

The missing early history of contingency tables (*)

STEPHEN STIGLER ⁽¹⁾

RÉSUMÉ. — L'apparente absence d'une histoire des tableaux de contingence avant 1900 est explorée, et trois hypothèses sont avancées pour expliquer cette absence de développements anciens sur le sujet : difficultés de calcul, manque de données, et obstacles conceptuels. Finalement, l'importance du modèle de quasi-symétrie dans cet historique est brièvement notée.

ABSTRACT. — The apparent lack of a history of contingency tables before 1900 is explored, and three explanatory hypotheses are considered for the absence of early development on this topic: computational difficulty, lack of data, and conceptual hurdles. Finally, the importance to this history of the model of quasi-symmetry is briefly noted.

1. Introduction

Any student of the history of contingency tables must have been struck by the lack of an early history. In some respects the study of the history of contingency tables is like a background check on a well-prepared secret agent: As one moves backwards in time from the present, the subject is remarkably well-documented – perhaps even too well accounted for – but a point is reached where suddenly all traces of a past disappear: it is as if the subject was at that point created out of whole cloth, born in a relatively mature state, with no trace of parents, schooling, or a struggle for

(*) Reçu le 18 septembre 2001, accepté le 18 septembre 2002

(1) Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A.
E-mail: stigler@galton.uchicago.edu

life. Since about 1900, at least the outline of the story is well-known: The Chi-square test was published in 1900, then there were a few works by Karl Pearson and G. Udny Yule (in the early 1900s), by Maurice Bartlett (1930s), by Deming and Stephan (1940s), by Lancaster and Roy and Kastenbaum (1950s), and then a brief quiet period until the 1960s, followed by an explosion. Birch, Caussinus, Darroch, Good, Goodman, Bishop, and Mosteller (1960s), then Haberman and Bishop, Fienberg, and Holland, and Goodman again (1970s), then McCullagh and Nelder and generalized linear models, and so forth. But unlike other branches of statistical analysis, there are apparently no early traces. Linear models, least squares, maximum likelihood, Bayesian inference, game theory, asymptotic theory – all of these have easily identified antecedents going back at least to the 18th century. Yet what many would consider to be the simplest of data summaries, say a 3×3 table of cross-classified counts, has apparently no papers that can prove its existence before about 1900. How could this branch of statistical analysis be so different from others? Was there really no early history, and if not, why not?

2. The exception: The 2×2 table

In fact, there was some early history, but it is remarkably limited. The 2×2 or fourfold table has a long and honorable ancestry in logic and in probability. It is based upon a dichotomous classification scheme that goes back at least to Aristotle. And there are early appearances of 2×2 classifications in probability: if n balls are drawn without replacement from an urn of M red balls and N black balls, and k of those drawn are red, $n - k$ black, this is essentially a 2×2 cross-classification (red-black, drawn-not drawn). In that respect the probabilistic background for the analysis of contingency tables could be said to go back hundreds of years. And the statistical use is fairly ancient as well: In the 19th century the hypergeometric analysis was refined by such statisticians as Bienaymé (Heyde and Seneta, 1977), and even subjected to a Bayesian analysis by Ostrogradsky (Seneta, 2001) and in a medical setting by Liebermeister (1877; see Seneta, 1994, for a good discussion of Liebermeister). Those analyses may be considered as concerned with 2×2 tables with fixed marginal totals; Laplace, Poisson, and Cournot all also treated 2×2 tables with only one margin fixed when they compared binomial proportions, and Gavarret and others applied these methods to medical data from 1840 on. There was also a scattering of late 19th century attempts to measure association for 2×2 tables, some involving appeals to probability. The redoubtable philosopher Charles Sanders Peirce published one such measure in 1884, as a contribution to an ongoing discussion of the ability to predict tornados. Goodman and Kruskal (1959) and Kruskal (1958) review Peirce's and a number of other early such examples.

3. Beyond the 2×2 table

But the exception is a curiously narrow one. When we look beyond the special case of a 2×2 table, with its unique ability to invite probabilistic thinking, with no scope for an ordinal analysis, with the data easily reducible to a single number and the marginal totals, we are again confronted with the dilemma: Why is there no history for contingency tables beyond this simple, almost trivial case? I cannot claim to have made a systematic search of the literature, but examples that even begin to address the question are few. In 1892 Francis Galton groped toward a methodology for larger tables in his study of fingerprints. In considering a 3×3 table, he defined a baseline from which to measure association in terms of the expected counts for cell (i, j) when the categories were independent, which he computed exactly as we might today from the marginal totals:

Expected count $(i,j) = (\text{marginal total } i) \times (\text{marginal total } j)/(\text{grand total}).$

But his analysis was rudimentary, consisting only of measuring where the separate diagonal entries fell, what percent of the distance they were between these expected counts and the highest feasible counts for the given marginal totals. He made no attempt to indicate when a discrepancy from the baseline was too large to be due to chance (Stigler, 1999, ch. 6).

