

EFFECTIVE CONDITIONING

CHRISTOPHER LLOYD¹

La Trobe University

Summary

The ultimate aim of conditioning is to obtain a more relevant estimate of estimation precision by separating experiments of higher precision from those of lower precision. The cost is the information, if any, contained in the conditioning variable. This paper considers the problem of quantifying the costs and benefits of conditioning. Several examples are studied and some of the asymptotic consequences explored.

Key words: Asymptotics; local inference; ancillarity; information; contingency table.

1. Introduction

Conditional arguments are an important tool in mathematical statistics. For instance, a rather complicated random variable may have a known and simple distribution conditional on some other variables, so its distribution may then be obtained by integration. Nuisance parameters, so called, can be eliminated by conditioning on statistics sufficient for them, leading to unconditional similar tests (Lehmann & Scheffé, 1950). Certain score functions are made unbiased by deducting their conditional expectations or, equivalently, by forming conditional likelihoods (Lindsay, 1982). However, so long as the final form of inference is unconditional, these are not true conditioning arguments but merely a mathematical artifice of a particularly elegant form. From a fundamental point of view they have no more inherent logic than performing all calculations in base eight.

In this paper we study conditioning. By this we mean not the mathematical details of pseudo-conditioning but rather the logic of true conditioning. True conditioning on a statistic A is achieved when the value of A is fixed at its observed value from the beginning to the end of the statistical analysis of a given model. Typically A is a component of the minimal sufficient statistic (MSS) distributed (approximately) free of θ , called an (approximate) ancillary for θ . In this paper, except for the last section, attention is confined to a single parameter θ , firstly because a scalar parameter requires a simpler notation and secondly to isolate the important features of conditioning without the further complicating issue of nuisance parameters. Cox (1988) gives a useful review of conditioning in an asymptotic context.

Received May 1990; revised June 1991.

¹Dept. Statistics, La Trobe University, Bundoora, Vic 3083, Australia.

The paper is in four sections. First, we introduce two criteria for judging a conditioning variable, called *utility* and *relevance*, and define quantitative measures of them in terms of conditional and unconditional information. The central contention of this paper boils down to three progressively more refined observations:

- (i) these two criteria are logically, conceptually, and mathematically distinct;
- (ii) both criteria must be considered in evaluating the effectiveness or otherwise of conditioning; and
- (iii) these two criteria can compete in the sense that a variable optimal in one may be far from optimal in the other.

In Section 3, we consider how best to combine utility and relevance in an overall measure of conditioning merit. Section 4 explores how utility and relevance may compete asymptotically and concludes that first order ancillaries are *not* necessarily the best conditioning variables. In the last section, the ideas are extended to the nuisance parameter case and we study the controversial example of the contingency table.

Most mathematical proofs are given in the Appendix to the paper.

2. Relevance and Utility of Conditioning Variables

Suppose that a sample is taken on a statistical model $f(x; \theta)$ with scalar θ but that the MSS S is *not* a scalar. Break S into an estimator $T(S)$ and an additional or *ancillary* statistic $A(S)$. This ancillary statistic A may be a vector, but for simplicity scalar notation is used. *Should T be evaluated by comparing it to its marginal (i.e. unconditional) distribution or to its A -conditional distribution?*

There are basically two approaches that justify conditioning on A . Firstly, imagine breaking up the experiment into two steps: observe A , and then observe T knowing the value of A . If the first step of the experiment happens to contain little or no information about θ then only the second step is of inferential importance. Such an uninformative statistic A is often called *ancillary*, the judgement of its informativeness or *utility* being based on its marginal distribution. The first approach that justifies conditioning is thus 'low utility statistics tell us little about θ so ignore the statistical experiment of observing them.' Additional weight is given to this argument from two sources. First, it is the same argument used to justify the sufficiency principle which is universally accepted. Second, when 'information' means Fisher information, the information in the second step is approximately equal to the total information in an average sense (see equation (2.1)).

A second justification for conditioning may be used when it can be argued that the value of A specifies certain important properties of the conditional experiment, say its precision. Conditioning on A then distinguishes between properties of the actual experiment and those of possible experiments that might have been conducted but were not. It fixes the *relevant* features of the experiment. The commonest example is the sample size of an experiment which is nearly always

considered fixed even though it is random in the sense that it cannot be predicted before the experimenter sets it. It is not necessary that A be even approximately ancillary to perform this function.

The following example is a modification to one of Cox (1958). Though artificial, it runs to the heart of the matter and justifies the itemised assertions about utility and relevance listed in the introduction. The main aim is to establish that non-ancillaries can be preferable to ancillaries as conditioning variables, a point perhaps appreciated but never before stressed in the literature.

Example 1. An unknown quantity θ is measured on one of four machines with normal errors and zero bias. The table below lists the probability of use of each machine and the variance of the measurement X as well as the values of two statistics A_1 and A_2 . The MSS for θ is (X, A_1, A_2) .

Machine i	M1	M2	M3	M4
Variance	99.9	100.1	10.1	9.9
$p_i(\theta)$	$.25 - \theta\lambda_1$	$.25 - \theta\lambda_2$	$.25 + \theta\lambda_1$	$.25 + \theta\lambda_2$
A_1	1	1	0	0
A_2	1	0	1	0

The λ_i are small but known constants. Either of A_1 and A_2 partially identifies which machine was used while the pair (A_1, A_2) fully identifies the machine. Which of these should we use as a conditioning variable in evaluating the measurement X ? We interpret this question as asking what variance we should quote: the variance of the actual machine used, the average variance across all machines, or something in between.

The statistic A_1 is not ancillary unless $\lambda_1 + \lambda_2 = 0$, but it distinguishes between a pair of good machines and a pair of bad machines and thus tells us much about the accuracy of the measurement X . On the other hand A_2 is an exact ancillary, but divides the four machines into two groups, both containing one good and one poor machine. Its value tells us almost nothing about the accuracy of the measurement X . Provided that λ_1 and λ_2 are small so that A_1 is not too informative about θ , it is surely the preferred conditioning variable despite the fact that it is not ancillary and an exact ancillary, A_2 , is available. However, if $|\lambda_1 + \lambda_2|$ is larger then A_1 may contain sufficient information about θ itself to make us question the appropriateness of conditioning on it. At what point would this occur?

