ANCILLARY HISTORY

STEPHEN M. STIGLER

University of Chicago

Ancillarity has long been a shadowy topic in statistical theory. — D.R. Cox (1982).

AMS subject classifications: 62-03.
Keywords and phrases: Conditional inference, Fisher, Galton, Edgeworth, Laplace.

1 Introduction

The origin of the term "ancillary statistics" is clear and well known. It was introduced in 1925 by Ronald A. Fisher in his paper "Theory of Statistical Estimation" (Fisher, 1925); it then lay dormant for nearly a decade until Fisher returned to the topic in his "Two new properties of mathematical likelihood," which was sent to the Royal Society of London in December 1933 and published as Fisher (1934). The term arose in these two papers in Fisher's characterization of statistical information and its relationship to the likelihood function. When a single sufficient statistic existed it would contain all of the information in the sample and serve as the basis for a fully efficient estimate, that estimate to be found from differentiating the likelihood function to find the maximum. When this was not the case, auxiliary or "ancillary" information was needed and could frequently be obtained from statistics arising from looking more closely at the likelihood in the neighborhood of the maximum, in particular at the second or higher order derivatives there.

Fisher expanded upon his earlier usage a year later, treating "ancillary" as a broader term of art not specifically wedded to local behavior of the likelihood function in "The Logic of Inductive Inference," read to the Royal Statistical Society on December 18, 1934 and published with somewhat acrimonious discussion as Fisher (1935). Partly as a result of this broadened view, the precise nature of the concept, and hence of its history both before and after the introduction of the term, has been elusive. In these early publications (and indeed also in later ones), Fisher explained the term most clearly by describing what "ancillary statistics" accomplished rather than what they were: They supplied auxiliary information to supplement the maximum likelihood estimate. In Fisher (1935) he wrote that when the best estimate fails to use all the information in the sample, when it "leaves a measurable amount of the information unutilized," he would seek to supplement the estimate to utilize that information as well. He asserted that "It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it. Their function is, in fact, analogous to the part which the *size* of our sample is always expected to play, in telling us *what reliance* to place on the result. Ancillary statistics are only useful when different samples of the same size can supply different amounts of information, and serve to distinguish those which supply more from those which supply less." No specific general guide was provided, although examples of their use were given, use that invariably involved conditional inference given the ancillary statistics.

In 1934 Fisher had included as a prime example the estimation of the location parameter of a double exponential distribution. There the maximum likelihood estimate, the sample median, is neither sufficient nor fully efficient. "The median is an efficient estimate in the sense of the theory of large samples, for the ratio of the amount of information supplied to the total available tends to unity as the sample is increased. Nevertheless, the absolute amount lost increases without limit." (Fisher, 1934, p. 300). By conditioning upon the sample spacings — what Fisher called the sample configuration — he was able to show in great detail that the median was conditionally efficient on average, and he noted that this conclusion extended to more general location-scale families (Hinkley, 1980).

A year later, Fisher (1935) illustrated the ancillarity idea through a new example, testing for homogeneity in a 2 × 2 table conditionally upon the marginal totals, an example that as we shall see introduced other complications to the discussion. In concluding that paper he indicated that ancillary statistics would be useful in the case, "of common occurrence, where there is no sufficient estimate." Then "the whole of the ancillary information may be recognized in a set of simple relations among the sample values, which I called the configuration of the sample." These statements were not clear to the audience at the time. The discussants who commented on this portion of his paper were distracted by other features of the example; only J. O. Irwin mentioned the term ancillary and then simply to say "it was not absolutely clear how one should define an ancillary statistic."

In a few scattered comments on the term in later writings, Fisher added little by way of elaboration. Some later writers, such as Cox (1958), Cox and Hinkley (1974, pp. 31-35), Lehmann and Scholtz (1992), and Welsh (1996, p. 383), have added clarity and specificity to the definition in cases such as where a minimal sufficient statistic exists; others, such as Basu (1964), Buehler (1982, with discussion), and Brown (1990, with discussion), have pointed to difficulties with the concept due to the non-uniqueness of ancillary statistics in some even well-structured parametric problems, or to paradoxes that can arise in a decision theoretic framework. Despite these misgivings and the vagueness of the definition, the notion has come to be key to powerful ideas of conditional inference: When an ancillary statistic can be found (usually taken to be a part of a sufficient statistic whose marginal distribution does not depend upon the parameter of interest), it is best (or at least prudent) to make inferences conditional upon the value of the ancillary statistic.

