Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **187**, 594–605
https://doi.org/10.1093/jrsssa/qnad138
Advance access publication 14 February 2024
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

D. R. Cox: aspects of scientific inference

H. S. Battey

Department of Mathematics, Imperial College London, London, UK

Address for correspondence: H. S. Battey, Department of Mathematics, Imperial College London, UK.
Email: h.battey@imperial.ac.uk

Abstract

The year 2022 marked 100 years since R. A. Fisher's landmark paper on the foundations of statistics, and 50 years since D. R. Cox's seminal work on the proportional hazards model. At a 1-day meeting in recognition of the latter, held at the London School of Hygiene and Tropical Medicine, I attempted to place some of D. R. Cox's most influential work in its appropriate historical context, reconstructing some of his ideas from unifying principles and providing a different exposition to that of the original papers. The present article is an elaboration of that talk. Additional material includes a geometric explanation of conditional inference in models with binary outcomes, relating the logistic analysis of matched comparisons presented by Cox (1958c, Two further applications of a model for binary regression. *Biometrika*, 45, 562–565) to the protracted and sometimes polemical discourse over the analysis of 2×2 contingency tables.

Keywords: 2 × 2 contingency tables, ancillarity, conditional inference, logistic regression, nuisance parameters, proportional hazards model

1 Introduction

The meeting at the London School of Hygiene and Tropical Medicine, organised by Ruth Keogh, was initially intended as a celebration of 50 years of the proportional hazards model and later converted to a more general scientific memorial meeting covering the work of D. R. Cox. A heavily condensed summary of some of his most important work can be found in the IMS obituary (Battey & Reid, 2022) and a more extensive account in the longer obituary by Davison et al. (2022). Personal reflections have appeared in other venues, e.g. issue 5.2 of the *Harvard Data Science Review*. The nature of the present article is different, focussing on a few key developments antecedent to two influential 1972 papers, and their historical origins and ramifications. Attempts are made to reconstruct conceptions from first principles, this aspect being purely speculative, based on the subject's earlier work and a degree of familiarity with his thought processes in other contexts.

2 Background to two influential 1972 papers

2.1 An overview of the Fisherian position

In core statistical thought, D. R. Cox was closely aligned with R. A. Fisher, whose contributions a generation earlier shaped the field's evolution in a profound way. In a remarkable departure from the existing conceptual base, Fisher (1922) introduced a large proportion of the statistical concepts covered in a mainstream undergraduate degree, including likelihood, sufficiency, consistency, efficiency, and information. At the most primitive level, as noted by Stigler (1976), it appears that Fisher was the first to use the term 'parameter' in the general modern sense. By Stigler's count, there are 57 occurrences of 'parameter' in Fisher's (1922) paper, to be contrasted with a single appearance in the works of Karl Pearson in the limited context of normal distributions (Pearson, 1894). For further historical detail, see Stigler (1976).

Received: July 14, 2023. **Revised:** November 8, 2023. **Accepted:** November 17, 2023
© The Royal Statistical Society 2024.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

Notable elements of commonality between R. A. Fisher's and D. R. Cox's scientific outlook, which to some extent characterise the Fisherian position are:

- emphasis on parameters representing important stable aspects of direct subject-matter relevance;
- presentation of the evidence in incisive form, as opposed to binary decisions;
- preference for an inferential strategy that respects sufficiency and ancillarity. (A statistic is sufficient for a parameter ψ if no other statistic that can be calculated from the same sample provides additional information as to the value of ψ. When the dimension of the minimal sufficient statistic exceeds that of ψ, it may be possible to isolate a statistic A, part of the minimal sufficient statistic, whose distribution does not depend on ψ; A, if it exists, is ancillary for ψ.)

A consequence of the latter point is an emphasis on exact, rather than asymptotic, conditional optimality and distribution theory when appropriate. Notwithstanding, Fisher was responsible for some of the most influential asymptotic results in statistics, including the asymptotic distribution of the maximum likelihood estimator (Fisher, 1922) and the extreme-value limit laws (Fisher & Tippet, 1928). Similarly, Cox recognised that the range of situations in which exact conditional inference applies is highly limited and sought to achieve the notional ideal analysis approximately (Section 3.3).

Fisher (1925) introduced the idea of conditioning on an ancillary statistic for recovery of information lost by using the maximum likelihood estimate in place of the full data. These ideas were developed over the second quarter of the 20th century, culminating with Fisher (1956) and Cox (1956, 1958a), who both saw the primary motivation for conditioning as ensuring relevance, as a means of distinguishing between samples of the same size that supply differing amounts of information on the parameter of interest. Intuitively, two samples of the same size can produce likelihood functions that differ appreciably in shape and yet are maximised at the same point.

A compelling familiar example where conditioning on an ancillary statistic is implicit and not usually questioned is in linear regression, where the error variance σ^2 and regression coefficient vector β specify the conditional density function given $X = x$. The normal-theory log-likelihood function, having discarded constants, is

$$-\frac{1}{2}n\log\sigma^2 - \frac{(y-x\beta)^{\mathrm{T}}(y-x\beta)}{2\sigma^2},$$

where

$$(y-x\beta)^{\mathrm{T}}(y-x\beta) = (y-x\hat{\beta})^{\mathrm{T}}(y-x\hat{\beta}) + (\hat{\beta}-\beta)^{\mathrm{T}}x^{\mathrm{T}}x(\hat{\beta}-\beta).$$

The data enter the log-likelihood only via the residual sum of squares, $\hat{\beta}$, and $x^{\mathrm{T}}x$. Thus, when X is treated as random, the dimension of the minimal sufficient statistic exceeds that of $(\beta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$, and $x^{\mathrm{T}}x$ by itself carries no information on either component. In other words, $A = X^{\mathrm{T}}X$ is ancillary for (σ^2, β). A Fisherian analysis conditions on the realised value $A = a$ even if the distribution of X is known, leading to a variance estimate for $\hat{\beta}$ of the form $\hat{\sigma}^2(x^{\mathrm{T}}x)^{-1}$, where $\hat{\sigma}^2$ is an estimate of error variance constructed from the residual sum of squares. A variance estimate for $\hat{\beta}$ based instead on $\mathbb{E}_X(X^{\mathrm{T}}X)$, an average over values of A that could have, but did not occur, would make the analysis less relevant to the data at hand.