Suppose that the λ_i are small and we condition on A_1 . Should we condition further on A_2 ? While this identifies the actual machine used very little extra information about the accuracy of the measurement X is obtained as the machines within each group are very similar. On the other hand, the statistic A_2 is *not* ancillary conditional on A_1 when $\lambda_1 \neq \lambda_2$. Under these conditions it is unclear if one would further condition on A_2 after A_1 ; rather little is gained and rather little lost.

Next we turn to the measurement of utility and relevance. In any study of conditioning the information in the conditional experiment, as opposed to the unconditional experiment, is central. We use Fisher information since it is widely accepted from both axiomatic and practical standpoints. Let ℓ_A^B denote the loglikelihood of the experiment of observing a statistic B conditional on a statistic A where the argument θ is suppressed. When B is sufficient it is dropped in the notation and we instead write ℓ_A . When A is trivial, corresponding to no conditioning at all, we write ℓ^B . The Fisher information in the distribution of B given A is

$$i_A^B = \text{var}((\ell_A^B)' | A) = E(-(\ell_A^B)'' | A),$$

where $'$ denotes derivative with respect to θ , and we assume mild conditions of regularity, as we do wherever necessary throughout the paper. By expanding $\ell^{B,A}$ into A -marginal and A -conditional terms we obtain the information equation

$$i^{B,A} = E(i_A^B) + i^A. \quad (2.1)$$

In particular when B is sufficient we obtain (Basu, 1964)

$$i = E(i_A) + i^A. \quad (2.2)$$

On average the difference between the conditional information and the full information is the marginal information contained in the conditioning variable. When the conditioning variable is exactly ancillary Fisher (1934) says that the 'information has been recovered' by conditioning and goes on to suggest that conditioning on derivatives of the likelihood function might achieve this to a high order of accuracy in the general case, though this has never been formalised. The extent of the information *loss* may be taken to measure the negative aspect of conditioning, something which is to be minimised. On the other hand, if i_A is sensitive to changes in the value of A then conditioning is strongly indicated. The extent of this *sensitivity* may be taken to measure the positive aspect of conditioning, something to be maximised. To the author's knowledge, these two complementary features of conditioning have been neither distinguished nor measured adequately in the vast literature on conditional inference.

Now i_A measures the information in the A -conditional experiment, but depends on the observed value of A , so that in order to measure overall utility one must take some location measure of i_A . Alternatively, we may take i^A to measure the information lost and note that in (2.2), i is fixed for a given statistical model. Either approach strongly suggests that we use

$$U(A) = E(i_A) = i - i^A \quad (2.3)$$

as a measure of utility. On the other hand, large variations in i_A with the observed value of A indicate a high degree of relevance. We take

$$R(A) = \text{SD}(i_A)$$

as a measure of the relevance properties of a statistic A . Cox (1971) uses $R(A)$ to distinguish between exact ancillary statistics, a restriction not imposed here. Both definitions need to be extended to conditional versions. For example, to evaluate a statistic B that is evaluated conditional on A , we apply the above definition to the A -conditional experiment to obtain

$$U(B | A) = E(i_{A,B} | A) = i_A - i_A^B, \quad R(B | A) = SD(i_{A,B} | A). \quad (2.4)$$

To a large extent the choice of expectation and standard deviation as respective measures of location and dispersion may be justified by the properties embodied in the following lemma and Theorem 2 of Section 3, though it is not claimed that any other measures would necessarily violate these properties.

Lemma 1. (i) $U(A) \geq U(A, B)$ with equality if and only if B is ancillary given A .

(ii) $E(U(B | A)) = U(A, B)$.

(iii) If B is ancillary given A or if A and B are independent then

$$R^2(A, B) = R^2(A) + E(R^2(B | A)) \geq R^2(A).$$

Remarks. Property (i) typically means that less information is utilised when we condition on a functionally larger statistic, the exception being when the last statistic is ancillary given the previous ones. Property (ii) is often convenient for calculations when the model for B given A is more transparent than the joint model. Relevance may be either increased or decreased when one further conditions on B (see Proof). Property (iii) both specifies conditions sufficient for relevance to be increased and defines the increase as the average of the square of the relevance of B given A . However, it is not necessary that B be ancillary conditional on A for relevance to be enhanced, and it is precisely those cases for which the optimal conditioning variable is not ancillary that are of major interest in this paper.

3. Measuring Conditioning Efficacy

A sufficient statistic, S , is the worst conditioning statistic possible: conditioning on it destroys the statistical model completely. In comparison, a constant statistic, C , is relatively harmless as a conditioning statistic: conditioning on it achieves precisely nothing at all. Since $i_S = 0$ and $i_C = i$ we have

$$U(S) = 0, \quad R(S) = 0, \quad U(C) = i, \quad R(C) = 0.$$

In between these extremes are non-degenerate ancillary statistics, A , for which $U(A) = i$ and $R(A) > 0$ and non-ancillary statistics, B , for which $U(B) < i$ and $R(B) > 0$, possibly greater than and possibly less than $R(A)$ depending on

the particular statistics A, B . In words, as we range across all possible statistics from C to S , U decreases from its maximum value i to zero while R begins and finishes at zero and presumably reaches some maximum at some statistic in between, possibly ancillary and possibly not. Here are the first fruits of carefully separating the roles of conditioning embodied in U and R ; when a statistic is too informative then it is poor as a conditioning statistic, not only because it leaves little information to be utilised, but also because it leaves insufficient 'space' for relevance. Notice also that U is far too heavy-handed to be used as a measure of conditioning merit on its own; it fails to distinguish between C and A or indeed any two ancillaries so, in the absence of the extra characteristic R , we would be forced to resort to functional maximality which is known to lead to non-uniqueness problems (Basu, 1964). Indeed, by Lemma 1(iii), R automatically leads us to prefer functionally maximal ancillaries to non-maximal ones, but it achieves much more than this.

It remains to combine $U(A)$ and $R(A)$ in some way to obtain an overall measure of conditioning efficacy which trades off utility against relevance. Let us call such a measure $D^2(A)$ with the interpretation that the ratio $D^2(A_1)/D^2(A_2)$ measures the *relative* merit of A_1 and A_2 as conditioning variables. This means that D^2 is defined only up to a multiplicative constant. We confine our attention to measures D^2 which depend on A only through the quantities $U(A)$ and $R(A)$. Using a flexible notation, D^2 denotes a function of either two arguments (U and R) or one argument (the statistic being evaluated) depending on the context.

