsewhere (Slutsky 1942/2005, p. 142): "In 1911, I became interested in mathematical statistics, and, more precisely, in its then new direction headed by Karl Pearson".

Slutsky possibly read some statistics at the Law faculty, but hardly much; he did not mention anything of the sort in his published works. So it seems that in about a year, all by himself, he mastered statistics and reached the level of a respected author!

1.2. A special publication: Slutsky's correspondence with Bortkiewicz, 1923 – 1926 (Wittich et al 2007). I describe some of Slutsky's letters.

Letter No. 3, 25.9.1923. Slutsky made 3000 statistical trials to study whether equally probable combinations occurred independently from the size and form of bean seeds, cf. § 42 of his translated book. He never heard that automatic registering devices were applied in such experiments and even invented something of the sorts "out of boredom".

Letter No. 7, 16.5.1926. Slutsky had to move to Moscow because of "some discord with the Ukrainian language", cf. § 1.1 above, most warmly mentioned the deceased Chuprov. He works as a consultant at the Conjuncture Institute "together with Chetverikov" (Chuprov's closest student and follower) and "had to become" consultant also at Gosplan (State Planning Committee), an extremely important and influential Soviet institution. I venture to suppose that the situation there also became difficult and real scientific work was even considered subversive. Anyway, nothing is known about Slutsky's work there so that he apparently soon quit it.

Letter No. 10, 14.6.1926. Slutsky discussed his paper of 1915 and stated

> *I would have now ended it in an essentially different manner. For uniqueness (to an additive constant) of the definition of the function of utility it is not necessary to demand that on each hypersurface of indifference there exists a pair of such benefits that*
>
> $$\frac{\partial^2 U(x_1, x_2, ..., x_n)}{\partial x_i x_j} = 0.$$
>
> *It is sufficient to be able to draw a line cutting a number of such hypersurfaces along which the marginal utility remains constant, and this is in principle always possible. This result can also be obtained by elementary considerations.*

Then Slutsky refers to his not yet published paper (1927); see also Chipman (2004).

2. The book on the theory of correlation

2.1. Opinions about it. The book was published, as stated on its title-page, in the *Izvestia* (*Annales*) of the Kiev Commercial Institute, and, as mentioned by several authors, appeared independently later the same year. Sections 25, 28 and 43 (these numbers conform to those adopted in the translation) contained "additions to the Pearson theories", see Slutsly's letter to Markov of 1912 (Sheynin 1990/1996, pp. 45 – 46). As mentioned out of place in a footnote to its Introduction, Slutsky reported on his work to the Kiev Society of Economists. Those "Pearson theories" are what the whole book is about, and it is hardly out of order to mention my future paper (2010) on that scientist.

2.1.1. Chuprov. He (Sheynin 1990/1996, p. 44) published a review of Slutsky's book stating that its author "gained a good understanding of the vast English literature" and described it "intelligently". He "most energetically" recommended the book to those having at least "some knowledge of higher mathematics". At the time, Chuprov was not yet critically inclined towards the Biometric school; he changed his attitude later, no doubt having been turned in the mathematical direction by his correspondence with Markov (Ondar 1977).

Apparently in 1916, Chuprov (Sheynin 1990/1996, p. 45) compiled Slutsky's scientific character which contained a phrase: in Slutsky's person "Russian science possesses a serious force", but he obviously did not imagine how correctly he assessed his new friend!

There also (p. 29) I published an archival letter written by N. S. Chetverikov to Chuprov at the end of 1926. He most favourably described the situation at the Conjuncture Institute (where he himself held a high position) and informed his correspondent, already terminally ill, that Kondratiev was inviting him to join their staff. He added, however, that the general situation in the Soviet Union was unclear.

2.1.2. Pearson. He rejected both manuscripts submitted by Slutsky (Sheynin 1990/1996, pp. 46 – 47). In 1913, Slutsky wrote Chuprov about that fact and asked his advice stating that at least in one instance the reason for the rejection "astonished" him. Chuprov did fulfil Slutsky's request and, accordingly, Slutsky successfully published one of his manuscripts (1914). I (Sheynin 2004, pp. 227 – 235, not contained in the original Russian paper) made public three of Slutsky's letters to Pearson of 1912.

2.1.3. Markov. Continental mathematicians and statisticians, and especially Markov utterly disapproved of the Biometric school and I myself have described vivid pertinent episodes (Sheynin 1990/1996, pp. 120 – 122; 2007). In his letters to Chuprov Markov (Ondar 1977/1981, letters 45 and 47, pp. 53 and 58) remarked that Slutsky's book (no doubt partly because of that general attitude) "interested" him, but did not "attract" him, and he did not "like it very much".

More can be added. A few years later, Markov (1916/1951, p. 533, translation p. 212) critically mentioned the correlation theory: it "simply" [?] aims to discover linear [?] dependences, and, when estimating the appropriate probable errors, "enters the region of fantasy […]". This statement was based on an unfortunate application of that theory by a Russian author, but Linnik (Markov 1951, p. 670; translation, p. 215), who commented on Markov's memoir, explained that the conclusions of the correlation theory depended on the knowledge of the appropriate general population. Slutsky, in 1912, did several times mention the general population (also see below) but certainly not on the level of mid-19th century. However, Markov could have well noted Slutsky's conclusion (§ 22) to the effect that the correlation method should not be applied when observations are scarce (which was the case discussed by Markov).

Markov's attitude shows him as a mathematician unwilling to recognize the new approaches to statistics and even to the theory of probability (and denying any optimal properties of the method of least squares), see Sheynin (2006). Markov had time to prepare the last edition of his treatise that appeared posthumously (1924). There, he somewhat softened his views towards the correlation theory and even included Slutsky's book in a short list of references to one of its chapters.

Upon reading Slutsky's book Markov asked Grave, a professor at Kiev university, about the new author. Dmitry Aleksandrovich Grave (1863 – 1939) was active in many branches of mathematics and he also published a treatise on insurance mathematics (in the same volume of the Kiev Commercial Institute *Izvestia* as Slutsky). In a letter toMarkov of 1912 Grave (Sheynin 1999/2004, p. 225) informed his correspondent that neither he himself, nor the lawyers, professors at that Institute, had understood Slutsky's report (see § 2.1 above), that they desired to acquaint themselves with the Pearson theories and asked him to explicate it properly. Grave, however, finds it "repulsive" to read Pearson.

Grave also told Markov about his conversation with an unnamed university professor of political economy who had explained that Slutsky was "quite a talented and serious scientist" not chosen to study as postgraduate "because of his distinct sympathy with social-democratic theories".

2.1.3. Slutsky explained himself in an apparently single extant letter to Markov of 1912 (Sheynin 1990/1996, p. 45 – 46). Improvements of his manuscript "were hindered by various personal circumstances" and he "decided to restrict myself [himself] to a simple concise description" the more so since it will help those Russian statisticians who are unable to read the original literature. He then prophetically stated that "the shortcomings of Pearson's exposition are temporary" and that his theories will be later based on a "rigorous basis" as it happened with mathematics of the 18th and 19th centuries. He added a most interesting phrase: "I consider it possible to develop all the Pearsonian theories by issuing from rigorous abstract assumptions".

Slutsky also mentioned Nekrasov: when his book (1912) had appeared, he began to think that

My [his] *work was superfluous; however, after acquainting myself* [himself] *more closely with Nekrasov's exposition, I* [Slutsky] *became convinced that he* [Nekrasov] *did not even study the relevant literature sufficiently.*

In § 31 (Note 31.1) Slutsky praised the same book; perhaps he did not yet read it "more closely": after ca. 1900, Nekrasov's contributions on the theory of probability and statistics became almost worthless (and utterly disgusted Markov), see Sheynin (2003).

In a letter to Chuprov of the same year Slutsky (Sheynin 1990/1996, p. 44) noted that Grave "actively participates" in the dispute (between Markov and him) and added that Markov "gave me [him] a good dressing-down". […] It was easy for Markov "to discover a number of weak points".

2.1.4. Kolmogorov (1948/2002) published Slutsky's obituary which clearly shows his personal ties with the deceased. He (p. 68) stated that the book of 1912 "became a considerable independent contribution to [mathematical statistics and] remains important and interesting". On the same page Kolmogorov listed "the main weakness[es] of the Biometric school:

Rigorous results on the proximity of empirical sample characteristics to the theoretical ones existed only for independent trials.

Notions of the logical structure of the theory of probability, which underlies all the methods of mathematical statistics, remained at the level of the 18th century results.

The third and last weakness concerned the incompleteness of the published statistical tables.

Kolmogorov indirect advice of applying Slutsky's book at least as a background was not, however, followed; even Slutsky's examples of statistically studying various problems had hardly ever been cited.

2.1.5. Some general remarks about the book. Information provided above, at the end of § 1.1, explains why Slutsky was unable to add a few pages about Pearson, his followers (and Galton!), or to be at least somewhat more critical. He certainly understood that the work of that great scientist was far from rigorous (see § 2.1.3 above), but on this point he only expressed himself about the method of moments (*Additional remarks*). Slutsky also felt that statistics ought to be based on the theory of probability; he said as much, although not quite generally, at the end of his § 32, and stated, in a letter to Markov (§ 2.1.3 above), that that approach was achievable.

On the other hand, the reader will not fail to note that Slutsky also became quite familiar with the practical side of statistics; his book abounds with pertinent remarks! And he also properly provided a lot of original examples of applying correlation theory.

Slutsky (the end of § 2.1.3 above) acknowledged that Markov had "discovered a number of weak points" in his book. For my part, I believe that he had succeeded by and large to provide a good general picture of his subject, but I ought to say the following.

1. He made a mistake in his reasoning on weighing observations, see my Note 28.1, in § 28 which contained his "additions to the Pearson theories", see § 2.1 above. I mentioned another mistake in Note 16.1.

2. His explanations were sometimes inadequate or even lacking, see Notes 3.1, 4.3, 16.2, 40.1 and 41.2.

3. An author ought to show readers not only the trees, but the wood as well, and I especially note that Slutsky had not stated expressly and simply that a zero correlation coefficient does not yet signify independence. His explanation (beginning of both §§ 19 and 29) is not quite sufficient, and in § 31 he only discusses correlation and causality.

4. He offered a faulty example (Note 31.3).

5. He introduced confusing notation (Note 18.5).

Slutsky's system of numbering the sections and formulas was not the best possible. Now, in the translation, sections are numbered consecutively (not separately for each part), and the numbering of the formulas allows to locate them quite easily; thus, formula (3.2) is the second numbered formula in § 3. The Notes (by Slutsky, signed E. S., and my own, signed O. S.) are numbered the same way.

I have omitted some pieces of the original text such as elementary explanations (even concerning the calculation of determinants), mathematical derivations and tables of data which after all can be looked up in the English literature described by Slutsky. Then, I have not included the numerous figures and, accordingly, had to modify their accompanying description.

3. Foreword to Slutsky (1960) by B. V. Gnedenko & N. V. Smirnov

The contents of the scientific heritage of the outstanding Soviet mathematician Evgeny Evgenievich Slutsky (1880 – 1948) are very diverse. In addition to mathematics and mathematical-statistical investigations proper, a number of his works are devoted to problems in mathematical economics, some problems in genetics, demography, physical statistics, etc. It seems unquestionable, however, that Slutsky will enter the history of our national mathematics as one of the founders of the theory of stochastic processes, of that branch of the theory of probability which is the main current channel of research stimulated by ever widening demands made by contemporary physics and technology.

Being absolutely specific both in their final goal and approach, and distinctively combining these qualities with rigour of mathematical treatment, Slutsky's fundamental contributions on the theory of random functions are an excellent introduction to this topical subject.

These *Selected Works* (1960) include all Slutsky's main writings on the theory of random functions and his most important investigations on statistics of connected series. Commentaries adduced at the end of the book trace the numerous links between his work and modern research. A complete [an almost complete] list of his scientific publications is appended. We take the opportunity to express our thanks to Yulia N. Slutsky and N. S. Chetverikov for the materials that they gave us.

Acknowledgements. Magister Guido Rauscher sent me his joint unpublished material (Rauscher & Wittich 2006) and photostat copies of Slutsky (1938; 1942), of the Contents of Slutsky (1910) and of the entire book translated below. From Dr Claus Wittich I received his unpublished contributions (2004; 2007).

Bibliography

Allen, R. G. D. (1950), The work of Eugen Slutsky. *Econometrica*, vol. 18, pp. 209 – 216.

Chetverikov, N. S. (1959, in Russian), The life and work of E. E. Slutsky. In Sheynin, O., translator, *Probability and Statistics. Russian Papers of the Soviet Period*. Berlin, 2005, pp. 146 – 168. Also at www.sheynin.de

Chipman, J. S. (2004), Slutsky's praxeology and his critique of Böhm-Bawerk. *Structural Change and Economic Dynamics*, vol. 15, pp. 345 – 356.

Chipman, J. S. & Lenfant, J.-S. (2002), Slutsky's 1915 article: How it came to be found and interpreted. *Hist. Polit. Economy*, vol. 34, No. 3, pp. 553 – 597.

Eliseeva, I. I., Volkov, A. G. (1999), Life and work of E. E. Slutsky. *Izvestia Sankt-Peterburgsk. Universitet Ekonomiki i Finansov*, No. 1, pp. 113 – 121. In Russian.

Kolmogorov, A. N. (1948, in Russian), Obituary: E. E. Slutsky. *Math. Scientist*, vol. 27, 2002, pp. 67 – 74.

Markov, A. A. (1900), *Ischislenie Veroiatnostei* (Calculus of Probabilities). Fourth edition: Moscow, 1924.

--- (1916), On the coefficient of dispersion. In Markov (1951). Translated in Sheynin, O. (2004), *Probability and Statistics. Russian Papers*. Berlin, pp. 206 – 215. Also at www.sheynin.de

--- (1951), *Izbrannye Trudy* (Sel. Works). Moscow.

Ondar, Kh. O., Editor (1977, in Russian), *Correspondence between Markov and Chuprov on the Theory of Probability and Mathematical Statistics*. New York, 1981. Russian original contained 90 significant mistakes, most of them necessarily retained in the translation (Sheynin 1990/1996, pp. 79 – 83).

Rauscher, G. & Wittich, C., Editors (2006), *E. E. Slutsky's Papers, Fond 21333, Russian State Archive of Literature and Art* [Moscow].

Seneta, E. (1992), On the history of the strong law of large numbers and Boole's inequality. *Hist. Mathematica*, vol. 19, pp. 24 – 39.

--- (2001), E. E. Slutsky. In *Statisticians of the Centuries*. New York, pp. 343 – 346.

Sheynin, O. (1990, in Russian), *Aleksandr A. Chuprov. Life, Work, Correspondence.* Göttingen, 1996.

--- (1998), Statistics in the Soviet epoch. *Jahrbücher f. Nationalökonomie u. Statistik*, Bd. 217, pp. 529 – 549.

--- (1999, in Russian), Slutsky: commemorating the 50th anniversary of his death. In author's *Russian Papers on the History of Probability and Statistics*. Berlin, 2004, pp. 222 – 240. Also at www.sheynin.de

--- (2003), Nekrasov's work on the central limit theorem. *Arch. Hist. Ex. Sci.*, vol. 57, pp. 337 – 353.

--- (2006), Markov's work on the treatment of observations. *Hist. Scientiarum*, vol. 16, pp. 80 – 95.

--- (2007), Integrity is just as important as scientific merits. *Intern. Z. f. Geschichte u. Ethik d. Naturwissenschaften, Technik u. Medizin*, Bd. 15, pp. 289 – 294.

--- (2010), Karl Pearson a century and a half after his birth. *Math. Scientist*, to appear.

Slutsky, E. E. (1910), *Teoria Predelnoi Poleznosti* (Theory of Marginal Utility). Kiev. Diploma thesis. Vernadsky Ukrainian Nat. Library. Ukrainian transl.: Kiev, 2006.

--- (1914), On the criterion of goodness of fit of the regression lines and on the best method of fitting them to the data. *J. Roy. Stat. Soc.*, vol. 77, pp. 78 – 84.

--- (1915, in Italian), On the theory of the budget of the consumer. In *Readings in Price Theory*. G. J. Stigler, K. E. Boulding, editors. Homewood. Ill., 1952, pp. 27 – 56.

--- (1916, in Russian), Statistics and mathematics. Review of Kaufman, A. A. (1916), *Teoria i Metody Statistiki* (Theory and Methods of Statistics). Moscow. Third edition. *Statistichesky Vestnik*, No. 3 – 4, pp. 104 – 120. Translation in *Studies in the History of Statistics and Probability*, pp. 89 – 105. Berlin, 2009. Also at www.sheynin.de

--- (1927, in German), A critique of Böhm-Bawerk's concept of value and his theory of the measurability of value. *Structural Change and Economic Dynamics*, vol. 15, 2004, pp. 357 – 369.

--- (1938, published 1999, in Russian), Autobiography. In Sheynin, O., translator (2005), *Probability and Statistics. Russian Papers of the Soviet Period*. Berlin, pp. 138 – 141. Also at www.sheynin.de

--- (1942, published 1999, in Russian), Autobiography. Ibidem, pp. 142 – 145.

--- (1960), *Izbrannye Trudy* (Sel. Works). Moscow. Economic publications not included. Omission without indication of (mainly foreign) references in footnotes, distortions in translations (comment by Wittich 2007).

Smirnov, N. V. (1948), E. E. Slutsky. *Izvestia Akad. Nauk SSSR*, ser. math., vol. 12, pp. 417 – 420.

Wittich, C. (2004), *E. E. Slutsky. Bio-Data*. Unpublished.

--- (2007), [Annotated] *Selected Sources: E. E. Slutsky, Economics and Vita*. Unpublished.

Wittich, C., Rauscher, G., Sheynin, O. B., Editors (2007, in Russian), Correspondence between E. E. Slutsky and V. I. Bortkevich. *Finansy i Biznes*, No. 4, pp. 139 – 154. Authors ordered in accord with Russian alphabet. We had not seen the proofs and the paper contains a number of misprints; in particular, four of my papers are attributed to Chipman.

0. Introduction

During the two latest decades, theoretical statistics has greatly advanced. Perfection of old methods; discovery and development of new ones; appearance of excellent works on biology and social sciences illustrating methods and proving their unquestionable scientific significance; finally, creation of a yet small personnel of scientists systematically applying and developing the new methods further, – all this, taken together, allows us to say that a new era has originated in statistics.

This movement had started and has been developed in England, and it is only beginning to penetrate other nations. Initiated by the recently deceased celebrated Francis Galton, it grew out of the demands of contemporary biology. Galton, however, was not a mathematician, and the merit of theoretically developing new ideas and establishing a school must almost solely be credited to Karl Pearson whose name will remain in the history of our science alongside those of Laplace, Gauss and Poisson[0.1]. In all fairness, the new school ought to be therefore called after Galton and Pearson.

The general awakening of interest in theoretical statistics allows us to expect that not in a very remote future the ideas of the new school will spread over all nations and all fields of their possible application, and I am humbly aiming at fostering that natural and inevitable process. The application of the new methods is comparatively easy and not difficult to learn. For making use of formulas, it is sufficient to understand their meaning and be able to calculate what they indicate, a task simplified by applying special tables also compiled on Pearson's initiative.

However, it is impossible to manage without breaking from routine. Unforeseen details can be encountered in each problem, and the boundaries of the applicability of a method, and the significance of the results obtained can perplex a student. Not only prescriptions for calculation are therefore needed, it is also necessary to comprehend the spirit of the theories and of their mathematical justification. Life itself thus raises a most important demand before those working at statistics: A statistician must be a mathematician because his science is a mathematical science[0.2].

It is for this reason that I had paid so much attention to formulas and mathematical proofs; nevertheless, one more point also played a certain role. Dry prescriptions are only good enough for being applied in old and firmly established spheres. I believe that no success can be expected in planting new methods in new soil without justifying them.

The sphere of mathematical knowledge needed for understanding most of the provided derivations and proofs is comparatively small. Most elementary information on analytic geometry and differential calculus as can be acquired in a few days is sufficient for understanding the elements of the theory of correlation. Further generalization in that area as well as the first part of my work dealing with curves of distribution demand somewhat wider mathematical knowledge.

I have attempted to satisfy different groups of possible readers and the proofs are therefore simplified as much as a rigorous description allowed it. Those mathematical derivations which I thought understandable to least prepared readers are provided in more detail than necessary for accomplished mathematicians. Finally, I attempted to elucidate the material in such a way that the reader, even after skipping a difficult place, will be able to pick up the lost thread and understand the meaning of the formulas and the manner of applying them. I do not however flatter myself by hoping to have solved that problem quite satisfactorily.

My main subject is the theory of correlation but I did not feel it possible to avoid the theory of the curves of distribution which I described far, however, from comprehensively and possibly even too concisely, in Part 1. I advise readers poorly acquainted with mathematics and only mainly interested in the method of correlation, to go over to Part 2 immediately after acquainting themselves with the first four sections of Part 1.

Part 1

Elements of the Doctrine of Curves of Distribution

1. General notion of curves of distribution or frequency curves

When considering any totality of items possessing a common and measurable indication, we perceive that not all of them have indications of one and the same magnitude. There was a time when statisticians ignored these differences and only concentrated on the arithmetic mean of the indications. Nowadays, it is not anymore necessary to struggle against this dated and self-imposed restriction. It is almost generally understood that the mean is reporting too little about the essence of the whole statistical group and that the aim of statistics comes to describing as completely and simply as possible the whole composition of totalities under consideration.

When solving that problem, the first step is a complete and detailed description of the distribution of the indication among the totality. An elementary, but invariably necessary form of such a description is well known to every statistician. It is a table in which the value of the indication is separated into intervals and the size of each subgroup thus occurring is shown[1.1].

The next step is the representation of the totality by a curve of distribution. From a formal mathematical point of view this is very simple. Mark off the values of the indication along the *x*-axis subdividing it into smallest possible intervals and represent the size of the subgroups by the areas of rectangles having the intervals as their bases. [The author explains in detail the transition to the continuous case and continues]

Thus, the area of the curve of distribution[1.2] shows the number of items having the considered indication contained within certain bounds and the ordinate of the curve gives their number per the unit difference of the magnitude of the indication. The curves of distribution are therefore also called *frequency curves*[1.3].

2. The moments of distribution

The curve of distribution empirically constructed by issuing from the data only provides us with what was contained there, but shows it more clearly. This, however, is not sufficient for a deeper study; we need numerical characteristics of the various properties of the totality. One of them we already know, it is the mean value of the indication. To find it, we ought to multiply the magnitude of the indication by the number of the appropriate items [in each interval, find the sum of these products], and to divide the product by the total number of items.

However, when that total number is very large, this method becomes inapplicable; first, because we do not perhaps know the exact value of the indication possessed by each of them; and second, even if we know it, the work becomes excessive, so that we need a method of approximately calculating of the arithmetic mean. [Denoting the total number of items by *N*, their number in interval *i* by n_i, the mean points of the intervals by x_i and the arithmetic mean by $\overline{x}$, he gets

$$N\overline{x} = \sum n_x x \qquad (2.1)$$

where the subscript *i* is dropped, and, in the continuous case,

$$N\overline{x} = \int yxdx].$$

where *y* (*x*) is the [frequency] curve […]

Statistical practice proved that, when the total number of items is not too small, formula (2.1) provides sufficiently precise results even for a small number (10 – 12) of groups.

We can also derive other numbers characterizing a studied totality by calculating in the same way the mean square, the cube, the fourth power etc of the indicator. Pearson calls the expressions thus found the *moments* (zero moment, mean zero power; first, second, … moment; arithmetic mean, mean square, cube, …), $k = 1, 2, 3, 4$:

$$\mu_0' = \frac{1}{N}\int yx^0 dx = 1,\ \nu_0' = \frac{1}{N}\sum n_x x^0 = 1,\ \mu_k' = \frac{1}{N}\int yx^k dx,\ \nu_k' = \frac{1}{N}\sum n_x x^k.$$

For the sake of convenience I will also denote the arithmetic mean by h, or h_x, h_y, … showing the appropriate variable.

Moments can be derived relative to any origin, but of most importance are those with origin at the mean (at the *centre* of the distribution). They may be called *central* moments and it is usual to denote them by the same letter but without the apostrophe.

Direct calculation of the central moments is inconvenient since it involves squaring, raising to the third and fourth power multidigit numbers ($\bar{x} - x$); only by chance will the arithmetic mean be an integer. It is therefore simpler to calculate the moments relative to any arbitrary origin, then to go over to the central moments. The transition is very simple. By definition of the p-th central moment

$$N\nu_p = \sum n_x (x - \bar{x})^p.$$

According to the formula of the Newton binomial

$$N\nu_p = \sum \{n_x[x^p - px^{p-1}\bar{x} + \frac{p(p-1)}{1\cdot 2}x^{p-2}\bar{x}^2 - \ldots \pm px\bar{x}^{p-1} \mp \bar{x}^p]\}$$

with the sign being plus or minus if p is even or odd respectively.

Summing the separate terms, and denoting for the sake of symmetry the mean by ν_1' we have

$$N\nu_p = \sum n_x x^p - p\bar{x}\sum n_x x^{p-1} + \ldots \pm p\bar{x}^{p-1}\sum n_x x \mp \bar{x}^p \sum n_x =$$

$$N[\nu_p' - p\nu_{p-1}'\nu_1' + \frac{p(p-1)}{1\cdot 2}\nu_{p-2}'(\nu_1')^2 - \ldots \pm p(\nu_1')^p \mp (\nu_1')^p].$$

Combining the last two terms we finally get

$$\nu_p = \nu_p' - p\nu_{p-1}'\nu_1' + \frac{p(p-1)}{1\cdot 2}\nu_{p-2}'(\nu_1')^2 - \ldots + (-1)^{p-1}(p-1)(\nu_1')^p. \tag{2.2}$$

In particular,

$$\nu_1 = 0,\ \nu_2 = \nu_2' - (\nu_2')^2,\ \nu_3 = \nu_3' - 3\nu_2'\nu_1' + 2(\nu_1')^3,$$
$$\nu_4 = \nu_4' - 4\nu_3'\nu_1' + 6\nu_2'(\nu_1')^2 - 3(\nu_1')^4. \tag{2.3}$$

When calculating, it is usual to choose as a conditional zero some magnitude of the indication corresponding to the midpoint of the interval nearest to the centre of the distribution, and to assume the length of the interval (call it k) as the unit. After calculation, the p-th moment should be multiplied by k^p.

3. The mean deviation and the coefficient of variation

Especially important in theoretical statistics and its applications is the square root of the second central moment, σ, called mean square error in the theory of observational errors, and otherwise *standard deviation*:

$$\sigma^2 = \nu_2 = \frac{1}{N}\sum n_x(x-\overline{x})^2. \qquad (3.1)$$

This formula is not altogether precise because ν_2 is an approximate magnitude which in most cases we can correct beforehand (§ 7). When the distribution follows the well-known Gaussian law, about 2/3 of the items of a totality deviates from the arithmetic mean in either direction not more than by a standard deviation. Therefore, σ can serve as a measure of variability determining how wide are the boundaries between which most items of a totality are situated.

The so-called *coefficient of variation* is

$$V = \frac{\sigma}{h}. \qquad (3.2)$$

For practical applications it is convenient to express it in percentage terms.

Example [variation in the mean monthly price of rye during 124 months in 1893 – 1903; the results are]

1. Moscow: $h_1 = 59.40 \pm 0.77$ copecks, $\sigma_1 = 12.64 \pm 0.54$ cop.
2. Elets: $h_2 = 52.64 \pm 0.65$ cop., $\sigma_2 = 10.74 \pm 0.46$ cop.
3. Samara: $h_3 = 47.04 \pm 0.84$ cop., $\sigma_3 = 13.84 \pm 0.59$ cop.

The three centres of commerce differ here not only in the mean price, but also in the fluctuation of these […]. Turning now to relative fluctuations, which characterize the stability of prices and are expressed by the coefficients of variation, we find

$V_1 = 21.28\% \pm 0.95$, $V_2 = 20.40\% \pm 0.91$, $V_3 = 29.42\% \pm 1.36$[3.1].

4. Probable errors

Along with means, standard deviations and coefficients of variation there are their probable errors calculated by formulas[4.1]

$$E_h = 0.67449\frac{\sigma}{\sqrt{N}},\ E_\sigma = 0.67449\frac{\sigma}{\sqrt{2N}},$$
$$E_V = 0.67449\frac{V}{\sqrt{2N}}\sqrt{1+2\,[\frac{V}{100}]^2}. \qquad (4.1, 2, 3)$$

The first two formulas were known long ago; for the last one, see Lee & Pearson (1897, p. 345). Tables (Gibson 1906; Pearl & Blakeman) essentially simplify calculations[4.2].

The boundaries of my work do not allow me to derive them. The most necessary material is included in one of the last sections of this book; here, I only provide a few remarks. The set of causes consisting of very many elementary, equally probable positive and negative influences may be called the complex of random causes. The more is the phenomenon repeated, the more completely do the random causes compensate each other and mutually destroy their influence on it[4.3]. If the Gaussian law [already mentioned in § 3] (see below) is realized, even if approximately, about a half of the random deviations will be smaller than the *probable error*, and about one half, larger. Deviations many times exceeding it (practically speaking, 6, 5 and only 4 times larger) are extremely unlikely.

The probable error is therefore the test separating the set of random influences from the complex of the influences of the main causes which determines the essence of the phenomenon. There exists no other test here. Suppose that, having a *very* large number of observations, we calculated the mean magnitude of a phenomenon and then derived its mean making use of a restricted part of the observations. Then the probable error of the second mean should serve as a test of whether the whole and the part differ from each other. For example, having determined the mean stature of a million Russians and of 200 men from the city of Yaroslavl, we could have stated that the statures differ if the *calculated* difference is several times greater than the probable error of the second, much less precise determination. A difference somewhat less or greater than the probable error may be explained by the influence of random causes.

Suppose now that we have two equally numerous groups, for example mean monthly prices in Moscow in 1891 – 1900 and 1901 – 1910. Then we may attribute any difference between them not exceeding, or negligibly greater than its probable error, by the influence of random causes. That error is equal to

$$\sqrt{\mathrm{E}_1^2 + \mathrm{E}_2^2},$$

i.e., to the square root of the sum of squares of both probable errors. We could have only stated that the level of prices had actually changed if the mean for the second period would have exceeded the first mean at least more than by five or six times the probable error of their difference.