A more difficult example is the analysis of the 2×2 contingency table. For a detailed account of the history, see Yates (1984). In brief, Fisher contended that the appropriate analysis is conditional on the marginal totals, leading to Fisher's exact test (Fisher, 1935). Barnard (1945, 1947) put forward a test which he claimed was more powerful, then withdrew the procedure (Barnard, 1949), stating that further reflection had led him to the same conclusion as Fisher. Yates wrote 'That this conclusion is still not accepted in many quarters, however, is very evident from numerous recent publications' (Yates, 1984). Indeed, this remains the state of affairs in 2023. For two explicit discussions, see Brown (1990) together with the rebuttal of Fraser and Reid (1990), and Buja et al.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

(2019) rebutted by Davison et al. (2019). Of the 2×2 table Cox (1984) wrote 'I accept three main theses [...], that the test should be conditional, that concentration on achieving preassigned magic levels like 0.05 rather than calculating p-values is misguided, and that by and large the power comparisons reported in the literature are irrelevant or worse'.

Section 4 discusses the role of conditioning in the analysis of the 2×2 table from a geometrical point of view, alongside an application of logistic regression presented by Cox (1958c), which is closely related.

2.2 Conditional inference: D. R. Cox's input

Conditioning from the standpoint of ensuring relevance was a topic that D. R. Cox emphasised continually, starting in 1956 at an invited address at a joint meeting of the Institute of Mathematical Statistics and the Biometrics Society. Cox (1958a) chronicles the lecture in what has become a core reference on the foundations of the subject. Through a deliberately oversimplified example, the paper gave a compelling demonstration of the need for conditioning in order to ensure scientific relevance and emphasised that such relevance is sometimes incompatible with notions of optimality, particularly that arising from the Neyman–Pearson theory of most powerful tests.

What is now called 'the weighing machine example', although Cox (1958a) did not use that terminology, is one of inference on the mean μ from one observation of

$$Z \sim \begin{cases} Y & \text{with probability } 1/2 \\ X & \text{with probability } 1/2 \end{cases}, \quad \begin{matrix} Y & \sim & N(\mu, \sigma_Y^2) \\ X & \sim & N(\mu, \sigma_X^2) \end{matrix},$$

where $\sigma_X^2 \gg \sigma_Y^2$ and where we know, for any given realisation, which of the two populations has been sampled. The standard one-sided rejection region at level α for the null hypothesis $\mu = 0$ is either $y > q_\alpha \sigma_Y$ or $x > q_\alpha \sigma_X$, the power of the test being

$$p_\bullet(\mu) = \mathrm{pr}_\bullet(\bullet > q_\alpha \sigma_\bullet) = 1 - \Phi\{(q_\alpha \sigma_\bullet - \mu)/\sigma_\bullet\}, \quad \bullet \in \{Y, X\}$$

depending on which population has been sampled. The unconditional test that treats the population indicator as random has rejection region $z > q_\alpha \sigma_Z$ with $\sigma_Z = (\sigma_Y^2/2 + \sigma_X^2/2)^{1/2}$ and power

$$p_Z(\mu) = \mathrm{pr}_Z(Z > q_\alpha \sigma_Z) = 1 - \frac{1}{2}\Phi\{(q_\alpha \sigma_Z - \mu)/\sigma_X\} - \frac{1}{2}\Phi\{(q_\alpha \sigma_Z - \mu)/\sigma_Y\}.$$

Although irrelevant once the data have been observed, $p_Z(\mu) > p_X(\mu)$ for small μ and $\mu > q_\alpha \sigma_Z$, appearing to show that the unconditional test has higher power than the conditional procedure in certain regions of the parameter space when the inaccurate weighing machine in used.

In the Cox (1958a) exposition, the ancillary statistic is the indicator of the experiment that has actually been performed. Its tangible form distinguishes it from the abstract constructions that are more commonly required. The simplest form of mathematical ancillary arises in location models, where the set of pairwise differences between observations specify the shape of the log-likelihood function but carry no information about its location (Fisher, 1934). A difficulty is that exact maximal ancillaries, should they exist, need not be unique, as illustrated by Basu (1964). Cox (1971) proposed a means of choosing between alternative ancillary statistics by attempting to partition most effectively the set of hypothetical realisations into those that supply relatively more information on the parameter of interest, and those that are relatively uninformative. Cox (1971) acknowledged a degree of arbitrariness in his proposed criterion, and later 'the disturbing possibility [...] that the choice might depend on the true and unknown value of [the interest parameter]' (Barndorff-Nielsen and Cox, 1994, p. 43).

By appeal to Cox's (1958) example, Fisher's argument for conditioning on appropriate reference sets is hard to refute. However, there has been considerable difficulty in pinning down a version of ancillarity that applies seamlessly to all problems, particularly in the presence of nuisance parameters. Lloyd (1992) pointed out that approximate ancillary statistics, i.e. ones whose distributions depend very slightly on the value of the parameter, may achieve the required

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

separation using the Cox (1971) criterion more effectively than a statistic that is exactly ancillary. Of Lloyd's example, Cox wrote 'The example is yet another warning of the dangers of overemphasising exact rather than approximate fulfilment of properties whenever competing requirements are involved' (Barndorff-Nielsen and Cox, 1994, p. 44). Much work throughout the 1980s and 90s sought to achieve the appropriate conditioning approximately. References are given in Section 3.3.