The goal here is not to explore the history of ancillarity subsequent to Fisher (1934, 1935), still less to attempt a rigorous and clear explication of the concept and its realm of appropriate application (for which see Fraser, 1979, Lehmann and Scholtz, 1992, Lloyd, 1992, and the recent book by Barndorff-Nielsen and Cox, 1994). Rather it is to present three earlier examples that bear on the understanding of the concept, examples which may help us better understand Fisher's idea as a not-fully crystallized recognition of a common thread in a variety of problems in statistical inference.

2 Laplace and the Location Parameter Problem, 1772-1777

It is common today, even where there is disagreement about the extent and usefulness of the idea of ancillarity, to adopt as sound statistical logic some of its consequences when considering location parameter problems. For example, in making inferences about μ in a random sample from a Uniform $[\mu - h, \mu + h]$ distribution with known h, where by inference we mean estimation and the assessment of the accuracy of the estimate of μ, we should condition on $D = X^{\max} - X^{\min}$, since the usual estimator $(X^{\max} + X^{\min})/2$ must invariably lie within $h - D/2$ of the unknown μ. Any assessment of the accuracy of this estimator that did not condition on the observed value of D could lead to absurd results (e.g. Welsh, 1996, p. 157). More generally (for other population distributions) we should assess accuracy conditional upon the residuals or the spacings between the observations. This practice has a long and distinguished provenance.

In subjecting the location parameter problem to formal treatment, notation is necessary, and the choices of notation will reflect, however imperfectly, conceptual understanding. One common choice today is to introduce a symbol for the target value, say μ, and then describe the n observations X_i in terms of μ and the errors of observation, say e_i, by $X_i = \mu + e_i$. The distribution of errors, a probability density, is represented by $\phi(e)$, and so the likelihood function is $\prod_{i=1}^{n} \phi(X_i - \mu)$.

This notation reflects in principle the approach taken by some early mathematical statisticians. For example, in 1755 Thomas Simpson worked with the errors and the error distribution in showing that an arithmetic mean would improve upon a single observation. Simpson's approach in terms of errors made the inverse step to theoretical statistical inference an easier one, as I have argued before (Stigler, 1986a, pp. 88ff.). Indeed, this approach underlies Fisher's fiducial probability and Neyman's confidence intervals. But it is not the only possible approach, nor, since the errors are not directly observable, is it even in practical matters necessarily the most natural. Others, and Laplace was a significant example, chose to frame the problem in a way where conditioning on ancillaries was much more tempting, in terms of the correction to be made to the first observation and the distances between the observations. This tendency was already present in Laplace's first serious memoir on mathematical statistics (Laplace, 1774; translated with commentary in Stigler, 1986b, see also Stigler, 1986a, pp. 105ff., and Hald, 1998, p. 176), but for present purposes it is clearer in a memoir Laplace read to the Académie des Sciences on March 8, 1777. The memoir remained unpublished until 1979 (Gillispie, 1979).

Laplace's memoir is unusual in presenting two approaches to the estimation problem, from two different, clearly delineated statistical vantage points. He explained that one might address the problem of choosing a mean from either an a priori perspective (before the observations have been made), or a posteriori (after they have been made). In the latter case — the one that concerns us here — he described the problem of choosing a mean as one of "determining a function of the observations a posteriori, that is to say taking account of the respective distances between the observations" (Laplace, 1777, p. 229). He provided interesting analyses from both perspectives leading to quite different results; we focus here upon the second.