When applying the same test to decide whether the difference between the mean prices of rye or their standard deviations in Moscow and Elets (§ 3) was essential, we will err because the formula above is only valid for *mutually independent phenomena* whereas the prices at two centres comparatively near to each other cannot be such. We can only admit that conclusion hypothetically with some subjectively estimated certainty if the difference *extremely* exceeded the probable errors of each magnitude taken separately. Concerning the same example of § 3, we may thus only decide about the difference between the coefficients of variation for Samara on the one hand and Moscow and Elets on the other hand. A rigorous test can only be derived by means of the correlation theory.

5. The Gaussian law and its generalization by Pearson

In some cases the number of items *N* in a group, the mean $\bar{x}$ and the standard deviation σ are quite sufficient for an exhaustive description of a totality. If the items obey the Gaussian law of distribution, these three magnitudes allow to determine in a purely theoretical way the size of any subgroup for which it is only necessary to apply any table of the integral of [the appropriate] probabilities[5.1].

The Gaussian law, however, is not sufficiently general. Indeed, however you derive it, the following assumptions are invariably admitted, explicitly or tacitly.

a) Deviations from the mean in both directions are equally probable.

b) An addition of a new deviation, either positive or negative, is equally probable independently of the sum of the already accumulated deviations (Pearson 1905a, p. 189).

Here is a possible elementary derivation resting on those assumptions (Ibidem, p. 179 note)[5.2]. Let there be *n* elementary causes, each leading to a deviation equal to ξ; suppose also that, taken separately, such a cause occasioned *r* positive and $(n - r)$ negative deviations. The total deviation will be

$$x_r = r\xi - (n - r)\xi = (2r - n)\xi, \qquad (5.1a)$$

and, similarly,

$$x_{r+1} = (r + 1)\xi - (n - r - 1)\xi = (2r + 2 - n)\xi. \qquad (5.1b)$$

Their difference is

$$\Delta x_r = x_{r+1} - x_r = 2\xi.$$

What will be the probability of x_r and x_{r+1}? The theory of probability tells us that the probability for equally likely events occurring *r* and $(n - r)$ times is equal to the $(r + 1)$-st term of the binomial $[(1/2) + (1/2)]^n$, that is, equal to $(1/2)^n C_n^r$. In the limit, frequencies are proportional to probabilities, so that out of *N* cases deviations of x_r and x_{r+1} will occur

$$y_r = N(1/2)^n C_n^r,\ y_{r+1} = N(1/2)^n C_n^{r+1}$$

times so that the second deviation will occur more often by

$$\Delta y_r = N(1/2)^n \frac{n!}{r!(n-r-1)!} \cdot \frac{n-2r-1}{(r+1)(n-r)}.$$

Then, the ordinate of the middle of the interval between y_r and y_{r+1} will be their half-sum

$$\Delta y_{r+1/2} = N(1/2)^n \frac{n!}{r!(n-r-1)!} \cdot \frac{1/2(n+1)}{(r+1)(n-r)}$$

and

$$\frac{\Delta y_r}{y_{r+1/2}} = \frac{(y_{r+1} - y_r)}{1/2(y_{r+1} + y_r)} = \frac{n-2r-1}{1/2(n+1)}.$$

The difference Δy_r corresponds to $\Delta x_r = 2\xi$ and

$$\frac{\Delta y}{y_{r+1/2}\Delta x} = \frac{n-2r-1}{1/2(n+1)2\xi}.$$

This expression can be presented in another form. The abscissa corresponding to ordinate $y_{r+1/2}$ is, see (5.1),

$$x_{r+1/2} = (1/2)(x_{r+1} + x_r) = (2r - n + 1)\xi,\ n - 2r - 1 = -(x/\xi)$$

and

$$\frac{\Delta y}{y\Delta x} = -\frac{x}{(n+1)\xi^2}. \qquad (5.2)$$

Consider now the limiting case. For any n, $x_r = 0$ at $r = n/2$. Here, exactly one half of the elementary deviations are positive and the other half, negative. If $r \neq n/2$, then, at $n = \infty$, see equations (5.1), $-\infty < x < +\infty$. For the limit

$$\lim(n + 1)\xi^2 = \infty \qquad (5.3)$$

we would have got for finite values of x

$$\Delta y/\Delta x = 0,\ y = \text{Const} = C$$

and the curve of distribution would have become a straight line parallel to the x axis. Since the total number of items is equal to the area of that curve which is now $C\infty$ and for this to be equal to N it is necessary that $C = 0$.

And so, if equality (5.3) takes place, there will be no items with the stated indication on any finite interval. And, if they actually exist, than, for $n = \infty$ and $\xi = 0$,

$$\lim\ (n + 1)\xi^2 = \text{a finite number} = a^2,$$

the differential equation of the curve of distribution will be

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x}{a^2},$$

$$y = y_0 \exp(-\frac{x^2}{2a^2}).$$

This is indeed the equation of the Gaussian curve, or, as Pearson named it, of the *normal* curve of distribution.

Abandoning the assumption *b* (above) that an elementary deviation is independent from the sum of the already accumulated deviations, that is, of x_r, and, on the contrary, supposing, which is Pearson's idea, that $\xi = f(x_r)$, we will get the most possible general dependence between the frequencies and the magnitude of the indication.

The limit (5.3) will not be equal to some constant a^2 anymore, but become a function of x, and the differential equation as derived from equation (5.2) will now be

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x}{F(x)},$$

or, for the origin chosen at an arbitrary point,

$$\frac{1}{y}\frac{dy}{dx}=-\frac{x-a}{F(x)}. \qquad (5.4)$$

In abandoning assumption *b*, we have thus freed ourselves from the supposition that positive and negative deviations are equally probable and derived the most possible general form of dependence.

If *F*(*x*) may be expanded in a MacLaurin series, we will have

$$\frac{1}{y}\frac{dy}{dx}=\frac{x-a}{b_0+b_1x+b_2x^2+...}. \qquad (5.5)$$

Any number of terms can be taken; in practice, we have to restrict the expansion by three terms. Indeed, for determining the equation of a curve fitting the statistical material, we ought, in accord with the Pearson method (see below), calculate the actual moments and equate them to their theoretical counterparts. In case of the equation (5.4) with three terms in the denominator of the right side, we have to know four moments of the actual distribution; otherwise, moments of higher orders are needed. However, Pearson (1905b, pp. 7 – 8) showed that the probable errors of the moments of those orders were very large and increased rapidly with the orders, so that the coefficients of a curve calculated by means of the moments of higher orders must also be unreliable.

In spite of that restriction, the experience of Pearson and his school showed that the Pearsonian curves [defined by equation (5.4)] almost always provided excellent results and described the special features of the data in cases in which the normal (the Gaussian) curve refused to serve statisticians.

6. Justification of the method of moments

How can the statistician apply a theoretical curve for showing his material? It should be shaped into its final form by applying the statistical data for calculating its coefficients[6.1]. Chronologically, the first solution was achieved by the method of least squares. Its idea consists in the following. Suppose that observations provided a number of points and that we wish to determine the coefficients of the equation in such a manner that the curve thus obtained as close as possible adjoins our points. […]

Its shortcoming is the need for very much work even for parabolic curves. In many other cases the method is either not applicable at all, or demands quite excessive toil. And in many instances the estimation of the probable errors of the calculated coefficients is also either impossible or very difficult.

Pearson proposed a modification of the method of least squares. Suppose that we have a continuous empirical curve rather than isolated points; such points should be joined by a parabolic curve as smoothly as possible and the coefficients sought determined by applying the condition of least squares to *all* the points of the empirical curve. Analytically this is expressed by replacing finite sums by integrals, and, as it can be shown, by equating the moments as specified in § 5.

Pearson (1902c, pp. 267 – 271) theoretically justified this modification in the following way; readers unfamiliar with higher mathematics may skip his considerations without any negative consequences. Suppose that a series of measurements or observations of a variable *y* are made corresponding to a series of values of another variable, *x*, $-l<x<l$. It is required to discover a good method of deriving a theoretical or empirical curve

$$y=\varphi(x;c_1;c_2;\ldots;c_n) \qquad (6.1)$$

fitting the data where $c_1, c_2, \ldots, c_n$ are arbitrary constants.

Let us assume that $\varphi(x)$ can be expanded in a MacLaurin series which moreover converges more or less rapidly:

$$y = \varphi(0) + x\varphi'(0) + (x^2/2)\varphi''(0) + \ldots = \alpha_0 + \alpha_1 x + \alpha_2(x^2/2) + \ldots$$

Here, $\alpha_0, \alpha_1, \alpha_2, \ldots$ are functions of the n parameters $c_1, c_2, \ldots, c_n$. It is therefore theoretically possible to determine all these parameters given the first n coefficients $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ and then to apply the calculated magnitudes for deriving all the rest coefficients $\alpha_n, \alpha_{n+1}, \ldots$ [only those necessary to be considered]. It follows that theoretically we can represent our curve as

$$y = \alpha_0 + \alpha_1 x + \alpha_2(x^2/2) + \alpha_3(x^2/6) + \ldots + \alpha_{n-1}[x^{n-1}/(n-1)!] +$$
$$\varphi^{(n)}(\alpha_0; \alpha_1; \ldots; \alpha_{n-1})(x^n/n!) + \ldots$$

Let Y be the ordinate of the empirical curve, then $(y - Y)$ will be the distance between the two curves at point x, and, in accord with the principles [with the principle] of least squares, we will set

$$\int (y-Y)^2 dx = \min. \qquad (6.2)$$

[Here is the essence of Slutsky's description. Let A and A' be the areas of the curves fitting the data and the empirical curve, and μ and μ', the respective moments. Then

$$A = A' - \int (y-Y)\frac{dR}{d\alpha_i}dx, \qquad (6.3.1)$$

$$A\mu_i = A'\mu_i' - \int (y-Y)\frac{dR}{d\alpha_i}dx,\ i = 1, 2, \ldots, n-1, \qquad (6.3.2)$$

$$R = \frac{x^n}{n!}\varphi^{(n)}(\theta x),\ 0 < \theta < 1$$

and, approximately,

$$A = A',\ \mu = \mu'. \qquad (6.4.1, 2)$$

Slutsky comments on the method of moments in his *Additional remarks*. Here, he continues:]

And so, we obtain the following rule. *For fitting a good theoretical curve (6.1) to an empirical curve it is necessary to equate its area and moments expressed by its parameters $c_1, c_2, \ldots, c_n$ to the area and moments of the empirical curve.*

The solution above provides a better approximation than can be possibly thought on the face of it because the corresponding higher moments will also be approximately equal and the nearer the larger is n. [The proof follows.]

We conclude from all the above that the equality of the moments is a good condition for fitting curves to data, and practice has shown that it is indeed not worse than the principle of least squares. In case of parabolic curves, the two methods coincide because the MacLaurin series are then finite. And, as mentioned above, the method of moments may be applied even when the first method is either not applicable at all, or demands quite excessive toil. In addition, the second method, whenever applicable, allows to estimate the probable errors of the calculated coefficients. For employing the method of moments a statistician must

1. Find the moments of any empirical system of observations.

2. Express the moments of the theoretical curve as a function of the parameters c_1, c_2, …, c_n in such a manner that equating the areas and the moments […] will not be excessively difficult.

I turn to the solution of the first problem.

7. Determining the empirical moments

Pearson calls the moments determined in accord with § 2 *raw* and denotes them by ν_1, ν_2, … if central, or by ν'_1, ν'_2, … otherwise. The inaccuracy involved consists in that we calculate them as though all items in an interval are situated in its middle. This method only leads to the true moments μ_1, μ_2, … for infinitely small intervals. Actually, the intervals are seldom small enough for the error to be negligible. We will distinguish three cases.

A. The empirical curve smoothly falls down on both sides to the *x* axis. The number of items in the extreme groups diminishes so gradually that, in mathematical language, the curve on *both* tails has *contact of an infinitely large order* with the *x* axis. Pearson calls such curves quasi-normal (because, in particular, the normal curve also has this property). The true moments are then easily calculated from the raw moments by applying the Sheppard corrections. Let us derive them (Pearson 1903b). A non-mathematician may skip the derivation and only turn attention to the result. [I am only providing it:]

$$\mu_0 = \nu_0 = 1,\ \mu_1 = \nu_1 = 0,\ \mu_2 = \nu_2 - 1/12,$$
$$\mu_3 = \nu_3,\ \mu_4 = \nu_4 - (1/2)\nu_2 + 7/240. \qquad (7.1)$$

[…]

The calculations are made for a unit interval, and the *p*-th moment is then multiplied by k^p where *k* is the actual length of the interval.

B. If the data indicates that the former case does not take place, and especially if the curve makes a finite angle when cutting the *x* axis, the Sheppard corrections must not be applied and another method is recommended (Pearson 1902c, pp. 282ff). For a non-mathematician this method will perhaps be difficult. In general, it only ought to be applied when the investigation demands high precision. In other cases, if the Sheppard corrections are not applicable, it is possible to employ the correction described below under **C**.

Let $y = \varphi(x)$ be, as previously, the curve showing the distribution. Mark off the intervals $[x_0; x_1]$, $[x_1; x_2]$ etc on the *x* axis. Here, *x* with a subscript denotes not the distances [from the origin] to the middle of intervals, but to their ends. Let also n_r be the number of the items in the interval $[x_{r-1}; x_r]$. The quantities n_1, n_2, … n_p are given by the data:

$$n_1 = \int_{x_0}^{x_1} ydx,\ n_2 = \int_{x_1}^{x_2} ydx,\ ...,\ n_p = \int_{x_{p-1}}^{x_p} ydx.$$

Denote also

$N = n_1 + n_2 + \ldots + n_p$.

For the *n*-th moment relative to the *y* axis we have

$$N\mu'_n = \int_{x_0}^{x_p} x^n y dx. \qquad (7.2)$$

Introduce a new variable

$$Z = \int_{x}^{x_p} y dx \qquad (7.3)$$

which is obviously a part of the area of our curve, i. e., the number of items having the value of the indication between some *x* and x_p.

Then

$$Z_0 = \int_{x_0}^{x_p} y dx = N,\ Z_1 = \int_{x_1}^{x_p} y dx = n_2 + n_3 + \ldots + n_p,\ \ldots,\ Z_p = \int_{x_p}^{x_p} y dx = 0. \qquad (7.4)$$

Differentiating the integral (7.3) [with respect to its lower limit] we get $dZ/dx = - y$. Formula (7.2), when substituting the derivative instead of $- y$, becomes

$$N\mu'_n = -\int_{x_0}^{x_p} x^n dZ.$$

Integrating by parts leads to

$$N\mu'_n = Z_0 x_0^n + n\int_{x_0}^{x_p} Zx^{n-1}dx,\ \mu'_n = x_0^n + \frac{n}{N}\int_{x_0}^{x_p} Zx^{n-1}dx.$$

We may measure the value of the indication by the difference between a given magnitude and some constant assumed as the origin which can coincide with the origin of the distribution. Then $x_0 = 0$ and

$$\mu'_n = \frac{n}{N}\int_{x_0}^{x_p} Zx^{n-1}dx. \qquad (7.5)$$

This is the main formula for deriving the true moments. The order of calculation is obvious. For

$x = x_0\ (= 0),\ x_1,\ x_2,\ \ldots,\ x_p$

calculate Z_0, Z_1, Z_2, … , see formulas (7.4). Introduce new magnitudes Y_i, ordinates of a new supplementary curve:

$$Y_0 = Z_0 x_0^{n-1} = 0,\ Y_k = Z_k x_k^{n-1}, k = 1, 2, ..., p\ (Y_p = 0).$$

Determine now the area *S* of the new curve. It is equal to the integral in formula (7.5) and therefore

$$\mu_n' = \frac{n}{N} S.$$

S can be calculated by means of any suitable formula of approximate integration. […] Pearson (1902c, p. 275) recommends the following very precise formula (Sheppard 1900): […]

C. The third case considers experimental numbers situated so disorderly that formulas representing curves could be thought superfluous. Indeed, all these bends of parabolic curves […] have no real meaning, they are occasioned not by the nature of the phenomenon, but by random irregularities in the experimental material. Here, it is more proper to consider the empirical curve just as it is derived, i. e., as a broken line, and the problem is reduced to calculating the moments of the area consisting of trapezoids. [Slutsky does not formally define this new concept but replaces *N*, the total number of items in a totality, by the area *under* the curve or broken line (retaining the previous notation).] Method **B** is applicable here also, but Pearson provided the final formulas allowing to go over at once from raw to true central moments.

In addition, since this method is much easier to apply, we may do so in all cases in which the Sheppard corrections cannot be used and the somewhat higher precision ensured by the method **B** seems unnecessary. [The author describes the derivation of Pearson's formulas (1896a, pp. 348 – 350) and gets]

$$\mu_n' = \nu_n' + \frac{n(n-1)}{12}\nu_{n-2}' + \frac{n(n-1)(n-2)(n-3)}{360}\nu_{n-4}' + \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{20160}\nu_{n-6}' + ..., \tag{7.6}$$

$$\mu_1' = \nu_1',\ \mu_2' = \nu_2' + 1/6,\ \mu_3' = \nu_3' + (1/2)\nu_1',\ \mu_4' = \nu_4' + \nu_2' + 1/15, \tag{7.7}$$

$$\mu_1 = 0, \mu_2 = \nu_2 + 1/6, \mu_3 = \nu_3, \mu_4 = \nu_4 + \nu_2 + 1/15. \tag{7.8}$$

The trapezoid method should be applied in this form. The interval is supposed to be unity and the raw moments ν'_1, ν'_2, … are calculated after which in accord with formulas (2.1) the raw central moments are determined, and, finally, either the Sheppard corrections are applied (in case **A**) or the transition to the true moments is accomplished by formulas (7.8). If returning to the initial units is desired, the *n*-th moment is multiplied by k^n.

8. Deriving parabolic curves fitting experimental data

Once the moments of the empirical curve are calculated, the problem is reduced to determining the coefficients of the theoretical curve having those moments. What kind of curve is chosen depends, as stated above, on general considerations for which no rule is possible. The best results for frequency curves are provided by the Pearsonian curves, but for many other aims [?] parabolic curves are often successfully applied. My goal here is to explain how to determine their coefficients by the method of moments (Pearson 1902c, pp. 12 – 16) so that we will have a complete example of its application.

Consider an empirical broken line on the base of the curve, i. e., on interval [– l; l] with ordinates y_1 and y_2 at its ends and area N *under* the line and denote the empirical moments relative to the ordinate passing through the origin by μ'_1, μ'_2, … We ought to determine the moments of the area and equate them to the empirical moments which will ensure the calculation of the coefficients of the curve

$$y = (N/2l)\ [e_0 + e_1(x/l) + e_2(x/l)^2 + \ldots + e_{n-1}(x/l)^{n-1}].$$

Multiplying both sides by $(x/l)^{2r}$ and integrating, we get

$$\frac{1}{l^{2r}}\int_{-l}^{l} yx^{2r}dx = N\frac{\mu'_{2r}}{l^{2r}} = N\ [\frac{e_0}{2r+1}+\frac{e_2}{2r+3}+\ldots+\frac{1-(-1)^n}{2}\frac{e_{n-1}}{2r+n}].$$

The terms including x in odd powers vanish. And, if both sides were multiplied by $(x/l)^{2r+1}$,

$$\frac{1}{l^{2r+1}}\int_{-l}^{l} yx^{2r+1}dx = N\frac{\mu'_{2r+1}}{l^{2r+1}} =$$

$$N\ [\frac{e_1}{2r+3}+\frac{e_3}{2r+5}+\ldots+\frac{1+(-1)^n}{2}\frac{e_{n-1}}{2r+n+1}].$$

Denote $\lambda_s = (\mu'_s/l_s)$, so that $\lambda_0 = 1$, then

$$e_0 + \quad (1/3)e_2 + (1/5)e_4 + \ldots = \lambda_0 = 1,$$
$$(1/3)e_0 + (1/5)e_2 + (1/7)e_4 + \ldots = \lambda_2,$$
$$(1/5)e_0 + (1/7)e_2 + (1/9)e_4 + \ldots = \lambda_4, \ldots,$$

$$(1/3)e_1 + (1/5)e_3 + (1/7)e_5 + \ldots = \lambda_1,$$
$$(1/5)e_1 + (1/7)e_3 + (1/9)e_5 + \ldots = \lambda_3,$$
$$(1/7)e_1 + (1/9)e_3 + (1/11)e_5 + \ldots = \lambda_5, \ldots$$

[Slutsky then derives working formulas for the theoretical curve being of the zero order (a straight line) and of the first, the second, …, the sixth order. For example, here are the coefficients of the parabola of the sixth order]

$$e_0 = \frac{35}{256}(35-315\lambda_2+693\lambda_4-429\lambda_6),\ e_2 = \frac{315}{256}(-35+567\lambda_2-1485\lambda_4+1001\lambda_6),$$

$$e_4 = \frac{3465}{256}(7-135\lambda_2+385\ \lambda_4-273\lambda_6),\ e_6 = \frac{3003}{256}(-5+105\lambda_2-315\lambda_4+231\lambda_6),$$

$$e_1 = \frac{105}{64}(35\lambda_1-126\lambda_3+99\lambda_5), \qquad e_3 = \frac{315}{32}(-21\lambda_1+90\lambda_3-77\lambda_5),$$

$$e_5 = \frac{693}{64}(15\lambda_1-70\lambda_3+63\lambda_5).$$

Pearson had thus solved this problem once and for all. In any practical applications, we may employ his formulas.

9. The normal frequency curve (the Gaussian curve).

Deviations from the normal type

Let us now better acquaint ourselves with the properties of the normal curve. When deriving its equation in § 4, I selected the centre of the distribution as the origin of the system of coordinates and measured the indication by its deviation from its arithmetic mean. When assigning any point as the origin, we will have now

$$y = y_0 \exp[-\frac{1}{2}\frac{(x-h)^2}{a^2}].$$

We will now find the dependence between the coefficients of this equation and the essential magnitudes of the distribution, of the number of items in the totality N, the arithmetic mean $\bar{x}$ and the standard deviation σ. We see first of all that y becomes ever smaller as $(x – h)$ increases; its maximal value corresponds to $x = h$. Then, $(x – h)^2$ is always positive , and the value of y does not change when x is more, or less than h by the same magnitude. This means that the curve is symmetric relative to its maximal ordinate; equal deviations from indication h occur equally often so that that parameter is the arithmetic mean:

$$\bar{x} = h.$$

When considering the frequency curve in general, the *centre* of the distribution does not always coincide with the point of the x axis corresponding to the maximal ordinate. In accord with Pearson’s proposal (1896a, p. 345 note) we will call this point the *mode*; by the same term he calls its abscissa. I think that according to the spirit of the Russian language it is more natural to call that abscissa the modal magnitude, and therefore be able to consider modal increase, modal wages etc.

We will call the interval *between the centre and the mode radius of asymmetry* (d), positive if the centre is to the right from the mode, and negative otherwise. The ratio of that radius to the standard deviation σ is called *skewness* and I denote it by α:

$$\alpha = \frac{d}{\sigma}.$$

In addition, the *median* is the value of the indication which divides the entire totality in two equal parts. I will call the corresponding point of the x axis the *middle* of the distribution. It is situated between the centre and the mode and Pearson (1896, pp. 375 – 376), also Pearson & Lee (1897, pp. 441 – 442) showed that in most cases there exists an approximate equality: the interval from the middle to the centre is one half of it to the mode. This property of the mode enables to determine it with a precision usually sufficient for practical applications.

For the normal curve all the three points (the mode, the arithmetic mean and the median) coincide. This is one reason why the statistician cannot be satisfied only by that curve but ought to master the Pearsonian asymmetric curves.

Let us go further. Again assigning the centre of the distribution as the origin, we will have the equation of the normal curve as

$$y = y_0 \exp(-\frac{x^2}{2a^2}).$$

We will now derive the standard deviation. By its definition, the second moment is

$$N\mu_2 = y_0 \int_{-\infty}^{\infty} x^2 \exp(-\frac{x^2}{2a^2})dx.$$

Integrating by parts according to the formula [… if providing that formula at all, Slutsky should have done it in § 7] we find […]

$$N\mu_2 = a^3\sqrt{2\pi}y_0.$$

The area of the curve can be easily determined because

$$N = \sqrt{2\pi}ay_0. \qquad (9.1)$$

Dividing the former equation (9.1) by the latter we get

$$\mu^2 = \sigma^2 = a^2, \sigma = a$$

and from (9.1) it follows that

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}}$$

so that finally, with the origin in the centre of the distribution or situated arbitrarily,

$$y = \frac{N}{\sigma\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2}),\ y = \frac{N}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-h)^2}{2\sigma^2}]. \qquad (9.2a, b)$$

Thus, issuing from N, the arithmetic mean h and the standard deviation σ, we can derive the equation of the normal curve corresponding to the data.

By applying the same method of integration to the equation (9.2a), we will determine that

$$\mu_3 = \mu_5 = \mu_7 = \ldots = 0.$$

For the Gaussian curve, all the odd moments vanish which indeed follows from its symmetric form. As to the even moments, there exist dependences between them. Restricting our considerations to the fourth moment, we will have

$$\mu_4 = 3\mu_2^2.$$

Pearson introduced notation

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3},\ \beta_2 = \frac{\mu_4}{\mu_2^2} \qquad (9.3a, b)$$

needed for his theory of asymmetric curves. The magnitudes β_1 and β_2 are derived from the moments calculated by issuing from the data. Evidently, no empirical data can be thought normal, if, allowing for probable errors, the following conditions are not satisfied

$$\mu_3 = 0,\ \mu_4 = 3\mu_2^2.$$

They can be written in another form. When studying the generalized equation of the distribution curve (5.4)

$$\frac{1}{y}\frac{dy}{dx} = \frac{x-a}{b_0 + b_1x + b_2x^2} \qquad (9.4)$$

Pearson (1905b) derived the radius of asymmetry and skewness

$$d = \frac{(1/2)\sqrt{\beta_1}(\beta_2+3)}{5\beta_2 - 6\beta_1 - 9}\sigma,\ \alpha = \frac{(1/2)\sqrt{\beta_1}(\beta_2+3)}{5\beta_2 - 6\beta_1 - 9}. \qquad (9.5a, b)$$

As he showed, for curves not essentially differing from the Gaussian these expressions can be simplified:

$$d\ =\ \frac{1}{2}\sigma\sqrt{\beta_1},\ \alpha = \frac{1}{2}\sqrt{\beta_1}. \qquad (9.6a, b)$$

Adducing the *coefficient of dispersion*

$$\eta = \beta_2 - 3 \qquad (9.7)$$

we will obtain the really needed formulas for determining how much our empirical curve differs from the Gaussian curve.

If the curve is asymmetric, we will calculate its peculiar features, the radius of asymmetry and the skewness by formulas (9.6a, b), or, in cases of more pronounced asymmetry, by formulas (9.5a, b). We will certainly consider these magnitudes meaningful only if they more or less considerably exceed their probable errors.

Denoting the probable errors by E with a proper subscript, we have for curves insignificantly differing from the normal type (Pearson & Filon 1898, pp. 276 – 277; Pearson 1902a, pp. 278 – 279)[9.1]

$$\mathrm{E}_d = 0.67449\sigma\sqrt{\frac{3}{2N}},\ \mathrm{E}_\alpha = 0.67449\sqrt{\frac{3}{2N}},\ \mathrm{E}_\eta = 0.67449\sqrt{\frac{24}{N}}.$$

It can also happen that a curve is sufficiently symmetric (both α and *d* are less than their double probable error) but it still cannot be called normal because β_2, see equality (9.3b), differs from 3.

The real significance of the coefficient η is this. If the extreme groups are represented more strongly than in the normal curve, the fourth moment will be increased in such a way that the dispersion becomes *supernormal* and $\eta > 0$. Otherwise, we will encounter *subnormal* dispersion with $\eta < 0$. […]

10. Calculating the coefficients of the Pearsonian curves

After satisfying ourselves, for example, by the methods described in § 9, that the normal curve does not fit the given data, and wishing to derive a theoretical model of the studied phenomena, we will be compelled to determine the equation of an asymmetric curve. In the extreme case it is possible only to derive the radius of asymmetry, the skewness and the coefficient of dispersion (§ 9).

For the calculation in case of a parabola see § 8 [...] which Pearson had however once and for all accomplished, and statisticians can apply his prescriptions. I will not dwell on the derivation of the formulas of the Pearsonian curves or equations for determining their coefficients; any mathematician will be able to do the necessary work by issuing from the equation (9.4) and following Pearson. [...] It only seems of some use to compare all the relevant formulas. First of all, we ought to determine as thoroughly as possible the moments of the empirical distribution. They serve to calculate the constants (9.3a, b),

$$s = \frac{6(\beta_2 - \beta_1 - 1)}{3\beta_1 - 2\beta_2 + 6} \text{ and } k = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}.$$

This k (Pearson 1896a, p. 368; 1901, p. 444) serves as a criterion of the type of the studied curve[10.1].

The Pearsonian curve of Type I

$$y = y_0(1 + \frac{x}{l_1})^{m_1}(1 - \frac{x}{l_2})^{m_2}.$$

The origin is in the mode, y_0 is the maximal (sometimes minimal [the case of an antimode]) ordinate. Radius of asymmetry

$$r = \bar{x} - x_{\text{mod}} = d,$$

$l_1 - l_2$ is the base and α is the skewness. The relevant formulas are

$$d = \frac{\mu_3}{2\mu_2}\frac{s+2}{s-2},\ \alpha = \frac{d}{\sigma}, \qquad (10.1)$$

$$l = \frac{\sigma}{2}\sqrt{\beta_1(s+2)^2 + 16(s+1)}, \qquad (10.2)$$

$$l_1 = \frac{1}{2}(l - ds),\ l_2 = \frac{1}{2}(l + ds), \qquad (10.3)$$

$$m_1 = \frac{l_1}{l}(s-2),\ m_2 = \frac{l_2}{l}(s-2), \qquad (10.4)$$

$$y_0 = \frac{N}{l}\frac{m_1^{m_1}m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}}\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}, \qquad (10.5)$$

where N is the number of items equal to the area of the curve. An approximate value of y_0 is

$$y_0 = \frac{N}{l}\frac{(m_1+m_2+1)\sqrt{m_1+m_2}}{\sqrt{2\pi m_1 m_2}}\exp\{\frac{1}{12}[\frac{1}{m_1+m_2} - \frac{1}{m_1} - \frac{1}{m_2}]\}. \qquad (10.6)$$

Formula (10.1) can be derived from Pearson (1896a, p. 370). Formulas (10.2, 5, and 6) are due to Pearson (1896a, p. 369) and formulas (10.3 and 4) to Davenport (1899 or 1904?, p. 32) and Pearson (1896a, pp. 369 – 370).