There are several elementary considerations of what we require of such a measure. These reduce the field of possible combinations of R and U to a manageable family. In what follows, we assume that $U(A)$ and $R(A)$ are *linear* measures of, respectively, location and dispersion of i_A .

(1) If A_1 and A_2 are two statistics on a given model then they should have the same relative efficacy when reparametrised from θ to ϕ . Since reparametrisation alters both $R(A)$ and $U(A)$ by a factor $\kappa = (\partial\theta/\partial\phi)^2$, the essential requirement of *parametrisation invariance* leads directly to the functional equation

$$D^2(\kappa x, \kappa y) = f(\kappa)D^2(x, y).$$

(2) Small changes in $U(A)$ or $R(A)$ should produce correspondingly small changes in the measured efficacy of A . This condition of *continuity* requires that D^2 be a continuous function of its two real arguments U and R .

(3) We require *validity* of D^2 as a measure of relative conditioning merit. At the very least we require

$$0 = D^2(S) < D^2(C) < D^2(A),$$

where S, C and A denote sufficient, constant and ancillary statistics. In addition, since large values of R and U imply a good conditioning variable, D^2 must be strictly increasing in both its arguments.

Theorem 1. *The only functions D^2 satisfying the requirements (1)–(3) of parametrisation invariance, continuity and validity are of the form*

$$D^2(A) = \gamma [E(i_A)^\beta + \alpha \text{SD}(i_A)^\beta]$$

where α, β are strictly positive constants and γ is an arbitrary non-negative constant.

Next let Δ^2 denote an *absolute* measure of conditioning merit. By this we mean that it may be used to compare a statistic to statistics defined on some other experiment and also to some absolute standard. What comprises a high or low value of Δ^2 ? We insist that $\Delta^2(C) = 1$; we may take this as the least value of Δ^2 that a statistic may have and still be considered as a conditioning variable. The rationale for this is that if $\Delta^2(A) < 1$ then it is a worse conditioning variable than a constant statistic, in which case conditioning on a constant statistic, i.e. not conditioning at all, is to be preferred to conditioning on A . In addition to (1)–(3) above Δ^2 should have a fourth property that we call *consistency*. Let A_1 be a statistic and C_2 a constant formally defined on some experiment independent of the first. The judgement of whether conditioning on A_1 is good or bad ought not to depend on the whether we take the first experiment or both experiments to generate the sample space. Thus $\Delta^2(A_1, C_2) >, =$ or < 1 according as $\Delta^2(A_1) >, =$ or < 1 respectively.

Here is a preparatory lemma and the main theorem concerning absolute measures.

Lemma 2. *Given n independent experiments, let A_j be some statistic on the j th experiment with associated quantities $U_j(A_j), R_j(A_j)$ for $j = 1, \dots, n$. Then*

$$(i) \quad U(A_1, \dots, A_n) = \sum_{j=1}^n U_j(A_j),$$

$$(ii) \quad R^2(A_1, \dots, A_n) = \sum_{j=1}^n R^2(A_j).$$

Proof. Use $i_{A_1, \dots, A_n} = \sum i_{A_j}$, $i^{A_1, \dots, A_n} = \sum i^{A_j}$ and $i = i_1 + \dots + i_n$ where i_j is the Fisher information in the j th experiment.

Theorem 2. *The only absolute measure of conditioning merit Δ^2 which depends only on U and R , and which is parametrisation invariant, continuous, valid and consistent, has the form*

$$\Delta^2(A; \alpha) = E\left(\frac{i_A}{i}\right) + \alpha \text{SD}\left(\frac{i_A}{i}\right) \quad (3.1)$$

where the constant $\alpha > 0$ and $i = E(i_A)$ is the full Fisher information.

The constant β in D^2 must equal 1 if D^2 is to lead to a consistent absolute measure and we henceforth take $\beta = 1$. From Lemmas 1–2 we deduce some important properties of D^2 , all of which translate directly into properties of Δ^2 .

Corollary 1. *If A' is ancillary conditional on A then $D^2((A, A')) \geq D^2(A)$. If the ancillary statistic B functionally contains the ancillary A then $R(B) \geq R(A)$ and $\Delta^2(B) \geq \Delta^2(A) \geq 1$.*

Proof. Proof follows from Lemma 1 with $B = (A, A')$.

Corollary 2. *Under the conditions and in the notation of Lemma 2,*

$$\begin{aligned} \sum_{j=1}^n D^2(A_j) &\geq D^2(A_1, \dots, A_n) = \sum U_j(A_j) + \alpha \sqrt{\sum R_j^2(A_j)} \\ &\geq D^2(A_1, \dots, A_{n-1}). \end{aligned}$$

It should be noted that Δ^2 may depend on θ ; however if θ is a conditional scale (or location) parameter then Δ^2 is free of θ . Unequivocal comparison of conditioning variables is then possible, without regard to the true value of the parameter. Measures D^2 , Δ^2 satisfy several essential criteria given in Theorems 1 and 2. We are still left with a family of measures indexed by α the choice of which must rest on further less objective criteria. The value we choose depends on how important we consider the relevance property, i.e. to what extent we are prepared to distinguish accurate estimates from inaccurate ones at the cost of perhaps underestimating the accuracy on average. *For the remainder of the paper we take $\alpha = 1$, for no better reason than that it gives equal weight to the natural quantities U and R for which we already have simple interpretations.* Together with $\beta = 1$, this specifies an unequivocal measure of conditioning merit. In what follows, we identify explicitly those points where this choice is important.