2.3 Logistic regression: a constructive derivation via sufficiency

The connection between the previous discussion of conditionality and the development of logistic regression (Cox, 1958b, 1958c) is clarified in Section 2.4.

In the introduction of the paper, Cox (1958b) wrote that the 'best form' for binary outcomes 'seems to be' the logistic formulation which 'has been extensively used in work on bioassays, notably by Berkson'. He went on to construct a theory of exact conditional inference for logistic regression, where 'exact' is in analogy with Fisher's exact test, the distribution of the conditioning statistic in both cases depending very slightly on the interest parameter.

It seems reasonable to surmise that D. R. Cox did not isolate logistic regression for detailed study because of its antecedence in bioassays. A possible genesis, more consistent with his distinctive way of thinking, is in a constructive derivation from first principles in which one starts from the Bernoulli likelihood function and matches up minimal sufficient statistics with those of a normal-theory linear model. The logistic form is recovered as the unique model that produces such unification and satisfies the boundary conditions. There are several reasons for forcing the sufficient statistics to match. Most importantly, simple sufficient statistics allow a conditional analysis, which both eliminates nuisance parameters and ensures relevance by attaching to conclusions the precisions actually achieved. A second reason is that a unified theory has some aesthetic appeal. Not only could binary and linear regression be framed in the same terms, also Fisher's exact test was recovered as a special case. Finally, the choice stabilises interpretation of the regression parameters to some extent, in a sense to be made clear.

The normal theory linear model has associated with it simple sufficient statistics for the regression coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and unknown error variance. These are $S = x^{\mathrm{T}}Y$, i.e. $(S_j)_{j=1}^p = (\sum_{i=1}^n x_{ij}Y_i)_{j=1}^p$ and the residual sum of squares. Suppose now that $(Y_i)_{i=1}^n$ are binary, with outcomes conveniently encoded as $\{0, 1\}$. There is no floating dispersion parameter. The following exposition reconstructs the logistic model using the same sufficient statistics $(S_j)_{j=1}^p$ as the foundation.

Parametrise the likelihood function in terms of the binomial success probabilities $(\theta_i)_{i=1}^n$, where 'success' corresponds to $y_i = 1$:

$$\ell(\theta_1, \ldots, \theta_n; y_1, \ldots, y_n) = \prod_{i=1}^n \theta_i^{y_i}(1-\theta_i)^{(1-y_i)}, \quad (y_i)_{i=1}^n \in \{0, 1\}^n$$

and consider which function $\theta_i(x_i^{\mathrm{T}}\beta)$:$\mathbb{R} \to [0, 1]$ produces sufficient statistics for β of the form $(\sum_{i=1}^n x_{ij}y_i)_{j=1}^p$. For the sufficient statistic to be a sum, either θ_i or $(1-\theta_i)$ must contain an exponential and, in order that the exponential of the sum be factorable in the likelihood, these probabilities must be equal up to the exponential term. Sufficiency of $(\sum_{i=1}^n x_{ij}y_i)_{j=1}^p$ for β thus requires

$$\theta_i = f(x_i^{\mathrm{T}}\beta)\mathrm{e}^{x_i^{\mathrm{T}}\beta} \in [0, 1]$$
$$1-\theta_i = f(x_i^{\mathrm{T}}\beta) = \frac{\theta_i}{\mathrm{e}^{x_i^{\mathrm{T}}\beta}} \in [0, 1],$$

for some function f to be chosen. Enforcing the $[0, 1]$ probability constraint leads to the logistic law

$$\theta_i = \frac{\mathrm{e}^{x_i^{\mathrm{T}}\beta}}{1+\mathrm{e}^{x_i^{\mathrm{T}}\beta}}, \quad 1-\theta_i = \frac{1}{1+\mathrm{e}^{x_i^{\mathrm{T}}\beta}}.$$

The argument is symmetric and these could be defined the other way round, as would be hoped since the encoding of the binary variables is arbitrary.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

2.4 An exact conditional analysis for logistic regression parameters

Let $\beta = (\beta_1, \ldots, \beta_p)$ be the vector of logistic regression coefficients and suppose β_p is the parameter of interest. The conditional distribution of $S_p = \sum_{i=1}^{n} x_{ip} Y_i$ given $S_1 = s_1, \ldots, S_{p-1} = s_{p-1}$ is (Cox, 1958b, 1970)

$$\mathrm{pr}(S_p = s_p \mid S_1 = s_1, \ldots, S_{p-1} = s_{p-1}) = \frac{c(s_1, \ldots, s_p)\mathrm{e}^{\beta_p s_p}}{\sum_u c(s_1, \ldots, s_{p-1}, u)\mathrm{e}^{\beta_p u}},$$

where $c(s_1, \ldots, s_p)$ is the number of possible realisations of $Y_1, \ldots, Y_n$ such that the values of $S_1, \ldots, S_p$ are equal to those actually observed. Cox (1958b) used this to construct conditional confidence sets for β_p. The exposition of Cox (1970) is both clearer and more general than that of Cox (1958b).

The statistics $S_1, \ldots, S_{p-1}$ are sufficient for $\beta_1, \ldots, \beta_{p-1}$ and near-ancillary for β_p. More specifically, $S_1, \ldots, S_{p-1}$ are ancillary in the same loose sense that Fisher specified in the context of the 2×2 table, i.e. if we are given only the values $s_1, \ldots, s_{p-1}$ no conclusions can be drawn about β_p without explicit constraints on the parameter space. By conditioning on $S_1, \ldots, S_{p-1}$, relevance is achieved. By sufficiency, conditioning also eliminates $p - 1$ nuisance parameter from the analysis. Thus, both justifications for conditioning lead to the same conclusion in the present case.