Tables for calculating the function Γ are in Leontovich. Pearson (Editorial 1908) provided an approximate but very precise formula

$$\lg\frac{\Gamma(x+1)}{x^x e^{-x}} = 0.3990899 + \frac{1}{2}\lg x + 0.080929\sin\frac{25°.623}{x}.$$

Its errors are 1/25,000, 1/50,000 1/900, 000 for $x + 1 = 2$, 3 and 7, and for $x + 1 = 11$ the formula provides 7 correct digits. It is therefore applicable for every value of x not included in other compiled tables.

Very precise is also the formula (Forsyth 1883), also Pearson (Editorial 1908, p. 118)

$$\Gamma(n+1) = \sqrt{2\pi}[\frac{\sqrt{n^2+n+1/6}}{e}]^{n+1/2}.$$

Its error is less than $1/240n^3$.

The Pearsonian curve of Type II

This is a particular case of Type I with $l_1 = l_2$ and $m_1 = m_2$. The equation of the curve is

$$y = y_0(1-\frac{x^2}{l^2})^m$$

where l is now half the base. The curve is symmetric and therefore $d = \alpha = 0$. Then

$$s = \frac{3(\beta_2-1)}{3-\beta_2} \text{ because } \beta_1 = 0,\ l = \sigma\sqrt{s+1},\ m = \frac{1}{2}(s-2),$$

$$y_0 = \frac{N}{l}\frac{\Gamma(m+1.5)}{\sqrt{\pi}\Gamma(m+1)},\ y_0 \approx \frac{N}{\sigma\sqrt{2\pi}}\frac{s-1}{\sqrt{(s+1)(s-2)}}\exp[-\frac{1}{4(s-2)}].$$

For the two last formulas see respectively Pearson (1896a, p. 372) and Davenport (1899 or 1904?, p. 33).

The Pearsonian curve of Type III

$$y = y_0(1+\frac{x}{l})^p e^{-\gamma x},\ p = \gamma l.$$

Theoretically, such curves demand that $k = \infty$, but even for its moderate positive values they provide good results [?]. The base of the curve is only limited in one direction, the origin of the system of coordinates is at the mode, and l is the interval from the left boundary to the mode. Only the first three moments are sufficient for calculating. Then, the radius of asymmetry and the skewness are

$d = \mu_3/2\mu_2$, $\alpha = d/\sigma$,

$l = (\mu_2/d) - d$, $\gamma = 1/d$, $p = l/d$, (10.7, 8, 9)

$$y_0 = \frac{N}{l}\frac{p^{p+1}}{e^p\Gamma(p+1)}. \qquad (10.10)$$

For the last formula see Pearson (1896a, pp. 373 – 374).

If the left boundary is given, then we know its distance from the centre,

$L = l + d$ and of course $l = L - d$

so that formula (10.7) provides

$$d = \frac{\mu_2}{l+d} = \frac{\mu_2}{L}.$$

Formulas (10.8, 9, 10) are needed for calculating γ, p and y_0. This method [?] can sometimes be applied for checking. Although less precise, it simplifies calculations because μ_3 is not necessary. However, without knowing the moments, we cannot be sure that that type will fit the studied curve.

The Pearsonian curve of Type IV

$$y = y_0(\cos\theta)^{2m}e^{-\nu\theta}, \quad \frac{\theta°}{180°}\pi = \text{arctg}\frac{x}{a}.$$

The curve is asymmetric, extends to infinity in both directions, and a is a positive constant. The origin is at point $x = 0$, $\theta = 0$, $y = y_0$ and does not coincide either with the mode or the centre.

We denote the distance from the mode and the centre to the origin by l and

$L = \mu'_1 = l + d$

respectively and introduce $r = -s$ instead of s:

$$r = -s = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6}.$$

For curves of this type r is always positive and larger than 3 (Pearson 1896a, p. 379). Then

$$m = \frac{1}{2}(r+2),\ d = \frac{\mu_3}{2\mu_2}\frac{r-2}{r+2},\ a = \frac{\sigma}{4}\sqrt{16(r-1) - \beta_1(r-2)^2}, \qquad (10.11,\ 12,\ 13)$$

$$\nu = -\frac{\mu_3}{4\mu_2}\frac{r(r-2)}{a},\ L = md,\ l = L - d,\ y_0 = \frac{Ne^{\pi\nu/2}}{a\int_0^\pi \sin^r\theta e^{\pi\nu}d\theta}, \ (10.14,\ 15,\ 16,\ 17)$$

$$y_0 \approx \frac{N}{a}\sqrt{\frac{r}{2\pi}}\frac{\exp[(\cos^2\varphi)/3r - (1/12r) - \varphi r\text{tg}\varphi]}{(\cos\varphi)^{r+1}},\ \text{tg}\varphi = \frac{\nu}{r}. \qquad (10.18)$$

All these formulas can be derived by simple transformations from Pearson (1896a, pp. 377 – 380). Formula (10.11) is on p. 378; formula (10.12) can be derived from the formula for skewness contained there

$$\text{skewness} = \frac{1}{2}\sqrt{\beta_1}\,\frac{r-2}{r+2}$$

when multiplying it by σ and putting μ_3^2 / μ_2^3 instead of β_1. The sign of *d* is determined by the sign of μ_3 which is directly seen in formula (10.12). We will arrive at formula (10.13) from Pearson's formulas on p. 378:

$$a = r\sqrt{\frac{\mu_2(r-1)}{z}},\ z = \frac{r^2}{1-\beta_1(r-2)^2/16(r-1)}.$$

On the same page Pearson gives formula

$$\nu = \sqrt{z-r^2}$$

from which, applying the value of *z* and making use of formula (10.13) we get formula (10.14). *Minus* is justified by Pearson's remark that the signs of ν and μ_3 are opposite, which, however, is seen from his expression on p. 377:

$$\mu_3 = -\frac{4a^3\nu(r^2+\nu^2)}{r^3(r-1)(r-2)}.$$

Issuing from Pearson's formula (same page)

$$\mu_1' = -\frac{a\nu}{r},$$

deriving – *a*ν/*r* from formula (10.14) and making use of formulas (10.11) and (10.12), I obtain formula (10.15). Formula (10.16) follows from the definition of the radius of asymmetry and, finally, formulas (10.17) and (10.18) are on pp. 378 and 380.

The aim of all these transformations is to simplify those formulas as much as possible so that in addition the signs of the magnitudes involved will be determined by the formulas themselves, and to arrange them in the order most suitable for calculations.

The Pearsonian curve of Type V

$$y = y_0 x^{-p} e^{-\gamma/x}.$$

The base is only limited in one direction, the origin of the system of coordinates is in the beginning of the base and the maximal ordinate is y_0. For curves of this type the supplementary magnitude *s* is always negative. Introducing as in the previous case the same $r = -s$, we will easily find (10.12) from the general formula (9.5a) and $\alpha = d/\sigma$, then

$p = r + 2$, $L = (1/2)dp$, $l = L - d$, $\gamma = lp$,

$$y_0 = \frac{N\gamma^{p-1}}{\Gamma(p-1)}.$$

Pearson (1901, p. 447) provides equation

$$(p-4)^2 - (16/\beta_1)(p-4) - (16/\beta_1) = 0$$

whose positive root determines *p*. He derived the general formula for *d* (9.5a) later, and *p* can be calculated by issuing from it as well without solving that quadratic equation. His formula (VIII) (Ibidem) allows to determine

$$\frac{\mu_3}{\mu_2} = \frac{4\gamma}{(p-2)(p-4)}$$

which we apply in formula (10.12) and equate the *d* to its other expression in his formula (XVI) (Ibidem, p. 448). The other formulas above can be easily derived from his formulas (Ibidem).

The Pearsonian curve of Type VI

$$y = y_0 \frac{(x-a)^{q_2}}{x^{q_1}}. \quad a = \sigma\sqrt{(1/4)\beta_1(r-2)^2 - 4(r-1)},$$

The base is only limited in one direction and the curve begins at distance *a* to the right from the origin. Here $r > 0$ and, as before, $s = -r$. Then

$$d = \frac{\mu_3}{2\mu_2}\frac{r-2}{r+2}, \ \alpha = \frac{d}{\sigma}, \qquad (10.19, 20)$$

$$a = \sigma\sqrt{(1/4)\beta_1(r-2)^2 - 4(r-1)}, \qquad (10. 21)$$

$$q_1, q_2 = \frac{1}{2}(r+2)\ (\frac{d}{a}r \ \pm\ 1), L = \frac{a}{r}(q_1 - 1),\ l = L - D,$$

$$y_0 = \frac{Na^{r+1}\Gamma(q_1)}{\Gamma(r+1)\Gamma(q_2+1)}. \qquad (10.22, 23, 24, 25)$$

Pearson’s *r* is my *s*. For q_1 and q_2 he (p. 450) provides an equation (in my notation)

$$Z^2 + rZ + \varepsilon = 0,\ \varepsilon = \frac{r^2}{4 + (1/4)\beta_1(r-2)^2/(1-r)}$$

whose roots are $(1 - q_1)$ and $(1 + q_2)$. Knowing *d* [?], we can solve this equation by making use of the properties of the roots of the quadratic equation.

Replacing $(1 - q_1)(1 + q_2)$ by ε in Pearson’s formula (XXIV) we derive formula (10.20) for *a*. Then Pearson (p. 450) provides

$$d = \frac{a(q_1 + q_2)}{(q_1 - q_2)(q_1 - q_2 - 2)}.$$

Putting *r* instead of $(q_1 - q_2 - 2)$ we find that

$$q_1 + q_2 = \frac{d}{a} r(r+2)$$

whence (10.21) and (10.22). Formula (10.23) corresponds to the first of Pearson’s formulas (XXII) on p. 449 and (10.25) is identical to (XXV).

These changes in formulas ought to simplify essentially the application of this type of curves since the solution of the quadratic equation with multidigit coefficients becomes not necessary anymore.

Notes

0.1. Where are Chebyshev, Markov, Liapunov? O. S.

0.2. In a few years Slutsky (1916) published a review of a Russian book written by a resolutely non-mathematical statistician, and there we find a somewhat contradictory (although not very definite) statement (p. 110/2009, p. 94) about the "theoretical considerations on which statistical methodology is built":

Isolating that which relates to the properties of, first, judgements and concepts, i. e., to logic [rather to philosophy] *and then of the properties of quantitative images upon which it* [logic] *is operating, i. e., of mathematics, we nevertheless obtain some remainder for which no acknowledged sanctuary is in existence, which remains uncoordinated and homeless until we perceive its special theoretical essence and provide it with the missing unity in the system of judgements fully deserving the name of theoretical statistics*. O. S.

1.1. These subgroups are regrettably often too large so that a proper idea about the distribution is difficult to obtain. Even if the grouping is rather accurate, the lowest subgroups and especially the highest subgroups are too wide, for example, the grouping of peasants according to the area under crops: 0 – 5, 5 – 10, 10 – 15, 15 – 25, 25 – 50 and more than 50 dessiatin [1 dessiatina = 2.7 acres O. S.]. It is hardly possible to treat rationally such a grouping and it should be insisted that the underlying indication be subdivided into equal intervals. Especially important is a detailed subdivision at the tails. E. S.

1.2. In such cases the author obviously has in mind the area under the curve. Cf. Laplace (1812/1886, p. 342): " […] l'ordonnée qui divise l'aire de la courbe en parties égales".O. S.

1.3. Slutsky often supplied the English term as well. O. S.

3.1. Slutsky had not explained either his calculation, or the essence of such presentation. True, it was commonly used, but at least the beginners would not understand the meaning of the additional terms preceded by the double sign. O. S.

4.1. The probable error had been in general usage, but to state, in the same section, that it provided the only test for some important conclusions was wrong. O. S.

4.2. Here, and many times below Slutsky refers to Leontovich, in particular as a source of statistical tables; below, in such cases, I am omitting these references.

4.3. This statement is at least ambiguous. Random errors actually accumulate with the number of such repetitions (proportionally to their square root). Boscovich (1758/1922, § 481) wrongly thought that "In circumstances that are fortuitous, […] the greater the number taken, the more the sum of the irregular inequalities [of velocities of a "particle"] decreases". However, perhaps he thought about the mean value. Even Helmert (1905, p. 604; 1993, p. 200) had to warn his readers that the sum of such errors did not tend to vanish. And because of unavoidable presence of systematic errors even the arithmetic mean does not approach certainty. Bayes, in an unpublished manuscript (Dale 2003, p. 385), effectively mentioned this circumstance apparently having in mind the celebrated Simpson memoir of 1756. Then, trials or measurements are not completely independent. Citing this fact, Chuprov (report of 1918, publ. 1926/2004, p. 80) stated: "Most statisticians are apt to rely blindly on the proposition that the random fluctuation of statistical numbers [though he hardly thought about the mean] must decrease when the number of trials increase […]". O. S.

5.1. Leontovich reprinted Sheppard's excellent tables (1903) where the values of the integral are a function of x/σ. The argument x in Markov (1900) and Chuprov (1909), for example, is divided by the modulus which is easily calculated as $\sigma\sqrt{2}$. In some cases (Encke's tables) it is necessary to know the ratio of x to the probable error equal to 0.67449 σ. The meaning of the probability integral and the method of extracting its values from the tables are explained in the quoted writings. E. S.

5.2. Here is a passage from the extant part of an unsigned and undated letter certainly written by Slutsky to Markov, likely in 1912 (Sheynin 1999, p. 132/2004, p. 226) describing the result of that derivation:

are not independent in magnitude from the sum of the already accumulated deviations or that the probabilities of equal deviations [of each sign] *are not constant, we shall indeed arrive at the formula* [(5.4) without the minus] […]. *Much material* [already shows that the Pearsonian curves are useful but] *it seems desirable also* […] *to provide a theoretical derivation which will put* [them] *in the same line as the Gauss curve on the basis of the theory of probability (hypergeometric series).*

In this § 5, Slutsky describes Pearson's derivation of the normal law and the case in which his assumptions are not obeyed and the Pearsonian curves appear instead. The demonstration is certainly unsatisfactory because the conditions of the central limit theorem are not fulfilled, nor are the assumptions necessary, and, indeed, the Pearsonian curves include the normal law as well. O. S.

6.1. What kind of equation ought to be chosen depends on many circumstances, and it is impossible to suggest here a general theoretical rule. In some cases a good result is provided by a parabola of an *n*-th degree, in other instances it is better to use a trigonometric or exponential curve. We ought to call that curve the best which adjoins the empirical line the closest and demands calculation of a lesser number of coefficients.

It should not be thought that an increase in that number is always essentially beneficial and that the choice a parabola of the sixth, eighth or tenth degree without fail secures a good fit; more important is the choice of the type of the curve. Pearson (1902c, pp. 16 – 19) indicates, for example, that in a case he considered a parabola of the sixth degree with seven parameters suited the empirical data worse than his curve of distribution with three parameters. On a rational degree of the conformity provided by a theoretical curve see Pearson (1900); also see the last chapter of this book. E. S.

9.1. It is convenient to arrange the calculations in the following way (Pearl 1906). Extract $0.67449/\sqrt{N}$ from the tables Gibson (1906), multiply that number by $\sqrt{3/2}$ to provide E_α, then calculate the other two errors making use of that E_α. E. S.

10.1. Slutsky provided a summary briefly describing the discussed six types of Pearsonian curves. For a modern summary covering all the 12 types see for example Dodge (2003, pp. 414 – 416). O. S.

Part 2. Theory of Correlation

Chapter 1. Correlation between Two Magnitudes

11. The notion of correlative dependence

The main type of dependence in the so-called exact sciences is the *single-valued functional dependence*. To each value of one magnitude corresponds one definite value of the second magnitude […].

We have to study relations of an absolutely different type. Suppose we wish to find the dependence between the statures of father, x, and son, y. Like previously, each pair of values of these magnitudes corresponding to the pairs of individuals can be represented by a point, but the points thus obtained will not lie on one and the same line, but rather provide a picture of a cloud. For the sake of clarity we subdivide the field by vertical and horizontal lines forming squares with side δx.

Enumerate the values of x and y corresponding to the middle of the intervals between adjacent lines and call them *versions*. The set of cases in which $x_i - \delta x/2 < x < x_i + \delta x/2$ is called an *array* corresponding to x_i. There is no question here about any single-valued functional dependence or a multivalued functional dependence of the usual type when several and sometimes even infinitely many values of the second variable correspond to a definite value of the first variable as in the case of the sine function. […]

All points are here isolated and only a more or less indefinitely outlined group of values of the second variable corresponds to each value of the first one. However, after calculating the arithmetic means of the values of y for each x-array we will see that they are located along some line. In our case [excluded from translation], it will be a broken line closely situated to a straight line. Repeating this procedure for the y-arrays, we will find another broken line closely situated to another straight line. These straight lines are called *regressions of y on x and of x on y*.

And so, we are unable to determine in each isolated case one magnitude given the other one, but we can indicate the mean value of one of them corresponding to a definite value of the other one. In addition, when considering some isolated array, we note that the points there are densest near the line of regression, that is, near their mean value. The farther from it are the points, the sparser they become, and beginning from some distance there are none or almost none. So, the frequency of each value of one of the magnitudes is a function of the other magnitude, or, expressed in a somewhat different manner, the frequency of the pair $(x_i; y_j)$ is a function of those magnitudes.

This function can be represented in a manner similar to that applied in Part 1 for showing the distribution of one magnitude. If the number of items in a subgroup corresponding to the i-th interval of x and, at the same time, to the j-th interval of y is n_{ij}, then n_{ij} is also the number of points in the corresponding square.

Imagine now a parallelepiped on each square whose volume is proportional to the size of the corresponding subgroup. Then its height will represent the number of items (cases) per unit area. The upper faces of the parallelepipeds when their number and the size of the totality increase unboundedly will merge and form a surface called surface of distribution[11.1] or of frequency. Its general equation will be $Z = f(x; y)$. We are now able to define correlation dependence generalizing it at once to any number of variables.

Several magnitudes are in correlation if to each totality of the values of all [of each] *of them except one there corresponds a whole complex of the values of that last one and* *the arithmetic mean of each variable changes depending on the values of the other ones and the frequency of each set of values of the variables is a function of those values.*

If the arithmetic mean of some variable remains constant in all the arrays arranged for the other variables, then we say that correlation does not exist (or is equal to zero). If the increase in one magnitude leads to the increase of the arithmetic mean of another one, the correlation is positive, and negative in the opposite case**[11.2]**. The closer the separate values of a magnitude adjoin the regression line, the smaller therefore is the difference between the arithmetic mean of a magnitude in each array and the separate values of that magnitude in that array, the more complete is the correlation.

Imagine that the points on the correlation diagram are grouping ever closer to a certain direction, then the lines of regression will have to approach each other ever closer, and when all the points become situated along a single line both regression lines coincide with it. We will have a picture of perfect correlation, or transition of the correlation connection to a usual functional connection.

A complete investigation of a correlation dependence ought to include: First, the derivation of the regression lines; second, the estimation of the degree of the correlation connection; and third, the derivation of the equation of the surface distribution which allows to calculate the probability of each value of any magnitude given the values of the other one. Up to now, the last-mentioned problem is only solved for the case of the so-called normal surface distribution corresponding to the normal Gaussian curve (Part 1).

12. The correlation table

Representing each separate case by its own point is quite suitable for ascertaining the essence of the examined relations, but not convenient in the practical sense. We should not forget that statistically studied phenomena are occasioned by the influence of innumerable causes and that one of the most important aims of research consists in discerning the main tendencies in the explored relations by freeing them as much as possible from the admixture of random elements.

Not a single characteristic of a mass phenomenon can be therefore thought sufficient without its probable error being indicated, and the results of any two methods differing less than by their probable error should be admitted on a par. Preferable is that which demands less calculations, and all those simplified methods applied by modern statistics are conditioned by this reasoning.

The main among them (Part 1) consists in that we combine separate cases and consider each thus derived group as consisting of identical items with value [of indication] corresponding to the middle of the appropriate interval. This is indeed done when constructing a correlation table. […]

If some point is situated on the line separating two intervals, it has to be halved and each half a case added to one of these intervals. Likewise, for points situated on the borderline of four intervals the cases are quartered. If the grouping is planned prior to measurement or calculation of magnitudes relating to separate cases, it is almost always possible to measure/calculate doubtful cases to such a degree of precision that will avoid the complication of splitting up the points (Yule 1899, p. 257).

Several correlation tables are adduced in the supplement to this book [excluded from translation] and the reader will see at once that such a table is nothing but a usual table well known to each statistician. To extract all possible from the raw and grouped data contained there is indeed the aim of the correlation method.

13. The regression lines

The first step of investigating correlation dependence is the derivation of the arithmetic means of the separate arrays. When representing these means by points and connecting these by straight lines, we obtain an empirical line of regression which at once provides valuable indications about the studied dependence.

Nevertheless, however valuable this is, a statistician cannot be satisfied by the picture being presented: it only suggests a number of questions the answer to which is only possible to get by mathematically treating the material further. In an example considered by Pearson & Lee (1903, pp. 362ff and Table XXII on p. 415) the measurements covered 1078 pairs of fathers and sons. […] Fathers with stature, in inches, [58.5; 59.5], had sons with mean stature 64.4; sons' mean stature 65.6 corresponded to stature 60 of the fathers etc. In general, increase in the stature of the fathers is connected with an essential increase in the sons' mean stature which is indeed the expression of the known fact of hereditary descent of quantitative traits.

On the face of it [?], the law of that descent is not simple. Following all the zigzags of the regression line we are compelled to conclude that for the group of the shortest fathers [59; 61] a mean increase of 1 inch in the sons' stature corresponded to the same increase in the fathers' stature. We also have, so to say, an anomalous interval [61; 62] of the fathers' stature whose increase is followed by a *decrease* in the sons' mean stature. [Discussion of the other intervals follows].

When analysing the regression line we could have obtained many more such *laws* which I do not adduce not because they have a proper place in a treatise on the theory of heredity rather than here: neither has that theory any use of them since they do not exist at all. Having measured another thousand of such pairs of individuals, we would have most certainly found no trace of our imaginary laws. The general direction of the regression line would not have changed, but its separate zigzags would have possibly become arranged in quite another way. Indeed, correlation connection is only expressed by mean values and a large number of measures is needed for revealing it because otherwise random causes can wholly conceal the common trend.

In the Pearson & Lee example the extreme zigzags of the regression line are easily explained by the small number of the appropriate cases […]. It is not necessary to be an eminent theoretician to feel at once that such details in the behaviour of that line are hardly reliable. The considerations above only concerned a particular case but they evidently have general meaning. When studying any dependence it is always necessary to abstract oneself from features peculiar to the particular material, to eliminate random deviations obscuring the action of general tendencies. In statistics, the only means for achieving this is to derive a numerical expression for each characteristic and to compare it with its probable error. Only this method secures trustworthy results.

In particular, the regression lines ought to be transformed, their zigzags smoothed and the main hidden tendency of each line revealed. This aim is attained by determining some smooth line which can be called the theoretical regression line and which adjoins the points of the empirical line as much as possible. If the deviations of the empirical line from the theoretical do not exceed in their totality [?] their probable random values, the theoretical line can be considered as quite adequately representing the real relations. Otherwise, it is also possible to make use of the theoretical line although not forgetting that it is only a more or less crude approximation.

A [theoretical] regression line adequate to the limits of probable errors [if these errors are allowed for] can serve as a criterion for establishing the type of regression. In this sense, we distinguish between linear regression whose geometric image is a straight line, and curvilinear regression, usually represented by some parabolic curve. Its theory is not yet sufficiently developed, and we will mostly deal with the former type of regression. At least to a certain approximation it happily occurs in most cases with which statistical practice has to deal.

14. Examples

[Slutsky introduces the regression coefficient, the slope of the regression line, and in cases of regression of y on x and of x on y respectively denotes them by $\rho_{y(x)}$ and $\rho_{x(y)}$ so that the equations of these lines passing through the origin are

$$y = \rho_{y(x)}x,\ y = \rho_{x(y)}x. \qquad (14.1,\ 2)$$

In the second case, however, the inclination is measured relative to 0y, call it β, and it will be equal to $(90° - \beta)$ relative to 0x, and $\tan \beta$ will be written as $\cot (90° - \beta)$. Slutsky denotes the inclination of the first regression line by α.

He considers two examples pertaining to 1901: expenditure on public education and on the maintenance of the administration itself concerning "all the 359 districts" [of Russia]; and the price [of rye] in commercial centres NNo. 2 and 3 [§ 3]. In the first example the dependence is weak, in the second it is much stronger with the regression coefficients being $\rho_{y(x)} = 0.14$ and 1.13 respectively. He continues:]

Of course, even a weak dependence is interesting, but we should first of all ascertain that it really exists. It is indeed possible that the inclinations of the regression lines were occasioned by random causes and that, had we compiled correlation tables for a number of years, the lines would then be inclined sometimes to one side, sometimes to another, and in the mean, assuming a long period of time, they would coincide with the coordinate axes thus indicating an absence of any dependence.

We are returning here to admitting the need to have the probable errors of the numerical characteristics of the studied phenomenon. Even an investigation extended over a number of years, as mentioned above, cannot replace them. Suppose we obtain a positive regression coefficient for one year, and a negative coefficient for another year, each of them ten times (say) exceeding its probable error. We will then be compelled to conclude with a very high probability even practically coinciding with certainty that the dependence between the [studied] phenomena did exist in each year, but that for some reason its type had changed[14.1]. […]

15. The correlation coefficient

We have seen that the correlation dependence between phenomena can be both more or less close. Beginning with complete independence and passing through a number of gradations it finally becomes a strict functional dependence between two magnitudes. As mentioned above, the degree of correlation dependence reflects on the inclination of the lines of regression. If correlation is absent, they ought to coincide with the coordinate axes, and in case the dependence becomes functional, they must merge into a single line, into a single straight line if the dependence is linear.

The examples above [excluded from translation] aimed at illustrating these propositions and at making them obvious to some extent. And now the reader will probably agree that the value of a separate regression coefficient cannot yet serve as a measure of the closeness of the correlation dependence (of the correlation). First, there are two regression coefficients; one of them can be rather large, the other near zero and the correlation will be yet far from a strict functional dependence. Second, regression coefficients are concrete numbers and therefore change with the choice of the units of measurement and scale. For example, when studying the correlation between the price of bread and mortality, these coefficients will take different values depending on whether the latter is measured in percents or thousandths, and whether the former is expressed in copecks per pound, per pood [16.4 *kg*] […].

The measure of correlation however must be an abstract number. Consider the square root of the product of regression coefficients. It will be such a number, independent from the choice of the units of measurement. […] [Slutsky explains that, in notation of § 14,

$$\sqrt{\rho_{y(x)}\rho_{x(y)}} = \sqrt{\tan\alpha \cot(90° - \beta)}$$

is indeed an abstract number. Then, in the absence of correlation the regression lines coincide with the respective coordinate axes and both multipliers in the right side vanish. If correlation is a linear functional dependence, those lines coincide with each other and the product in the right side becomes equal to unity.]

The indicated properties make the geometric mean of the regression coefficients a convenient measure of the degree of correlation dependence. This magnitude is important in the correlation theory; it is designated by letter r with appropriate subscripts and called *correlation coefficient*:

$$r = \pm\sqrt{\rho_{y(x)}\rho_{x(y)}}. \qquad (15.1)$$

The regression coefficients always have identical signs [so that their product is always positive]. We can agree that r is positive when they both are positive, and negative otherwise. In the sequel, we will derive another formula for the same correlation coefficient which can, and usually is assumed as its main expression with its sign determined without involving any further reasoning. In general, we may consider all the statements made until now as preliminary, aimed at ascertaining the main notions of correlation theory. In our next section, we turn to their rigorous proofs and a derivation of a number of propositions and formulas.

16. Formulas for the regression coefficients and the correlation coefficient

A straight regression line ought to indicate some mean direction of the empirical line and in general it can be obtained by different methods. For example, it can conform to the minimal sum of the distances of the empirical points from it, all of them considered positive. It is also possible to determine a straight line for which the sum of the squares, or of the fourth powers of those distances, or the sum of their third powers taken independently from their signs, will be minimal.

To some extent, we are free to choose; each of these methods is good enough if it provides a comparatively simple result and if everyone will *agree* to apply it. These remarks are necessary for stressing the conventionality inherent in the *method of least squares* widely applied in science[16.1]. In any case it is important to indicate that, when turning to that method, we do not assume anything about the essence of the distribution of the separate values of our magnitudes so that all the formulas of the correlation theory persist under any "law" of distribution.

We already know some notation pertaining to correlation tables. The general number of cases is N; the size of the i-th array of the x's [the frequency of the x's] (of the vertical column of the table) is n_{xi}, and in a similar way we introduce n_{yj}. The subgroup belonging at the same time to the x-array and y-array will be n_{xiyj}, or, shorter, n_{xy} or n_{ij}. The arithmetic means for all the totality will be $\bar{x}$ and $\bar{y}$, or h_x and h_y, and σ_x and σ_y, the standard deviations. Constants of distributions can also be found for the totalities comprising separate arrays: the arithmetic means y_{xi}, or shorter y_x, and standard deviations σ_{nxi} or σ_{nx} and x_{yj} or x_y and σ_{nyi} or σ_{ny}.

Let us now derive the equation of the regression line for y on x (Yule 1897b). Suppose it is

$$Y = a + bX \qquad (16.1)$$

where $b = \rho_{y(x)}$ is its slope, call the line P_1P_2, and the origin is at the centre of the distribution […]. Denote the vertical distance between the points [with identical abscissas] situated on the regression line and P_1P_2 by d. If a point on the former is the centre of distribution of an x-array, then

$$d = y_x - Y$$

and the straight line P_1P_2 should be determined by the sum of the squares of such magnitudes being minimal. [Slutsky derives the condition sought:

$$\sum[y-(a+bx)^2] = \text{min} \qquad (16.2)$$

where the sum covers the partial sums taken over the i-th arrays of x.]