Laplace began as we might now with the observations (he wrote a, a', a'', $\ldots$, where we write $X_1, X_2, X_3, \ldots$), but in one section of the memoir he re-expressed these data in a different notation. He let x denote the correction that would be applied to the first observation to arrive at the true value; in our notation $x = \mu - X_1$, so $X_1 + x = \mu$. And he let $q^{(1)}, q^{(2)}, q^{(3)}, \ldots$ represent the distances of the second and subsequent observations from the first. We could write these as $q^{(i)} = X_{i+1} - X_1$. The likelihood function would then become $\phi(-x)\phi(q^{(1)} - x)\phi(q^{(2)} - x)\cdots$. Laplace quoted his 1774 "principe général" for reasoning to inverse probabilities — what we would now describe as Bayes Theorem with a uniform prior distribution; see Stigler (1986a, pp.100ff.; 1986b). He concluded that the probabilities of the different values of the correction x given the respective distances between the observations $q^{(1)}, q^{(2)}, q^{(3)}, \ldots$ would be proportional to this same function, $\phi(-x)\phi(q^{(1)} - x)\phi(q^{(2)} - x)\cdots$. This agrees with the result that Fisher obtained in 1934 for the case of the double exponential or Laplace density $\frac{1}{2}e^{-|u|}$, as the conditional distribution of the difference between the median and the location parameter given the spacings. Fisher had noted that in general this "frequency distribution ... is the mirror image of the likelihood function." (Fisher, 1934, p. 303).

As an example Laplace considered a sample from a Uniform $[\mu-h,\mu+h]$ distribution, h known. He wrote,

> Suppose for example that the law of facility of errors is constant and equal to K, that it is the same for all observations, and that the errors are each taken between $t=-h$ and $t=h$; $a^{(n-1)}$ being the time fixed by the last [largest] observation, we set $a^{(n-1)}-M=h$ and $N-a=h$ [that is, $M=a^{(n-1)}-h$ and $N=a+h$, where a is the minimum observation]. It is clear that the true time of the phenomenon falls necessarily between the points M and N; further that the probability that each of the intermediate points will be this instant is proportional to K^n; ... and that the mean we need to choose, X, is evidently the midpoint of the line segment $(a,a^{(n-1)})$, and so in this case, to take the mean among n observations it is necessary to add to the smallest result half the difference between the smallest and the largest observations." (Laplace, 1777, p. 241)

He thus concluded that the posterior distribution for the true value was Uniform $[X^{\max}-h,X^{\min}+h]$, leading him to suggest the midrange (that is, the posterior mean) as a posterior estimate. Some of Laplace's language was suggestive of Fisher, particularly his conditioning upon the spacings between the observations ("en ayant égard aux distances respectives des observations enter elles"), which was echoed by Fisher's "configuration of a sample." Laplace's perspective was closer to a Bayesian analysis than a Fisherian fiducial one, but then perhaps so was Fisher's in his initial foray into likelihood-based inference in 1912, before he took great pains (not always successfully) to distinguish his approach from others from 1922 on; see Zabell (1989, 1992), Edwards, (1997a,b), Aldrich, (1997).

3 Edgeworth, Pearson, and the Correlation Coefficient

Another area in which the idea of ancillarity has been appealed to is in inference about the parameters of a bivariate normal distribution,where the values of (say) X may be treated as ancillary with respect to inference about $\mathrm{E}(Y \mid X)=aX+b$, justifying conditioning upon the X's (or sufficient statistics for the distribution of the X's) whether the X's are random or assigned by experimental design (see, for example, Cox and Hinkley, 1974, pp. 32-33). There is interesting historical precedent for this. In 1893 Francis Edgeworth considered the estimation of the correlation ρ of n bivariate normal pairs (X_i,Y_i), assumed centered at expectations and measured in standard units, effectively marginally $N(0,1)$ (Edgeworth, 1893; Stigler, 1986a, pp. 321-322). Of course in this case $\mathrm{E}(Y \mid X)=\rho X$. Edgeworth considered the pairs with the X's "assigned", that is he conditioned upon the X's, so that for X not equal to zero the conditional expected value of Y/X would be ρ, and the conditional variance of Y/X would be proportional to $1/X^2$. He then found the optimal weighted average of the Y/X's to be weighted by the X^2's, and he gave that as the "best" value for ρ:

$$\frac{\sum(X^2 \cdot Y/X)}{\sum(X^2)} = \frac{\sum XY}{\sum(X^2)}.$$

Three years later, Karl Pearson attacked the problem of estimating the parameters of a bivariate normal distribution directly as a bivariate estimation problem. Approaching the problem from the standpoint of inverse probability (but in a manner mathematically equivalent to maximum likelihood estimation), he was led to the estimate of the correlation $\sum(XY)/n\sigma_1\sigma_2$, where he had $\sigma_1^2 = \sum(X^2)/n$ and $\sigma_2^2 = \sum(Y^2)/n$, in the process blurring the distinction between these as statistics and as parameters (Pearson, 1896; Stigler, 1986a, pp. 342-343). Had Edgeworth similarly blurred this distinction (and to a degree he did, see Stigler, 1986a, p. 322), these estimates would seem to agree. But while Edgeworth noted this identity on several occasions, he stopped short of claiming priority. I have a reprint of Edgeworth's 1893 paper to which Edgeworth added a manuscript note after he had seen Pearson's work. He wrote,

> The value of ρ which I give at p. 101 is the most accurate on the assumption that the best value is a weighted mean of $y_1/x_1, y_2/x_2, \ldots$; Prof. Karl Pearson obtains the same result without that arbitrary assumption. I have proceeded like one who having to determine the most probable value of the modulus [i.e. standard deviation], for given observations, ranging under an ordinary Probability-curve [i.e. a normal density], assumes that the quaesitum [what is desired] is a function of *some* mean power of errors and then proves that the most accurate result is afforded by the *second* power; Prof. Karl Pearson has proceeded without any such assumption. F. Y. E. 1896.

Edgeworth made a similar, briefer and less specific, comment in print that same year (Edgeworth, 1896, p. 534).

Edgeworth had approached the estimation of ρ conditionally, conditioning upon the ancillary X's, but his method of inference was not Fisherian inference: he estimated ρ by a weighted average (effectively using least squares conditionally given the X's) rather than conditionally employing maximum likelihood. And there is a good reason why he would not have used maximum likelihood: For his specification of the problem, with marginal means equal to zero and marginal variances equal to one, the maximum likelihood approach leads to algebraic problems; neither the Pearsonian product moment estimator nor Edgeworth's version is maximum likelihood. For that restricted setting, the maximum likelihood estimator of ρ is the solution of a cubic equation that resists closed form expression (Johnson and Kotz, 1972,

p. 105). The same is true whether one proceeds conditionally given the X's (as may be sanctioned by appeal to Cox and Hinkley, 1974, pp. 34-35) or unconditionally. The difficulty stems from the fact that conditionally given the X's, not only is $\mathrm{E}(Y \mid X) = \rho X$, but the conditional variance $1 - \rho^2$ depends upon ρ as well. Edgeworth had avoided this problem (as he noted) by restricting the form of his estimator to a weighted average; Pearson (perhaps inadvertently) had avoided it by allowing the marginal variances to vary freely in his calculation. In any case, Edgeworth seemingly took conditional inference here for granted.

4 Galton and Contingency Tables

As I noted earlier, Fisher had in his 1935 paper enlarged upon his broadened descriptive definition of ancillary statistics with a quite different example, one that involved testing, not estimation: the application of the concept of ancillary statistics to 2 × 2 tables. He presented a cross-classification based upon 30 sets of twins (Table 1), where in each pair one twin was a known criminal and the remaining twin was then classified as convicted or not. He supposed for the purposes of the example that the data were "unselected" and asked if there was evidence here that the "causes leading to conviction" had been the same for the monozygotic as for the dizygotic twins.

	Convicted	Not Convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Table 1. Convictions of Like-sex Twins of Criminals. Lange's data, from Fisher (1935).

Fisher wrote,

> To the many methods of treatment hitherto suggested for the 2 × 2 table the concept of ancillary information suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies. (Fisher, 1935)

He went on to develop his conditional test, showing that the distribution of the table entries given the marginal totals was a hypergeometric distribution, independent of the probability of conviction under the hypothesis this is the same for both types of twin.