In trying to account for the lack of an early history of contingency tables, these three hypotheses present themselves:

1. The development was hindered by computational difficulties; since all modern contingency analyses require extensive computation with statistical packages, perhaps the absence of computers was sufficient to hinder development of statistical methods.
2. There were no appropriate data available; if multi-way classifications of counts were not made, early analysts would have lacked the incentive to develop methods for their analysis.
3. Conceptual difficulties barred the way, more so than in other areas of statistical analysis.

I shall consider these hypotheses in sequence.

4. Computational problems?

The relatively heavy computation involved in modern contingency table analysis makes this hypothesis initially attractive. How could early researchers have coped with multivariate likelihood surfaces requiring iterative

solutions? But even a cursory study of the history of other areas shows the error of this way of thinking. The Newton-Raphson algorithm and Gaussian elimination were not devised for modern computers but to compensate for their lack. Daniel Bernoulli was using iteratively reweighted means in 1769 (Stigler, 1999, ch. 16), and by the 1850s Gaussian elimination methods had been streamlined to the point where some brave geodesists were tackling problems requiring solution by least squares with 60 unknowns (Stigler, 1986, p. 158). Computation was burdensome, but it was not a prohibitive barrier in any other area. Had the need arisen I have no doubt that an enterprising statistician in the 19th century would have employed iterative proportional fitting without the helpful instruction of Deming and Stephan (1940).

5. No data available?

On the face of it, this sounds absurd – how could there ever have been a shortage of simple counts? We learned to count before we developed higher analysis, and the urge to count is ubiquitous – surely the compiling of tables of counts, arrayed in various ways, some of them in cross-classifications, must predate all other forms of statistical analysis. And of course this is correct, and evidence of it is available from the clay tablets of Mesopotamia to the statistical registers of 19th century Prussia. But it is also misleading. Anyone who expects to be able to pull down a random volume from a 18th or 19th century scientific journal, say of any of the European academies of science or statistical societies, and to find easily a table of cross-classified counts, will be sorely disappointed. Such tables exist, but they are quite rare and hard to find, whether one looks in published scientific papers or in compendia of the data of a sovereign state. The examples that can be found are curiously revealing; however, and they will point to an explanation, indeed to an answer to the question I posed at the beginning.

Consider, for example, the work of the leading social statistician of the mid-nineteenth century, Adolphe Quetelet. Quetelet pioneered in the fitting of normal curves to grouped data on counts. His most famous data set was a collection of measurements of the chest circumferences of a large group of Scottish soldiers; see Table 1.

Table 1. — Quetelet’s data on the chest circumferences of Scottish soldiers, as presented in 1846 (Stigler, 1986, p. 207).

Inches	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	Total
Counts	3	18	81	185	420	749	1073	1079	934	658	370	92	50	21	4	1	5738

As it stands, this is a fine set of one-dimensional data; furthermore, it fits a normal curve quite well, as Quetelet demonstrated. But where did he get these data? He gave as his source the *Edinburgh Medical and Surgical Journal* for 1817, and indeed the data are given on pages 260 – 264. But the form in which those data are given is a surprise! The article reports both chest measurements and heights of the soldiers, cross-classified, for each of 11 different local militias. They were effectively given as a three-way contingency table with two ordinal dimensions and one nominal dimension: chest by height by regiment. Table 2 gives one regiment’s data.

Table 2. — Data from the East Stirlingshire Regiment of Local Militia of 647 Men, as given by the *Edinburgh Medical and Surgical Journal* (1817). This table, one of eleven for different regiments that were given, was compiled by an army contractor in the course of taking measurements for uniforms. The table gives that regiment’s cross-classification by height and chest circumference.

		33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5’4” & 5’5”	67		1	1	9	16	14	15	6	4	1						
5’6” & 5’7”	208			1	5	16	52	65	38	22	7	2					
5’8” & 5’9”	207					1	4	31	57	42	47	14	10	1			
5’10” & 5’11”	125				1	4	4	12	24	43	24	7	4	1	1		
6’0” & 6’1”	40						1	1	1	9	7	6	9	2	4		

In a later era, a Galton would have related such data to a bivariate normal distribution; Quetelet was content to sum the columns and then sum over the seven tables to get the marginal totals for chest circumferences. (Or that was his goal – he made minor errors and seems to have slightly inflated the number of soldiers, increasing it from 5732 to 5738, although the contractor’s figures are not entirely consistent either; see Stigler, 1986, p. 206.)