This result throws new light on the condition to which we have subordinated the regression line (16.1). If all the points of the totality are situated on it, we will be able to calculate y for a given x from equation

$$y = a + bx.$$

However, a number of cases with differing values of y correspond to one value of x; in other words, each time we will arrive at a more or less wrong result with error

$$y - (a + bx).$$

The condition of least squares […] is tantamount, as we see now, to another condition (16.2): *Determine such a linear dependence between x and y that, when applying it for calculating y from a given x, the sum of squares of errors thus encountered is minimal.*

We derive the second empirical regression line, not coinciding with the first one, in the same way […]

We will now determine the final form of the equation of the straight regression line for y, that is, derive its coefficients a and b from condition (16.2). [Slutsky derives

$$\sum x[y-(a+bx)] = 0, \; a = 0.] \qquad (16.3)$$

This result is very important; it indicates that with $x = 0$ y also vanishes in the mean. That is, when the first magnitude takes its mean value, the second one (in the mean of a number of cases) also coincides with its mean. Since the same takes place for the case of many variables (§ 37), we conclude that for linear regression the concept of *typical* as a combination of arithmetic means is quite admissible. The Average man of a given age (introduced by Quetelet) having mean stature, mean size of various organs, mean abilities, etc does not represent anything unreal[16.2].

A number of statistical investigations in anthropology [anthropometry]$^{\mathbf{16.3}}$, and especially those made by the Pearson school, indicated that linear formulas can be applied with insignificant error to most various indications. However, the dependence of stature on age is absolutely non-linear, see for example Powys (1901, p. 47). It is quite possible that in cases of non-linear regression a mean value of one indication will be associated, in the mean, not with the mean of another indication, and vice versa. In this case an individual possessing all indications of mean size can be extremely unlikely, therefore not typical (Pearson 1905b, p. 29).

From equation (16.3) with $a = 0$ we have

$$b = \frac{\sum xy}{\sum x^2}. \qquad (16.4)$$

Replacing $(x - \bar{x})$ by x in formula (3.1), that is, assuming that x is the deviation from the mean, and, instead of multiplying x^2 by n_x, simply repeating it a necessary number of times, we will obtain the denominator in (16.4):

$$\sum x^2 = N\sigma_x^2$$

and

$$\rho_{y(x)} = \frac{\sum xy}{N\sigma_x^2},\ \rho_{x(y)} = \frac{\sum xy}{N\sigma_y^2},\ r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y}. \qquad (16.5, 6, 7)$$

The last-written expression is *the main formula of the correlation method.*

Replacing now the sum in formulas (16.5) and (16.6) by its expression following from (16.7), we obtain the generally applied and simple formulas

$$\rho_{y(x)} = \frac{\sigma_y}{\sigma_x} r_{xy},\ \rho_{x(y)} = \frac{\sigma_x}{\sigma_y} r_{xy} \qquad (16.8)$$

for the regression coefficients. The equations of the straight regression lines will be

$$Y = \frac{\sigma_y}{\sigma_x} r_{xy} x,\ X = \frac{\sigma_x}{\sigma_y} r_{xy} y. \qquad (16.9)$$

17. Other formulas for the correlation coefficient

Before analyzing the obtained expressions, we will dwell somewhat on the mathematical aspect of the matter and derive a number of formulas occurring in the sequel.

A. In a certain respect the sum $\sum xy$ is an expression inconvenient for calculating. First of all, having a large number of cases, it is extremely burdensome to calculate every product xy for each pair of values separately. Nevertheless, this method of calculating ought to be recommended when the total number of cases in the table is not very large, 20, 30 or 50, say. For facilitating calculations the table should otherwise be subdivided into squares as was described above, and all the values of x and y in each such subgroup (cell) should assumed to be invariably equal to the appropriate *x* and *y* versions [see beginning of § 11], and then […]. We will have

$$r = \frac{\sum n_{xy} xy}{N\sigma_x \sigma_y}.$$

B. The order of calculation can differ. [Slutsky derives the following formulas:

$$\sum n_{xy} xy = \sum n_x y_x x = \sum n_y x_y y.] \qquad (17.1)$$

They indicate the comparatively most convenient order of addition; their theoretical application will be encountered in the sequel.

C. Let us recall that, when deriving the formulas of the correlation coefficient, we measured our magnitudes by their deviations from their means. For returning to the usual method of measurement, we ought to replace *x* and *y* by $(x - \overline{x})$ and $(y - \overline{y})$ respectively. Then the formula for the correlation coefficient will be

$$r_{xy} = \frac{\sum n_{xy} (x - \overline{x})(y - \overline{y})}{N\sigma_x \sigma_y}. \qquad (17.2)$$

Denoting deviations in the numerator by δ*x* and δ*y*, we obtain another often applied form […] which is also written as

$$N\sigma_x \sigma_y r_{xy} = \sum \delta x \delta \mathrm{y}.$$

The regression equations (16.9) will assume their most generally applied form

$$Y - \overline{y} = \frac{\sigma_y}{\sigma_x} r_{xy} (x - \overline{x}),\ X - \overline{x} = \frac{\sigma_x}{\sigma_y} r_{xy} (y - \overline{y}).$$

D. The expression (17.2) is yet inconvenient for calculating since it includes products of numbers $(x - \overline{x})$ and $(y - \overline{y})$, multidigit because the arithmetic means $\overline{x}$ and $\overline{y}$ only by chance and rarely are integers. For determining a convenient formula we will multiply $(x - \overline{x})$ by $(y - \overline{y})$ and calculate the sums thus obtained. We will have

$$\sum n_{xy} (x - \overline{x})(y - \overline{y}) = [...] = \sum n_{xy} xy - N\overline{xy}. \qquad (17.3)$$

Inserting this in formula (17.2) we get

$$r = \frac{(1/N)\sum n_{xy} xy - \overline{xy}}{\sigma_x \sigma_y}. \qquad (17.4)$$

This is indeed the most convenient formula for calculating the correlation coefficient, especially if one of the expressions (17.1) is inserted in the numerator.

18. The mean square error of the regression equation[18.1]

If two magnitudes are connected by a strict linear functional dependence, the equation [its equation]

$$y = a + bx$$

enables to determine one of them given the other one. Suffice it to glance at any correlation table to become convinced that for a correlation dependence this is, however, impossible. We will then see that the sons' stature is far from being determined by the stature of the fathers, that […]. It is not difficult to note also that the boundaries inside which one of the magnitudes fluctuates *when the other one takes a definite value* are narrower than in the general case in which that *other magnitude takes every possible value*.

This narrowing can be so great that, pursuing some goals, we may completely neglect in the former case the difference between the values of the first magnitude. The regression formula will then serve to determine the value of one magnitude given the value of the other one. The only difference here as compared with a strict functional dependence consists in that, independently from the precision of the measurement itself, the result of calculation is more or less approximate.

These considerations can be extended to include not only correlation *close* to functional dependence, but *all* the cases in general. The regression formula provides the mean value of one magnitude given the value of the other one. In a separate case, the value of a magnitude will deviate from that mean, but, knowing the mean value and the law of distribution of these deviations, we will be able to apply the regression formula in particular cases. We would then reason in the following way. If the fathers' stature is x, the sons' stature will be y ± a certain mean square error. If the distribution obeys the Gaussian law, or some known to us law, our forecast can be made more definite; we would then be able to indicate that, for example, in a half, in three quarters, in 90% of all cases the sons' stature will differ from y not more than by a certain magnitude.

The error which we make when applying the regression formula to a separate case will obviously be equal to

$$y - (a + bx)$$

and the mean square error of all such determinations will be

$$\sum_y = \sqrt{\frac{\sum[y-(a+bx)\ ]^2}{N}}.$$

In § 16, the regression equation was derived in such a way that the numerator of the square root was minimal. Therefore, when applying that equation for determining y in separate particular cases, we will make errors whose sum of squares is minimal. The regression equation is thus the best of all possible formulas of a linear dependence.

It is not difficult to calculate the value itself of the mean square error. Suppose that x and y are deviations from their mean values, then the regression equation will take the simple form

$$y = \frac{\sigma_y}{\sigma_x} rx$$

so that

$$(\sum_y)^2 = \sigma_y^2(1-r^2),\ \sum_y = \sigma_y\sqrt{1-r^2}. \qquad (18.1, 2)$$

Expression (18.1) allows to formulate a number of important conclusions about the correlation coefficient; one of them is on p. 70[18.2], but the proof provided there cannot be considered rigorous. The sum (18.1) is a sum of squares, always positive, as well as σ_y^2. Therefore,

$$1 - r^2 \geq 0,\ r^2 \leq 1,\ -1 \leq r \leq 1.$$

Thus, *the absolute value of the correlation coefficient cannot exceed unity*. Then, if $r = 1$, $1 - r^2 = 0$ and $(\sum_y)^2 = 0$ which is only possible if each of the appropriate terms is zero. Therefore, for each value of y we have

$$[y - \frac{\sigma_y}{\sigma_x}rx]^2 = 0,\ y = \frac{\sigma_y}{\sigma_x}rx.$$

The correlation coefficient only equals unity with either sign if the regression equation is satisfied by each pair of the correlated magnitudes; that is, when the correlation becomes strictly functional, and, in addition, linear.

We have provided the appropriate geometric interpretation (end of § 11): as correlation approaches strict functional dependence, the points of the correlation diagram group ever closer around a single direction and are finally situated (in case of a linear regression) along a straight line. The mean square error of determining y given x will then vanish.

Formula (18.2) indicates that, for an insignificant value of the correlation coefficient the mean square error ($\sum$) will barely differ from the standard deviation (σ). This means that the distribution of y in each array will little differ from its distribution over the whole totality and that, therefore, the forecast which we are able to make about y given x must be very imperfect[18.3]. In any case, its precision will barely differ from that of a judgement which we may formulate by issuing from the arithmetic mean and mean square deviation, and thus to indicate under the normal distribution (say) with a certain probability the boundaries between which we may expect to encounter the values of the studied magnitude. Examples: Table VII and Fig. 17 [excluded from translation].

Another picture emerges if the correlation coefficient is near unity. Suppose for example that $x = y = 200$, $\sigma_x = \sigma_y = 50$, $r_{xy} = 0.999$ and that the distribution is Gaussian. Then each magnitude will vary between rather wide boundaries: about 2/3 cases will be situated in the interval [150; 250], but if one of the magnitudes is fixed, the distribution of the other one will be very compressed. Indeed, formula (18.2) provides 2.25: the variation of this magnitude will be reduced to 4.5% of its initial spread. Calculating the mean value of y given, for example, $x = 210$, we find from the regression equation that

$$y_x = 200 + 0.999(210 - 200) = 209.99$$

with all the separate values of y situated in such a way that in 2/3 cases they will be in comparatively narrow boundaries [209.99 – 2.25; 209.99 + 2.25].

Formula (18.2) furnishes the mean square error of determining y for *all* the cases included in the totality, but that error will generally differ from one array to another.

This circumstance depends on whether the standard deviations are identical in all the arrays or not. In the first case the totality is *homoscedastic* and otherwise *heteroscedastic* (Pearson 1905b, p. 22). The distribution can be of either type in spite of the nature of regression, although usually the former is accompanied by linear regression, and the latter, by curvilinear regression[18.4].

If the regression is linear and the standard deviations σ_{nx} are the same in all arrays, the mean square error for all of them will also be the same. Denoting by $\sum_i$ the addition of magnitudes belonging to array i, and by ${}_i\sum_y$ the mean square error in that array[18.5], we will indeed have

$$({}_i\sum{}_y)^2 = \frac{1}{n_{xi}}\sum{}_i[y-\rho_{y(x)}x]^2 = [...] = \frac{1}{n_{xi}}\sum{}_i(y-y_x)^2 = \sigma^2_{nxi}$$

since [cf. (14.1)], because the regression is linear,

$$y_x - \rho_{y(x)}x = 0.$$

The mean square error for all the totality will be equal to the same magnitude because the sum $\sum\limits_i$ is extended over magnitudes belonging to the i-th array and

$$N(\sum_y)^2 = n_{x1}\cdot{}_1(\sum_y)^2 + n_{x2}\cdot{}_2(\sum_y)^2 + ... + n_{xp}\cdot{}_p(\sum_y)^2.$$

But, as proved, all the mean square errors of separate arrays are identical and equal to σ_{nx}. Therefore

$$\sum_y = \sigma_{nx}$$

and on the strength of (18.2), homoscedastic distribution and linear regression the standard deviation of each array is equal to

$$\sigma_{nx} = \sigma_y\sqrt{1-r^2}.$$

If, in addition, the distribution does not much deviate from the normal law, we can also determine the probable error and then, for deriving an *individual* value of one magnitude given the correlatively connected other one, arrive at formula

$$y = \overline{y} + \rho_{y(x)}(x-\overline{x}) \pm 0.67449\sigma_y\sqrt{1-r^2}.$$

I provide an example of applying these formulas below.

19. The straight lines of regression

The regression equation can be presented as

$$\frac{Y}{\sigma_y} = r\frac{x}{\sigma_x}$$

where Y and x are deviations from mean values. When measuring these by their standard deviations,

$$\frac{Y}{\sigma_y} = \eta, \ \frac{x}{\sigma_x} = \xi,$$

the formula for the regression of y on x and of x on y becomes extremely simple:

$$\eta = r\xi, \ \xi = r\eta. \qquad (19.1, 2)$$

Geometrically, in each case r is the slope of the regression line with inclination measured from its own axis. When choosing the standard deviations as units of measure, these inclinations will be identical […]. This is a corollary of the remarkable expressions (19.1) and (19.2). Consider a numerical example. Let the correlation coefficient be 0.5, then the deviation of x from its mean by σ_x will lead to the deviation *in the mean* of y by $\sigma_y/2$ and vice versa. And if x deviates by $\sigma_x/4$, the mean deviation of y will be $\sigma_y/8$ etc.

If the correlation coefficient is zero, the regression lines coincide with their respective axes. The mean value of the deviation of a magnitude will be zero whatever be the deviation of the other magnitude. If that coefficient is 1, then, see formulas (19.1; 19.2), $\eta = \xi$, the deviations of both magnitudes measured in units of their standard deviations will be equal, the regression lines will coincide and have inclination 45° (as measured from either axis).

Returning now to the usual units of measurement of each magnitude, we will only encounter such a symmetric relation between the deviations and a symmetric arrangement of the regression lines if the standard deviations of both lines are equal to each other. This is what approximately occurs in the realm of heredity since the standard deviations of the magnitudes of an indication of parents and offspring only differ insignificantly.

Neglecting that nevertheless observed difference, we will obtain a very simple relation; for example, if the correlation coefficient is 0.5 (which is its typical mean value in heredity), the sons' mean deviation of the indication will be twice less than that of their fathers. Thus, a group of fathers whose stature is 20 *cm* higher than its mean value, will have sons only higher by 10 *cm* than the mean [of the entire population, as Slutsky added in his next similar example excluded from translation].

In general, selecting a group of fathers, we will observe their sons, whose stature deviates from the mean level in the same direction, but remains nearer to it as though *regressing* to that level. In the English literature, this phenomenon was initially called *regression*, a term that later acquired a more general meaning.

With unequal mean deviations, the regression coefficients also differ. Thus (Pearson & Lee 1903, pp. 370 and 378), the correlation coefficients of the stature of mother and daughter was 0.507, $\sigma_x = 2.39$ (mother), $\sigma_y = 2.61$ (daughter) inches so that daughters were more variable[19.1]. And, on the face of it, we have here a strange relation: daughters resemble mothers stronger than vice versa (Pearson 1896b, p. 276).
Indeed, the regression coefficient for the daughters and mothers was, respectively,

$$(2.61/2.39)\cdot 0.507 = 0.55; \ (2.39/2.61)\cdot 0.507 = 0.46.$$

Therefore, a group of mothers with stature higher by 10 *cm* than the mean of all mothers has daughters whose stature is 5.5 *cm* higher than the mean of all daughters, but the figures for the inverse case are 10 and only 4.6 *cm*. In short, daughters, in the mean, are closer to the mothers and farther from the general mean level than could be stated about the mothers of a certain group of daughters.

20. Calculating a correlation coefficient, an example

For facilitating the use of formulas, we illustrate the method of computing the correlation coefficient[20.1]. All the figures in our example are imagined in such a manner that the arithmetical operations are as simple as possible, and the number of groups is less than usual. [In this section, I am only translating the essence of Slutsky's detailed example in my own wording and leaving out minute explanations.]

For calculating the *raw* moments all magnitudes "as a first approximation" are thought to belong to their versions (§ 11), and we suppose that all versions of the same column or row are identical. If the total number of cases is large, the errors thus made will partly compensate one another. No correction is needed to the first moments and to the moments of the product $\sum xy$ whereas the second moments should be corrected either in accord with the Sheppard system, formulas (7.1), if the figures gradually diminish to zero (this is difficult to understand, see § 7), or by applying the method of trapezoids, formula (7.6).

The *second raw central moments* are calculated by formula (2.2). In this case, they are corrected by the Sheppard corrections. Thus the true central moments $\mu_{2(x)}$ and $\mu_{2(y)}$ are obtained. Extracting the square root we will have σ_x and σ_y. For calculating the *correlation coefficient* we determine the sum $\sum n_{xy}xy$, twice, in order to check the calculations, by formulas (17.1). Note that y_{xi} is the mean value of y in the i-th vertical column, and therefore equal to

$$\sum_i \frac{n_{xiy}y}{n_{xi}}, \text{ so that } n_{xi}y_{xi} = \sum_i n_{xiy}y.$$

Then the correlation coefficient and the coefficients of regression are calculated by formulas (17.4) and (16.8) respectively.

Formulas for the probable errors of the obtained magnitudes are in the next section. And, when wishing to have the final results in the *usual system,* the transition from the assumed units is more conveniently done before calculating the probable errors. To achieve this, we need to multiply $\bar{x}$ and σ_x, $\bar{y}$ and σ_y by the assumed units, k_x and k_y respectively, and we have to multiply $\rho_{(y)x}$ and $\rho_{(x)y}$ by k_y/k_x and k_x/k_y respectively.

It is possible to manage without calculating the regression coefficients in the assumed units at all, but derive them by issuing from the standard deviations after expressing these in the usual system. This, however, is not always possible because, when wishing to draw the graph of regression in the assumed scale, we ought to have the regression coefficients in the same scale as well.]

21. The general population and the random sample

Suppose we have a very large totality of cases each of which is characterised by a pair of magnitudes, x and y and that it is impossible to enumerate the whole totality called *general*. Our aim is to find out its main features by studying its random sample. The general totality is characterised by mean values $\bar{h}_1$ and $\bar{h}_2$, $\bar{\sigma}_1$ and $\bar{\sigma}_2$, regression coefficients $\bar{\rho}_{1(2)}$ and $\bar{\rho}_{2(1)}$, correlation coefficients $\bar{r}_{12}$ etc.

Since the composition of the sample is random, we certainly cannot expect that in each separate case constants of its distribution (h_1, h_2, σ_1, σ_2, $\rho_{1(2)}$, $\rho_{2(1)}$, r_{12} etc) coincide with the respective magnitudes of the general population. Only when choosing ever more numbers of samples more and more values will be provided for each constant and their means will approach the values which they have in the general population. For infinitely many random samples all the values of each constant, for r, say, will comprise a totality with mean $\overline{r}$ and deviations in each case equal to $\delta r = r - \overline{r}$. As a first approximation we may assume that the distribution of the magnitudes δr, $\delta\rho$, $\delta\sigma$, δh will obey the Gaussian law (Pearson & Filon 1898): lesser deviations will occur oftener than large ones, and their probabilities could be found from usual tables of the probability integral if only we know the standard deviation of the given magnitude (for example, of δr)[21.1].

If many random samples be indeed chosen out of one and the same general population, we could have empirically obtained the standard deviations for each of the errors δh, $\delta\sigma$, $\delta\rho$, and δr by means of calculations indicated in Part 1. This approach is however too difficult, and the theory of errors attempts to derive these magnitudes a priori, by various theoretical considerations. Such derivations are partly of a general nature, another part, being based on the assumption that the general population itself obeys the Gaussian law, is only approximately correct[21.2].

However, that premise but little depreciates the value of the results because the standard deviations and the probable errors of the studied magnitudes are usually very small, and a special precision of their determination does not play a large role in estimating the results. Once we know the theoretically derived standard deviation of some error, we are also able to calculate its probable error, for example

$$\mathrm{E}_r = 0.67449\sum{}_r,\ \mathrm{E}_\rho = 0.67449\sum{}_\rho.$$

One more circumstance demands attention. The theory indicates that in most cases the errors of separate constants are not independent but correlated. When selecting from our random samples those in which, for example, h_1 is larger than its mean, the mean of all the h_2′s for the same samples will not coincide with the mean $\overline{h}_2$ for all the samples; it will be greater or smaller depending on the sign of the correlation coefficient R_{h1h2}.

When actually having a large number of random samples we will be able to calculate that coefficient in the usual way:

$$M\sum{}_{h1}\sum{}_{h2}R_{h1h2} = \sum \delta h_1 \delta h_2.$$

Here, M is the number of samples, the sums in the left side are the standard deviations of h_1 and h_2 which vary from sample to sample because of random causes. The same formula is employed when theoretically deriving the correlation coefficient of errors, see example below.

22. Probable errors and coefficients of correlation between constants for the normal distribution

The derivation of probable errors is too complicated and we have to abandon it and only to provide the most important pertinent results[22.1]. For the sake of comprehensiveness I repeat some formulas from Part 1.

Probable errors

$$\mathrm{E}h = 0.67449\frac{\sigma}{\sqrt{N}},\ \mathrm{E}_\sigma = 0.67449\frac{\sigma}{\sqrt{2N}},\ \mathrm{E}_r = 0.67449\frac{1-r^2}{\sqrt{N}},\quad (22.1, 2, 3)$$

$$E_{\rho 12} = 0.67449\frac{\sigma_1}{\sigma_2}\sqrt{\frac{1-r^2}{N}},\ E_{\rho 21} = 0.67449\frac{\sigma_2}{\sigma_1}\sqrt{\frac{1-r^2}{N}}.$$

Correlation coefficients

$$R_{h1h2} = r_{12},\ R_{\sigma1\sigma2} = r_{12}^2,\ R_{\sigma1r12} = R_{\sigma2r12} = \frac{r_{12}}{\sqrt{2}},$$
$$R_{h1\sigma1} = R_{h2\sigma2} = R_{h1\sigma2} = R_{h2\sigma1} = R_{h1r12} = R_{h2r12} = 0. \qquad (22.4, 5, 6, 7)$$

Some remarks are here necessary. First of all, it is obvious that the probable error of each magnitude diminishes as the size of the population increases. In addition, the probable error of the coefficients of regression and correlation diminishes with the increase of *r*. Therefore, the stronger the correlation, the less can be the number of cases sufficient for determining with certainty the presence of a correlation connection and its magnitude. If $r = 0.9$ and $N = 25$, the probable error $E_r = 0.026$ and does not amount to 3% of the magnitude itself. For the correlation coefficient of 0.1, the same ratio will only be obtained with $N \approx 100{,}000$. And if, for determining a magnitude with certainty it is necessary that it exceeds by at least five times its probable error, it will not be difficult to find out that then, for $r = 0.1$, *N* will still be not less than 1000.

As noted above, the formulas provided are approximate. Student (1908) stated that formula (22.3) for the probable error of the correlation coefficient may be already applied when $N = 30$. For lesser groups it should not be relied upon, and we have to apply another method of calculation. It is complicated and I do not describe it, readers can look up Student. As a rule of thumb, it can be assumed that to have some certainty about the very existence of correlative connection for $20 < N < 30$ the correlation coefficient should not be less than 0.5.

Student concluded that in the absence of correlation in the general population and $N = 21$ the correlation coefficient can by chance only twice in a hundred random samples take a value exceeding $|0.5|$[22.2]. After all, we ought to recognize that, regrettably, until the theory be further developed (or at least until tables based on Student's formulas be compiled) statisticians should not apply the correlation method to groups consisting less than of 20 cases.

We turn now to the formulas for the coefficients of correlation between the constants of distribution. To provide an idea of their importance, we will touch on their relation to the theories of heredity and selection. Assume that the distribution of the indications of individuals of a certain biological species obey the normal law (for many, if not for all indications and species this is not far from the truth), then formulas (22.4) – (22.7) will at once ensure a number of important conclusions. Let subscripts 1 and 2 denote the size of organs and r_{12} be the correlation coefficient between them. We can easily determine that correlation by measuring a few hundred individuals. How will selection act (for example, natural selection, when an individual with the organ of a size conforming to new and different conditions of life has more chances of surviving) if directed towards changing the mean size of one of the organs?

Formulas (22.7) show that the absolute variability of a given indication in the species does not change (because $R_{h1\sigma1} = 0$) and neither does the variability of other indications ($R_{h1\sigma2} = 0$) or the correlation coefficients between the given and the other indications. However, the mean size of other organs will have to change and, knowing r_{12}, r_{13}, …, which can always be determined, we can say beforehand by how much.

It will be different if the selection is directed to the magnitude of the standard deviation; for example, when under changed conditions the previous most favourable mean size of some organ persists, but deviations from it become more harmful. The mean size of the organ will not change ($R_{\sigma 1h1} = 0$), the same applies to other organs ($R_{\sigma 1h2} = 0$), but their standard deviations will have to change ($R_{\sigma 1\sigma 2} = r_{12}^2$) as well as the correlation coefficient between them ($R_{\sigma 1r12} = r_{12}/\sqrt{2}$).

And because r^2 is a comparatively small magnitude and rapidly decreases with r, it is obvious that, first, the influence of the selection of the standard deviation of one organ on that of another one is less than in the case of the selection of mean sizes. Second, that this influence can only become somewhat noticeable for organs with a comparatively high correlation connection. Yet it ought to be remarked that, however weak is that influence in some cases, the correlation coefficient ($R_{\sigma i\sigma j} = r^2$) is always positive so that an increase in the variability of one organ is always connected with an increase, and a decrease, with a decrease in the variability of all the other organs.

For example, if random circumstances (or artificial selection) isolate a group whose members in a certain way resemble each other more strongly, they will more resemble each other in every other way (Pearson & Filon 1898, p. 241 note).

Although we did not dwell on more complicated theoretical considerations, we have far from exhausted all even most direct possible conclusions from the formulas above. But even that seems to be enough for the reader to feel to what kind of important problems in this field does the correlation theory lead us.

23. The probable error of the difference

Knowledge of the correlation coefficients between the constants of distribution enables us to derive further formulas concerning probable errors. First of all, we illustrate the general principle here by an important case of the *probable error of the difference*.

Let $z_0 = x_0 - y_0$ be the difference between two constants of a general population. Their values in some random sample are *z, x, y* differing from their true values by δz, δx, and δy. Obviously,

$$\delta z = \delta x - \delta y,\ \sum(\delta z)^2 = \sum(\delta x)^2 - 2\sum\delta x\delta y + \sum(\delta y)^2$$

where M random samples are considered. These sums are known and can be expressed by the standard deviations and correlation coefficients

$$M\sigma_z^2 = M\sigma_x^2 - 2M\sigma_x\sigma_y r_{xy} + M\sigma_y^2,\ \sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y r_{xy}}. \qquad (23.1a, b)$$

For independent x and y, that is, for $r_{xy} = 0$,

$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}. \qquad (23.2)$$

When multiplying both parts of equalities (23.1) and (23.2) by 0.67449, we change each standard deviation into the respective probable error. The formulas are therefore also valid for these errors. Note that for r_{xy} greater/less than zero, the probable error as provided by formula (23.1) will be less/greater than that given by formula (23.2). Therefore, statisticians who apply formula (23.2) in case of dependent magnitudes run the risk of failing to recognize a sufficient difference, or, even worse, of admitting as significant an inessential difference in cases of positive and negative correlation respectively.

In particular, the probable error of the difference between arithmetic means and standard deviations, see formulas (22.4) and (22.5), will be

$$E_{h_1-h_2} = \sqrt{E^2_{h_1} + E^2_{h_2} - 2E_{h_1}E_{h_2}r_{12}},$$

$$E_{\sigma_1-\sigma_2} = \sqrt{E^2_{\sigma_1} + E^2_{\sigma_2} - 2E_{\sigma_1}E_{\sigma_2}r^2_{12}}. \qquad (23.3, 4)$$

.

For finding out whether *the difference between two correlation coefficients* is significant, we should calculate a particular form of expression (23.1)

$$E_{r_a-r_b} = \sqrt{E^2_{r_a} + E^2_{r_b} - 2E_{r_a}E_{r_b}R_{r_a r_b}} \; .$$

E_{r_a} and E_{r_b} are calculated by formula (22.3) and $R_{r_a r_b}$, the correlation coefficient between correlation coefficients, by formulas (Pearson & Filon 1898, pp. 259, 262)

$$R_{r_{12}r_{13}} = r_{23} - (1/2)r_{12}r_{13}\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{(1-r_{12}^2)(1-r_{13}^2)}, \qquad (23.5)$$

$$2(1-r_{12}^2)(1-r_{34}^2)R_{r_{12}r_{34}} = (r_{13} - r_{12}r_{23})(r_{24} - r_{23}r_{34}) +$$
$$(r_{14} - r_{13}r_{34})(r_{23} - r_{12}r_{13}) + (r_{13} - r_{14}r_{34})(r_{24} - r_{12}r_{14}) + (r_{14} - r_{12}r_{24})(r_{23} - r_{24}r_{34}). \qquad (23.6)$$

The first of these, as the subscripts show, is applied when examining the difference between the correlation of a magnitude with two others and the second one, the difference between the correlation coefficients of two different pairs of magnitudes. Expression (23.6) becomes essentially simpler if some correlation coefficients vanish. The former formula (23.5) can be applied without great difficulties since the entering expressions should be calculated for other purposes [as well], when studying correlation between three magnitudes, see below.

Example. In § 3, data concerning the price of rye in three commercial centres were quoted. We can now estimate the differences between the arithmetic means and the standard deviations. We will not calculate the probable errors of the former: it is obvious that they are essential. [Slutsky calculated r_{12}, r_{13} and r_{23} and the probable errors of the latter differences by formula (23.4). His conclusion: a [real] difference between centres 1 and 3 may be assumed probable, and believed certain between centres 2 and 3, cf. § 14.][23.1]

24. Probable errors in case of a normal distribution

If the distribution does not obey the Gaussian law, the formula for probable errors in § 22 can only be considered as approximate with error depending on the closeness of the distribution to the normal law. In particular, the expression for the probable error of the arithmetic mean (22.1) persists for any distribution. The general expressions for the probable error of the standard deviation are

$$E_\sigma = 0.67449\sqrt{\frac{\mu_4 - \mu_2^2}{4N\mu_2}} = 0.67449\frac{\sigma\sqrt{1+\eta/2}}{\sqrt{2N}} \qquad (24.1, 2)$$

where η is the coefficient of dispersion (9.7). For a small η

$$\sqrt{1+\eta/2} \approx 1+\eta/4$$

and it is easy to conclude that, with $\eta > 0.2$ or < -0.2 (that is, with $\beta_2 > 3.2$ or < 2.8) and an error of E_σ not exceeding 5% the usual formula (22.2) should be replaced by (24.2). However, *if* the probable error is small as compared with the standard deviation, less precision will in most instances be also sufficient. Therefore, even in those comparatively rare cases in which $\beta_2 = 4$ or 2, the error ensuing when the formula (22.2) is applied will only amount to 19 and 29% respectively of the true probable error (Pearl 1908, p. 117).