To obtain some idea of what constitutes a practically high value of Δ^2 consider again the scheme of Example 1, but with only two machines with respective variances σ^2 and σ^2/ψ . Let A indicate the machine used and suppose $\Pr\{A = 1\} = \pi$ so that A is an exact ancillary. In this case

$$(i_A, A) = \begin{cases} (1/\sigma^2, 1) & \text{with probability } \pi, \\ (\psi/\sigma^2, 2) & \text{with probability } 1 - \pi. \end{cases}$$

Then, since A is ancillary, $U(A) = E(i_A) = (\psi + (1 - \psi)\pi)/\sigma^2$ and

$$\frac{R^2(A)}{i^2} = \frac{\text{var}(i_A)}{E^2(i_A)} = \frac{(\psi - 1)^2 \pi(1 - \pi)}{(\psi + (1 - \psi)\pi)^2}$$

with maximum value $(\psi - 1)^2/4\psi$ when $\pi = \psi/(\psi + 1)$. Notice that this is unchanged if ψ is replaced by $1/\psi$. If we take $\psi > 1$ then

$$\Delta^2(A) = 1 + \alpha \sqrt{(\psi - 1)^2/4\psi}. \quad (3.2)$$

For example taking $\alpha = 1$, if a certain statistic in a problem has $\Delta^2 = 1.3535$ then it may be compared to a two machine experiment with $\psi = 2$. It is easy to invert (3.2) to obtain

$$\psi = 1 + 2(\Delta^2 - 1)^2 \pm 2|\Delta^2 - 1|\sqrt{1 + (\Delta^2 - 1)^2}. \quad (3.3)$$

These two values are reciprocal. This equation may be used to convert a given value of Δ^2 into an equivalent variance ratio for two machines.

To conclude this section, we examine two of Fisher's examples that are often used to illustrate the conditionality argument.

Example 2. This example is discussed in Fisher (1948). Suppose that X, Y are independent variables with exponential distributions and respective means $1/\theta$ and θ . The MSS for θ from identical, independent samples is $(X', Y') = (\sum X_i, \sum Y_i)$ and the full Fisher information is $i = 2n/\theta^2$. Consider the variables

$$T = \sqrt{Y'/X'}, \quad A = \sqrt{X'Y'},$$

where A is ancillary for θ and T is the MLE of θ . This model is often called the gamma hyperbola since regions for which A is constant are hyperbolic in the original variables. The density of T conditional on A is

$$f_{T|A=a}(t; a, \theta) = \frac{1}{t K_0(a)} \exp \left[-a \left(\frac{t}{\theta} + \frac{\theta}{t} \right) \right] \quad (t \geq 0).$$

Here, $K_0(a)$ is a conditional normaliser closely related to a Bessel function and θ is a conditional scale parameter. The conditional information is

$$i_A^T(\theta) = 2A \frac{E(T | A)}{\theta^3} = \frac{2A}{\theta^2} E \left(\frac{T}{\theta} \mid A \right)$$

which is free of θ when divided through by i . Both the conditional normaliser and conditional expectation may be readily approximated. To calculate Δ^2 we must find the variance of i_A^T with respect to the θ -free density of A ,

$$f_A(a) = \frac{a^{2n-1} K_0(a)}{[(n-1)!]^2}.$$

A model of a similar nature is given in Fisher (1959, p.134) where bivariate unit scale normal variables (X_i, Y_i) are centred at an unknown point on the unit circle. The ancillary in this case is the length of the resultant data vector $(\sum X_i, \sum Y_i)$ and the conditional information is another complex expression involving Bessel functions. In Table 1, values of $\Delta^2(A)$ with $\alpha = 1$ have been numerically approximated for both these models and several values of n . The equivalent proportional difference in machine variance, $\psi - 1$, is also given. This

TABLE 1
Effectiveness of conditioning in Example 2

N	Circle		Hyperbola	
	Δ^2	$100(\psi - 1)$	Δ^2	$100(\psi - 1)$
1	1.397	228	1.518	281
2	1.221	148	1.348	207
3	1.152	114	1.270	171
4	1.116	95	1.211	144
5	1.096	84	1.176	126
6	1.078	74	1.152	114
8	1.062	64	1.116	95
10	1.048	55	1.094	83
15	1.031	35	1.064	65
20	1.024	26	1.046	53

quantity depends neither on α , since A is ancillary, nor on θ , as it is a location parameter. In both cases, the ancillary is a 'leverage' index since the larger its value the further the resultant vector from the origin and the more powerfully may one infer the position of the mean on the hyperbola/circle. Conditioning is apparently somewhat more effective for the circular than for the hyperbolic model. As the sample size increases, this effectiveness decreases as one would expect of any model for which the MLE is asymptotically sufficient.

4. Asymptotically Effective Conditioning

In this section we study conditioning for large sample sizes. By calculating the quantities R and U in an expansion about θ_0 we can find a variable which has the desired property locally and to a given order in the sample size n . Moreover, we study so called first order ancillaries that have information $O(1)$, and hence their utility is $1 + O(n^{-1})$. It is shown that the relevance of such statistics is typically $O(n^{-1/2})$ and that consequently Δ^2 is not necessarily maximised to $O(n^{-1})$ by a first order ancillary: we can find a linear combination with an informative statistic that has larger Δ^2 to this order. Higher order calculations are difficult but there is certainly no reason *a priori* to suppose that minimising information to a given order produces a statistic with maximum relevance, particularly in view of Example 1. We begin by returning to Example 1.

Example 3. Most of the salient features of the asymptotic expansions studied later in this section can be easily brought out by considering n i.i.d. observations on Example 1. For each observation the machine used and the measurement, X_j , are recorded. Then $(\bar{X}^*, \alpha_1, \alpha_2, \alpha_3)$ is the MSS for θ where \bar{X}^* is the variance inverse weighted mean of the X_j s, $\alpha_3 = \#(A_1 = 1 = A_2)$, and $\alpha_j = \#(A_j = 1)$ ($j = 1, 2$). As described in Example 1, α_1 is more relevant than α_2 but at $\theta = 0$ it contains information $4n(\lambda_1 + \lambda_2)^2$ which, although small in proportion to the total information available, grows unboundedly with n . For n large enough this information loss might outweigh the enhanced relevance of treating it as fixed at

its observed value a when $\overline{X^*}$ is a combination of a observations on M1 or M2 (with variance around 100) and $n - a$ observations on M3 or M4 (with variance around 10) so that

$$\text{var}(\overline{X^*} \mid \alpha_1 = a) \approx \frac{100}{n} \frac{a}{n} + \frac{10}{n} \left(1 - \frac{a}{n}\right),$$

which varies greatly as a varies. The inverse of this is the conditional Fisher information from which we may easily calculate the quantities of interest in $D^2(\alpha_1)$:

$$\frac{E(i_{\overline{X^*}|\alpha_1})}{i} = 1 - 220(\lambda_1 + \lambda_2)^2, \quad \frac{SD(i_{\overline{X^*}|\alpha_1})}{i} = \frac{9}{11\sqrt{n}} \left(1 + \frac{18\theta(\lambda_1 + \lambda_2)}{11}\right),$$

in both cases with error $O(n^{-1})$. Since the first term is less than one there is loss of asymptotic information in conditioning on α_1 whereas the relevance decreases with the sample size as $n^{-1/2}$. Finally, consider

$$\alpha = \alpha_1 + (\lambda_1 + \lambda_2)\overline{X^*}$$

which is a statistic with expectation free of θ and variance $\frac{1}{4} + O(\lambda^2/n)$ and so second order ancillary in the terminology of Cox (1980). Since the estimate $\overline{X^*}$ is a linear function of α_1 conditional on α , we evaluate the conditional information in $\overline{X^*}$ to be firstly free of the value conditioned on, and secondly uniformly low. Apparently, this conditioning does not enhance the relevance of our inference.

In the remainder of this section, we derive an asymptotic expansion for $\Delta^2(A)$ of the form

$$\Delta^2(A) = \delta_1 + \frac{\delta_2}{\sqrt{n}} + O(n^{-1})$$

where δ_1, δ_2 depend on various cumulants and their derivatives for the particular case of (2,1)-exponential families. We then seek to maximise this within some class of conditioning statistics including the first order ancillary of Cox (1980).

Theorem 5. *In a (2,1)-exponential family, let S_1, S_2 be standardised versions of the MSS described in Cox (1980). Let A_0 be the first order ancillary and S_0 the combination of S_1, S_2 orthogonal to A_0 . Then there exists a linear combination A of A_0 and S_0 , not necessarily equal to A_0 , with $\Delta^2(A) > \Delta^2(A_0)$ to order n^{-1} .*

Proof. As in Cox (1980) consider values of $\theta = \theta_0 + \delta/\sqrt{n}$ in a neighbourhood of a point of interest θ_0 . Let κ_{ij} denote mixed cumulants of (S_1, S_2) and

$$\kappa_{\ell m}(\theta) = \kappa_{\ell m} + \alpha_{\ell m}\delta/\sqrt{n} + \beta_{\ell m}\delta^2/n + O(n^{-3/2})$$

where $\kappa_{01} = \kappa_{11} = \kappa_{10} = 0$, $\kappa_{20} = \kappa_{02} = 1$. The first order ancillary of Cox (1980) has $\alpha_{10} = 0$ and the second order ancillary has $\beta_{10} = \alpha_{20} = 0$. This

is based on minimising the information (i.e. maximising the utility) of $A(c) = cS_1 + (1 - c)S_2$ with respect to c to first/second order. However, this takes no account of relevance. To study the effect of relevance we need an asymptotic expression for the information in S_2 given S_1 . This is obtained in Lloyd (1991) by substituting expansions for conditional moments of S_2 into a general expansion for Fisher information. The resulting rather complicated expression simplifies when (S_1, S_2) have standardised cumulants and is given by

$$i_{S_2|S_1}(\theta) = n\alpha_{01}^2 + \sqrt{n}S_1(2\alpha_{01}\alpha_{11} - \alpha_{01}^2\kappa_{12}) + O(1). \quad (4.1)$$

Since the total information in the model is $i = n(\alpha_{10}^2 + \alpha_{01}^2) + O(1)$, on taking expectation and standard deviation of (4.1) we obtain

$$\frac{U(S_1)}{i} = \frac{\alpha_{01}^2}{\alpha_{10}^2 + \alpha_{01}^2} + O(n^{-1}), \quad \frac{R(S_1)}{i} = \frac{|2\alpha_{01}\alpha_{11} - \alpha_{01}^2\kappa_{12}|}{\sqrt{n}(\alpha_{10}^2 + \alpha_{01}^2)} + O(n^{-1}). \quad (4.2)$$

Define a class of statistics indexed by c and which includes the first order ancillary A_0 and the linear function of S_1, S_2 orthogonal to it, S_0 . Explicitly, let

$$A(c) = \bar{c}A_0 + cS_0, \quad S(c) = cA_0 - \bar{c}S_0 = A(-\bar{c}) \quad (c \in (-1, 1)),$$

where $\bar{c} = \sqrt{1 - c^2}$ and note that $A(c), S(c)$ are standard scale, orthogonal and jointly sufficient for θ . To obtain an expansion for $F_n(c) = D^2(A(c))$ we need their joint cumulants in terms of those of A_0, S_0 . These are

$$\begin{aligned} \alpha_{10}^* &= c\alpha_{01}, & \alpha_{01}^* &= -\bar{c}\alpha_{01}, & \alpha_{11}^* &= \alpha_{11}(c^2 - \bar{c}^2) + (\alpha_{20} - \alpha_{02})c\bar{c}, \\ \kappa_{12}^* &= \kappa_{12}\bar{c}^3 + (\kappa_{03} - 2\kappa_{21})c\bar{c}^2 + (\kappa_{30} - 2\kappa_{12})c^2\bar{c} + \kappa_{21}c^3. \end{aligned}$$

Substituting these into the functions at (4.2) gives an expression for Δ^2 and leads directly to the form

$$i^{-1}F_n(c) = g(c) + \frac{h(c)}{\sqrt{n}} + \frac{k(c)}{n} + O(n^{-3/2})$$

with $g(c) = \alpha_{01}^2(1 - c^2)$ and

$$\begin{aligned} h(c) = \alpha_{01}^2 \left| \frac{2\alpha_{20} - 2\alpha_{02}}{\alpha_{01}} c\bar{c}^2 + \frac{2\alpha_{11}}{\alpha_{01}} \bar{c}(c^2 - \bar{c}^2) + \kappa_{12}\bar{c}^5 + (\kappa_{03} - 2\kappa_{21})c\bar{c}^4 \right. \\ \left. + (\kappa_{30} - 2\kappa_{12})c^2\bar{c}^3 + \kappa_{21}c^3\bar{c}^2 \right|, \end{aligned}$$