Over four decades earlier, Francis Galton had faced a similar table, and his analysis sheds interesting light upon Fisher's. In his study of fingerprints, Galton had inquired as to the propensity for related individuals to have similar patterns. As part of this study he presented the data in Table 2 on relationships between the patterns on the right fore-fingers of 105 sibling pairs (Galton, 1892, p. 172-176; Stigler, 1995; 1999, Chapter 6). To investigate the degree to which sibling pairs shared the same general pattern of fingerprint, Galton needed to test these data for evidence of association, to measure the degree to which the diagonal entries of this table exceed what they would be, absent any heritable link.

	A children			
B children	Arches	Loops	Whorls	Totals in B children
Arches	5	12	2	19
Loops	4	42	15	61
Whorls	1	14	10	25
Totals in A children	10	68	27	105

Table 2. Observed fraternal couplets (Galton, 1892, p. 175). The A sibling was distinguished from the B sibling in being the first "that happened to come to hand" (Galton, 1892, p. 172; presumably no pun was intended).

Recall that this was eight years before Karl Pearson introduced the Chi-square test, and 12 years before he applied it to testing independence in cross-classifications. Focussing entirely upon the diagonal, Galton constructed his own measure by first determining what the counts would be if the prints were paired at random. Thus for the first diagonal entry he found the number 19 × 10/105 = 1.7, for the second, 61 × 68/105 = 37.6, and for the third, 27 × 25/105 = 6.2. He labeled these "Random", and considered them as the baseline for comparison with the "Observed"; see Table 3. All of the "Observed" exceeded the "Random", but was the difference to be judged large enough to reject the "Random" hypothesis? Galton constructed a scale using "Random" as the baseline and measuring how large the "Observed" were in degrees on a centesimal scale, essentially as a percent of the distance to the "Utmost feasible" as determined from the marginal totals (this being the minimum of the two corresponding marginal totals). For these data the degrees are 40^o, 19^o, and 20^o. He made no attempt to assign a probability to such discrepancies.

Galton's procedure had one element in common with Fisher's, and it was an important one. His measure was, like Fisher's, conditional upon the marginal totals. The baseline values were, in common with all analyses since Karl Pearson, computed as the expected counts under the hypothesis of random assignment — independence between each of the pair of sibling's

	A and B both being		
	Arches	Loops	Whorls
Random	1.7	37.6	6.2
Observed	5.0	42.0	10.0
Utmost feasible	10.0	61.0	25.0

Table 3. Galton's test of independence for the fingerprint patterns of fraternal couplets. On Galton's centesimal scale, these observed counts are 40°, 19°, and 20° degrees above the random, higher than in other examples that were based upon a finer classification (Galton, 1892, p. 176).

patterns. Indeed, I do not know of an earlier example of this calculation of expected values, at least for tables larger than 2 × 2, although I have not made an extensive search. But there was one point where Galton departed from Fisher's program: he expressed a principled reservation about the appropriateness of one aspect this conditioning on the margins.

When Galton introduced this approach earlier in his book he had qualified it as follows: "Now consider the opposite extreme of the closest possible relationship, subject however, and this is the weak point, to the paramount condition that the average frequencies of the A. L. W. classes may be taken as *pre-established*." (Galton's italics, Galton, 1892, p. 126). To Galton there was a "self-contradiction" in the assumption that the analysis proceed conditionally on the observed marginal frequencies, a contradiction that constituted a "grave objection" to his procedure. The problem was that if the relationship were perfect and all the counts fell on the diagonal, the marginal totals should agree, but they did not. The problem was particularly apparent in Galton's example, where the row and column categories were the same; indeed, they were based upon the same population and — absent sampling variation — should have agreed. But the problem holds more generally. Even in Fisher's 2 × 2 table the fact that the row totals do not equal the column totals is prima facie evidence that the relationship is not a perfect one: The margins do contain information about the degree of association! Plackett (1977) has noted this with specific reference to Fisher's data, but there is a suggestion in Fisher's wording that he realized it as well. His statement was conditional is a way that is technically correct even though misleading: "*If it be admitted* that these marginal frequencies by themselves supply no information ..., we may recognize the information they supply as wholly ancillary" (emphasis added). An unsuspecting reader would read this as suggesting the supposition clearly held, and would be lured into granting the premise and so accepting the conclusion of ancillarity. For was that not the point of the example? As Plackett has shown, however, the amount of information in the margins is slight, so this conclusion is not seriously misleading in practice. On this point see Plackett (1977), and particularly

Barnard (1984) and Cox (1984). It is an extremely subtle point, and the fact that Galton picked up on it in 1892 is remarkable.