It is clear that the absence of any discussion of Newton–Raphson and maximum likelihood fitting from Cox (1958b) was deliberate. Fisher proposed Fisher scoring (an application of the Newton–Raphson algorithm) for solving the maximum-likelihood estimating equations in 1925 and D. R. Cox would certainly have been aware of it in 1958.

The conditional analysis of logistic regression was not widely taken up, partly because calculation of the combinatorial quantity $c(s_1, \ldots, s_p)$ was difficult with 1950s computation, and also because the problem can be degenerate when p is moderately large relative to the sample size n, yet another motivation for seeking to achieve the idealised conditioning approximately.

2.5 An exact conditional analysis for canonical exponential-family regression

Another major insight, which is likely to have influenced the proportional hazards work in 1972, was introduced in a little-known paper (Cox, 1968). In the original notation:

> The justification of maximum likelihood methods is asymptotic but sometimes analogues of at least a few of the 'exact' properties of normal-theory linear models can be obtained. The simplest case is when the *i*th observation on the dependent variable has a distribution in the exponential family (Lehmann, 1959, p. 50)
>
> $$\exp\{A_i(y)B(\theta_i) + C_i(y) + D(\theta_i)\},$$
>
> where θ_i is a single parameter and there is a linear model
>
> $$B(\theta_i) = \sum_r x_{ir}\beta_r,$$
>
> where the β's are unknown parameters and the x's are known constants. Special cases are the binomial, Poisson, and gamma distributions when the linear model applies to the logit transform, to the log of the Poisson mean and to the reciprocal of the mean of the gamma distribution. Sufficient statistics are obtained and in very fortunate cases useful 'exact' significance tests for single regression coefficients emerge.
>
> (Cox, 1968)

The above construction is essentially the generalised linear model with canonical link. A similar discussion appears as an exercise in Cox (1970), where alternative link functions for the binary case are also discussed on p. 20. The canonical versions of Poisson and exponential regression had been proposed much earlier (Cox, 1955, 1964). Both sets of conclusions from Section 2.4, about conditioning on $S_1, \ldots, S_{p-1}$ and about the interpretation of ratios of coefficients, apply in this broader setting of canonical exponential-family regression.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

3 1972

3.1 Elimination of nuisance parameters by partial likelihood

The main connection of the foregoing to the proportional hazards paper (Cox, 1972) is that D. R. Cox's lifelong study of inferential separations, both for the accomplishment of relevance and the elimination of nuisance parameters, was surely a major aspect in the inception of the model and its analysis by partial likelihood (Cox, 1972, 1975). The latter, which uses a data-based factorisation of the likelihood function, achieves the same result, namely elimination of the nuisance parameter. In the case of the proportional hazards model the nuisance parameter is the baseline hazard function, thus partial likelihood eliminates an infinite-dimensional nuisance parameter. I am not aware of any other setting in which such a feat has been performed.

3.2 Incompatibility of the logistic and proportional hazards models

Suppose that lifetimes $Y(x)$ are generated from a distribution with proportional hazards determined by covariates x but that only the {alive, dead} = {0, 1} indicator at the end of the study is retained, that is $\mathring{Y}(x; t) = \mathbb{1}\{Y(x) \leq t\}$, where t is an arbitrary cut-off for observation, fictitious if survival times are uncensored. Let β_{PH} and β_{L} be the coefficient vectors (without intercept) in the log-linear proportional hazards and linear logistic models, respectively. Specifically, with $\theta_t(x) = \mathbb{E}(\mathring{Y}(x; t))$, the proportional hazards (Cox, 1972) and logistic regression (Cox, 1958b) models are

$$\log \frac{h(t; x)}{h_0(t)} = \alpha + x^{\mathrm{T}}\beta_{\mathrm{PH}}, \tag{3.1a}$$

$$\log \frac{\theta_t(x)}{1 - \theta_t(x)} = \alpha_t + x^{\mathrm{T}}\beta_{\mathrm{L}}, \tag{3.1b}$$

where $h(t; x)$ is the instantaneous probability of failure at time t and $h_0(t)$ is the corresponding quantity with covariates at baseline, the baseline hazard function. That models (3.1a) and (3.1b) are in contradiction is seen most clearly through the equivalent expressions

$$\begin{aligned} \mathrm{pr}(Y(x) \leq t) &= 1 - \exp\Big(-\exp(\alpha + x^{\mathrm{T}}\beta_{\mathrm{PH}}) \textstyle\int_0^t h_0(u)\mathrm{d}u\Big), \\ \mathrm{pr}(Y(x) \leq t) &= \frac{\exp(\alpha_t + x^{\mathrm{T}}\beta_{\mathrm{L}})}{1 + \exp(\alpha_t + x^{\mathrm{T}}\beta_{\mathrm{L}})}, \end{aligned} \tag{3.2}$$

for the proportional hazards and logistic regression models, respectively. If time-to-event data are discretised into more than two categories by taking a sequence of terminal times, then equation (3.1b) at times $t_1, \ldots, t_{\max} - 1$ is the proportional odds model for ordinal outcomes (McCullagh, 1980), which entails the constraint $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{t_{\max}-1}$ to ensure non-negative probabilities.

For binary outcomes, the model

$$\log[-\log\{1 - \theta_t(x)\}] = \gamma_t + x^{\mathrm{T}}\beta_{\mathrm{LL}},$$

$\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{t_{\max}-1}$, replaces the logistic transform of probabilities by a double-logarithmic transform called the complementary log–log link (McCullagh, 1980). This sacrifices the simple sufficient statistics and consequently the exact conditional significance tests developed for the logistic model by Cox (1958b, 1970).

The four models are related in pairs according to $\beta_{\mathrm{L}} = \beta_{\mathrm{PO}}$ and $\beta_{\mathrm{LL}} = \beta_{\mathrm{PH}}$, where β_{PO} is the coefficient parameter of the proportional odds model.