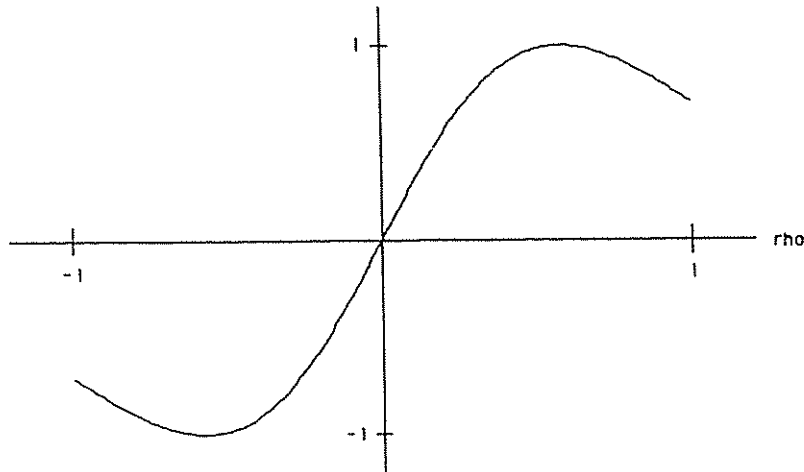
taking $\alpha = 1$. Now $g(c)$ is maximised at $c_0 = 0$ and $g''(0) = -2\alpha_{01}^2$, and

$$h'(0) = \text{sgn}(\kappa_{12} - 2\alpha_{11}/\alpha_{01}) (2\alpha_{01}(\alpha_{20} - \alpha_{02}) + \alpha_{01}^2\kappa_{03} - 2\alpha_{01}^2\kappa_{21}),$$

so the optimising value of c is

$$\hat{c}_n = \text{sgn}(2\alpha_{11}/\alpha_{01} - \kappa_{12}) \frac{1}{\sqrt{n}} \left(\frac{\alpha_{20} - \alpha_{02}}{\alpha_{01}} + \frac{\kappa_{03}}{2} - \kappa_{21} \right)$$

(see Lemma 3). The choice $\beta = 1$ is essential to these results. The chosen value of α enters directly as a scale parameter for \hat{c}_n but has no effect on the order of the correction in the sample size n .

Fig. 1.— $\sqrt{n} \hat{c}_n(\rho)$

Example 4. Let U and V be standard normal random variables with correlation ρ . For a sample of n independent observations the MSS for ρ is

$$S_1 = \frac{1}{n} \sum_{i=1}^n (U_i^2 + V_i^2), \quad S_2 = \frac{1}{n} \sum_{i=1}^n U_i V_i,$$

and the first order ancillary is just the standardised version of S_1 . The standardised MSS in the notation of this section is then

$$A_0 = \frac{1 - \frac{1}{2}S_1}{\sqrt{1 + \rho_0^2}}, \quad S_0 = \left(S_2 - \rho_0 - \frac{\rho_0(S_1 - 2)}{1 + \rho_0^2} \right) \frac{\sqrt{1 + \rho_0^2}}{1 - \rho_0^2}.$$

The moments of A_0 and S_0 are

$$E(A_0) = 0, \quad \text{var}(A_0) = \frac{1 + \rho^2}{n(1 + \rho_0^2)}, \quad \text{cov}(A_0, S_0) = \frac{2(1 - \rho_0\rho)(\rho - \rho_0)}{n(1 - \rho_0^4)}$$

$$E(S_0) = \frac{\sqrt{1 + \rho_0^2}}{1 - \rho_0^2} \frac{\rho - \rho_0}{\sqrt{n}},$$

$$\text{var}(S_0) = \frac{1 - \rho^2}{n(1 - \rho_0^2)} + \frac{2(\rho - \rho_0)[(\rho - \rho_0)(1 + \rho_0^2) - 2\rho_0(1 - \rho_0\rho)]}{n(1 + \rho_0^2)(1 - \rho_0^2)^2}.$$

The required derivatives and third order cumulants are thus

$$\alpha_{10} = 0, \quad \alpha_{20} = \frac{2\rho_0}{1 + \rho_0^2}, \quad \alpha_{11} = \frac{2}{1 + \rho_0^2}, \quad \alpha_{01} = \frac{\sqrt{1 + \rho_0^2}}{1 - \rho_0^2}, \quad \alpha_{02} = \frac{-2\rho_0(3 + \rho_0^2)}{1 - \rho_0^4},$$

$$\kappa_{21} = \frac{2\rho}{(1 + \rho^2)^{3/2}}, \quad \kappa_{03} = \frac{-2\rho(3 + \rho^2)}{(1 + \rho^2)^{3/2}}.$$

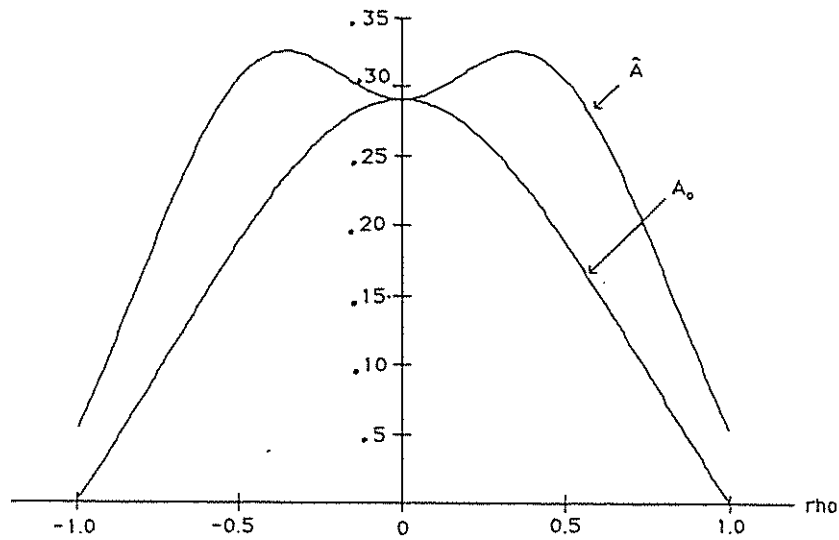


Fig. 2.— Conditioning effectiveness of A_0 and \hat{A} measured by $\sqrt{\psi} - 1$ ($n = 20$)