The correlation between the arithmetic mean and standard deviation will not be equal to zero since in the general case**24.1**

$$\sum{}_h\sum{}_\sigma R_{h\sigma} = \frac{\mu_3}{2N\sigma}. \qquad (24.3)$$

The sign of that expression depends on the sign of the third moment. If that is positive, the increase in σ is connected with an increase in *h* which in turn is connected with the increase in σ, and in the opposite case the dependence is inverse. As an example of applying that formula let us determine the probable error of the coefficient of variation

$$V = 100\frac{\sigma}{h}$$

[cf. formula (3.2)]. Taking logarithms and […] we find that

$$\frac{\delta V}{V} = \frac{\delta\sigma}{\sigma} - \frac{\delta h}{h}.$$

Adding up the squares of this expression for all the random samples and dividing by their number we get

$$[\frac{\delta V}{V}]^2 = [\frac{\delta\sigma}{\sigma}]^2 + [\frac{\delta h}{h}]^2 - 2\frac{\delta\sigma\delta h}{\sigma h},$$
$$\frac{1}{V^2}(\sum\delta V)^2 = \frac{1}{\sigma^2}(\sum\delta\sigma)^2 + \frac{1}{h^2}(\sum\delta h)^2 - \frac{2}{\sigma h}\sum\delta\sigma\delta h \qquad (24.4)$$

and therefore

$$\frac{1}{V^2}(\sum{}_V)^2 = \frac{1}{\sigma^2}(\sum{}_\sigma)^2 + \frac{1}{h^2}(\sum{}_h)^2 - \frac{2}{\sigma h}\sum{}_\sigma\sum{}_h R_{\sigma h}.$$

For the normal distribution, see formulas (22.1, 22.2, 22.7),

$$R_{\sigma h} = 0,\ \sum{}_h = \frac{\sigma}{\sqrt{N}},\ \sum{}_\sigma = \frac{\sigma}{\sqrt{2N}}$$

and we arrive at formula (4.3) for E_V.

For a non-normal distribution $R_{\sigma h} \neq 0$ and we ought to calculate the entire expression above. Issuing from formulas (22.1), (24.2) and (24.3) and performing simple transformations, we easily get

$$\mathrm{E}_V = 0.67449\frac{V}{\sqrt{2N}}\sqrt{1 + 2[\frac{V}{100}]^2 + [\frac{\eta}{2} - 2\frac{\mu_3}{h\sigma^2}]}.$$

If the distribution is not very distinct from the normal law, the last term under the square root is in all cases usually small so that the simpler formula (4.3) can be widely applied.

We still have to derive the rather complicated expression for the probable error of the correlation coefficient for non-normal distributions. Denote

$$p_{qs} = \sum \frac{n_{xy}(x-\bar{x})^q(y-\bar{y})^s}{N}$$

whose particular cases

$$p_{20} = \sigma_x^2,\ p_{02} = \sigma_y^2,\ p_{11} = \sigma_x\sigma_y r_{xy}$$

we have met above.

The most general expression for the probable error of the correlation coefficient valid for *any* distribution (Sheppard 1898) in a somewhat simplified form (Pearson 1907, p. 25)[24.2] is

$$\mathrm{E}_r = 0.67449\frac{r}{\sqrt{N}}\sqrt{\frac{p_{22}}{p_{11}^2} + \frac{p_{22}}{2p_{20}p_{02}} + \frac{p_{40}}{4p_{20}^2} + \frac{p_{04}}{4p_{02}^2} - \frac{p_{31}}{p_{11}p_{20}} - \frac{p_{13}}{p_{11}p_{02}}}. \qquad (24.5)$$

The calculation here is difficult, and that formula is not usually applied. The experience gained when employing it happily ascertained that even in cases in which the distribution far deviated from the normal law it provides results sufficiently close to those obtained by the usual formula (22.3). This latter can therefore be assumed sufficiently reliable for application in all useful cases[24.3].

Example. [Slutsky considers a simple imaginary example with distribution "absolutely dissimilar" to the normal law. He calculates

σ_x, σ_y, r_{xy}, r^2, $1 - r^2$, $\mathrm{E}r'$ (formula (22.3)) and $\mathrm{E}r$ (formula (24.5)),

notes that $\mathrm{E}r'/\mathrm{E}r = 1.34$ and concludes that this is "not much" considering that the distribution involved was much farther from the normal law than in most cases encountered in practice.]

25. The difference method of determining the correlation coefficient

Formula (23.1) leads to

$$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2\sigma_x\sigma_y r_{xy}.$$

In a somewhat changed form this expression will be convenient and in many cases can essentially shorten the determination of the correlation coefficient.

Let *x* and *y* be the values of the magnitudes entered in the correlation table and measured, each from some assumed zero which generally does not coincide with the corresponding centre of distribution. Then

$$\sum (x-y)^2 = \sum x^2 + \sum y^2 - 2\sum xy,\ \sum x^2 = N\nu'_{2(x)},\ \sum y^2 = N\nu'_{2(y)}. \qquad (25.1)$$

where ν' as always stands for raw non-central moments. Then,

$$\sum xy = \sum (x-\overline{x})(y-\overline{y}) + N\overline{xy} = N\sigma_x\sigma_y r_{xy} + N\overline{xy}$$

follows from formula (17.3). Inserting these magnitudes in formula (25.1) we get

$$\sum (x-y)^2 = N\nu'_{2(x)} + N\nu'_{2(y)} - 2N\overline{xy} - 2N\sigma_x\sigma_y r_{xy},$$
$$r_{xy} = \frac{\nu'_{2(x)} + \nu'_{2(y)} - 2\overline{xy} - (1/N)\sum (x-y)^2}{2\sigma_x\sigma_y}. \qquad (25.2)$$

From the magnitudes entering here, $\overline{x}, \overline{y},$ σ_x and σ_y should be determined, even if we do not wish to calculate the correlation coefficient, since they are the main constants of a statistical group; $\nu'_{2(x)}$ and $\nu'_{2(y)}$ are obtained as supplementary magnitudes along with σ_x and σ_y. To determine the correlation coefficient it is only needed to calculate the last term in the numerator which in most cases will be easier than finding the sum of products.

It is easiest to show the course of this work by a numerical example [the imaginary example in § 20 excluded from the translation. Slutsky finally stated:] Since the expression (25.2) was derived from a number of identities, the magnitude determined by applying it should also be always identical with that calculated by the method of products, and the formula for the probable error therefore persists[25.1].

26. Curvilinear regression

The straight regression line can only be applied as quite a suitable theoretical model of a phenomenon until the deviations of the empirical regression line are so insignificant as to be thought possibly random. Although great many phenomena can quite satisfactorily be represented by linear formulas, cases of curvilinear regression are not seldom. Then, if the researcher does not wish to restrict his considerations by the empirical regression line, i. e., by simply establishing the actual state of his data, but attempts to reveal the main, the *non-random* features of the studied dependence, he will have to treat his material further.

The first method, the oldest and crudest, consists in drawing by hand a smooth curve as closely as possible adjoining the isolated empirical points. A s*mooth curve* means that its curvature ought to change as gradually as possible with least possible points of inflexion. Elementary as it is, this method is being applied even now, see for example Pearson (1902b), and in many cases it can provide sufficiently fair results, especially with a small number of cases in the totality, so that more perfect methods of deriving the curve do not ensure precise results either[26.1].

In addition, a higher precision is often not demanded, and, consequently, applying such methods accompanied by much calculations would have only been a waste of time. The second method consists in fitting theoretical curves to separate parts of the empirical line of regression. The benefit here is that the suitable curves are simpler, and is especially felt when the regression is complicated and parabolic curves of the second and third degree provide a poor fit.

It is hardly advisable to apply parabolas higher than the third, or, in extreme cases, the fourth degree because the calculation of their coefficients is accompanied with determining moments of the higher orders with large probable errors and in addition demands much time which just the same can possibly be pointless. Better, after sensibly discussing the case, to separate the correlation table into parts and apply simple curves for each of these, see for example Powys (1901, p. 49), also Powys (1905).

Example. [Slutsky considers the cost of public education for a local administration in Russia compared with its receipts from industry and commerce. He concluded that for a group of localities with receipts from industry amounting to 15 – 55% of their budget an increase of 10% in that part is connected by an increase in the expenditure for education of 3.5%, cf. one of his examples in § 14.]

27. Calculating the coefficients of the regression curve

The simplest method consists in following the rules of § 17 and, more precisely, the **C** version for calculating the moments, and then in determining the coefficients of the parabolic curve, see formulas of § 18.

As compared with § 17, our case is somewhat peculiar. [In § 17 Slutsky considered a broken empirical line with zero extreme ordinates. As previously, he assumed that the adjacent empirical points are separated one from another by distances $\Delta x = 1$. The extreme empirical points, A $(-l; y_0)$ and B $(l; y_m)$, are situated at distances $|\Delta x| = l$ from the origin; Segments AR and BS are drawn with points R and S being on Ox and A and R, B and S separated by $\Delta x = 1$. Slutsky considers the new area "under" the extended empirical broken line in the same way as previously, then allows for the moments of the two fictitious triangles (Pearson 1902c, pp. 7 – 9). He states:] Our formulas will become simpler of we introduce some new notation.

Previously [in § 2] we called expressions of the type

$$v'_p = \frac{1}{N}\sum n_x x^p$$

raw moments with n_x being the size of the subgroup and x, the distance of the *middle* of the appropriate interval from the origin. Now, as in § 18, we have to do not with size, but with areas and determine their moments, so that instead of n_x we must insert the ordinate y multiplied by the length of the interval (assumed to be unity), and the raw moment of the area (S) will be

$$v'_p = \frac{1}{S}\sum yx^p.$$

We may call this the *relative moment* as compared with the *absolute moment* equal to the same sum not divided by S. Denoting the absolute raw and real moments by Ψ'_p and ω'_p we have[27.1]

$$\psi'_p = S\nu'_p = \sum yx^p,\ \omega'_p = S\mu'_p.$$

Returning to our problem, we denote the absolute moments by ψ and ω, and by Ψ and Ω the absolute moments of the old and new (larger) area. The raw moments are obviously the same for both areas:

$$\Psi'_0 = \psi'_0 = \sum y_x,\ \Psi'_1 = \psi'_1 = \sum y_x x,\ ...,\Psi'_p = \psi'_p = \sum y_x x^p.$$

Then we determine the true moments of the new area by formula (7.6) which, as it is easy to perceive, is also valid for that goal**27.2**:

$$\Omega'_1 = \psi'_1,\ \Omega'_2 = \psi'_2 + 1/6,\ \Omega'_3 = \psi'_3 + (1/2)\psi'_1,\ \Omega'_4 = \psi'_4 + \psi'_2 + 1/15,$$
$$\Omega'_5 = \psi'_5 + (5/3)\psi'_3 + (1/3)\psi'_3,\ \Omega'_6 = \psi'_6 + (5/2)\psi'_4 + \psi'_2 + 1/28 \text{ etc.}$$

For determining the true absolute moments of the previous area we now ought to subtract the moments of the areas of two additional triangles. According to Pearson's calculations (1902c, p. 8), we have, for odd and even moments respectively,

$$\omega'_n = \Omega'_n - L_n(y_m - y_0), \qquad (27.1)$$

$$L_n = \frac{(l+1)^{n+2} - (n+2)l^{n+1} - l^{n+2}}{(n+1)(n+2)}.$$

The previous area is

$$S = \sum y - (1/2)(y_m + y_0)$$

[notation as at the beginning of § 27] and the relative true moments will be derived by dividing the absolute moments (27.1) by *S*:

$$\mu'_n = \frac{\omega'_n}{S}.$$

Then, as previously in § 18, we calculate the auxiliary magnitudes

$$y_0 = \frac{S}{2l},\ \lambda_n = \frac{\mu'_n}{l^n}$$

and easily determine the coefficients of a parabolic curve of regression, see § 18. For facilitating this work I adduce a table of L_n for $n = 1, 2, \ldots, 5$ and $l = 3, 4, \ldots, 20$ [with three significant digits after the decimal point; excluded from translation] (Pearson 1902c p. 9).

28. Calculating the coefficients of the regression curve (continued)

The method described in § 27 is comparatively simple and can therefore be applied in many cases which is what for example Powys (1905, p. 236) did. However, as far as our problem is concerned, it has a serious shortcoming in that all the ordinates equally influence the result whereas some of them are more, the other ones less reliable. For this reason the described method is not applied when the equation of a regression straight line is derived. Instead, the formulas based on the principles [on the principle] of least squares is made use of (Pearson 1905b). Pearson developed that other method (more precisely, the method of moments) also for being applied to curvilinear correlation. It is, however, complicated, and therefore cannot be here described.

Nevertheless, we can approach the matter simpler. Indeed, abandoning the method of moments and applying the traditional method of least squares, we will obtain quite a rational and comparatively simple solution at that. The only problem consists in assigning proper weights to the separate ordinates of the empirical regression line. Let, as previously, n_{xi} be the size of the i-th array of x's; y_{xi}, the arithmetic mean of y's (of ordinates of the regression line) in that array; the standard deviation, again for that array, σ_{nxi}, then the probable error of y_{xi} will be

$$\frac{0.67449\sigma_{nxi}}{\sqrt{n_{xi}}}.$$

The precision with which we know the separate ordinates is the higher, the less is that probable error, and the greater the denominator of that fraction. We will therefore assume that the weights of those ordinates (p_i) are proportional to the square root of n_{xi} and inversely proportional to the standard deviations of the arrays. Consequently, we suppose the weights of the *squares* of the differences proportional to the squares of the previous magnitudes**28.1**:

$$p_i = \frac{n_{xi}}{\sigma_{nxi}^2}.$$

If the standard deviations in different arrays are equal, or almost so, the weights will be simply proportional to the sizes [the frequencies] (as Pearson assumed them to be):

$$p_i = n_{xi}.$$

Then the generalized principle of least squares

$$\sum p_i(y_{xi} - a_0 - a_1x_i - a_2x_i^2 - ... - a_mx_i^m)^2 = \text{ min}$$

ought to be fulfilled. [Slutsky derives the appropriate normal equations in (m + 1) unknown coefficients of the parabola sought**28.2**

$$Y = a_0 + a_1x + a_2x^2 + ... + a_mx^m.]$$

In any case, this method should provide results not worse than those obtained by the Pearson method, and even better results for heteroscedastic distributions.

29. Correlation ratio

Curvilinear correlation demands its own special measure of correlation dependence. As shown in § 18, the correlation coefficient can only be equal to 1 when correlation is perfect and regression, strictly linear. When only the first condition is satisfied, the coefficient will still be less than unity. Then, the vanishing of the correlation coefficient can only testify to the absence of correlation under linearity. Indeed, since [cf. (15.1)]

$$r = \sqrt{\rho_x\rho_y},$$

$r = 0$ if one of the factors is zero, i. e., when one of the regression lines is horizontal. This, however, is also possible with correlation (and even perfect correlation identical with strict functional dependence) being present. Let for example the dependence between two magnitudes be represented by a parabola or any other symmetrical curve with a vertical axis and two branches, one of them ascending, the other one descending. A straight line closest to the points of such a curve will be horizontal, and the correlation coefficient vanishes.

The following considerations enable to establish a measure of correlation also for curvilinear regression (Pearson 1905b, pp. 9 – 11). If there is no correlation dependence between two indications, groups corresponding to one of them should show that the distribution for the second one is identical to that of the general totality. Neglecting random deviations, i. e., assuming that the totality is very large so that all the probable errors are sufficiently small and can be neglected, we arrive at

Proposition I. *In the absence of correlation the arithmetic mean of indications in each array is equal to that for all the totality*

$$y_x = \overline{y}$$

and the standard deviation calculated separately for each array is equal to that for all the totality

$$\sigma_{nx} = \sigma_y.$$

Consider the case in which the correlation becomes perfect, that is, transforms into strict functional dependence. Then one definite value of the second variable y corresponds to a definite value of the variable x. The deviations of y from its single value are equal to zero (we assume that the intervals are infinitely narrow) and therefore the standard deviation of y in each array of x's also vanishes. And so we have

Proposition II. *In cases of perfect correlation, i. e. when the correlation dependence converts into strict functional dependence, all the standard deviations of separate arrays vanish*

$$\sigma_{nx} = 0.$$

Let σ_{nx} be a particular standard deviation, and σ_y the general standard deviation [it appears somewhat below], then

$$\sigma_a = \sqrt{\frac{\sum n_x \sigma_{nx}^2}{N}} \qquad (29.1)$$

can be called the *mean square particular standard deviation*. It is equal to the product […].

Consider now the ratio σ_a^2 / σ_y^2. In the absence of correlation it becomes equal to unity since each $\sigma_{nx} = \sigma_y$ and consequently

$$\sigma_a^2 = \frac{1}{N}\sum n_x \sigma_{nx}^2 = \sigma_{nx}^2 \frac{\sum n_x}{N} = \sigma_y^2. \qquad (29.2)$$

When correlation is perfect, each $\sigma_{yx} = 0$, so that $\sigma_a = 0$ and that ratio also vanishes. It is, however, more convenient to assume a measure of correlation increasing with the degree of the latter and Pearson proposed as a measure not the ratio itself, but

$$\eta^2 = 1 - \frac{\sigma_a^2}{\sigma_y^2}$$

and called it the *correlation ratio*. We may now assume the following proposition proven:

Proposition III. *In the absence of correlation the correlation ratio vanishes and it equals unity when correlation is perfect*[29.1].

The correlation ratio can be represented in another and very convenient form. However, before transforming it, we ought to prove a supplementary proposition. Call *the difference* ($y_x - \bar{y}$) *between the arithmetic means of a separate array and of the whole totality the deviation of the regression line from the central axis of the distribution*[29.2]. Multiply its square by the size of the appropriate array, add up such products and divide the new product by the size of all the totality. Then the magnitude

$$\sigma_m = \sqrt{\frac{n_x(y_x - \bar{y})^2}{N}} \tag{29.3}$$

can be called the *mean square deviation of the regression line from the central axis*. We will now prove

Proposition IV:

$$\sigma_y^2 = \sigma_a^2 + \sigma_m^2.$$

We know that

$$\sigma_y^2 = \frac{1}{N}\sum n_x(y - \bar{y})^2 = \frac{1}{N}\sum (y - \bar{y})^2$$

(instead of multiplying the difference by n_x it is possible to repeat it n_x times). Obviously,

$$\sigma_y^2 = \frac{1}{N}[\sum_1 (y - \bar{y})^2 + \sum_2 (y - \bar{y})^2 + \ldots + \sum_p (y - \bar{y})^2] \tag{29.4}$$

Each of these sums, for example the i-th, can be represented as

$$\sum_i (y - \bar{y})^2 = \sum_i [(y - y_x) + (y_x - y)]^2 =$$

$$\sum_i (y - y_x)^2 + \sum_i (y - \bar{y})^2 + 2\sum_i (y - y_x)(y_x - \bar{y}) \tag{29.5}$$

where y_x is the arithmetic mean of the considered array and $(y - y_x)^2$ is the square of the deviation of a separate magnitude from the mean of that array, so that

$$\sum_i (y_x - \overline{y})^2 = n_x \sigma_{nx}^2.$$

For a given array ($y_x - \overline{y}$) is constant; it is represented in the sum a number of times equal to the number of individuals in that array. Consequently,

$$\sum_i (y_x - \overline{y})^2 = n_x (y_x - \overline{y})^2$$

and the last sum in (29.5) vanishes since the difference ($y_x - \overline{y}$) does not depend on *i*. Therefore, (29.5) becomes

$$\sum_i (y - \overline{y})^2 = n_x \sigma_{nx}^2 + n_x (y - \overline{y})^2.$$

Inserting this in (29.4), that is, determining the sum of such expressions for all the arrays, we get, because of definitions (29.1) and (29.3),

$$\sigma_y^2 = \frac{1}{N}[\sum n_x \sigma_{nx}^2 + \sum n_x (y_x - \overline{y})^2] = \frac{1}{N}[N\sigma_a^2 + N\sigma_m^2]$$

so that Proposition IV is proved and

$$\frac{\sigma_m^2}{\sigma_y^2} = 1 - \frac{\sigma_a^2}{\sigma_y^2} = \eta^2,\ \eta = \frac{\sigma_m}{\sigma_y}. \qquad (29.6)$$

This expression is more convenient for calculating the correlation ratio and it also ensures the possibility to prove a theorem converse of Proposition III:

Proposition V. *If* $\eta = 0$, *there is no correlation; if* $\eta = 1$, *the correlation is perfect.*

Indeed, [see formulas (29.1) and (29.6)],

$$\eta^2 = \frac{(1/N)\sum n_x (y_x - \overline{y})^2}{\sigma_y^2}.$$

The numerator of this fraction is a sum of positive numbers, so η can only vanish when each term of that sum is zero, i. e., when there is no correlation. Then, according to definition (29.6), when $\eta = 1$, σ_a^2 ought to vanish, but then, see formula (29.1), this is only possible when all $\sigma_{nx} = 0$, that is, when correlation is perfect.

Proposition IV means that σ_y^2 is a sum of two essentially positive numbers, therefore $\sigma_m = \sigma_y$ only if $\sigma_a = 0$ which is the case of perfect correlation, otherwise $\sigma_m^2 < \sigma_y^2$ and η is always less than unity and never exceeds it.

As to the probable error of the correlation ratio, Pearson (1905b, pp. 11 – 19) had also derived it. Its complete expression is too complicated, but the approximate formula

$$\mathrm{E}_\eta = 0.67449\frac{1-\eta^2}{\sqrt{N}}$$

is sufficiently precise.

Example. [Slutsky considers the dependence between mean monthly prices of rye and cast iron in Germany during 1879 – 1900 providing its source in a special section (*Tables*). He concludes that the "correlative dependence" between them "many times exceeds its probable error and is not therefore random" and that the dependence of the price of cast iron on that of rye is more than twice higher [stronger] than the inverse dependence, but reasonably does not attempt to explain this fact.]

30. Dependence between the correlation ratio η and the correlation coefficient *r*

We saw that the straight regression line is a straight line which most closely adjoining the empirical regression line. Its equation [see formula (17.2)]

$$Y = \overline{y} + \frac{\sigma_y}{\sigma_x} r(x - \overline{x}),\ r = \frac{\sum n_{xy}(x-\overline{x})(y-\overline{y})}{N\sigma_x\sigma_y} \qquad (30.1)$$

or, as was proved,

$$= \frac{\sum n_x(y_x - \overline{y})(x-\overline{x})}{N\sigma_x\sigma_y}.$$

Multiplying both parts by $N\sigma_y^2 r$ we have

$$N\sigma_y^2 r^2 = \frac{\sigma_y}{\sigma_x} r \sum n_x(y_x - \overline{y})(x - \overline{x}). \qquad (30.2)$$

Applying formula (29.6) we get

$$N\sigma_y^2\eta^2 = \sum n_x(y_x - \overline{y})^2$$

and, subtracting (30.2),

$$N\sigma_y^2(\eta^2 - r^2) = \sum n_x(y_x - \overline{y})^2 - \frac{\sigma_y}{\sigma_x} r \sum n_x(y_x - \overline{y})(x - \overline{x}) =$$

$$\sum \{n_x(y_x - \overline{y})[y_x - \overline{y} - \frac{\sigma_y}{\sigma_x} r(x - \overline{x})\} = \sum [n_x(y_x - \overline{y})(y_x - Y)]. \qquad (30.3)$$

Here, see formula (30.1), *Y* is the ordinate of the straight regression line. When replacing in (30.3) ($y_x - \overline{y}$) by an identical expression ($y_x - Y$) + ($Y - \overline{y}$) we arrive at

$$N\sigma_y^2(\eta^2 - r^2) = \sum \{n_x[(y_x - Y) + Y - \overline{y}](y_x - Y)\} =$$
$$\sum n_x(y_x - Y)^2 + \sum n_x(Y - \overline{y})(y_x - Y).$$

Inserting *Y* from formula (30.1) into the last sum, we have

$$N\sigma_y^2(\eta^2 - r^2) = \sum n_x(y_x - Y)^2. \tag{30.4}$$

This expression is extremely important. It shows that $\sigma_y^2(\eta^2 - r^2)$ is the mean square deviation of the regression line from the straight line most closely adjoining it, and it also allows us to formulate some conclusions about the magnitude of the correlation ratio. Since its right side only consists of positive magnitudes, the inequality $\eta^2 > r^2$ must always hold, and since we ought to consider the correlation ratio positive, it follows that $\eta > |r|$, i. e., that *the correlation ratio is always greater than the absolute value of the correlation coefficient. These magnitudes can only be equal*, see formula (30.4), *when each $y_x = Y$, which means, when the regression is strictly linear.*

All the previous considerations were based on the assumption that our totality is so large, that probable errors can be neglected. Because of random deviations inherent in our invariably limited sources, we will never encounter an absolutely linear regression. After gaining some experience, we can certainly feel whether a regression is sufficiently linear, but when in doubt it is important to know the probable error of linearity $(\eta^2 - r^2)$ denoting it

$$\varsigma = \eta^2 - r^2. \tag{30.5}$$

Blakeman (1905, pp. 337 and 339) derived the precise value of that probable error but it is admissible to apply his approximate formula

$$E\varsigma = 0.67449 \cdot 2\sqrt{\frac{\varsigma}{N}}. \tag{30.6}$$

When higher precision is needed, we may apply, along with (30.5), the difference

$$\theta = \eta - |r|.$$

For better approximation Blakeman provided the formulas

$$\frac{\varsigma}{E\varsigma} = \frac{\sqrt{\varsigma N}}{2 \cdot 0.67449\sqrt{1 + (1-\eta^2)^2 - (1-r^2)^2}}, \tag{30.7}$$

$$\frac{\theta}{E\theta} = \frac{\sqrt{4\theta\eta N\,|\,r\,|}}{2 \cdot 0.67449(\eta + |\,r\,|)}[1 + \frac{|\,r\,|(1-\eta^2)^2 - \eta(1-r^2)^2}{\eta + |\,r\,|}]^{-1/2}. \tag{30.8}$$

Usually we may apply formula (30.6) or its corollary, formula (30.7), without the square root in its denominator. If it does not exceed 2 or 2.5, regression can be considered linear; and in doubtful cases, we can turn to formulas (30.7) and (30.8) and be satisfied if the results provided are close to each other. Or, if the discrepancy is considerable, we have to apply the precise formulas for the probable errors ς and θ, see Blakeman (above). They are complicated, and I do not reprint them.

31. Correlation and causal dependence

The issue, with which we intend to conclude this Chapter, certainly deserves much more attention than we are able to spare, and we are compelled to restrict our considerations to a few remarks.

Obviously, neither correlation, nor strictly functional dependence are identical with causal connections between phenomena. This at least follows from the fact that both can take place in such spheres where discussing cause and effect would have been meaningless; we bear in mind the realm of ideal geometric and similar constructions.

The logical nature of these concepts essentially differs, which we may note even without touching on the theory of knowledge. First of all, any functional dependence, whether strict or not[31.1] (correlational) is mutual in the sense that we arbitrarily decide which variable should be considered independent, and which one dependent. However, concerning cause and effect, we are tied to the actual state of things and must submit to results of [the appropriate] study. Then, even when assessing one and the same material, the essence of judgement about the presence of a functional connection and of a cause is different.

Consider for example the dependence between the mass (*m*), volume (*v*), temperature [in centigrades] (*t*) and pressure (*p*) of a gas[31.2]

$$\frac{pv}{t+273} = cm$$

where *c* is some constant. This equation ties four magnitudes together, and each can be determined given the other ones which is its main meaning. If two magnitudes remain constant, and a third varies, the fourth will also vary in a known way. All four magnitudes are functionally tied up, and it is pointless to ask which is the cause, and which is the effect.

The same question, however, becomes quite legitimate when concerned not with dependence of magnitudes *in general*, but with concrete physical processes of change. Thus, suppose that *v* and *m* are constant; then, given *p*, we can find *t* and vice versa. From a purely mathematical viewpoint each of these magnitudes can be considered as an independent variable, but they play essentially differing physical parts. Under the same assumption we can only change the pressure by changing the state of temperature which will be the cause, and the former will always only be the effect. Compressing the gas is a more complicated phenomenon since then both its resiliency and temperature will generally have to change.

Considering for the sake of simplicity the change of the former under constant temperature and only from the point of view of a purely functional mutual relations of magnitudes, we will determine that it is a very simple function of the volume. The same change of resiliency from the angle of causality will be the result of a complicated combination of the input of work on decreasing the volume and the outflow of energy for preserving the previous level of temperature.

Correlation is still more complicated. Neither of the two correlatively connected phenomena can be generally considered as a complete cause of the other one, otherwise we would have a strict functional dependence. Correlation only testifies that one of the two phenomena is either a partly cause or a partly effect of the other one, or that they both are brought about by partly common causes.

Thus, the correlation between the prices of rye in Moscow and Samara [in centres 1 and 3, see §§ 3, 14 and 23] is established by multiple observation of facts which are based [which depend] on complicated processes. [Slutsky mentions two causes.] The correlation coefficient extinguishes all this lively play of economic forces and only transfers the result of all the clashing mutual influences to the language of exact numbers. It is for this reason that a statistician as such, being quite competent in establishing correlation between any magnitudes belonging to whichever realm, is not qualified to judge causal connections. To do this, depending on the branch of knowledge concerned, he should also be a biologist, or physician, a meteorologist, an economist, etc.

If two phenomena of the outer world are functionally connected, it does not yet mean that they are [also] causally linked directly or indirectly. The former can be purely accidental, or, more precisely, purely ideographic[31.3]. For example, there is no causal connection, either direct or not, between the movements of the Earth and Sirius along their orbits. However, after deriving from observations a regularity in the motion of Sirius, an astronomer brings it in a functional dependence with the Earth's motion because time is determined by the Earth's position (by the apparent position of the Sun), and the location of Sirius is determined as a function of that solar time.