5 Conclusion

Fisher seems to have initially, in 1925, conceived of the ancillary statistics of a parametric inference problem as being that part of the likelihood function that varied from sample to sample but was not captured by the location of the maximum, more specifically as the second and higher derivatives of the likelihood at the maximum. By 1934 and 1935, with two quite different and vivid examples in hand that did not fit so easily (if at all) with his earlier conception, he broadened the definition and made it less specific — almost qualitative. Fisher had a powerful statistical intuition that worked best from exceedingly well chosen examples, and in this case his intuition led him to postulate a concept that indubitably worked well in his examples but resisted rigorous codification, just as fiducial probability has, and even as some aspects of maximum likelihood have. Laplace preceded Fisher down one of his lines, but in a different time and with a different statistical intuition he did not attempt to abstract from the location problem to more general considerations. Edgeworth may have had the best appreciation of the subtleties of statistical theory of anyone between Laplace and Fisher, but while he found it natural to use conditional inference given the ancillary X's, the problem he faced did not in his formulation yield a manageable answer without the expedient step of restricting the form of the estimator. If he had treated the more general problem it is tempting to think he might have reasoned to the Pearsonian estimator without restriction and been inspired to investigate how far the idea might be generalized. But he did not. Galton too had a statistical mind of the very first order, and he clearly noted a problem that Fisher barely hinted at, if that.

Ancillary statistics were an unusual product of an extraordinary statistical mind. The breadth of the conception exceeded (or has so far exceeded) what is mathematically possible. No single, crisply rigorous mathematical definition delivers all that Fisher promised. But if his reach exceeded his (or anyone's) grasp in this case, it was still very far from a failure. Savage has called the idea "of more lasting importance than fiducial probability" (Savage, 1967, p. 467), and while that smacks of faint praise, it need not have been. Ancillarity has led to a broad collection of procedures that travel together under the banner of conditional inference; it is an idea that has been with profit invoked to simplify, to sharpen, to improve inferences in an even broader list of applications than Fisher envisioned, and can, despite misgivings about how and when to apply it, be expected to continue to serve these roles for an indefinite future.

REFERENCES

[1] Aldrich, J., (1997). R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science* **12**, 162–176.

[2] Barnard, G., (1984). *Contribution to discussion of Yates (1984).*

[3] Barndorff-Nielsen, O.E. and Cox, D.R., (1994). *Inference and Asymptotics.* Chapman and Hall, London.

[4] Basu, D., (1964). Recovery of ancillary information. *Sankhya (A)* **26**, 3–16.

[5] Brown, L.D., (1990). An ancillarity paradox which appears in multiple regression (with discussion). *Annals of Statistics* **18**, 471–538.

[6] Buehler, R.J., (1982). Some ancillary statistics and their properties (with discussion). *Journal of the American Statistical Association* **77**, 581–594.

[7] Cox, D.R., (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics* **29**, 357–372.

[8] Cox, D.R., (1982). *Contribution to discussion of Buehler (1982).*

[9] Cox, D.R., (1984). *Contribution to discussion of Yates (1984)..*

[10] Cox, D.R. and Hinkley, D., (1974). *Theoretical Statistics.* Chapman and Hall, London.

[11] Edgeworth, F.Y., (1893). Exercises in the Calculation of Errors. *Philosophical Magazine (Fifth Series)* **36**, 98–111.

[12] Edgeworth, F.Y., (1896). Supplementary notes on statistics. *Journal of the Royal Statistical Society* **59**, 529–539.

[13] Edwards, A.W.F., (1997a). Three early papers on efficient parametric estimation. *Statistical Science* **12**, 35–47.

[14] Edwards, A.W.F., (1997b). What did Fisher mean by "inverse probability" in 1912-1922?. *Statistical Science* **12**, 177–184.