Notice that

$$\Delta^2(\rho, A^*(c)) = \Delta^2(-\rho, A^*(-c)),$$

so $\hat{c}^*(-\rho) = -\hat{c}^*(\rho)$. This may be inferred from the symmetry properties of the model under sign changes. Figure 1 shows a graph of the numerator of \hat{c}_n as a function of ρ . Note that

$$\frac{2\alpha_{11}}{\alpha_{01}} - \kappa_{12} = \frac{2(1 - \rho^2)}{(1 + \rho^2)^{3/2}}$$

is always positive. When $\rho = 0$ is the point of interest, the correction to the first order ancillary A_0 is zero to this order. For all other values of ρ the correction is non-zero and is actually a maximum (of ± 1.0) when $\rho = \pm 1/\sqrt{3}$. $\Delta^2(A_0)$ is given by

$$\Delta^2(A_0) = 1 + \frac{2(1 - \rho^2)}{(1 + \rho^2)^{3/2}\sqrt{n}} + O(n^{-1}) = \Delta^2(A(\hat{c}_n)) + O(n^{-1}).$$

However, the effectiveness of $\hat{A} \equiv A(\hat{c}_n)$ actually exceeds that of A_0 by an amount $2\hat{c}_n^2/n$. Provided the neglected $O(n^{-1})$ term is not too large compared to this, one should be able to make a rough comparison of the effectiveness of these two statistics via

$$\Delta^2(\hat{A}) = \Delta^2(A_0) + \frac{2\hat{c}_n^2}{n}$$

where both estimates of Δ^2 are only correct to $O(n^{-1/2})$. Even more usefully, one can convert these values into equivalent machine variances using formula (3.3) for ψ . Figure 2 presents plots of $100(\sqrt{\psi} - 1)$ against ρ for A_0 and \hat{A} at $n = 20$. One can interpret $100(\sqrt{\psi} - 1)$ as the percentage difference in standard deviation of two machines. For instance when $\rho = \pm 0.55$, conditioning on A_0 compares with conditioning in a machine example with a difference in standard deviation of 17% while conditioning on \hat{A} compares to a figure of around 29%.

The graph indicates that for $.2 < |\rho| < .9$ there is considerable difference in the conditioning efficacy of the two statistics. Note also that the efficacy of conditioning is least when $|\rho|$ is closer to 1, despite there being a larger amount of total information available ($i(\rho) = n\alpha_{01}^2$). Optimal conditional inference would be based on the conditional distribution of $S(\hat{c}_n)$ given $A(\hat{c}_n)$ which may be approximated by

$$f_{\hat{S}|A=a}(s) = \phi(s) \left(1 + \frac{1}{6\sqrt{n}} H^T \rho^{*[3]}(a, s) \right) + O(n^{-1})$$

in the notation of Barndorff-Nielsen & Cox (1979) where the $O(n^{-1})$ term is also given. This involves calculating the third and fourth order cumulants κ_{ij}^* .

5. Conditioning in the Presence of Nuisance Parameters

In addition to the parameter of interest θ , models often contain parameters of no direct interest in themselves, usually termed nuisance parameters. Typically, inference on θ is required, valid regardless of the value of the nuisance parameters henceforth labelled ϕ . A common device for achieving such inference is to identify sufficient statistics, S , for ϕ . It is often recommended that inference be based on the conditional distribution of the data. While it is true that this produces ϕ -free inference this is not necessarily the only way to proceed.

Taking the testing problem as an illustration, one can obtain a UMPU critical region of size α by 'glueing together' S -conditional regions of size α (Lehmann & Scheffé, 1950). The conditional power of this test can be calculated regardless of ϕ . In contrast, its unconditional power depends on ϕ but may be bounded as ϕ varies, in which case a ϕ -free bound for the power may be given. Do we evaluate the test conditionally or unconditionally?

From one point of view, conditioning out nuisance parameters with a sufficient statistic S is quite a different process from conditioning in the single parameter framework. It is a device for separating θ -inference from irrelevant details such as the value of ϕ . However, we may still apply exactly the same criteria in evaluating S as a conditioning variable, namely,

- (1) how much information is lost in conditioning on S , and
- (2) to what extent does the value of S index the actual precision attained?

Put another way, even though the original motivation for conditioning on S is to obtain inference about θ alone, S still performs much the same function as a conditioning variable in the single parameter case, over and above its function of eliminating nuisance parameters. How may we measure the efficacy of this function?

An analogue of equation (2.2) is easily developed for this purpose. Define $I^X(\theta; \phi)$ by

$$I^X(\theta; \phi) = E \left(\frac{-\partial^2 \log \ell(X; \theta, \phi)}{\partial \theta \partial \phi} \right)$$

and the information $I^X(\theta | \phi)$ in the statistic X about θ ignoring ϕ by

$$I^X(\theta | \phi) = I^X(\theta) - \frac{I^X(\theta; \phi)^2}{I^X(\phi)}.$$

$I^X(\theta | \phi)$ is often used, in the presence of nuisance parameters, to measure θ -information, though generalisation is possible (Godambe, 1984). If we denote conditional information in the obvious way, for instance

$$I_S(\theta; \phi) = E\left(\frac{-\partial^2 \log \ell_S(\theta; \phi)}{\partial \theta \partial \phi} \mid S\right),$$

then it is easy to show that the equation

$$I(\theta | \phi) = E(I_S(\theta)) + I^S(\theta | \phi) \quad (5.1)$$

holds, where we have used the fact that S is sufficient for ϕ so that $I_S(\theta; \phi) = 0$. We thus define

$$\Delta^2(S) = E\left(\frac{I_S(\theta)}{I(\theta | \phi)}\right) + \alpha \text{SD}\left(\frac{I_S(\theta)}{I(\theta | \phi)}\right) \quad (5.2)$$

as a measure of the conditioning merit of S . This differs from (3.1) only in the use of $I(\theta | \phi)$ rather than $I(\theta)$ as a divisor.