Just the same, the presence of correlation does not at all by itself indicate the presence of a causal connection, either direct or indirect. When a correlation coefficient many times exceeds its probable error, it only serves as a criterion helping to isolate the main features of a phenomenon by eliminating the influence of an indefinite set of causes which are apt to compensate each other in a large number of trials. A small value of the probable error does not ensure us against either systematic errors or a possible random coincidence in time of two independent series of causes[31.4].

For example, had we wished to study the dependence between the motions of the wings of any two birds, and somehow noted them, perhaps by a cinematographic camera, we could have obtained results of two kinds depending on the interval of observation. Registering the inclinations of the wings each second during a more or less long period, we would have found correlation equal to zero; however, registering them a thousand times during one second, the ensuing correlation would have been rather essential, positive, had our observations *accidentally* begun at the moment when both birds lifted or lowered their wings, and negative had these movements been in the opposite directions.

A similarity of sorts occurs when studying the correlation of various social phenomena moving in time. We often observe here, first, a slow long-term change of magnitudes (of prices, births, mortality, etc), and, in addition, yearly, monthly and even daily fluctuations around this slowly changing level. It can happen that the rise or fall of the level of each of these magnitudes was occasioned by two series of independent causes, that is, that, without allowing for this circumstance, the calculated correlation coefficient will only mislead us by indicating a causal connection. [Slutsky returns here to his study of the prices of cast iron and rye (§ 29)]. Both experienced periodic fluctuations of a great amplitude. An approximate coincidence of the periodicities is, however, already sufficient for correlation to be present, but the coincidence could have been accidental. Previously, the oscillations perhaps did not coincide, and after a certain period of time they can diverge, but, having [only these] two periods of observation, we cannot say anything about this[31.5].

And so, our problem consists in separating and independently studying two points: the mutual dependence of slow changes of the level of magnitudes; and the mutual dependence of their fluctuations around that level. The first problem is not yet properly solved, and we leave it apart[31.6]. Some methods, however, are already developed for treating the second one.

32. Methods of instantaneous average and successive differences

The first one is due to Hooker (1901a). He noted that a positive dependence between the curves of the rate of marriages and foreign trade (import and export) was certainly striking, but that it likely only exists between the deviations of these magnitudes from some slowly changing level[s]. Although separate zigzags of the curves essentially correspond, i. e., the marriage rate increases with the increase in export and vice versa, these magnitudes generally move in opposite directions. During the latest 40 years, export had increased, and the rate of marriages decreased. Therefore, the correlation coefficient proved insignificant, only equal to 0.18 with probable error 0.9.

Both phenomena indicate some periodicity with period being approximately 9 years, and Hooker proposed to measure the "instantaneous level", the arithmetic mean for a number of years with the given [the chosen] year being in the middle. He had thus calculated the mean export taking into account the given year and the other years of the period. He determined in the same way the curves of the movement of the levels of all the other phenomena: of import, turnover of clearing houses, rate of marriage, etc. The following treatment reduces to calculating the correlation coefficient not between absolute magnitudes, but between deviations from the curves of levels.

There also Hooker (1901b, p. 604; and, in more detail, in 1905) applied another interesting method. He determined, as described above, the correlation coefficient between the rate of marriage and exports for the same year, then for half a year, a year, for a year and a half previously, half a year later, etc. Having treated the other phenomena in the same way, he determined the "curves of the correlation coefficients" for the rate of marriage and, in turn, exports, its complete turnover [?], turnover of clearing houses. The maximal value of that coefficient (the peak [the mode] of the appropriate curve) indicates the interval of time after which the studied phenomenon manifests its greatest influence on the marriage rate. The appended figures [excluded from translation] show that the influence of changes in the export and import on that rate become most noticeable after about five months, whereas those in the turnover of the clearing houses, after a year and 21/2 months.

These results in any case represent something new which is quite impossible to discover by simply comparing and considering the curves[32.1]. To Hooker is also due the method of successive deviations, as it is called in this section. Now, the correlation coefficient is determined not between the magnitudes themselves, but between the differences of their successive values. Let

$x_0, x_1, x_2, \dots, x_n$ and $y_0, y_1, y_2, \dots, y_n$

be two series of observations uniformly spaced in time (for example, yearly prices in two markets). Denote the differences

$$d_{x1} = x_1 - x_0,\ d_{x2} = x_2 - x_1,\ \dots,\ d_{xn} = x_n - x_{n-1},$$
$$d_{y1} = y_1 - y_0,\ d_{y2} = y_2 - y_1,\ \dots,\ d_{yn} = y_n - y_{n-1},$$

and very simple expressions will follow for the arithmetic means

$$\bar{d}_x = \frac{1}{n}\sum d_{xi} = \frac{x_n - x_0}{n},\ \bar{d}_y = \frac{1}{n}\sum d_{yi} = \frac{y_n - y_0}{n},$$

which are often close to zero.

Standard deviations and the correlation coefficient $r_{d_x d_y}$ are then calculated in the usual way. This method indicates that a large difference can exist between the correlation coefficients of the magnitudes themselves and between their successive differences. Thus, the price of maize at the farms in Iowa and the total yield of maize in the US only show a correlation coefficient of – 0.28 ± 0.14 which, considering the probable error, ought to be judged close to zero. However, the dependence between successive changes of these magnitudes from one year to another is characterized by a very large correlation coefficient of – 0.84. Hooker (1905, p. 703) believes that the changes in the total yield is one of the most important factors determining the price paid in Iowa.

In any case, the application of the correlation method to phenomena changing in time is yet on the level of first attempts, often made purely empirically by groping around. This problem awaits to be systematically developed, and a vast field of study is opening here. First of all, it is important to ascertain how, under which conditions, is it possible to move here from simply establishing a correlation connection to judging about one or another type of causality. Then, also in connection with the previous question, the entire problem should be considered stochastically. Only thus can we find the key to solving a number of particular pertinent questions[32.2].

Chapter 2. Correlation between Three Or More Magnitudes

33. The main theorem of the theory of linear regression

For the sake of brevity we will call a group of values

$$x_1, x_2, \ldots, x_p$$

an *individual of a totality*. It can represent the sizes of various organs of one and the same being; or, x_1 can be the size of an organ of a progeny with the other x's denoting the sizes of that, or of other organs of his successive ancestors. They can represent prices of the same commodity in different places of a market at the same moment or prices of various commodities in the same place, etc.

The methods of uniting definite values of magnitudes in a group vary infinitely but throughout a study this procedure should be certainly done in the same way. Each x_i can take differing values in different groups, or, as we say, in different individuals of a totality. Had the connection between these magnitudes been strictly functional, and had we known the values of all of them except one, we would have been able to determine it as well.

In case of correlation the connection is freer. Not one, but a set of values of x_1 corresponds to a totality of definite values of $x_2, \ldots, x_p$, and the arithmetic mean of this set is a function of those values. We denote this mean for the *array* of a definite complex of values of $x_2, \ldots, x_p$ by x_{1m}, the size of the array, $n_{x2\ldots xp}$, or, shortened, $n_{(x1)}$, the standard deviation of x_1 in that array, σ_{1m}.

As previously, we will call regression equation that, which connects the mean value of one variable with the values of the other ones. If all the variables entering the equation are of the first degree, we will have a linear regression and a curvilinear otherwise. Because of random causes we will never actually encounter a strictly linear regression and have only to manage with approximation. Therefore, we will as previously distinguish between empirical and theoretical regressions. In case of three magnitudes the geometrical image conforming to linear regression is a plane, and some other surface for curvilinear regression.

For the case of a greater number of magnitudes a vivid geometrical representation is impossible, but mathematicians have an ersatz of obviousness in a multivariate space with various conceivable images in it similar to those representable in the three-dimensional space. Nevertheless, we do not need to turn to this concept since the empirical mutual relations between magnitudes can be represented in a number of tables and we will analytically describe the theoretical regression by the equation

$$X_1 = a_{11} + a_{12}\,x_2 + \; a_{13}\,x_3 + \ldots + a_{1p}\,x_p \qquad (33.1)$$

and in a similar way for the other variables.

We turn now to the main theorem of the theory of linear regression proved in § 16 for the particular case of two variables[33.1]. Denote the difference between the empirical and theoretical values [of the appropriate magnitude] by d_1:

$$d_1 = x_{1m} - X_1$$

and determine the coefficients of equation (33.1) under the condition

$$\sum n_{(x1)} d_1^2 = \min.$$

Consider the following sum for the i-th array and extend it over all the values of the x_1-th array

$$\sum_i [x_1 - (a_{11} + a_{12}x_2 + a_{13}x_3 + ... + a_{1p}x_p)]^2 = \sum_i (x_1 - X_1)^2 = \sum_i [(x_1 - x_{1m}) + (x_{1m} - X_1)]^2 =$$
$$n_{(x1)}\sigma_{1m}^2 + n_{(x1)}(x_{1m} - X_1)^2 = n_{(x1)}\sigma_{1m}^2 + n_{(x1)}d_1^2,$$

$$n_{(x1)}d_1^2 = \sum_i [x_1 - (a_{11} + a_{12}x_2 + ... + a_{1p}x_p)]^2 - n_{(x1)}\sigma_{1m}^2$$

and that extended sum will be

$$\sum n_{(x1)}d_1^2 = \sum [x_1 - (a_{11} + a_{12}x_2 + ... + a_{1p}x_p)]^2 - \sum n_{(x1)}\sigma_{1m}^2.$$

Since the last term on the right side depends on the essence of the totality itself rather than on $a_{11}, a_{12}, ..., a_{1p}$, the entire right side will be minimal[33.2] when

$$\sum_i [x_1 - (a_{11} + a_{12}x_2 + a_{13}x_3 + ... + a_{1p}x_p)]^2 = \text{ min.} \tag{33.2}$$

Thus, the main condition [the only condition] *of the method of least squares*, as in the case of two variables, *is reduced to determining such a linear function* (33.1) *that, when applying it for deriving in each individual case the magnitude* [the value] *of the first variable given all the other ones, we obtain errors with the least sum of squares.*

34. The case of three variables

The regression equations will now be

$$X_1 = a_{11} + a_{12}x_2 + a_{13}x_3$$
$$X_2 = a_{21} + a_{22}x_2 + a_{23}x_3$$
$$X_3 = a_{31} + a_{32}x_2 + a_{33}x_3$$

We will determine the coefficients of the first one; those of the other two can be written down according to symmetry. A regression coefficient ought to obey the condition (33.2). Differentiating consecutively with respect to a_{11}, a_{12}, a_{13}, we have

$$\sum x_1 - \sum a_{11} - \sum a_{12}x_2 - \sum a_{13}x_3 = 0$$
$$\sum x_1x_2 - \sum a_{11}x_2 - \sum a_{12}x_2^2 - \sum a_{13}x_2x_3 = 0 \tag{34.1}$$

$$\sum x_1x_3 - \sum a_{11}x_3 - \sum a_{12}x_2x_3 - \sum a_{13}x_3^2 = 0$$

Now, $\sum x_1 = \sum x_2 = \sum x_3 = 0$ because we again assume that x_1, x_2, x_3 are the deviations of the appropriate magnitudes from their arithmetic means and, therefore, the first equation immediately provides $a_{11} = 0$. Then,

$$\sum a_{12}x_2^2 = a_{12}N\sigma_2^2,\ \sum a_{13}x_3^2 = a_{13}N\sigma_3^2,$$

$$\sum x_1x_2 = N\sigma_1\sigma_2 r_{12},\ \sum x_1x_3 = N\sigma_1\sigma_3 r_{13},\ \sum x_2x_3 = N\sigma_2\sigma_3 r_{23}$$

where r_{12}, r_{23} and r_{13} are the correlation coefficients for the variables considered in pairs and calculated by methods known to us. [After elementary operations excluded from translation] the equations (34.1) are reduced to

$$a_{12}\sigma_2^2 + a_{13}\sigma_2\sigma_3 r_{23} = \sigma_1\sigma_2 r_{12},\ a_{12}\sigma_2\sigma_3 r_{23} + a_{13}\sigma_3^2 = \sigma_1\sigma_3 r_{13}$$

so that

$$a_{12} = \frac{\sigma_1}{\sigma_2}\frac{r_{12} - r_{23}r_{13}}{1 - r_{23}^2},\ a_{13} = \frac{\sigma_1}{\sigma_3}\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}. \qquad (34.2)$$

Here, a_{12} and a_{13} are the regression coefficients. The equation of theoretical regression will be

$$X_1 = a_{12}x_2 + a_{13}x_3 \qquad (34.3)$$

or, if x_1, x_2 and x_3 are measured from the usual zero rather than being deviations from arithmetic means,

$$X_1 = \bar{x}_1 + a_{12}(x_2 - \bar{x}_2) + a_{13}(x_3 - \bar{x}_3). \qquad (34.4)$$

In case of linear regression this equation furnishes arithmetic means x_{1m} for the arrays of x_1 formed by x_2 and x_3. If regression diverges from the linear type, this equation still provides some indications about the dependence between the magnitudes because it represents some mean linear dependence and its application to separate particular cases will result in errors whose sum of squares is minimal.

The mean (square) magnitude of that error [of those errors], call it $\sum_1$, is of some interest. Its square, see (33.2), is equal to

$$(\sum{}_1)^2 = \frac{1}{N}\sum(x_1 - a_{12}x_2 - a_{13}x_3)^2.$$

Inserting the values of a_{12} and a_{13} from (34.2) and performing some simple algebraic operations [Slutsky adduces them in a long footnote], we will obtain

$$(\sum{}_1)^2 = \sigma_1^2\frac{1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}. \qquad (34.5a)$$

Other formulas can be written down according to symmetry:

$$(\sum{}_2)^2 = \sigma_2^2 \frac{1-r_{12}^2-r_{23}^2-r_{13}^2+2r_{12}r_{23}r_{13}}{1-r_{13}^2}, \qquad (34.5b)$$

$$(\sum{}_3)^2 = \sigma_3^2 \frac{1-r_{12}^2-r_{23}^2-r_{13}^2+2r_{12}r_{23}r_{13}}{1-r_{12}^2}. \qquad (34.5c)$$

These expressions are very important since we can apply them, when using formulas (34.3) and (34.4), for estimating the mean [square] value of the errors of determining the magnitude [the value] of the indication in particular cases. If the distribution obeys the normal (the Gaussian) law, the standard deviations in all arrays will be identical with the mean square errors (34.5) becoming the standard deviations of separate arrays

$$\sum{}_1 = \sigma_{1m},\ \sum{}_2 = \sigma_{2m},\ \sum{}_3 = \sigma_{3m}.$$

Making use of the table of the integral of probability [for the normal distribution] we can also determine the probability of any deviation from the value provided by the regression equation.

There exists some degree of dependence between the three correlation coefficients: since the left side of each of the expressions (34.5) is a sum of squares of errors, and cannot therefore be negative, the common numerator of the fractions there ought to be positive, and the coefficients are linked by inequality

$$1-r_{12}^2-r_{23}^2-r_{13}^2+2r_{12}r_{23}r_{13} > 0.$$

Isolating a complete square of ($r_{23} - r_{12}r_{13}$), we get

$$(r_{23}-r_{12}r_{13})^2 < 1-r_{12}^2-r_{13}^2+r_{12}^2r_{13}^2,$$

and therefore

$$r_{12}r_{13}-\sqrt{1-r_{12}^2-r_{13}^2+r_{12}^2r_{13}^2} < r_{23} < r_{12}r_{13}+\sqrt{1-r_{12}^2-r_{13}^2+r_{12}^2r_{13}^2}.$$

Now we can calculate the boundaries for r_{23} given some values of r_{12} and r_{13}, with those inequalities providing a number of interesting indications:

$r_{12} = r_{13} = 0$, **then** $-1 \le r_{23} \le 1$[34.1]; $r_{12} = r_{13} = \pm 1$, **then** $r_{23} = 1$;
$r_{12} = 1$, $r_{13} = -1$, **then** $r_{23} = -1$; $r_{12} = 0$, $r_{13} = \pm 1$, **then** $r_{23} = 0$;
$r_{12} = 0$, $r_{13} = \pm r$, **then** $-\sqrt{1-r^2} < r_{23} < \sqrt{1-r^2}$;
$r_{12} = r_{13} = \pm r$, **then** $2r^2 - 1 < r_{23} < 1$;
$r_{12} = r$, $r_{13} = -r$, **then** $-1 < r_{23} < 1 - 2r^2$ [34.2];
$r_{12} = r_{13} = \pm\sqrt{0.5}$, **then** $0 < r_{23} < 1$;
$r_{12} = \sqrt{0.5}$, $r_{13} = -\sqrt{0.5}$, **then** $-1 < r_{23} < 0$

It could be thought that, supposing that x_1 is positively correlated with x_2 and x_3, these two magnitudes should be also correlated positively, but this is not necessarily so. For example, if $r_{12} = r_{13} = 1/4$, r_{23} can take any value in the interval $[1; 2(1/4)^2 - 1] = [1; -0.875]$; if $r_{12} = 7/10$ and $r_{13} = -7/10$, r_{23} takes any value in interval

[– 1; 1/50]. Only if r_{12}, equal in absolute value to r_{13}, exceeds $\sqrt{0.5}$, r_{23} invariably takes a definite sign (positive if the signs of r_{12} and r_{13} coincide, and negative otherwise).

Is the precision of determining a magnitude by a regression equation always higher when passing from correlation between two magnitudes to that of between three? In the first instance the mean [square] error of determining x_1 will be

$$\sigma_1\sqrt{1-r_{12}^2},$$

cf. formula (18.2), and, in the second case, equal to the square root of the fraction of the formula (34.5a).

For precision to become higher, the square of that fraction should be less than $1 - r_{12}^2$ or greater than r_{12}^2. Since the denominator of that square is positive, the obtained inequality can be written as

$$(r_{13} - r_{12}r_{23})^2 > 0.$$

The left side can still vanish, so that, after adding the third variable, precision cannot become less, but can persist rather than increase. This occurs when the regression coefficient with numerator $(r_{13} - r_{12}r_{23})$, see formula (34.2), vanishes. Let for example $r_{12} = 0.8$, $r_{23} = 0.5$, $r_{13} = 0.4$. Then, only considering x_2, we have x_1 with mean [square] error

$$x_1 = \overline{x}_1 + \frac{\sigma_1}{\sigma_2}\cdot 0.8(x_2 - \overline{x}_2),\ \sum{}_1 = \sigma_1\sqrt{1-0.8^2} = 0.6\sigma_1.$$

It could be thought that the addition of the third variable rather substantially correlatively connected with the first two of them will heighten the precision of that determination. However, just as previously [see formula (34.2)],

$$\rho_{1(2)} = a_{12}(\sigma_1/\sigma_2) = 0.8(\sigma_1/\sigma_2),\ \rho_{1(3)} = a_{13}(\sigma_1/\sigma_3) = 0.$$

As proven, the mean [square] error ought to persist. Indeed [cf. formula (34.5a)] $\sum_1 = 0.6\sigma_1$.

Thus, the correlation connection of the first two magnitudes with the third one (with x_3) had not in the least influenced our formulas and did not heighten the certainty or precision of our conclusions.

Another extreme case is presented by the transition of correlation connection into a strict functional dependence. It only takes place if *only one* value of x_1 will correspond to each pair of x_2 and x_3, i. e., when the *mean square error* of determining x_1, given the other two variables, vanishes. For understanding this case clearer we will simplify it by supposing that the correlation of the first magnitude with the second one is the same as with the third: $r_{12} = r_{13} = r$. Assume also that $\rho = r_{23}$, then the mean square error $\sum_1$ will be exactly zero if the fraction in formula (34.5a) vanishes, which means that

$$1 - \frac{2r^2 - 2r^2\rho}{1-\rho^2} = 0,\ 1 - \frac{2r^2}{1+\rho} = 0,\ r = \sqrt{\frac{1+\rho}{2}}.$$

For example, if $\rho = 0.7572$, correlation becomes a strict functional dependence at $r =$

0.9373.

35. Examples

Now, 0.7572 is the correlation coefficient between barometrically measured air pressure (AP) at Southampton (S), southern coast of England, and Laudale (L), western coast of Scotland (Pearson & Lee 1897, pp. 458 – 459)[35.1] the distance between them being 444 miles, and the meteorological station Stonyhurst (St) situated almost in the middle between them. According to Pearson's rough calculation, the correlation coefficient between points St and each of the other locations should almost coincide: both are near to 0.94 – 0.95. When moving away from the line L – S but leaving equal the correlation coefficients between the moving position and each of these stations, we will come to a point in which these coefficients decrease to 0.9373 (see above end of § 34). A meteorological station, if established there will be, as Pearson (pp. 458 – 459) supposes, somewhere near Whitby, and will have a remarkable feature in that the air pressure measured there will be a precise linear function of the pressures at S and L.

The treatment of meteorological data by the correlation method could have [then] advanced weather forecasting, it would only need much calculations for determining the correlation coefficient between the factors of weather in some points for a previous moment and these factors in other points for subsequent moments. With properly chosen points and intervals of time interesting results would have probably be ensured. Pearson & Lee only attempted to turn the meteorologists' attention to the new method. Calculation of tens and hundreds of correlation coefficients from daily meteorological data collected during a number of years is only possible for an institution rather than for separate individuals.

However, to show how close the calculated magnitudes can coincide with the really existing, provided the stations were chosen properly, Pearson offered a simplified method. The regression equation is linear. Assuming a linear dependence between the pressure at St on the one hand and S and L on the other hand, he issued from the equation

$$\mathrm{AP}_{\mathrm{St}} = x\mathrm{AP}_{\mathrm{S}} + y\mathrm{AP}_{\mathrm{L}} + z.$$

For determining the constants *x, y* and *z* 12 observations for the 15th day of each month were chosen [p. 460] and treated by the method of least squares. Then $\mathrm{AP}_{\mathrm{St}}$ was calculated given the measurements at S and L for 50 moments spaced 14 days apart [p. 460][35.2]. [Slutsky provided these results and their comparison with observations.] We see that the differences are very small, "fairly evenly" distributed "positively and negatively" and that their mean [absolute] value is only ca. 1/40 inches = 0.635*mm.* 38 errors are less than 1*mm*, 11 are greater but less than 2*mm*, and only one error is still greater [equal to 0.19 inches = 4.8*mm*]. Pearson has grounds to consider these results as confirming that not very far from St there should exist in Lancashire a point in which correlative dependence becomes strictly functional.

We can attempt also to estimate the attained degree of approximation by deriving the value of a fictitious correlation coefficient for determining the air pressure at St just as precisely not from two, but from one magnitude. The mean square error as calculated by the data in the provided table is 0.041 inches, the standard deviation of the pressure at St (p. 435 of their paper) is 0.3503 and the dependence between these magnitudes should be such that

$$0.041 = 0.3503\sqrt{1-r^2},\ r = 0.993$$

which is already an essential approximation to a strict functional dependence.

Not only in the field of heredity, where $r = 0.5$ is a normal value, but even in the correlation of organs we do not encounter such close ties. Thus, among organs [with reference to Pearson (1899b, p. 181) Slutsky notes that the maximal value of r was 0.89]. For various aims (forensic medicine, criminal investigations, scientific goals, etc) it is possible to determine with considerable success the stature and size of organs of a human body by the size of one or two organs [cf. § 38]. However, as the reader sees, in these fields we are yet far from the precision achieved by Pearson in forecasting air pressure.

Our last example pertains to prices of rye in Russia [§ 3, also discussed in §§ 14 and 23]. The correlation coefficient between the mean monthly price of rye in a given and a previous month in Samara is $r_{\text{s, s–1}} = 0.93292$, for the mean price in Elets and Samara in the same month, $r_{\text{ES}} = 0.87796$, and, finally, for the mean price in Elets in a given month and the price in Samara in the previous month, $r_{\text{E,S–1}} = 0.85593$. The arithmetic means and standard deviations are given in the Supplement [excluded from translation]. Issuing from these figures, the regression coefficient for the mean monthly price in Samara will be (in copecks)

$$x_{\text{s}} = 47.04 + 0.383(x_{\text{E}} - 52.64) + 0.674(x_{\text{S}-1} - 47.16).$$

The mean monthly prices in Samara for 36 months in 1894 – 1896 calculated by this equation are [the appended table is excluded from translation]. The reader will see that they are close to the actual prices with discrepancies only amounting to 1.9 copecks in the mean which is even less than expected from the theory: theoretically, the mean square error of our calculation equals [cf. formula (34.5a)] 4.5 copecks, actually however, for those three years, only 2.4 cop[35.3].

This compels us to suggest that, first, other years must provide a greater error; and, second, that the distribution of prices does not obey the Gaussian law, see § 40 below. In any case, it would have been a rewarding problem to study the fluctuations of prices on a large scale and the correlation between them for various cities and differing periods. It is certainly very difficult to allow theoretically for all the factors of pricing which the businessmen are considering when striking spot deals. Still, the extreme precision of our calculations does not allow to consider such a problem impossible. And forecasts can perhaps be essentially simplified if the businessmen themselves were to be considered as sensitive instruments for studying the correlation dependence between the prices struck for spot deals and the actual prices at the appropriate future time. It will thus be probably possible to heighten considerably the already attained level of accuracy of empirical forecasts.

36. Partial correlation coefficients

We have already noted that correlation connection, although not identical to causality, can provide valuable indications about the existence of the latter. Regression equations connecting several magnitudes allow us to analyze deeper this point by separating the influence of various factors one from another.

Let phenomena B and C be supposed partial causes of phenomenon A. For convincing ourselves in that fact we should have studied the change in A only with B, then only with C, each time under an invariable state of all the other possible causes. That condition is however impossible to realize, we can only study the change in A under an invariable state of *som*e definite factors whose influence we wish to eliminate and in many cases this alone is sometimes enough for making very valuable conclusions.

Indeed, the main disadvantage for a researcher of social phenomena only consists in that he is just an observer unable to experiment, to arbitrarily ensure the constancy of one or another group of conditions. The correlation theory provides him with such a possibility. Consider a regression equation

$$X_A - h_A = \rho_{A(B)}(x_B - h_B) + \rho_{A(C)}(x_C - h_C)$$

from this viewpoint. Here, h_A is the mean value of x_A for *all possible* values of B and C. X_A is the mean of x_A for *their definite* values. For studying the change in A only brought about by B under invariable C we ought to assume here that x_C is constant. Denoting for the sake of brevity

$$h_A + \rho_{A(C)}(x_C - h_C) = h'_A = \text{Const},$$

we have

$$X_A - h'_A = \rho_{A(B)}(x_B - h_B).$$

Here, h'_A is the arithmetic mean of all the values taken by x_A with a given value of x_C and any value of x_B. We see now that with an invariable C the deviations of x_A from its mean are proportional to those of x_B from its own mean with $\rho_{A(B)}$ being the coefficient of proportionality.

Quite similarly, if phenomenon B is invariable, the deviations of x_A from its mean are proportional to those of x_C with $\rho_{A(C)}$ being that coefficient:

$$X_A - h''_A = \rho_{A(C)}(x_C - h_C).$$

The regression coefficients $\rho_{A(B)}$ and $\rho_{A(C)}$ thus *separately* measure the dependence between A and B and C.

Let us consider the following example. The correlation coefficient between the mean monthly prices of rye in Moscow and Samara [§ 3, also §§ 14, 23 and 35] is rather large, $r_{MS} = 0.77$. When wishing to know whether the causal connection between the prices in Moscow and Samara at the same time is really so tight, we must study the change in the former with *all* other factors except the latter remaining invariable.

This is impossible and we will only exclude one of them, the mean price, again in Samara, but for the previous month. We have

$$r_{MS} = 0.77;\ r_{M,S-1} = 0.79;\ r_{S,S-1} = 0.93.$$

Inserting this in formula (34.4) (the standard deviations are provided in the *Tables* [excluded from translation]) we arrive at

$$X_M - h_M = 0.49(x_{S-1} - h_{S-1}) + 0.23((x_S - h_S). \qquad (36.1)$$

We see that the price in Samara for the previous month influences the price in Moscow stronger than that for the same time. If choosing cases in which the previous price in Samara was the same [as in the given month], its excess over its mean by 10 copecks will be connected with a mean increase in the Moscow price of only 2.3 cops. On the other hand, if the prices in the two cities and at the same time were invariable, the increase in the Samara previous price amounting to 10 cop. above its mean level will be connected with a mean increase in the Moscow price of 4.9 cop.

We had not eliminated other influences, and without further studies it is hardly possible to think here about a causal dependence; it would be more cautious to say that the mean Samara price for the previous month indicates more strongly influences on the Moscow price than the mean price there at the same time.

We conclude that the correlation coefficient r_{MS} is only relatively large because the prices in the two cities, directly influencing each other in a comparatively weak way, are at the same time strongly influenced by other factors. For separately determining the degree of the dependence between the prices in both cities we should eliminate the influence of all these other factors by studying cases in which they are invariable. For example, to exclude the influence of the previous monthly Samara price we will reason thus:

The regression equation for Moscow, when considering the previous and the given month in Samara, is (36.1), and for Samara, when considering Moscow and the previous month in Samara (derived in a similar way) it is

$$X_S - h_S = 0.86(x_{S-1} - h_{S-1}) + 0.10((x_M - h_M).$$

For an *invariable* previous Samara price the regression coefficient of Moscow on Samara ${}_{S-1}\rho_{M(S)} = 0.23$ and the coefficient of Samara on Moscow under the same condition, ${}_{S-1}\rho_{S(M)} = 0.10$. If x_{S-1} = Const, the correlation coefficient between them, called the *partial correlation coefficient*, is equal to

$${}_{S-1}r_{MS} = \sqrt{{}_{S-1}\rho_{M(S)} \cdot {}_{S-1}\rho_{S(M)}} = \sqrt{0.23 \cdot 0.10} = 0.15.$$

The general formula for that coefficient will therefore be

$${}_3r_{12} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad (36.2)$$

and, considering symmetry, it is easy to write down the other ones.

The probable error of the partial correlation coefficient is represented by the formula

$$E_{3r12} = 0.67449\frac{1 - {}_3r_{12}^2}{\sqrt{N}}$$

identical with that of the complete coefficient [cf. (22.3], see Yule (1907, p. 182ff); for a more elementary derivation see Heron (1910).

I borrow my second example from Hooker (1907), an extremely interesting paper rich in content. His immediate aim was to ascertain the period during which meteorological conditions most strongly influence the future harvest. Here, I only consider wheat. His chosen period was eight weeks partly overlapping each other, weeks 9 – 16, 13 – 20, 17 – 24, etc of the year. It was necessary to determine the correlation coefficients between the meteorological conditions of each period with the future harvest. The most important of those were rain (the period precipitation) and heat (the so-called *accumulated* temperature above/below 42°F [= 5°.6C][36.1]).