[15] Fienberg, S.E. and Hinkley, D.V., (1980). *R. A. Fisher: An Appreciation. Lecture Notes in Statistics* **1**. Springer-Verlag, New York.

[16] Fisher, R.A., (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41**, 155–160. (Reprinted in: *Fisher (1974) as Paper 1; reprinted in Edwards (1997a).*)

[17] Fisher, R.A., (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**, 700-725. (Reprinted in: *Fisher (1950) as Paper 11; reprinted as Paper 42 in Fisher (1974).*)

[18] Fisher, R.A., (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London (A)* **144**, 285–307. (Reprinted in: *Fisher (1950)as Paper 24; reprinted as Paper 108 in Fisher (1974).*)

[19] Fisher, R.A., (1935). The logic of inductive inference. *Journal of the Royal Statistical Society* **98**, 39–54. (Reprinted in: *reprinted as Paper 26 in Fisher (1950); reprinted as Paper 124 in Fisher (1974).*)

[20] Fisher, R.A., (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

[21] Fisher 1974, R.A.. *The Collected Papers of R. A. Fisher* (eds: J. H. Bennett). U. of Adelaide Press, Adelaide.

[22] Fraser, D.A.S., (1979). *Inference and Linear Models.* McGraw-Hill, New York.

[23] Galton, F., (1892). *Finger Prints.* Macmillan, London.

[24] Gillispie, C.C., (1979). Mémoires inédits ou anonymes de Laplace sur la théorie des erreurs, les polynomes de Legendre, et la philosophie des probabilités. *Revue d'histoire des sciences* **32**, 223–279.

[25] Hald, A., (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York.

[26] Hinkley, D.V., (1980). Fisher's development of conditional inference. *In Fienberg and Hinkley (1980)*, 101–108.

[27] Johnson, N.L. and Kotz, S., (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

[28] Laplace, P.S., (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de mathématique et de physique, presentés à l'Académie Royale des Sciences, par divers savans, & lu dans ses assemblées* **6**, 621–656. (Translation: *Stigler (1986b).*)

[29] Laplace, P.S., (1777). Recherches sur le milieu qu'il faut choisir entre les résultats de plusieurs observations. *In Gillispie (1979)*, 228–256.

[30] Lehmann, E.L. and Scholz, F.W., (1992). Ancillarity. *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (eds: Malay Ghosh and P. K. Pathak). *IMS Lecture Notes Monograph Series* **17**, 32–51. Institute of Mathematical Statistics, California.

[31] Lloyd, C., (1992). Effective conditioning. *Australian Journal of Statistics* **34**, 241-260.

[32] Pearson, K., (1896). Mathematical contributions to the theory of evolution, III: regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London (A)* **187**, 253–318. (Reprinted in: *Karl Pearson's Early Statistical Papers, Cambridge: Cambridge University Press, 1956, pp. 113-178.*)

[33] Plackett, R.L., (1977). The marginal totals of a 2 × 2 table. *Biometrika* **64**, 37–42.

[34] Savage, L.J., (1976). On rereading R. A. Fisher. *Annals of Statistics* **4**, 441-500.

[35] Stigler, S.M., (1986a). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press, Cambridge, Mass..

[36] Stigler, S.M., (1986b). Laplace's 1774 memoir on inverse probability. *Statistical Science* **1**, 359–378.

[37] Stigler, S.M., (1995). Galton and Identification by Fingerprints. *Genetics* **140**, 857–860.

[38] Stigler, S.M., (1999). *Statistics on the Table.* Harvard University Press, Cambridge, Mass.

[39] Welsh, A.H., (1996). *Aspects of Statistical Inference.* Wiley, New York.

[40] Yates, F., (1984). Tests of significance for 2 × 2 tables (with discussion. *Journal of the Royal Statistical Society (Series A)* **147**, 426–463.

[41] Zabell, S.L., (1989). R. A. Fisher on the history of inverse probability. *Statistical Science* **4**, 247–256.

[42] Zabell, S.L., (1992). R. A. Fisher and the fiducial argument. *Statistical Sciencevol* 7, 369–387.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, IL 60637
USA
stigler@galton.uchicago.edu