This was not the only time Quetelet confronted a cross-classification. In his classic 1835 book *Sur l’homme* he had investigated propensities for crime by looking at conviction rates, classified by sex, by type of crime, by education level, by age, by year. His investigation of these rates was as close to an analysis of a multi-way contingency table as is to be found before 1900 (see Stigler, 1986, pp. 174 – 182, for a full account of his methodology), but it too fell short in concentrating on marginal characteristics, and in reducing the analysis to dichotomous contrasts. Indeed, Quetelet’s works abound in data that had at one time (or could have at one time) been cross-classifications, but which are presented in reduced form, as frequency distributions or as lists of ratios or rates. Just so in other works of that

century – in the works of Lexis or von Bortkiewicz. At the end of the century, Bortkiewicz did give a cross-classification in presenting one of the most famous data sets of all time, the deaths by horse kick in the Prussian cavalry. His table (Bortkiewicz, 1898) gave the deaths classified by year and corps (20 years and 14 corps), but his analysis did not exploit that cross-classification, rather it famously looked at the unclassified counts and presented their frequency distribution compared to what might be expected from a Poisson distribution. And the table he did give was itself a drastically marginalized version of the published Prussian military data, where each year's data were given in a separate large volume, and in the relevant table on accidental deaths, more than 3 dozen causes were tabled in addition to horse kick, and the deaths were also broken down by place of occurrence, and by victim's duty status, religion, and place of birth.

These examples could be multiplied. Published tables of cross-classified counts were quite rare before 1900, but published data that either were derived from cross-classified counts or were derived from data that could have been cross-classified, were not. The derivation typically took the form of reduction to marginal totals or rates or ratios. The lesson seems to be clear: There was no shortage of data restricting the development of contingency tables. In fact, I suggest that quite the contrary was true: The lack of models and analyses for multiply-classified data restricted the data that were considered worthy of discussion and publication. Students in elementary statistics courses frequently conclude from the course presentation that the form of the data determines the analysis, when the reverse is more often the case: *The contemplated analysis determines the data.*

6. Conceptual hurdles!

An apparent lack of interest in a statistical question may simply mean that no one has asked the question. Or it may mean that either no one has put the question in a way that it can be answered, or that the tools were not evident that could solve the question – the problem was too hard for the available tools. I would refer to either of these as conceptual hurdles, and both seem to have played a role in the slow development of contingency theory both before and after 1900.

In simple 2×2 cases, the question of scientific interest was frequently clear, and after the tools of Laplace and Poisson were developed there was no hurdle to be overcome. For example, in 1840, Gavarret presented tables in the form of Table 3.

Faced with what amounted to a simple 2×2 table,

	Legitimate Children	Illegitimate Children	Totals
Male	939641	71661	1011302
Female	877931	68905	946836
Totals	1817572	140566	1958138

which clearly presented the question as one of comparing the two fractions of male births $0.51697 - 0.50980 = 0.00717$, Gavaret felt no compunction in adopting Poisson's methods and comparing this difference with $2\sqrt{\frac{2mn}{\mu^3} + \frac{2m'n'}{\mu'^3}} = 0.00391$. We would now describe this comparison value as $2\sqrt{2} = 2.828$ times the estimated standard deviation of the difference of the ratios, a procedure that would only differ essentially from a typical Chi-squared test in that the standard deviation was not estimated under the null hypothesis of equal ratios (this has but a trifling effect in this case) and that a rather severe significance level was effectively being used (essentially $\alpha = 0.46\%$). But Gavaret's conclusion, that the observed difference 0.00717 was larger than could be attributed to experimental variation, was consistent with modern methods.

Table 3. — Gavaret's presentation of data on the sex ratio for legitimate and illegitimate births (Gavaret, 1840, p.274).