It occurred that the period 37 – 44 weeks, i. e. the time of sowing and the nearest weeks after it, was especially important because the excess of rain during it negatively and very appreciably influenced the harvest. The correlation coefficient for that period between rain and harvest was the greatest, $r_{wr} = -0.66$. The coefficients for the adjacent periods 33 – 40 and 41 – 48 weeks, – 0.55 and – 0.47, were also substantial. Among other periods only one, for weeks 1 – 8, had that coefficient equal to – 0.47, whereas the rest were considerably smaller.

When restricting our attention to the main two periods (Hooker 1907, p. 30), we note that the correlation coefficient between the accumulated temperature and harvest was 0.36 and 0.52 for weeks 37 – 44 and 1 – 8 respectively. Although the probable errors were large (Hooker only considered 21 years), 0.36 nevertheless is a value which seems to suggest, not with certainty, but with some probability, that *in addition* to rain the temperature accumulated up to the period of sowing also influences the future harvest.

This assumption would however be unfounded. The positive coefficient 0.36 only tells us that there does exist a correlation connection between the temperature conditions during sowing and harvest, but it does not empower us to conclude that it will exist under constancy of other factors; and we have expressed this idea by stressing *in addition*. It is indeed possible that temperature conditions only indicate other circumstances, for example the greater or lesser rainfall in the autumn, but do not at all influence the harvest when the rainfall is usual.

We have thus come to the conclusion that it is necessary to separate the influence of both factors on the harvest and to calculate the partial correlation coefficients between harvest and rain given constant temperature ($_ar_{wr}$) and harvest and temperature given constant rainfall ($_rr_{wa}$).

According to formula (36.2) we have for weeks 37 – 44

$$(_ar_{wr}) = -\,0.59,\ (_rr_{wa}) = 0.006.$$

It occurs that in themselves the temperature conditions during sowing are of no consequence, the only important factor is the rainfall. For weeks 1 – 8 both factors are essential because

$$(_ar_{wr}) = -\,0.55,\ (_rr_{wa}) = 0.49.$$

37. The general case: correlation between *n* variables

An acquaintance with the theory of determinants is here necessary. Quite sufficient is the elementary information, for example in Lorentz (1907) [Slutsky refers to its Russian translation].

In this present case the equation of linear regression can be written as

$$X_1 = a_{11} + a_{12}x_2 + \ldots + a_{1n}x_n. \tag{37.1}$$

Here, X_1 is the probable mean value of all the x_1's given the values of $x_2, x_3, \ldots, x_n$; all these magnitudes are measured by their deviations from the respective arithmetical means for all the totality. As previously, we similarly denote this means by $h_1, h_2, \ldots, h_n$. And similar equations can be written down for each of the other variables; they are easily derived by considering symmetry.

I proved in § 33 that the coefficients of equation (37.1) ought to satisfy the condition

$$\sum(x_1 - a_{11} - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n)^2 = \min. \tag{37.2}$$

[Slutsky derives here the normal equations in unknowns $a_{11}, a_{12}, \ldots, a_{1n}$ with $a_{11} = 0$, see quite similar reasoning in § 34] and, applying determinants, gets

$$b_{1i} = a_{1i}\frac{\sigma_i}{\sigma_1} = -\frac{R_{1i}}{R_{11}},\ i = 2,\ 3,\ ...,\ n,$$

$$R_{1i} = (-1)^{i+1}\begin{vmatrix} r_{21} & 1 & r_{23} & ... & r_{2,i-1} & r_{2,i+1} & ... & r_{2n} \\ r_{31} & r_{32} & 1 & ... & r_{3,i-1} & r_{3,i+1} & ... & r_{3n} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ r_{n1} & r_{n2} & r_{n3} & ... & r_{n,i-1} & r_{n,i+1} & ... & 1 \end{vmatrix},$$

$$R_{11} = \begin{vmatrix} 1 & r_{23} & ... & r_{2n} \\ r_{32} & 1 & ... & r_{3n} \\ ... & ... & ... & ... \\ r_{n2} & r_{n3} & ... & 1 \end{vmatrix}.$$

Both determinants are minors of

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & ... & r_{1n} \\ r_{21} & 1 & r_{23} & ... & r_{2n} \\ r_{31} & r_{32} & 1 & ... & r_{3n} \\ ... & ... & ... & ... & ... \\ r_{n1} & r_{n2} & r_{n3} & ... & 1 \end{vmatrix}.$$

The regression equation will then be

$$X_1 = -\frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2}x_2 - \frac{R_{13}}{R_{11}}\frac{\sigma_1}{\sigma_3}x_3 - ... - \frac{R_{1n}}{R_{11}}\frac{\sigma_1}{\sigma_n}x_n\ , \qquad (37.3)$$

or, if the magnitudes themselves rather than their deviations from arithmetic means are considered,

$$X_1 - h_1 = -\frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2}(x_2 - h_2) - \frac{R_{13}}{R_{11}}\frac{\sigma_1}{\sigma_3}(x_3 - h_3) - ... - \frac{R_{1n}}{R_{11}}\frac{\sigma_1}{\sigma_n}(x_n - h_n).$$

As proved in § 33, when applying this equation for determining x_1 given x_2, x_3, …, x_n, we make the least possible error (more precisely, a number of errors having the least possible sum of squares). The derivation of a general formula for the mean square error is also of interest. To achieve this, we ought to insert the values of a_{11}, a_{12}, …, a_{1n} as derived above into expression (37.2), divide it by N and extract a square root. We will have

$$N(\sum{}_1)^2 = \sum[x_1 + \frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2}x_2 + \frac{R_{13}}{R_{11}}\frac{\sigma_1}{\sigma_3}x_3 + ... + \frac{R_{1n}}{R_{11}}\frac{\sigma_1}{\sigma_n}x_n]^2 =$$

$$\frac{1}{R_{11}^2}\sum[R_{11}x_1 + R_{12}\frac{\sigma_1}{\sigma_2}x_2 + R_{13}\frac{\sigma_1}{\sigma_3}x_3 + ... + R_{1n}\frac{\sigma_1}{\sigma_n}x_n]^2.$$

[Performing a number of elementary operations Slutsky gets]

$$\sum{}_1 = \sigma_1\sqrt{\frac{R}{R_{11}}}, \qquad (37.4)$$

a very simple and very important expression.

38. The case of four variables

Expressions (37.3) and (37.4) include all particular formulas derived above for two and three variables. Thus, for correlation between two magnitudes

$$R = \begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix} = 1 - r_{12}^2;\ R_{11} = 1;\ R_{12} = -r_{12}$$

and the regression equation will be

$$X_1 - h_1 = -\frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2}(x_2 - h_2) \ = r_{12}\frac{\sigma_1}{\sigma_2}(x_2 - h_2)$$

and the mean square error of determining x_1 given x_2 is

$$\sum{}_1 = \sigma_1\sqrt{\frac{R}{R_{11}}}, \ = \sigma_1\sqrt{1 - r_{12}^2}.$$

We will now apply the general formulas for deriving equations for the case of four variables. Keeping to the previous notation, we have

$$X_1 - h_1 = -\frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2}(x_2 - h_2) \ - \ \frac{R_{13}}{R_{11}}\frac{\sigma_1}{\sigma_3}(x_3 - h_3) \ - \ \frac{R_{14}}{R_{11}}\frac{\sigma_1}{\sigma_4}(x_4 - h_4),$$

$$\sum{}_1 = \sigma_1\sqrt{\frac{R}{R_{11}}}.$$

[Slutsky explains how to calculate the determinant *R*.]

As an example, we consider the reconstruction of a man's stature given the size of his various organs (Macdonell 1902)**[38.1]**. The data below concern the prison population in the main penitentiaries of England and Wales and are based on 3000 forms taken from the Scotland Yard's anthropometric bureau. In England, criminals are distinguished between *habituals* and *non-habituals* and Macdonell's material concerns the latter group characterized by comparatively insignificant crimes and punishment.

[Slutsky selected three sizes (left middle finger, F, left elbow, E, left sole, S) out of the six considered by the English author for calculating stature, H, and writes them down in centimetres to four or five places after the decimal point of which I retain only one, and I am only providing the end results. Slutsky notes that anthropometric data "sufficiently closely" obey the Gaussian law. "Therefore, as I prove below, the standard

deviation of each magnitude is the same in all its arrays and the mean square error of all the determinations are equal to the standard deviation of an array, so that
$E_H = 0.67449\sigma_H$".

Table 1 (Macdonell 1902)
Measurements (*cm*)

	St. deviation	Arithm. mean
F	0.5	11.5
E	2.0	45.0
S	1.1	25.7
H	6.4	166.4

Table 2 (Macdonell 1902)
Correlation coefficients
See explanation in text

	F	E	S	H
F	1	0.85	0.76	0.66
E	0.85	1	0.80	0.80
S	0.76	0.80	1	0.74
H	0.66	0.80	0.74	1

1. H = 166.4 + 7.8(F – 11.5); 2. H = 166.4 + 2.6(E – 45.0);
3. H = 166.4 + 4.0(S – 25.7);
4. H = 166.4 – 0.7(F – 11.5) + 2.8(E – 45.0)

The last-written equation "reveals an original dependence […]: an individual with a *given* E whose F is shorter has a greater stature and vice versa". Slutsky continues:]

5. H = 166.4 + 2.8(F – 11.5) + 3.0(S – 25.7);
6. H = 166.4 + 1.9(E – 45.0) + 1.5(S – 25.7);
7. Reconstruction of H given F, E, and S: Slutsky only provides the appropriate regression coefficients. He then indicates the probable errors of H. In all cases they amount to 2.5 – 3.3*cm* and, in addition, he provides that error, 2.4, for the case of all the six sizes considered by Macdonell whose additional sizes were length and width of head and width of face. Slutsky comments: "In itself, the increase in the number of independent variables improves the precision but little, much more important is their appropriate choice […]".

He then adduces Macdonell's table comparing actual and calculated (by issuing from one of the six sizes in turn) statures for 10 randomly chosen criminals. Out of the 60 discrepancies 53 were less than |5|*cm*, the greatest amounted to 8.1*cm* in excess; 32 calculated statures were greater, and 28 shorter than the actual figures, and the mean value of the appropriate probable errors persisted (2.5 – 3.3*cm* as above). Slutsky concludes:]

The results of the reconstruction are only approximate, and could only be such because the regression formulas provide not the individual stature but a mean value for a group with a given size of other organs. As proved in § 33, these formulas ensure the least mean square error among all other possible formulas of linear dependence. The inaccuracy of determination is thus based on the variability inherent in individuals and on the incomplete correlation between the organs. For many aims (for example, in forensic medicine, criminal investigations, scientific goals) that precision is, however, more or less sufficient.

39. Normal correlation. The equation of distribution[39.1]

The correlation theory as explicated above is independent from some special laws of distribution, and the applicability of the regression formulas is only conditioned by linearity of the distribution of the empirical data. In its initial stage, the correlation theory, however, was only an extension of the Gaussian law onto many variables. As an extremely important particular case, it certainly deserves to be at least briefly described. Let n magnitudes be somehow joined together and denote by $x_1, x_2, \ldots, x_n$ their deviations from their appropriate arithmetic means. Assume that their values are determined by a set of m causes with m being much larger than n. Let the deviations of the intensities of these causes from their mean values be $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ so that $x_1 = x_2 = \ldots = x_n = 0$ if $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_m = 0$.

Our main assumptions will be, first, that the changes in the intensities of each separate cause are so insignificant that their squares can be neglected. Second, that their distribution obeys the Gaussian law, and third, that the probabilities of their various values are mutually independent. According to the first assumption, the dependence of any x_i from the ε's will be approximately expressed by a linear function; we suppose here that such functions can be expanded into a Taylor's series and, as we assumed, their squares and higher degrees can be neglected. Then

$$\begin{aligned} x_1 &= \alpha_{11}\varepsilon_1 + \alpha_{12}\varepsilon_2 + \ldots + \alpha_{1m}\varepsilon_m \\ x_2 &= \alpha_{21}\varepsilon_1 + \alpha_{22}\varepsilon_2 + \ldots + \alpha_{2m}\varepsilon_m, \ldots, \\ x_n &= \alpha_{n1}\varepsilon_1 + \alpha_{n2}\varepsilon_2 + \ldots + \alpha_{nm}\varepsilon_m \end{aligned} \qquad (39.1)$$

The probability that the deviation of the intensity of the j-th cause from its mean will be restricted to interval $[\varepsilon_j; \varepsilon_j + \delta\varepsilon_j]$ is (second main assumption)

$$C_j \exp[-\frac{\varepsilon_j^2}{2k_j^2}]\delta\varepsilon_j$$

where k_j is the standard deviation.

Such probabilities are mutually independent (third main assumption), therefore the probability that the deviations of the intensities of the causes will at the same time be restricted to intervals $[\varepsilon_1; \varepsilon_1 + \delta\varepsilon_1]$, $[\varepsilon_2; \varepsilon_2 + \delta\varepsilon_2]$, …, $[\varepsilon_m; \varepsilon_m + \delta\varepsilon_m]$ is

$$P = C\exp[-\frac{1}{2}(\frac{\varepsilon_1^2}{k_1^2} + \frac{\varepsilon_2^2}{k_2^2} + \ldots + \frac{\varepsilon_m^2}{k_m^2})]\delta\varepsilon_1\delta\varepsilon_2\ldots\delta\varepsilon_m. \qquad (39.2)$$

Equations (39.1) allow to express any n (for example, the first n) of the ε's through n x's and the rest ε's. We will have a fractional expression for every ε_j with numerator being linear in respect of all the variables, and the denominator in all the expressions will be the same and only contain constants. Inserting the determined values of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ in equation (39.2) we will receive an expression for the probability that each x_i is contained in interval $[x_i; x_i + \delta x_i]$:

$$P = \text{Const} \exp[-(U + V + W)/2]\ \delta x_1\delta x_2 \ldots \delta x_n\delta\varepsilon_{n+1}\delta\varepsilon_{n+2} \ldots \delta\varepsilon_m. \qquad (39.3)$$

Here, U is a quadratic function of variables x_i, a sum of their squares and products taken in pairs with constant coefficients; V is the same function of ε_{n+1}, ε_{n+2} ,…, ε_m; W is the sum of terms containing products of one of the two x's and one of the two ε's.

What is the meaning of the expression (39.3)? It was derived from expression (39.2) which was the probability of a certain combination of causes whose intensities were contained within certain boundaries. Equations (39.1) indicate, however, that a certain totality of values of x_i corresponds to each such combination, but the number of causes exceeds that of the x's so that not one, but many combinations of causes can correspond to each given combination of the x's. However, if these latter are known as well as ε_{n+1}, ε_{n+2},…, ε_m, the equations (39.1) will determine the other n causes ε_j so that formula (39.3) provides the probability of the combination of the x's and (m – n) causes.

The latter do not, however, interest us since they are to remain unknown; what we need to know is the probability of a definite combination of the x's independently from any values of ε_{n+1}, ε_{n+2},…, ε_m, and these magnitudes we now have to eliminate from (39.3). We begin with ε_{n+1}. Given definite values of the x's, ε_{n+1} can take values [ε_{n+1}; $\varepsilon_{n+1}+ \delta\varepsilon_{n+1}$], [$\varepsilon_{n+1}+ \delta\varepsilon_{n+1}$; $\varepsilon_{n+1}+ 2\delta\varepsilon_{n+1}$], … The probability that one of these cases will occur is equal to the sum of the appropriate probabilities, and the addition is accomplished by integrating the expression (39.3) with respect to ε_{n+1} between limits (–∞; + ∞). Infinite limits have appeared because of the need to know the probability of the combination of the x's given any arbitrary value of ε_{n+1}.

When integrating, the essence of U in equation (39.3) persists, ε_{n+1} vanishes from V and W, and a new factor not including either any x_i or any ε_j appears in the constant term. Reasoning similarly in respect to all the magnitudes ε_{n+2},…, ε_m, we will have, after (m – n) integrations, an expression of the previous type as far as the x's are concerned but lacking the magnitudes ε_{n+1}, ε_{n+2},…, ε_m. Therefore, the probability that the x's will be contained within intervals [x_1; $x_1+ \delta x_1$], [x_2; $x_2+ \delta x_2$], …, [x_n; $x_n+ \delta x_n$] is

$$P = C\exp(-\frac{a_{11}x_1^2 + a_{22}x_2^2 + ... + 2a_{12}x_1 x_2 + ...}{2})\ \delta x_1\, \delta x_2 ... \delta x_n . \qquad (39.4)$$

The law of large numbers ensures that, given a large number of trials, the frequency of a phenomenon approaches [is near to] its probability. Formula (39.4) is thus an expression for the *frequency* of the phenomenon as well; it indicates the relative number of cases in which all the x's are contained at the same time in those intervals.

Suppose that there are N cases in all, then the frequency of distribution will be

$$Z = C\exp(-\frac{a_{11}x_1^2 + a_{22}x_2^2 + ... + 2a_{12}x_1 x_2 + ...}{2})\ . \qquad (39.5)$$

The right side *is the normal distribution function*[39.2] *of n magnitudes correlatively connected one with another.*

40. The main properties of the normal distribution function. The Edgeworth theorem

Let x_2, x_3, …, x_n take definite values; x_1 can take various values, but, as we shall see, their arithmetic mean will be a function of the other x's. Indeed, isolating in (39.5) all the terms containing x_1, we will have

$$Z = C\exp\{-\frac{a_{11}[x_1^2 + 2(a_{12} / a_{11})x_1 x_2 + 2(a_{13} / a_{11})x_1 x_3 + ... + 2(a_{1n} / a_{11})x_1 x_n] + V}{2}\}\ .$$

Here, *V* is a quadratic function of those other *x*'s, constant according to our condition. Denote terms not containing x_1 by *W*, then

$$Z = C\exp\{-\frac{a_{11}[x_1 + (a_{12}/a_{11})x_2 + (a_{13}/a_{11})x_3 + ... + (a_{1n}/a_{11})x_n]^2 + W}{2}\} =$$

$$Ce^W \exp\{-\frac{a_{11}[x_1 + (a_{12}/a_{11})x_2 + ...]^2}{2}\} .$$

However, for given x_2, x_3, …, x_n *W* is constant, and under the conditions stated the distribution of x_1 is Gaussian**40.1**:

$$Z = C\exp\{-\frac{a_{11}[x_1 + (a_{12}/a_{11})x_2 + ... + (a_{1n}/a_{11})x_n]^2}{2}\}.$$

As proved in § 9, the arithmetic mean of all the values of x_1, given all the other *x*'s, will be

$$X = -\frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - ... - \frac{a_{1n}}{a_{11}}x_n \qquad (40.1)$$

and the standard deviation

$$\sigma_{1m} = \frac{1}{\sqrt{a_{11}}}.$$

We immediately arrive at an important result:

A. *Given the normal distribution, regression is strictly linear.*

B. The particular *standard deviation is the same for all the arrays.*

Under these conditions (see § 18) the mean [square] error of determining one variable through the other ones is equal to a particular standard deviation, so that

$$\sum\nolimits_1 = \sigma_{1m} = \frac{1}{\sqrt{a_{11}}}. \qquad (40.2)$$

Let us apply the obtained results for deriving the Edgeworth theorem that establishes the dependence between the coefficients of the exponential function in the equation of the normal distribution on the one hand, and correlation coefficients and standard deviations on the other hand.

We proved, see § 37, that any equation of linear regression can be written down as (37.3). Comparing it with the equation of linear regression given a normal distribution (40.1) we will find that

$$\frac{R_{12}}{R_{11}}\frac{\sigma_1}{\sigma_2} = \frac{a_{12}}{a_{11}}, \quad \frac{R_{13}}{R_{11}}\frac{\sigma_1}{\sigma_3} = \frac{a_{13}}{a_{11}}, \text{ etc.}$$

Then, comparing the expression for the mean [square] error (37.4) with (40.2) we will have

$$\frac{1}{a_{11}}=\sigma_1^2\frac{R}{R_{11}},\ a_{11}=\frac{R_{11}}{R}\frac{1}{\sigma_1^2},\ a_{12}=\frac{R_{12}}{R}\frac{1}{\sigma_1\sigma_2},\ a_{13}=\frac{R_{13}}{R}\frac{1}{\sigma_1\sigma_3},\ \text{etc},$$

$$a_{ii}=\frac{R_{ii}}{R}\frac{1}{\sigma_i^2},\ a_{ij}=\frac{R_{ij}}{R}\frac{1}{\sigma_i\sigma_j}.$$

Inserting the obtained magnitudes in formula (39.5), we will have it in an elegant form

$$Z=C\exp[-\frac{1}{2}(\frac{R_{11}}{R}\frac{x_1^2}{\sigma_1^2}+\frac{R_{22}}{R}\frac{x_2^2}{\sigma_2^2}+...+2\frac{R_{12}}{R}\frac{x_1}{\sigma_1}\frac{x_2}{\sigma_2}+...)] \qquad (40.3)$$

which is indeed the Edgeworth theorem. It remains to derive *C*. [After very much work Slutsky gets

$$C=\frac{N}{(2\pi)^{n/2}\sigma_1\sigma_2...\sigma_n\sqrt{R}}.]^{\mathbf{40.2}} \qquad (40.4)$$

41. On the probability of a system of deviations correlatively connected with each other

Here, I wish to explicate one of the most splendid and at the same time important applications of the theory of normal correlation (Pearson 1900). Suppose we have a system of magnitudes correlatively connected with each other with arithmetic means h_1, h_2, …, h_n, standard deviations σ_1, σ_2, …, σ_n and deviations from means x_1, x_2, …, x_n. Let the deviations, as it actually often occurs, follow the Gaussian law so that their distribution will obey equation (40.3) with *C* provided by (40.4).

Various combinations of these magnitudes occur with differing frequencies and therefore have different probabilities whose values depend on the exponent of the function in (40.3). If that function is constant, the probability of a certain combination of the *x*'s will not change, nor will it depend on any changes of the individual *x*'s. Therefore, if χ^2 = Const, the equation

$$\chi^2=\sum\frac{R_{ii}}{R}\frac{x_i^2}{\sigma_i^2}+2\sum\frac{R_{ij}}{R}\frac{x_i x_j}{\sigma_i\sigma_j} \qquad (41.1)$$

will provide all the possible equally probable values of the deviations**41.1**.

For the sake of brevity we will now write (40.3) as

$$Z = Z_0\exp[-(\chi^2/2)].$$

To recall, *Z* expresses the *frequency* with which all the combinations of x_1, x_2, …, x_n occur satisfying equation (41.1); or, otherwise, all the equally probable (equally often occurring) combinations characterized by a definite value of χ^2. The *number* of cases in which the first magnitude is contained in interval [x_1; $x_1 + dx_1$], the second, in [x_2; $x_2 + dx_2$] etc is equal to

$$dN = Z_0\exp[-(\chi^2/2)]dx_1dx_2 \;\ldots.\; dx_n. \qquad (41.2)$$

Let us find now the probability that the *x*'s have any values of the same or lower probability than in the given case. Our question reduces to determining the probability of all the combinations of the values of *x*'s for which χ^2 is the same as the given one, or larger than it. The required probability is equal to the number of all cases in which χ^2 is equal or greater than its given value divided by the number of all cases in the totality. We will find the former number by adding up the expressions similar to (41.2) for all the values of *x*'s beginning with such that provide an χ equal to its given value and taking all those in which χ is greater; that is, for the given χ to $\chi = \infty$. The latter number is determined by adding up the same expressions leading to whichever values of χ from 0 to ∞. The probability is thus

$$P = \frac{\int\int \ldots\int \exp(-\frac{\chi^2}{2})dx_1dx_2\ldots dx_n}{\int\int \ldots\int \exp(-\frac{\chi^2}{2})dx_1dx_2\ldots dx_n} \qquad (41.3)$$

with the limits of integration mentioned above.

For simplifying this expression we turn to the geometric representation, see Note (41.1). The product of $dx_1dx_2 \;\ldots.\; dx_n$ is the elementary volume; it should be multiplied by the exponential function, let us say by the density exp $(-\chi^2/2)$, and all such expressions from the surface of the ellipsoid χ to infinity added up. This addition should begin by determining the sum of such expressions from a thin ellipsoidal layer [ring] between ellipsoids χ and $[\chi + d\chi]$. Call its volume *dV*. Then the mass of the layer will be exp $(-\chi^2/2)dV$ and (41.3) is reduced to

$$P = \frac{\int_\chi^\infty \exp(-\chi^2/2)dV}{\int_0^\infty \exp(-\chi^2/2)dV}. \qquad (41.4)$$

If (41.1) is divided by χ^2, the quotient will enter the denominator of each term of each sum. The *linear* dimension of the ellipsoid is proportional to χ, its volume proportional to χ^n and is therefore**41.2** equal to $V = C\chi^n$ so that

$$dV = Cn\chi^{n-1}d\chi$$

and (41.4) becomes

$$P = \frac{\int_\chi^\infty \chi^{n-1}\exp(-\chi^2/2)d\chi}{\int_0^\infty \chi^{n-1}\exp(-\chi^2/2)d\chi}. \qquad (41.5)$$

This is indeed the probability that, because of random circumstances, there can occur a system of deviation as, or less probable then the given one. [Simplifying (41.5) Slutsky gets for odd and even *n* respectively]

$$P=\sqrt{\frac{2}{\pi}}\int_{\chi}^{\infty}\exp(-\frac{\chi^2}{2})d\chi+\sqrt{\frac{2}{\pi}}\exp(-\frac{\chi^2}{2})[\frac{\chi}{1}+\frac{\chi^3}{1\cdot 3}+...+\frac{\chi^{n-2}}{1\cdot 3\cdot 5...(n-2)}], \qquad (41.6)$$

$$P=\exp(-\frac{\chi^2}{2})[1+\frac{\chi^2}{2}+\frac{\chi^4}{2\cdot 4}+...+\frac{\chi^{n-2}}{2\cdot 4\cdot 6...(n-2)}]. \qquad (41.7)$$

In calculating these formulas [probabilities] the only difficulties are of a purely arithmetical nature, but in most cases even they can be avoided by applying Elderton's tables (1902)[41.3] that provide P with sufficient precision, and, for most applications, quite sufficient completeness. However, they were compiled under somewhat different assumptions.

Until now, we assumed that all the n magnitudes, being correlatively connected, were mutually independent in the strict functional meaning, whereas the Elderton tables are compiled for n_1 magnitudes connected by an equation, so that only $(n_1 - 1)$ of those magnitudes are independent. [We ought therefore to equate $n_1 = n + 1$.]

42. A test for conformity of a theoretical with an empirical distribution

Suppose (Pearson 1900) we have a totality whose items are distributed into $(n + 1)$ groups according to the value of some indication, and out of each N items let $\mu_1, \mu_2, \ldots, \mu_{n+1}$ in the mean be included in these groups. However, when selecting a sample group of N items from the parent population the size of the groups will in general be different: $m_1, m_2, \ldots, m_{n+1}$. In the mean, for a set of such samples, m_i will be equal to μ_i, but in individual cases we will encounter errors [deviations]

$$e_1 = m_1 - \mu_1,\ e_2 = m_2 - \mu_2,\ \ldots,\ e_{n+1} = m_{n+1} - \mu_{n+1}.$$

Only n of them are independent because obviously

$$e_1 + e_2 + \ldots + e_n + e_{n+1} = 0$$

so that the last one can be calculated given the other ones. Therefore, when applying the formulas of the previous section, we should only consider n variables.

It is not difficult to prove that the standard deviation of e_i, supposing that m_i are subjected to random variations, will be[42.1]

$$\sigma_i=\sqrt{N\frac{\mu_i}{N}(1-\frac{\mu_i}{N})} \qquad (42.1)$$

and that, according to the main formula of the correlation theory [(16.7)], the correlation coefficient between the errors is determined from[42.2]

$$\sigma_i\sigma_j r_{ij}=-\frac{\mu_i\mu_j}{N}. \qquad (42.2)$$

Now, the moments $r_{12}, r_{13}, \ldots, r_{ij}, \ldots$, see formulas (42.1) and (42.2), should be inserted in the expression for the determinant R (§ 37) which is to be simplified and all of its minors R_{ii} and R_{ii} determined[42.3]. The magnitudes thus obtained ought to be inserted into formula (41.1).

Pearson had done all that and arrived at a very simple result:

$$\chi^2 = \sum \frac{e^2}{\mu}.$$

Here, the addition should be extended over all the $(n + 1)$ errors. Denote the number of groups by n_1, then determine the probability sought by χ^2 and n_1 from the Elderton table (1902). If these arguments, χ^2 or n_1, exceed the boundaries of that table, P corresponding to $n = n_1 - 1$ is found by formulas (41.6) or (41.7).

Examples. 1. The so-called law of large numbers was repeatedly checked by various experiments[42.4]. Thus, Weldon (Pearson 1900, p. 167) reports that 12 dice were thrown 26 306 times, and each time the number of dice showing 5 or 6 points was registered, see Table[42.5]. Let us ask ourselves now, how good does the experiment conform to theory; is it possible to explain the deviations only by random causes? This question cannot be answered by the naked eye, it demands the application of the Pearson test. The calculation of χ^2 is arranged in the next table[42.6] leading to

$$\chi^2 = 43.87241, \chi = 6.623625.$$

I note that the number of significant digits was obviously larger than necessary; it is usually sufficient to calculate χ^2 to two, or at most three places after the decimal point. There were 13 groups, and 12 independent variables; with the last term on the right side $\chi^{10}/(2{\cdot}4{\cdot}6{\cdot}8{\cdot}10)$ being allowed for, (41.7) provides $P = 0.000\ 016$.

This means that the probability of a system of deviations not more probable than the observed is extremely insignificant. Had the experiment been repeated very many times under ideal conditions, we would have 62 499 times obtained lesser, and only once the same or larger deviations. We can therefore bet 62 499 against 1, i. e., with almost absolute certainty, that the dice in that experiment were not precisely made, that the results of various throws were not equally probable.

The reader sees that the *check* of a stochastic theorem[42.7] was actually a check of the physical properties of the dice, which would have probably demanded less time if performed by methods of measuring applied in physics.

2. I have mentioned many times that the sizes of various parts of the human body are usually distributed sufficiently close to the theoretical normal distribution (to the Gaussian *law*), and I wish therefore to adduce one more pertinent example, the distribution of statures of 1052 mothers (Pearson & Lee 1903) [for calculating the test of conformity]. Here is the table[42.8] providing all the data, necessary for calculating that test. A remark should be added here. Each theoretical frequency curve represents a *continuous* function and thus deviates from reality since the size of any group cannot be less than unity. It is also conditional because the size of some individuals is located exactly on the boundary of two groups and a half of an individual has to be added to each of those groups. The theoretical size of a group, however, can be 1/10, 1/100, 1/1000 and less.