3.3 Closely related developments

McCullagh's (1980) complementary log–log model is a special case of the more flexible extension of Cox's (1968, 1970) canonical exponential family regression models usually termed generalised

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

linear models following the paper by Nelder and Wedderburn (1972), which introduced the terminology of link functions and emphasised maximum-likelihood fitting by the Newton–Raphson algorithm applied to the full likelihood function.

Exponential families are very special in their structure. A body of work, perhaps starting with Fraser (1964) sought to achieve the appropriate conditioning approximately, beyond exponential-family canonical form. Important subsequent contributions were due to Barndorff-Nielsen and Cox (1979), Cox (1980), Barndorff-Nielsen (1983), McCullagh (1984), Cox and Reid (1987), Fraser and Reid (1988), Barndorff-Nielsen (1990), and Fraser (1990). For a helpful review of the Fraser and Reid line of work, see Davison and Reid (2023). In a recent summary of some of D. R. Cox's lesser-known papers, Reid wrote:

> David raises several points that he feels need further work: e.g. he writes that 'the conceptual basis for the choice between alternative test statistics needs clarification'. In my view this was essentially solved by the development of the so-called r^*-approximation by Barndorff-Nielsen (1990) and Fraser (1990), but I must say that David firmly disagreed with me about this.
>
> (Reid, 2024)

4 A geometric illustration of conditional inference

4.1 Pearson's and Fisher's analyses for the 2×2 table

From at least one point of view, the simplest type of setting to which logistic regression applies was presented by Cox (1958c) in the form of a matched-pair assessment of treatment effect on a binary outcome. For this, a geometric representation is feasible, cast alongside an older formulation for multivariate binary outcomes, advocated by Pearson and forcefully criticised by Fisher.

The term *pure contingency table* refers in the 2×2 case to a situation with two binary outcomes (Z_1, Z_2), each encoded as $\{0, 1\}$ and treated on an equal footing. In the classical Pearsonian examples, the two variables have equal status and the problem is bivariate. In Section 4.2 an arguably more relevant situation is considered in which one variable is an outcome and the other considered potentially explanatory.

From a total of n randomly sampled individuals, a pure contingency table records the number in each of the four combinations of levels in a 2×2 contingency table of the form

$$\begin{array}{cc|c} N_{00} & N_{01} & N_{0\cdot} \\ N_{10} & N_{11} & N_{1\cdot} \\ \hline N_{\cdot 0} & N_{\cdot 1} & n \end{array}, \tag{4.1}$$

where $N_{j\cdot} = N_{j0} + N_{j1}$ $(j = 0, 1)$ are the marginal totals at each level for variable 1, $N_{\cdot k} = N_{0k} + N_{1k}$ are those for variable 2 and $N_{\cdot 0} + N_{\cdot 1} = N_{0\cdot} + N_{1\cdot} = n$. The table can always be normalised such that the entries are proportions. Thus write $\hat{\pi} = (\hat{\pi}_{00}, \hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{11})$ for the vector of empirical probabilities for the four cells, which belongs to a three-dimensional subspace $\mathcal{S}_3$ of $\mathbb{R}^4$ due to the unit-sum constraint on the proportions; the same constraint holds for the vector of true probabilities $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$.

The classic statistical problem, dating back at least to Pearson (1900), is to assess independence of (Z_1, Z_2), for which a necessary and sufficient condition is that the probability of each joint event is equal to the corresponding product of marginal probabilities, i.e. $\pi_{jk} = (\pi_{j0} + \pi_{j1})(\pi_{0k} + \pi_{1k})$ for all four combinations of j and k. The independence condition holds if ψ, the cross-product ratio, is equal to 1, where

$$\psi = (\pi_{00}\pi_{11})/(\pi_{01}\pi_{10}) \tag{4.2}$$

(e.g. Cox and Hinkley, 1974, p. 393). This independence constraint defines a two-dimensional subspace $\mathcal{S}_\perp \subset \mathcal{S}_3$, depicted as the curved manifold in Figure 1. This geometric representation is due to Fienberg and Gilbert (1970).

Fisher argued that, for assessment of the null hypothesis $\psi = 1$, it is the relative probabilities of the different configurations subject to the observed row and column totals $N_{j\cdot} = n_{j\cdot}$ and $N_{\cdot k} = n_{\cdot k}$ that characterise the reference distribution against which the observed data should be calibrated.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

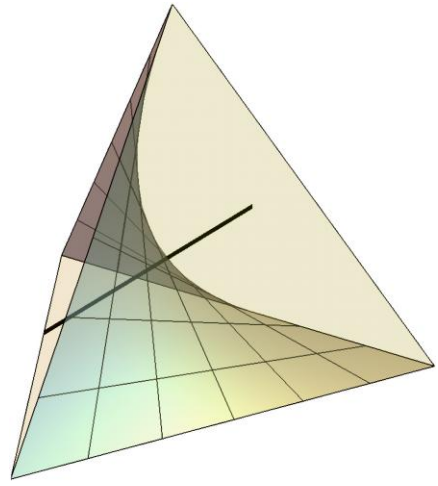


Figure 1. The two-dimensional surface is the constraint $\psi = 1$ in $\mathcal{S}_3$ (viewed as the parameter space for π), imposed by the independence hypothesis. The black line is the constraint on $\mathcal{S}_3$ (viewed as the sample space for $\hat{\pi}$) imposed by the marginal totals $\hat{\pi}_{1\cdot} = 0.6$, $\hat{\pi}_{\cdot 1} = 0.4$.

The basis for this is that the row and column totals are ancillary for ψ in the sense that knowledge of these by themselves carry no information on ψ. The qualifier 'by themselves' covers the situation in which explicit constraints can be made on the parameter space for ψ, allowing a degree of information about its value to be recovered. It is in fact only necessary to condition on one of each of the row and columns totals, which determines the other two.