1824–1825	
<i>Enfants légitimes</i>	<i>Enfants illégitimes</i>
<p>$m = 939641 =$ le nombre de garçons. $n = 877931 =$ le nombre des filles. $\mu = 1817572 =$ le nombre des naissances.</p> <p>D'où résulte que la chance moyenne de naissance d'un garçon en France dans l'état de mariage, est représentée par le rapport</p> $\frac{m}{\mu} = \frac{939641}{1817572} = 0, 51697$ <p>En poussant l'approximation jusqu'à la cinquième décimale.</p>	<p>$m' = 71661 =$ le nombre de garçons. $n' = 68905 =$ le nombre des filles. $\mu' = 140566 =$ le nombre des naissances.</p> <p>D'où résulte que la chance moyenne de naissance d'un garçon en France hors l'état de mariage, est représentée par le rapport</p> $\frac{m'}{\mu'} = \frac{71661}{140566} = 0, 50980$ <p>En poussant l'approximation jusqu'à la cinquième décimale.</p>

With larger tables, though, there were serious difficulties. The greatest of these was conceptual, but one was technical: How do you assess the significance of deviations in a list of counts? Earlier work on the binomial (going back to De Moivre) pointed the way to dealing with binomial counts

and proportions, even, as Laplace and Poisson had showed, with pairs of binomial proportions, but what about multinomial proportions? There the multiple comparison problem raised its head; the higher the dimension, the more discrepancies were being evaluated. The struggles with this technical problem in the 1890s culminated with the publication of Karl Pearson's Chi-squared test in 1900 (Stigler, 1986, p. 329; Plackett, 1983), and this led in short order to the application of the test to contingency tables by 1904. Except for the early troubles with understanding and computing degrees of freedom (Stigler, 1999, ch. 19), the subject stood by 1915 the way it does now in all widely used general introductory statistical texts.

7. The largest hurdle

The greatest conceptual problem was a very basic one: With multiply-categorized tables of counts, what questions should be asked? The answers of the 19th century were basically “look at margins of the table and/or dichotomize the classifications,” and this did not carry the enterprise far. With Pearson's Chi-squared test the possibility of comparing an array of counts with a simple baseline corresponding to the hypothesis of independence presented itself immediately (or even earlier: as we have seen Galton had considered that baseline in 1892). But aside from the introduction of the concept of degrees of freedom by R. A. Fisher in the 1920s, the situation languished for about half a century much the way it had been in 1904. In the 1930s Fisher made explicit the gains to be had from making the analysis conditional upon the marginal totals, considering them as “ancillary” statistics, although the area remained somewhat mysterious, since as Galton sensed in 1892 and Plackett showed in 1977 (and Fisher no doubt realized), the marginal totals were not entirely innocent of information about the association in the body of a table (Stigler, 2001). In 1935 Bartlett brought some clarity to the understanding of the differing types of partial independence that could be considered in high dimensional tables, with a test for second order interaction.

The problem was similar to one faced in the development of time series. In high dimensions it is easy to describe the hypothesis of independence, but that hypothesis is seldom of interest except as a baseline – multivariate data are usually considered as multivariate precisely because their correlations or associations are known or suspected to be substantial and of interest. But once you leave the hypothesis of independence behind there is seemingly no limit to the possibilities. The vast mathematical space between independence and complete dependence can be filled in too many ways, most of them of no interest to either mathematician or applied scientist. The challenge is to find intermediate models that are sufficiently general to

adequately encompass a broad class of applications, yet sufficiently specific to be both amenable to mathematical analysis and to permit interesting questions to be addressed and the results interpreted. In time series the temporal structure of the series helped point the way relatively early on, to the autoregressive models of G. Udny Yule and Eugeny Slutsky in the 1920s, for example. Lacking that structure, the problem faced with contingency tables was in some ways much more difficult.

8. Conclusion

An examination of how this hurdle was overcome in the 1960s and 1970s is beyond the scope of this paper, but for this occasion some remarks on the role of the model of quasi-symmetry are warranted. Early in the 20th century some advances had been made, even by Karl Pearson, in building models that incorporated special structure, such as tables of triangular form, with models now called “quasi-independence” (Stigler, 1999, ch. 19). In the 1960s what has proved to be an important step was taken with the formulation of models for a different special situation with the hypothesis of quasi-symmetry for square tables, particularly in Henri Caussinus’s 1965 dissertation, published in 1966, but also in early work of Leo Goodman and William Kruskal and in Goodman’s (1965) loglinear analyses of social mobility. Quasi-symmetry helped to fill the mathematical space beyond independence, by tailoring its questions to the particular structure of the most frequently encountered class of square tables, those where rows and columns have essentially the same labels. The model is remarkable for many reasons, including the breadth of its applications and its connections with other approaches. In 1976 Stephen Fienberg and Kinley Larntz showed it encompassed the paired comparison model of Ralph Bradley and Milton Terry, which has been traced back to Ernst Zermelo (1929).