This weak point of theoretical curves is indeed revealed in the extreme groups of any distribution; theoretical frequencies such as 0.5, 0.3, 0.2 in a number of extreme groups should therefore be understood as stating that, for 0.5, in a number of pertinent random samples 1 or 0 individuals should occur equally often. For 0.3, […] And each random sample ought to include in the mean one individual (0.5 + 0.3 + 0.2) in all the three extreme subdivisions taken together.

Because of this purely conditional meaning of the fractional magnitudes in the extreme groups, it would be a gross mistake to apply the Pearson test considering each of them separately. To ensure a comparability of the theoretical and the empirical distributions it is obviously necessary, when applying that test, to unite individuals in such groups whose theoretical size will be not less than unity or at least close to it (Pearson 1900, p. 164).

When applying this consideration to the table above, we will have 17 groups, then $\chi^2 = 14.47$ and the Elderton table (1902) will provide $P = 0.56$. Having the arithmetic mean stature 62.484 inches with standard deviation 2.3904 inches, the function

$$Z = \frac{N}{2.3904\sqrt{2\pi}}\exp[-\frac{(x-62.484)^2}{2\cdot 2.3904^2}]$$

determined by Pearson expressed the size of the group provided that mothers' statures were really distributed according to the Gaussian law.

If that formula conformed to reality, in 56 random samples out of 100 we will have purely random divergences greater than the observed. The conformity of the empirical and the theoretical distributions ought to be thought very satisfactory[42.9].

43. A test for conformity of theoretical with an empirical regression line

If some magnitudes are not correlated, all the correlation coefficients are zero, the determinant R (see beginning of § 37) and all its minors of the type of R_{ii} are unity and minors R_{ij} are zero. The equation of distribution [unusual term, also in title of § 39] will then be

$$Z = Z_0 \exp[-\frac{1}{2}\sum\frac{x_i^2}{\sigma_i^2}], \text{ and } \chi^2 = \sum\frac{x_i^2}{\sigma_i^2}.$$

The probability of all probable [possible] systems of deviations not higher than that of the given one will be determined by the same formulas (41.6) and (41.7) when issuing from that value of χ^2 and n or from values χ^2 and $n_1 = n + 1$ in the Elderton table (1902). These considerations directly bear on the regression curve. Indeed, Pearson (1905b, p. 13) proved that the errors of the arithmetic mean of an array are not correlated with those of another array. It follows that, for deriving the probability of the conformity of a theoretical with an empirical regression line we should calculate the deviations

$$e_1 = y_{x1} - Y_1, e_2 = y_{x2} - Y_2, \ldots$$

and the standard deviations of y in separate arrays,

$$\sigma_{nx1}, \sigma_{nx2}, \ldots$$

The standard deviation of e_i is that of the arithmetic mean, and, according to the known formula (22.1), equal to

$$\sigma_{ei} = \frac{\sigma_{nxi}}{\sqrt{n_{xi}}} \text{ so that } \chi^2 = \sum\frac{n_{xi}e_i^2}{\sigma_{nxi}^2}. \qquad (43.1, 2)$$

We see now that the derivation of the regression curve (§ 28) is the most theoretically proper. Indeed, from all curves of a given type the most probable regression curve for which χ^2 (43.2) will be the least is that which satisfies the condition

$$\sum \frac{n_{xi} e_i^2}{\sigma_{nxi}^2} (y_{xi} - Y)^2 = \min,$$

i. e., that which we proposed for deriving its coefficients.

Additional remarks

1. On terminology

[Slutsky provides considerations mostly concerning the translation of English terms into Russian. He believes that Nekrasov's attempt (1912, pt. 3) to transplant the notion of *correlation* without mentioning that term "does not seem to deserve imitation", and that the term *regression* is somewhat doubtful: "for a Russian ear, it rings rather strange, and even in England itself it cannot fail to seem slightly artificial because of its accidental biological origin […]. Future will show whether someone will not be able to replace it by a sufficiently apt Russian expression". (No, it is still with us.)]

2. On the method of moments

Pearson's justification of the method of moments can hardly be recognized as quite rigorous. To say nothing about the declared rather than substantiated insignificance of the abandoned terms in formula (6.3), the very method of proof based on applying a Taylor (or Mac Laurin) series is doubtful. Lakhtin's attempt (1903, pp. 483 – 488) of a more rigorous justification of the method of moments is therefore of interest.

For more or less approximately satisfying the condition (6.2) Lakhtin expands functions y and Y in infinite series of Legendre polynomials (of spherical functions). Such series are more general than the Taylor series: for their convergence it is sufficient that these functions taken between given limits were continuous and did not have infinitely many extreme points. Statistical curves usually satisfy those conditions.

Equating the first n terms of the expansion to zero, Lakhtin derives the main equation of the method of moments (6.4.2) for $i = 0, 1, 2, \ldots, (n - 1)$:

$$\mu_0 = \mu'_0,\ \mu_1 = \mu'_1,\ \mu_2 = \mu'_2,\ \ldots,\ \mu_{n-1} = \mu'_{n-1}.$$

[Slutsky does not explain his notation; see, however, § 2.] The abandoned terms can be neglected because of the proved convergence of the series. [Just the same, a "declared rather than substantiated" statement.]

Tables

Tables I – VI show the correlation between the mean monthly prices of rye in Moscow, Elets and Samara including also the previous monthly price in Samara [§ 3]. The data are those for the years 1893 – 1903, but because of gaps they only concern 124 months […].

Tables VII – VIII are compiled by issuing from the data in Veselovsky (1909, Suppl. 4, §§ 2 and 5, pp. 674 – 682) […]. It is my pleasant duty to thank Mr. Dobryden for helping me to transfer the data on cards.

Table IX is compiled by issuing from materials grouped in Schmitz (1903, pp. 65 – 69 and 217 – 221) [§ 29].

[The Tables themselves are omitted from the translation.]

Notes

11.1. An unusual term, repeated below. O. S.

11.2. This statement concerns two variables rather than *n.* O. S.

14.1. At the end of § 13 Slutsky applied the term *type of regression* in another sense. O. S.

16.1. This is wrong: the method of least squares possesses certain optimal properties. Slutsky possibly followed Markov who, strangely enough, had up to the end of his life defended Gauss's mature justification of that method contradicting himself by denying its properties (Sheynin 2006, pp. 81 and 84). O. S.

16.2. "Given age" is an important restriction lacking in Quetelet's reasoning. However, Slutsky's examples should have considered the impossibility of an individual having mean weight and mean stature. O. S.

16.3. *Anthropometry* is a term introduced on Humboldt's advice by Quetelet (1870, p. 670) and in 1871 he also published a French book entitled *Anthropométrie.* However, for quite a few decades many authors continued to use the much more general term *anthropology*. Macdonell (1902) whose paper Slutsky described in § 38 was possibly an exception. O. S.

18.1. Slutsky wrote *mean error* and in general he often omitted the *square*. Many authors made the same mistake: *mean error* is a definite term of the classical error theory. Note, however, that mean square error was indeed introduced in that theory, but not by Gauss. In 1823, in §§ 7 and 8, Gauss had introduced variance calling it "medium metuendum, sive simpliciter errorum medium". Newcomb (1908, p. 540) remarked that astronomers "generally designated" the mean square error "as the mean error".

18.2. Slutsky referred to his § 15, apparently to formula (15.1), as I called it. Slutsky had not numbered it although he assigned numbers almost to all displayed formulas (consecutively throughout the book). His system was at the same time awkward and incomplete. O. S.

18.3. An explanation would have been appropriate. O. S.

18.4. According to Slutsky, linear regression was accompanied by homoscedastic totalities [homoscedastic distributions] which seems to be methodically wrong. O. S.

18.5. In itself, this notation is acceptable, but Slutsky greatly complicated it by denoting mean square errors in the same way; for example, $\sum_r$ often meant the mean square error of the estimate *r*. I experienced difficulties when distinguishing these two cases one from another and in any case I replaced $\sum_i$ by its usual modern symbol. O. S.

19.1. There could have been various causes. For example, mothers are less numerous than daughters, not all of whom become mothers, and this circumstance could have influenced the result. E. S.

20.1. The arithmometer renders an irreplaceable service. In all calculations connected with applying the correlation method it is rarely possible or necessary to retain more than two or three significant digits in the final results, but the magnitudes playing an intermediate role and needed in further work ought to be determined more precisely, so that the errors of calculations will not accumulate and become comparable with, or even greater than the probable error of the result. When applying the arithmometer, it is not difficult to retain five or even six places after the decimal point and thus to ensure a precise coincidence of the results of calculation made by two researchers of the same data. If possible to be satisfied by a lower precision, it is advisable to make vast calculations by a good slide rule for reducing the work that could have otherwise become excessive. E. S.

21.1. The restrictions mentioned just below are not sufficient for the realization of the normal law. O. S. The authors of this remarkable memoir arrived at results which they themselves (p. 234) consider only approximate. This circumstance ought to become clear even for a layman at least in respect of *r* when understanding that under the Gaussian (the normal) distribution all deviations from + ∞ to – ∞ are possible whereas *r* can only change from 1 to – 1. Its law of distribution will be different. For random samples of large size this is of no consequence because considerable deviations are extremely unlikely. For very small samples and in problems where large deviations ought to be allowed for, this fact should be taken into consideration, see Student (1908). E. S.

21.2. Beginners could have falsely concluded that the theory of errors was based on the normal law. O. S.

22.1. Apart from Pearson & Filon (1898) mentioned above, the following papers ought to be cited: Sheppard (1898); Pearson (1902a; 1905b) and Pearson (1903a), a small editorial, indispensable because of being written for a broader circle of readers. E. S.

22.2. Student(1908, p. 308), see also Hooker (1907, p. 6) and comments on this latter by Edgeworth and Yule […]. E. S.

23.1. By applying Encke's table, we find that the probability of a deviation *not* exceeding, as in the first case, more than 2.4 times its probable error, is 0.89450. The probability of a deviation larger in absolute value is 0.10550, and that of one larger positive deviation is 0.05275. The probability of the contrary event is 0.94725 and the sought ratio of chances is 0.94725/0.05275 = 18/1.

The third case exceeds the boundaries of Encke's table and we will reason thus. If a magnitude 8.2 times exceeds its probable error, it will 5.5 (= 0.67449·8.2) times exceed its standard deviation. However, according to the Sheppard table (1903) the probability that that magnitude will not deviate in the *positive* direction more than 5.5 times its standard deviation will be 0.999 999 9810, and the ratio will be 0.999 999 9810/0.000 000 0190 = $53{\cdot}10^6/1$. E. S.

24.1. What did Slutsky mean by *general case*? For the normal distribution the sample variance and the arithmetic mean are independent, this being the Student – Fisher celebrated theorem anticipated by Helmert (Sheynin 1995, p. 98) and even Laplace, in the First Supplement to his *Théorie analytique des probabilités* (Sheynin 1977, p. 36). O. S.

24.2. There is a misprint in the last term, immediately noticeable since it corrupts symmetry: p_{20} should be replaced by p_{02}. E. S.

24.3. It is well worth to apply the general formula (24.5) only when asymmetry is very strong, the usual formula provides doubtful results and it is extremely important to estimate rigorously the value of the obtained correlation coefficient.

The moments (the products) p_{qs} ought to be calculated the same way as the usual moments by expanding $(x-\overline{x})^q$ and $(y-\overline{y})^p$ into a binomial series, multiplying and adding up the results thus reducing the central moments p_{qs} to non-central in respect to any axes, of the type

$$\pi_{qs}=\frac{\sum n_{xy}x^qy^s}{N}.$$

First, the non-central moments π_{qs} are calculated, then the central moments by applying the deduced formulas. E. S.

25.1. The explicated method is a modification of the difference method applied by Pearson for replacing the main and most reliable method of products. In its previous form it was not free from a shortcoming since in general it provided somewhat differing values of r_{xy} as compared with the latter. See Wright, Lee & Pearson (1907) and Harris (1909). The modified form eliminated the mentioned shortcoming. E. S.

26.1. Tutubalin (1973, p. 27) reported that a number of mathematicians had experimentally confirmed Pearson's opinion, but he did not refer to anyone. O. S.

27.1. I have dropped Slutsky's Gothic letters replacing them by more usual notation. O. S.

27.2. Here, Slutsky added the two last formulas lacking in the expression (7.6). O. S.

28.1. Slutsky understands *weight* (p) exactly in the sense of the theory of errors (and least squares). This means that

$$p_\xi=\frac{C}{\text{var}\,\xi}\text{ and he assumes that }(p_\xi)^2=(\frac{C}{\text{var}\,\xi})^2.$$

In other words, he believes that varξ^2 = (varξ)^2 which is wrong. O. S.

28.2. The coefficients of the normal equations are very large numbers, and it seems that these are most conveniently solved in the following way. At first we only take into account not too many digits and calculate a'_0, a'_1, …, a'_m approximately, insert ($a'_0 + \Delta a'_0$), ($a'_1 + \Delta a'_1$), … into the equations and find the corrections. Estimate their approximate size by the residual free terms; if needed (usually it is not) a next step is made […]. E. S.

29.1. Slutsky calls both η and η^2 correlation ratio, see formula (29.6) below. According to modern definition, it is the latter magnitude. O. S.

29.2. Slutsky thus indirectly defined the central axis. O. S.

31.1. The term *not strictly definite* (or *not perfect*) dependence, or common, mutual relations is due to Nekrasov (1912), see for example pp. 427, 439. I was unable to make use of this extremely interesting work since I have received it during the printing of my own contribution. E. S. [See Foreword, § 2.1.3.]

31.2. The same example is in a recent treatise (Smirnov & Dunin-Barkovski 1959, § 9.1.1). O. S.

31.3. An ideographic system of writing means that characters (ideograms) signify whole words or their significant parts (Chinese hieroglyphics, say). Did many readers understand Slutsky? And his example concerning Sirius astonishes. Slutsky should have discussed its proper motion rather than the mysterious orbit. More: for us, that motion is connected with the orbit (yes, orbit) of the Sun, but certainly not with the Earth's motion. O. S.

31.4. One of Aristotle's example of a chance event was an intersection of two such chains. Cournot (1843, § 40) repeated that explanation, see also his later book (1872/1973, pp. 9 – 10) where he also stated that "The idea of hazard […] is the key of (de la) statistics".

31.5. Another method: Hooker (1901b, p. 603). E. S.

31.6. I allow myself to remark that this problem yet awaits a researcher who will attempt to solve it by applying the methods of *harmonic analysis* that rendered great services in other fields. E. S.

32.1. Hooker's method of "smoothing" the fluctuations is certainly imperfect. Social phenomena are not distinguished by strict periodicities so that having, for example, 9 years as the mean period, during some nine-years intervals there can occur two maxima or two minima lifting or lowering without sufficient cause the level of a given year. A curve thus derived can be smoothened further, by the naked eye, say, but a subjective element best avoided will then enter.

The best method of "smoothing" is to apply a suitable curve, a parabola for example, whose coefficients are not difficult to calculate (§§ 7 – 8 and 27). When determining the correlation coefficients by the Hooker method and when applying a curve smoothened by the naked eye, it is in any case *absolutely* necessary to publish a figure of that *smoothened* curve in a sufficiently large scale without which it is impossible to assess the correctness of the work done. E. S.

32.2. For references to (yet scarce) attempts to apply the correlation method to economic and social problems see Yule (1909). For more complete bibliography, mostly theoretical and biological, see Leontovich (1911, pt. 2, pp. 191 – 214). E. S.

33.1. Edgeworth (1892) was the first to study the case of n variables with Pearson (1896b) developing it further. Our explication is based on Yule's original paper (1897b) who freed the theory of linear regression from unnecessary assumptions. Still, he restricted his study to the cases of two, three and four variables. E. S.

33.2. Slutsky wrote: "that term will be minimal when…" O. S.

34.1. Yule (1897b) mistakenly provided the value 0 which perhaps was a misprint. E. S.

34.2. There also $(2r^2 - 1)$ is mistakenly stated. E. S.

35.1. In the sequel, Slutsky several times only mentioned the first author, Pearson. O. S.

35.2. Pearson & Lee have not explained the possible slight discrepancy between *15th day*… and *14 days*. These periods possibly corresponded to different years. O. S.

35.3. According to my calculation, the mean square error was 4.1 cop. O. S.

36.1. According to the so-called law of the sums of temperatures (Réaumur, in 1738), leaves, flowers and fruits come out on plants of a given species after that sum attains certain values. In 1846, Quetelet reasonably proposed sums of squares of temperatures instead (Sheynin 1980, pp. 326 – 327). O. S.

38.1. Another important work on this subject is Pearson (1899). E. S.

39.1. Unusual term. O. S.

39.2. *Distribution function* is a modern term having another meaning. O. S.

40.1. The value of C has obviously changed. O. S.

40.2. In case of each particular density function that constant can be derived by demanding that the area *under* it be equal to unity. O. S.

41.1. From the geometrical viewpoint, (41.1) is an equation of a generalized ellipsoid of equal probabilities in an n-dimensional space. To form an appropriate idea it is sufficient to restrict that picture to the case of three variables, and, in the sequel, to imagine always a usual ellipsoid. The values of x_1, x_2, x_3 (and, in general, $x_1, x_2, \ldots, x_n$) corresponding to any point on the surface of the ellipsoid are equally probable. E. S.

41.2. This explanation is hardly sufficient. It is extremely simple, however, to note that a sphere with radius R, $x^2 + y^2 + z^2 = R^2$, has volume proportional to R^3. O. S.

41.3. Slutsky sometimes refers to Elderton's table in the plural (*tables*). O. S.

42.1. Slutsky explains the derivation of the formula (42.1) below and refers to Chuprov (1909). The formula is now generally known. O. S.

42.2. The main formula is

$$N\sigma_i\sigma_j r_{ij} = \sum \delta\mu_i \delta\mu_j.$$

However, if the deviations are random, the excessive number of items in the i-th group should be *in the mean* proportionally distributed among the other groups, so that approximately, since the deviations are not considerable,

$$\delta\mu_j = - \ \delta\mu_i \frac{\mu_j}{N - \mu_i}$$

and

$$\sigma_i\sigma_j r_{ij} = -\frac{1}{N}\sum[\delta\mu_i^2\frac{\mu_j}{N-\mu_i}] = -\sigma_i^2\frac{\mu_j}{N-\mu_i} = -\mu_i(1-\frac{\mu_i}{N})\frac{\mu_j}{N-\mu_i} = -\frac{\mu_i\mu_j}{N}.\text{ E. S.}$$

42.3. Slutsky explains Pearson's calculation of *R* and its minors. O. S.

42.4. A check of a mathematical theorem, if proved rigorously, can only mean checking that its assumptions are being fulfilled. See also Slutsky's own adjoining explanation below. O. S.

42.5. With the number of dice (0, 1, 2, …, 12) showing 5 or 6 points (or groups 0, 1, …, 12); the respective number of cases, theoretical (μ) and observed (*m*), and the deviations $e = m - \mu$. O. S.

42.6. It shows groups 0, 1, …, 12; e^2; e^2/μ. O. S.

42.7. Slutsky mentions the *theory of probability* (singular) whereas the proper Russian term is *theory of probabilities*. O. S.

42.8. The table shows stature in inches 52 – 53, 53 – 54, …, 70 – 71; and the number of mothers, actual and theoretical. O. S.

42.9. An unsubstantiated statement. O. S.

Bibliography

It does not cover the Foreword

K. Pearson

(1896a), Skew variation in homogeneous material. *Phil. Trans. Roy. Soc.*, vol. A186, pp. 343 – 414.

(1896b), Regression, heredity and panmixia. *Phil. Trans. Roy. Soc.*, vol. A187, pp. 253 – 318.

(1899), On the reconstruction of the stature of prehistoric races. *Phil. Trans. Roy. Soc.*, vol. A192, pp. 169 – 244.

(1900), On a 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. *London, Edinburgh and Dublin Phil. Mag.*, vol. 50, pp. 157 – 175.

(1901), Supplement to the memoir on skew variation. *Phil. Trans. Roy. Soc.*, vol. A197, pp. 443 – 459.

(1902a), On the mathematical theory of judgement with special reference to the personal equation. *Phil. Trans. Roy. Soc.*, vol. A198, pp. 235 – 299.

(1902b), On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, vol. 1, pp. 261 – 264.

(1902c), On the systematic fitting of curves to observations and measurement. *Biometrika*, vol. 1, pp. 265 – 303; vol. 2, pp. 1 – 23.

(1903a), On the probable errors of frequency constants. *Biometrika*, vol. 2, pp. 273 – 281.

(1903b), On an elementary proof of Sheppard's formulae for correcting raw moments and other allied points. *Biometrika*, vol. 3, pp. 308 – 312.

(1905a), "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson". A rejoinder. *Biometrika*, vol. 4, pp. 169 – 212.

(1905b), *On the General Theory of Skew Correlation and Non-Linear Regression. Drapers' Co. Res. Mem.*, Biometric ser. 2, 54 pp.

(1907), *On Further Methods of Determining Correlation. Drapers' Co. Res. Mem.*, Biometric ser. 4, 39 pp.

(1948, 1956), *Early Statistical Papers*. Cambridge. Contain papers 1896b, 1899, 1905b and joint paper Pearson & Filon (1898).

K. Pearson, Joint Author

Lee, Alice & **Pearson, K.** (1897), On the relative variation and correlation in civilized and uncivilized races. *Proc. Roy. Soc.*, vol. 61, pp. 343 – 357.

Pearson, K. & Filon, L. N. G. (1898), On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Phil. Trans. Roy. Soc.*, vol. A191, pp. 229 – 311.

Pearson, K. & Lee, Alice (1897), On the distribution of frequency (variation and correlation) of the barometric heights at divers stations. *Phil. Trans. Roy. Soc.*, vol. A190, pp. 423 – 469.

--- (1903), On the law of inheritance in man. *Biometrika*, vol. 2, pp. 357 – 462.

Wright, A., Lee, Alice & Pearson, K. (1907), A cooperative study of queens, drones and workers in Vespa vulgaris. *Biometrika*, vol. 5, pp. 407 – 422.

Other Authors

Blakeman, J. (1905), On tests for linearity of regression in frequency distributions. *Biometrika*, vol. 4, pp. 332 – 350.

Boscovich, R. J. (1758), *Philosophiae naturalis theoria*. Chicago – London, 1922. Latin – English edition.

Chuprov, A. A. (1909, 1910), *Ocherki po Teorii Statistiki* (Essays on the Theory of Statistics). Moscow, 1959.

--- (report,1918; 1926, in Swedish), The theory of stability of statistical series. In author's book (2004), *Statistical papers and Memorial Publications*, pp. 74 – 90. Berlin. Also at www.sheynin.de

Cournot, A. A. (1843), *Exposition de la théorie des chances et des probabilités*. Paris, 1984.

--- (1872), *Considérations sur la marche des idées. Oeuvr. Compl.*, t. 4. Paris, 1973.

Dale, A. I. (2003), *Most Honourable Remembrance. The Life and Work of Thomas Bayes*. New York.

Davenport, C. B. (1899, 1904, 1914), *Statistical Methods*. New York – London.

Dodge, Y., Editor (2003), *Oxford Dictionary of Statistical Terms*. Oxford.

Edgeworth, F. J. (1892), Correlated averages. *London, Edinburgh and Dublin Phil. Mag.*, vol. 34, pp. 190 – 204. Not included in author's *Writings in Probability, Statistics and Economics*, vols 1 – 3. Cheltenham, UK – Brookfield US, 1996.

Editorial (1908), On a formula for determining Γ(*x* + 1). *Biometrika*, vol. 6, pp. 118 – 119.

Elderton W. P. (1902), Tables for testing the goodness of fit of theory to observation. *Biometrika*, vol. 1, pp. 155 – 163.

Forsyth, A. R. (1883), On an approximate expression for *x*! *Rept. Brit. Assoc. Advancement of Science.*

Gibson W. (1906), Tables for facilitating the computation of probable errors. *Biometrika*, vol. 4, pp. 385 – 393.

Harris, J. A. (1909), A short method of calculating the coefficient of correlation in case of integral variates. *Biometrika*, vol. 7, pp. 214 – 218.

Helmert, F. R. (1905), Über die Genauigkeit der Kriterion des Zufalls bei Beobachtungsreihen. *Sitgz.-Ber. Preuss. Akad. Wiss. Berlin*, Hlbbd 1, pp. 594 – 612. Reprinted in author's *Akademie Vorträge*, pp. 189 – 208. Frankfurt/Main, 1993.

Heron, D. (1910), On the probable error of a partial correlation coefficient. *Biometrika*, vol. 7, pp. 411 – 412.

Hooker, R. H. (1901a), Correlation of the marriage-rate with trade. *J. Roy. Stat. Soc.*, vol. 64, pp. 485 – 492.

--- (1901b), The suspension of the Berlin produce exchange and its effect upon the corn prices. Ibidem, pp. 574 – 604 + discussion.

--- (1905), On the correlation of successive observations. *J. Roy. Stat. Soc.*, vol. 68, pp. 696 – 703.

--- (1907), Correlation of the weather and crops. *J. Roy. Stat. Soc.*, vol. 70, pp. 1 – 42.

Kovalevsky, G. (1911), *Osnovy Differenzialnogo i Integralnogo Ischislenii* (Elements of Differential and Integral Calculus). Odessa.

Lakhtin, L. K. (1903), On the Pearson method in the applications of the theory of probability to problems in statistics and biology. *Matematich. Sbornik*, vol. 34, pp. 481 – 500. In Russian.

Laplace, P. S. (1812), *Théorie analytique des probabilités. Oeuvr. Compl.*, t. 7. Paris, 1886.

Leontovich, A. (1909 – 1911), *Elementarnoe Posobie k Primeneniu Metodov Gaussa i Pirsona pri Otsenke Oshibok v Statistike i Biologii* (Elementary Manual on Application of the Methods of Gauss and Pearson for Estimating Errors in Statistics and Biology). Kiev. Tables, to which Slutsky referred many times, are in pts 2 and 3 (1911).

Lorentz, H. A. (1900, 1907), *Lehrbuch der Differenzial- und Integralrechnung*. Leipzig.

Macdonell, W. R. (1902), On criminal anthropometry. *Biometrika*, vol. 1, pp. 177 – 227.

Markov, A. A. (1900), *Ischislenie Veroiatnostei* (Calculus of Probability). Later editions: 1908, 1913, 1924. German translation: Leipzig – Berlin, 1912.

Nekrasov, P. A. (1912), *Teoria Veroiatnostei* (Theory of Probability). Moscow. 2nd edition.

Newcomb, S. (1908), Considerations on the form and arrangement of new tables of the Moon. *Monthly Notices Roy. Astron. Soc.*, vol. 68, pp. 538 – 544.

Pearl, R. (1906), The calculation of the probable errors of certain constants of the normal curve. *Biometrika*, vol. 5, p. 190.

--- (1908), On certain points concerning the probable error of the standard deviation. *Biometrika*, vol. 8, pp. 112 – 117.

Powys, A. O. (1901 – 1905), Data for the problems of evolution in man. *Biometrika*, vol. 1, pp. 30 – 49; vol. 4, pp. 233 – 285.

Quetelet, A. (1870), Des lois concernant le développement de l'homme. *Bull. Acad. Roy. Sci., Lettr., Beaux Arts Belg.*, 39e année, t. 29, pp. 669 – 680.

Schmitz, O. (1903), *Die Bewegung der Warenpreise in Deutschland von 1851 bis 1902*. Berlin.

Sheppard, W. F. (1898), On the application of the theory of error to cases of normal distribution and normal correlation. *Phil. Trans. Roy. Soc.*, vol. A192, pp. 101 – 167.

--- (1900), Some quadrature-formulas. *Proc. London Math. Soc.*, vol. 32, pp. 258 – 277.

--- (1903), New tables of the probability integral. *Biometrika*, vol. 2, pp. 174 – 190.

Sheynin, O. (1977), Laplace's theory of errors. *Arch. Hist. Ex. Sci.*, vol. 17, pp. 1 – 61.

--- (1995), Helmert's work in the theory of errors. Ibidem, vol. 49, pp. 73 – 104.

--- (1999), Slutsky: fifty years after his death. *Istoriko-Matematich. Issledovania*, vol. 3 (38), pp. 128 – 137. Engl. translation: in author's *Russian Papers on the History of Probability and Statistics*. Berlin, pp. 222 – 240. Also at www.sheynin.de

--- (1980), On the history of the statistical method in biology. *Arch. Hist. Ex. Sci.*, vol. 22, pp. 323 – 371.

--- (2003), Nekrasov's work on the central limit theorem. Ibidem, vol. 57, pp. 337 – 353.

--- (2006), Markov's work on the treatment of observations. *Hist. Scientiarum*, vol. 16, pp. 80 – 95.

--- (2010), Karl Pearson one and a half century after his birth. *Math. Scientist*, to appear.

Smirnov, N. V. & Dunin-Barkovski, I. W. (1959, in Russian), *Mathematische Statistik in der Technik.* Berlin, 1969, 1973.

Student (1908), On the probable error of a correlation coefficient. *Biometrika*, vol. 6, pp. 302 – 310.

Tutubalin, V. N. (1973), *Statisticheskaia Obrabotka Riadov Nabliudenii* (Statistical Treatment of Series of Observations). Moscow.

Veselovslky, B. (1909), *Istoria Uezdnykh Zemstv za Sorok Let* (History of District Zemstvo during Forty Years), vol. 1. Petersburg.

Yule, G. U. (1897a), On the theory of correlation. *J. Roy. Stat. Soc.*, vol. 60, p. 812.

--- (1897b), On the significance of Bravais' formulae for regression in the case of skew correlation. *Proc. Roy. Soc.*, vol. 60, pp. 477 – 489.

--- (1899), An investigation into the causes of changes in pauperism in England. *J. Roy. Stat. Soc.*, vol. 62, pp. 249 – 295.

--- (1907), On the theory of correlation for any number of variables, treated by a new system of notation. *Proc. Roy. Soc.*, vol. 79, pp. 182 – 193.

---(1909), The applications of the method of correlation to social and economic statistics. *J. Roy. Stat. Soc.*, vol. 72, pp. 721 – 730.

--- (1971), *Statistical Papers*. London. Contain paper 1897a.