Figure 1 overlays the marginal total constraint in the unconstrained sample space for $\hat{\pi}$ and the independence constraint $\psi = 1$ in the unconstrained parameter space for π. By construction, the observed $\hat{\pi}$ belongs to the one-dimensional subspace indicated by the thick black line, which crosses the $\psi = 1$ subspace $\mathcal{S}_\perp$ at a single point. Fisher argued that the null distribution constrained to the line should be used to assess compatibility of the observed data with the null hypothesis, as opposed to the null distribution in the larger space $\mathcal{S}_3$.

4.2 Cox's logistic analysis in the context of the 2×2 table

Two broad forms of conditioning are that implicit in the formulation of a model, which should incorporate known physical constraints in the data generating process, and the more abstract form arising from inferential considerations. The discussion in Section 4.1 concerned the latter and was free of modelling assumptions.

In the application of logistic regression to the analysis of a matched-pair design (Cox, 1958c), the sample space is constrained at once to $\hat{\pi}_{0\cdot} = 0.5$, $\hat{\pi}_{1\cdot} = 0.5$ by the balance of the design, even without considerations of ancillarity. This subspace of the sample space, $\mathcal{S}_{\mathrm{MP}}$ say, is depicted alongside $\mathcal{S}_\perp$ from Section 4.1 in Figure 2.

Since $\mathcal{S}_{\mathrm{MP}}$ is also a subspace of the parameter space for π, a geometric representation of the statistical model is also feasible. The left panel of Figure 2 parameterises the probabilities of the two outcomes as $p_0, (1 - p_0)$ for the controls and $p_1, (1 - p_1)$ for the treated individuals. Division by 2 is needed to convert these to the scale of π. Then the algebraic form of the cross-product ratio coincides with the odds ratio

$$\psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{p_{11}/p_{10}}{p_{01}/p_{00}},$$

where the notation on the right hand side is for ease of comparison to (4.2). The two systems of straight lines on the plane correspond to regions of the parameter space for π over which p_j is

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

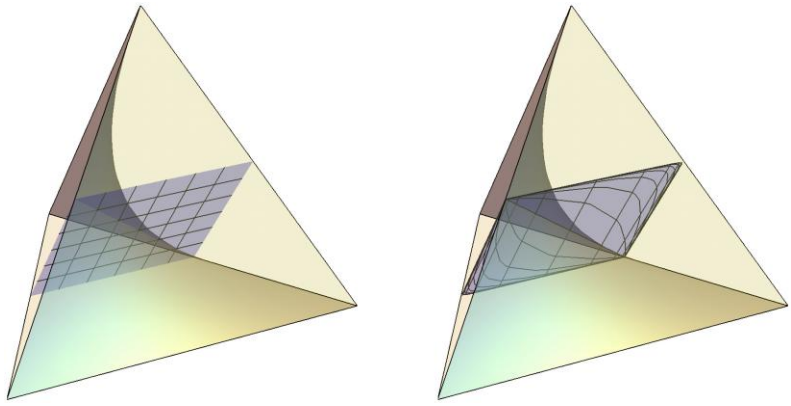


Figure 2. The subspace $\mathcal{S}_{\mathrm{MP}} \subset \mathcal{S}_3$ associated with the matched pair design in two parameterisations. Left: no parametric model for the effect of treatment $(p_0, p_1) \in [0, 1]^2$; right: logistic model for the effect of treatment $(\alpha, \beta) \mapsto \exp(\alpha + \beta)/\{1 + \exp(\alpha + \beta)\}$ for $(\alpha, \beta) \in (-\infty, \infty)^2$.

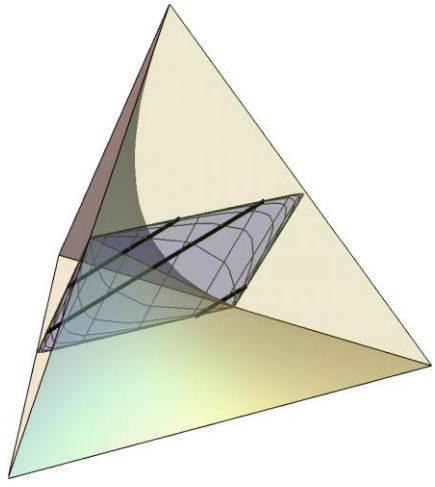


Figure 3. Subspaces $\mathcal{S}_1 \subset \mathcal{S}_{\mathrm{MP}}$ of the matched-pair plane determined by conditioning on the column totals for three different pairwise tables in n observations with $\hat{\pi}_{\cdot 1} = 0.1$, $\hat{\pi}_{\cdot 1} = 0.8$, and $\hat{\pi}_{\cdot 1} = 0.6$.

constant ($j = 0, 1$). The intersection $\mathcal{S}_{\mathrm{MP}} \cap \mathcal{S}_{\perp}$ is the one-dimensional subspace $p_0 = p_1$. For comparison, the right panel uses the logistic parametrisation

$$\alpha \mapsto \frac{e^{\alpha}}{1 + e^{\alpha}} = p_0, \qquad (\alpha, \beta) \mapsto \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} = p_1,$$

the curved contours on the plane indicating the regions of the parameter space for π over which β is constant while α varies. At $\beta = 0$, $p_0 = p_1$ as before but in the (α, β) parameterisation, β has the interpretation of a treatment effect.