Square tables with either symmetry or marginal homogeneity are sufficiently uncommon that it would perhaps not be expected that the simple step of moving to quasi-symmetry, the step of supposing that, but for the lack of marginal homogeneity, the table of mean counts would be symmetric, would greatly enlarge the sphere of applications. And yet that is exactly what has been found. The relationship with paired comparisons models affords simple ways of interpreting the parameters – in effect the $n \times n$ table can be characterized by a set of scores attached to the row (or column) labels, reducing the high dimensional table to ordered scores, a remarkable simplification where it works. And there is more: Accompanying this reduction is an avoidance of inconsistencies when categories are combined through aggregation (Stigler, 1994, p. 102)!

Quasi-symmetry does not answer all questions – far from it – but its unexpectedly broad set of applications and the elegance of the interpretations available when it does apply, must count it as one of the great success stories in the analysis of contingency tables in the century since the invention of the Chi-squared test.

Bibliography

- [1] BARTLETT (M. S.). — Contingency table interactions, *Supplement to the Journal of the Royal Statistical Society*, 2 (1935), 248–252.
- [2] BISHOP (Y. M. M.), FIENBERG (S. E.), HOLLAND (P. W.). — *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.
- [3] VON BORTKIEWICZ (L.). — *Das Gesetz der Kleinen Zahlen*, Teubner, Leipzig, 1898.
- [4] CAUSSINUS (H.). — Contribution à l'analyse statistique des tableaux de corrélation, *Annales de la Faculté des Sciences de Toulouse*, 29 (année 1965), (1966), 77–183.
- [5] DEMING (W. E.), STEPHAN (F.). — On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, 11 (1940), 427–444.
- [6] *Edinburgh Medical and Surgical Journal*. — Statement of the sizes of men in different counties of Scotland, taken from the local militia, 13 (1817), 260–264.
- [7] FIENBERG (S. E.), LARNTZ (K.). — Loglinear representation for paired and multiple comparison models, *Biometrika*, 63 (1976), 245–254.
- [8] GALTON (F.). — *Finger Prints*, Macmillan, London, 1892.
- [9] GAVARRET (J.). — *Principes généraux de statistique médicale*, Béchét Jne et Labé, Paris, 1840.
- [10] GOODMAN (L. A.). — On the statistical analysis of mobility tables, *American Journal of Sociology*, 70 (1965), 564–585.
- [11] GOODMAN (L. A.), KRUSKAL (W. H.). — Measures of association for cross classifications II: Further discussion and references, *Journal of the American Statistical Association*, 54 (1959), 123–163.
- [12] HEYDE (C. C.), SENETA (E.). — *I.J. Bienaymé. Statistical Theory Anticipated*, Springer-Verlag, New York, 1977.
- [13] KRUSKAL (W. H.). — Ordinal measures of association, *Journal of the American Statistical Association*, 53 (1958), 814–861.
- [14] LIEBERMEISTER (C.). — Ueber Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik, *Sammlung Klinischer Vorträge in Verbindung mit deutschen Klinikern*, 110 (20th Heft der 4th Serie) (1877), 935–962.
- [15] PEIRCE (C. S.). — The Numerical measure of the success of prediction, *Science*, 4 (1884), 453–454.
- [16] PLACKETT (R. L.). — The marginal totals of a 2×2 table, *Biometrika*, 64 (1977), 37–42.
- [17] PLACKETT (R. L.). — Karl Pearson and the Chi-squared test, *International Statistical Review*, 51 (1983), 59–72.
- [18] QUETELET (A.). — *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*, Bachelier, Paris, 1835.
- [19] SENETA (E.). — Carl Liebermeister's hypergeometric tails, *Historia Mathematica*, 21 (1994), 453–462.

- [20] SENETA (E.). — M. V. Ostrogradsky as probabilist, in *Mykhailo Ostrohrads'kyi (Mikhail Ostrogradsky), Honoring his bicentenary*, pages 69–81, Institute of mathematics, Kyiv, Ukraine, 2001.
- [21] STIGLER (S. M.). — *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, Mass, 1986.
- [22] STIGLER (S. M.). — Citation patterns in the journals of statistics and probability, *Statistical Science*, 9 (1994), 94–108.
- [23] STIGLER (S. M.). — *Statistics on the Table*, Harvard University Press, Cambridge, Mass, 1999.
- [24] STIGLER (S. M.). — Ancillary history, in M. C. M. de Gunst, C. A. J. Klaassen and A. W. van der Vaart , editors, *State of the Art in Probability and Statistics; Festschrift for Willem R. van Zwet*, Lecture Notes–Monograph Series, pages 555–567, Institute of Mathematical Statistics, 2001.
- [25] YULE (G. U.). — *An Introduction to the Theory of Statistics*, Charles Griffin, London, first edition, 1911, and many later editions.
- [26] ZERMELO (E.). — Die Berchnung der Turnier-Ergrbnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift*, 29 (1929), 436–460.