The situation can be considered as a single contingency table in n observations with row totals $n/2$, as depicted in Figure 3 for particular values of the column totals. The corresponding thick black lines represent one-dimensional subspaces $\mathcal{S}_1 \subset \mathcal{S}_{\mathrm{MP}}$ containing the observed value of $\hat{\pi}$, for which the distribution of the relevant statistic, viewed as a random variable constrained to $\mathcal{S}_1$, does not depend on α. Now suppose, as in the Cox (1958c) formulation, that there are

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

pair-specific nuisance parameters representing, for instance, genetic differences between the twins. Had data on covariates been available, these might have been modelled as $\alpha_i = x_i^{\mathrm{T}}\gamma$, leading to the usual logistic formulation, but the present approach treats them more flexibly as fixed arbitrary constants. Thus for the ith matched pair with binary outcomes (Y_{i0}, Y_{i1}) the probabilities conditional on treatment group are

$$p_{00}^{(i)} = 1/(1+\mathrm{e}^{\alpha_i}) \quad p_{01}^{(i)} = \mathrm{e}^{\alpha_i}/(1+\mathrm{e}^{\alpha_i})$$
$$p_{10}^{(i)} = 1/(1+\mathrm{e}^{\alpha_i+\beta}) \quad p_{11}^{(i)} = \mathrm{e}^{\alpha_i+\beta}/(1+\mathrm{e}^{\alpha_i+\beta}).$$

In contrast to the notation in Section 4.1, Y_{i0} and Y_{i1} for pair index i represent the same binary outcome, and the numeric subscript refers to conditioning on the value of the treatment variable. Since the α_i are distinct, the situation is best considered as $n/2$ separate contingency tables in 2 observations with row totals 1. In this case, only four tables are possible for each pair, and there is no continuous path between extreme tables with the same marginal totals on which a physically-achievable table can reside. The column total. $N_{\cdot 1}^{(i)}$ in the single-pair analogue of (4.1) is the pair total $Y_{i0} + Y_{i1}$, which is sufficient for α_i.

5 Question and answer

An audience member asked whether I was aware of David's position on the Birnbaum (1962) paper. A slightly more considered answer than that given in my lecture follows.

Birnbaum (1962) purportedly showed that the sufficiency principle, which states that two data sets having the same value of the minimal sufficient statistic should yield the same inference, and the conditionality principle, which states that one should condition on an ancillary statistic if this exists, jointly imply the likelihood principle. The latter is problematic because the likelihood principle is typically rejected by non-Bayesians, and indeed is incompatible with many of the standard methods of non-Bayesian inference. The implication of this result, if it is accepted, is that Fisherian conditional inference is self-contradictory, and adoption of the Fisherian position inevitably leads to the Bayesian paradigm.

I asked David what he made of this in 2020, after reading a paper by Mayo (2014) and several discussions of it (e.g. Dawid, 2014; Evans, 2014; Fraser, 2014), some of which appeared at first sight to be logically sound yet mutually contradictory. David's somewhat hazy recollection was that Birnbaum had ultimately recanted on his earlier position. Evans (2014) attempts to explain some of the apparent contradiction, highlighting discrepancies among the definitions of ancillarity used in the different proofs. One aspect of this, although only a partial depiction, is whether one reduces by sufficiency first, taking as ancillary statistics only components of the minimal sufficient statistic, which was David's position. In the Fraser and Reid line of work, the initial reduction by sufficiency is not explicit. Instead, the approximate ancillary conditioning is achieved through a projection onto the space tangent to the ancillary manifold at the point where the data are observed, the ancillary manifold being the subspace of $\mathbb{R}^n$ on which the ancillary statistic is fixed at its observed value. Conveniently, this tangent space can be constructed without explicitly constructing the ancillary manifold. Furthermore, the projection is unique to the order of approximation considered, apparently resolving any ambiguity in the choice of ancillary statistic. See Davison and Reid (2023) for an accessible introduction to the original papers.

A referee has pointed out a note in Cox (2006, p. 62), in which David cites confusions between the weak and strong likelihood principles among the sources of paradox.

Acknowledgments

I am grateful to the editors for their invitation to produce this article and to Nancy Reid for discussion of these issues over many years and for comments on a draft.

Conflict of interest: None declared.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

Funding

The work was supported by a UK Engineering and Physical Sciences Research Fellowship (EP/T01864X/1).

Data availability

There are no new data associated with this work.

References

Barnard G. A. (1945). A new test for 2 × 2 tables. *Nature, 156*, 177.
Barnard G. A. (1947). Significance tests for 2 × 2 tables. *Biometrika, 34*, 123–138.
Barnard G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society. Series B*, 11, 115–139.
Barndorff-Nielsen O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, *70*(2), 343–365. https://doi.org/10.1093/biomet/70.2.343
Barndorff-Nielsen O. E. (1990). Approximate interval probabilities. *Journal of the Royal Statistical Society. Series B*, *52*(3), 485–496. https://doi.org/10.1111/j.2517-6161.1990.tb01802.x
Barndorff-Nielsen O. E., & Cox D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *Journal of the Royal Statistical Society. Series B*, *41*(3), 279–312. https://doi.org/10.1111/j.2517-6161.1979.tb01085.x
Barndorff-Nielsen O. E., & Cox D. R. (1994). *Inference and asymptotics*. London: Chapman and Hall.
Basu D. (1964). Recovery of ancillary information. *Sankya, Series A, 26*, 3–16.
Battey H. S., & Reid N. (2022). Obituary: David Cox. *IMS Bulletin*, *51*(3), 14–16. https://imstat.org/wp-content/uploads/2022/02/Bulletin51_3.pdf
Birnbaum A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*(298), 269–306. https://doi.org/10.1080/01621459.1962.10480660
Brown L. D. (1990). An ancillary paradox which appears in multiple regression. *Annals of Statistics, 18*(2), 471–493.
Buja A., Brown L., Berk R., George E., Pitkin E., Traskin M., Zhang K., & Zhao L. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science, 34*(4), 523–544.
Cox D. R. (1955). Some statistical methods connected with series of events (with discussion). *Journal of the Royal Statistical Society. Series B*, *17*(2), 129–164. https://doi.org/10.1111/j.2517-6161.1955.tb00188.x
Cox D. R. (1956). *Joint meeting of the IMS and biometric society*. Invited address.
Cox D. R. (1958a). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, *29*(2), 357–372. https://doi.org/10.1214/aoms/1177706618
Cox D. R. (1958b). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society. Series B*, *20*(2), 215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
Cox D. R. (1958c). Two further applications of a model for binary regression. *Biometrika*, *45*(3-4), 562–565. https://doi.org/10.1093/biomet/45.3-4.562
Cox D. R. (1964). Some applications of exponential ordered scores. *Journal of the Royal Statistical Society. Series B*, *26*(1), 103–110. https://doi.org/10.1093/10.1111/j.2517-6161.1964.tb00544.x
Cox D. R. (1968). Notes on some aspects of regression analysis (with discussion). *Journal of the Royal Statistical Society: Series A*, *131*(3), 265–279. https://doi.org/10.2307/2343523
Cox D. R. (1970). *The analysis of binary data*. Methuen.
Cox D. R. (1971). The choice between alternative ancillary statistics. *Journal of the Royal Statistical Society. Series B*, *33*(2), 251–255. https://doi.org/10.1111/j.2517-6161.1971.tb00876.x
Cox D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B*, *34*, 187–220. https://www.jstor.org/stable/2985181
Cox D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276. https://doi.org/10.1093/biomet/62.2.269
Cox D. R. (1980). Local ancillarity. *Biometrika*, *67*(2), 279–286. https://doi.org/10.1093/biomet/67.2.279
Cox D. R. (1984). Discussion of 'Tests of significance for 2 × 2 contingency tables' by F. Yates. *Journal of the Royal Statistical Society. Series A*, *147*(3), 451.
Cox D. R. (2006). *Principles of statistical inference*. Cambridge University Press.
Cox D. R., & Hinkley D. V. (1974). *Theoretical statistics*. Chapman and Hall.
Cox D. R., & Reid N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society. Series B*, *49*(1), 1–39. https://doi.org/10.1111/j.2517-6161.1987.tb01422.x
Davison A. C., Isham V., & Reid N. (2022). Sir David Cox: 1924–2022. *Journal of the Royal Statistical Society: Series A*, *185*(4), 2295–2306. https://doi.org/10.1111/rssa.12964
Davison A. C., Koch E., & Koh J. (2019). Comment: Models are approximations! *Statistical Science, 34*(4), 584–590.
Davison A. C., & Reid N. (2023). The tangent exponential model. In M. Xie, N. Reid, J.O. Berger, & X.-L. Meng, (Eds.), *Bayes, frequentist, fiducial*. Chapman & Hall/CRC Press.

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024

Dawid A. P. (2014). Discussion of "On the Birnbaum argument for the strong likelihood principle". *Statistical Science*, *29*(2), 240–241. https://doi.org/10.1214/14-STS470
Evans M. (2014). Discussion of "On the Birnbaum argument for the strong likelihood principle". *Statistical Science*, *29*(2), 242–246. https://doi.org/10.1214/14-STS471
Fienberg S. E., & Gilbert J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, *65*(330), 694–701. https://doi.org/10.1080/01621459.1970.10481117
Fisher R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309–368.
Fisher R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725. https://doi.org/10.1017/S0305004100009580
Fisher R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A, 144*, 285–307.
Fisher R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society: Series A*, *98*(1), 39–54. https://doi.org/10.2307/2342435
Fisher R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd.
Fisher R. A., & Tippet L. H. C. (1928). Limiting forms of the frequency distribution of the largest of smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180–190. https://doi.org/10.1017/S0305004100015681
Fraser D. A. S. (1964). Local conditional sufficiency. *Journal of the Royal Statistical Society. Series B*, *26*(1), 52–62. https://doi.org/10.1111/j.2517-6161.1964.tb00538.x
Fraser D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, *77*(1), 65–76. https://doi.org/10.1093/biomet/77.1.65
Fraser D. A. S. (2014). Discussion: On arguments concerning statistical principles. *Statistical Science*, *29*(2), 252–253. https://doi.org/10.1214/14-STS473
Fraser D. A. S., & Reid N. (1988). On conditional inference for a real parameter: A differential approach on the sample space. *Biometrika*, *75*(2), 251–264. https://doi.org/10.1093/biomet/75.2.251
Fraser D. A. S., & Reid N. (1990). Discussion: An ancillary paradox which appears in multiple regression. *Annals of Statistics, 18*(2), 503–507.
Lehmann E. L. (1959). *Testing statistical hypotheses*. Chapman & Hall.
Lloyd C. J. (1992). Effective conditioning. *Australian Journal of Statistics*, *34*(2), 241–260. https://doi.org/10.1111/anzs.1992.34.issue-2
Mayo D. G. (2014). On the Birnbaum argument for the strong likelihood principle (with discussion). *Statistical Science*, *29*(2), 227–239. https://doi.org/10.1214/13-STS457
McCullagh P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, *42*(2), 109–142. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x
McCullagh P. (1984). Local sufficiency. *Biometrika*, *71*(2), 233–244. https://doi.org/10.1093/biomet/71.2.233
Nelder J. A., & Wedderburn R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, 135*, 370–384.
Pearson K. (1894). III. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, *185*, 71–110.
Pearson K. (1900). On the criticism 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. *Philosophical Magazine, 50*(5), 157–175.
Reid N. (2024). On some publications of sir David Cox. *Scandinavian Journal of Statistics*, to appear.
Stigler S. (1976). Discussion of "On rereading R. A. Fisher" by Leonard J. Savage. *Annals of Statistics*, *4*, 441–500.
Yates F. (1984). Tests of significance of 2 × 2 contingency tables (with discussion). *Journal of the Royal Statistical Society: Series A*, *147*(3), 426–463. https://doi.org/10.2307/2981577

Downloaded from https://academic.oup.com/jrsssa/article/187/3/594/7607715 by Imperial College of Science Technology and Medicine user on 09 October 2024