A common situation where nuisance parameters are easily eliminated is the (2,2)-exponential family characterised by a density of the form

$$f_{X,S}(x, s) = \frac{g(x, s)e^{x\theta + s\phi}}{h(\theta, \phi)}.$$

The nuisance parameter ϕ can be eliminated by conditioning on S . The resulting conditional density is an exponential family so that UMPU tests of θ are easily constructed. It is straightforward to show that

$$I(\theta) = \text{var}(X), \quad I(\theta; \phi) = \text{cov}(X, S), \quad I(\phi) = \text{var}(S),$$

so that $I(\theta | \phi) = \text{var}(X)(1 - \text{corr}^2(X, S))$. Moreover, $I_S(\theta) = \text{var}(X | S)$ so to calculate $\Delta^2(S)$ we must find the expectation and variance (with respect to S) of

$$\frac{\text{var}(X | S)}{\text{var}(X)(1 - \text{corr}^2(X, S))}.$$

These ideas are illustrated in the following example.

Example 5. Comparative trials are one of the simplest statistical experiments. Suppose that we have two independent binomial variables

$$X_i = \text{Bi}(p_i, N_i) \quad (i = 1, 2)$$

with interest being in whether or not $p_1 = p_2$. The natural parametrisation of this problem is in terms of

$$\theta = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right), \quad \phi = \log\left(\frac{p_2}{1-p_2}\right).$$

Then $\theta = 0$ corresponds to $p_1 = p_2$ and ϕ is a nuisance parameter. The joint distribution of X_1, X_2 is of the exponential family form and we find that $T = X_1 + X_2$ is sufficient for ϕ . There is still disagreement over whether or not the hypothesis $\theta = 0$ should be tested conditional on the observed value of T (Barnard, 1945; Berkson, 1978; Yates, 1984; Lloyd, 1988). Conditional on T , X_1 has a generalised hypergeometric distribution leading to the Fisher-Yates exact test. The information about θ ignoring ϕ is

$$I(\theta | \phi) = \frac{N_1 N_2 p_1 p_2 (1-p_1)(1-p_2)}{N_1 p_1 (1-p_1) + N_2 p_2 (1-p_2)}$$

expressed in terms of p_1, p_2 for convenience. The information about θ ignoring ϕ in the statistic T is

$$I^T(\theta | \phi) = \text{var}(\mathbb{E}(X_1 | T; \theta)) - \frac{(N_1 p_1 (1-p_1))^2}{N_1 p_1 (1-p_1) + N_2 p_2 (1-p_2)}$$

for which $I^T(\theta | \phi) > 0$ unless $\theta = 0$. When $\theta = 0$, $\mathbb{E}(X_1 | T) = N_1 T / (N_1 + N_2)$, and so $I^T(\theta | \phi)$ vanishes. Hence, if $\theta = 0$ then

$$\mathbb{E}\left(\frac{I_T(\theta)}{I(\theta | \phi)}\right) = 1.$$

The expression for $I(\theta | \phi)$ also simplifies to $N_1 N_2 p(1-p) / (N_1 + N_2)$, where p is the assumed common value of p_1, p_2 . Since X_1 has a conditional hypergeometric distribution,

$$\frac{I_T(\theta)}{I(\theta | \phi)} = \frac{T(N_1 + N_2 - T)}{(N_1 + N_2)(N_1 + N_2 - 1)p(1-p)};$$

we require the variance of this expression with respect to the $\text{Bi}(p, N_1 + N_2)$ distribution of T . Some algebra leads to

$$\Delta^2(T) = 1 + \left(\frac{(2p-1)^2}{(N_1 + N_2)p(1-p)} + \frac{2(p^2 - p + 1)}{(N_1 + N_2)(N_1 + N_2 - 1)^2 p(1-p)} \right)^{1/2}$$

for $N_1 + N_2 > 1$. This equation is symmetric in p and $1-p$ as required by the invariance properties of the model. It is evident that the closer p is to 0 or 1 the more favourably we judge T as a conditioning variable. Note that there is less total information available for extreme values of p . Hence, T must index

experiments of (on average) worse precision more effectively. In fact, when $p = \frac{1}{2}$, $\Delta^2 = 1 + O((N_1 + N_2)^{-3/2})$ but otherwise exceeds unity by $O((N_1 + N_2)^{-1/2})$. In the correlation model of Example 4 we found a similar feature of conditioning being effective for those cases where less total information is available, but there is no reason why this should be true in general. A case of some interest is $N_1 + N_2 = 2$ when

$$\Delta^2(T) = 1 + \sqrt{\frac{3}{2p(1-p)}} - 3.$$

This has a minimum value of $1 + \sqrt{3}$ when $p = \frac{1}{2}$, corresponding to $\psi = 13.9$. Contingency tables with $N_1 + N_2 = 2$ arise in the analysis of binary responses for matched pairs and are analysed by conditioning on T since otherwise the MLE of θ is inconsistent as the number of matched pairs, and hence tables, increases. The above calculation shows that even for a single table, T is extremely effective as a conditioning variable and conditioning on it is strongly indicated. For tables with larger values of $N_1 + N_2$, $\Delta^2(T)$ is smaller. For example if $p = 0.1$ and $N_1 + N_2 = 50$ then $\Delta^2(T) = 1.377$ and $\psi = 2.09$. Clearly, conditioning on T is still strongly indicated.

It would be of interest to extend these calculations to arbitrary linear logistic models or, indeed, to generalised linear models (GLMs). Davison (1988) has given formulae for approximating conditional significance probabilities to order $n^{-3/2}$ in GLMs. What is not established in any generality is whether or not the conditioning is important when n is large enough for these approximations to be adequate. The example of the 2×2 table suggests that in many cases it will be.

Appendix

Proof of Lemma 1. Using (2.1),

$$U(A, B) = i - i^{A,B} = i - (i^A + E(i_A^B)) = U(A) - E(i_A^B), \quad (\text{A.1})$$

and taking the expectation of definition (2.4) gives (ii). Equation (A.1) shows that $U(A, B) = U(A)$ if and only if $i_A^B = 0$, i.e. that B is ancillary given A . To show (iii) we have

$$\begin{aligned} E(R^2(B | A)) &= E(\text{var}(i_{A,B} | A)) = \text{var}(i_{A,B}) - \text{var}(E(i_{A,B} | A)) \\ &= \text{var}(i_{A,B}) - \text{var}(i_A - i_A^B), \end{aligned}$$

using (2.4) for the last equality. We thus obtain the equality of (iii) except for a term $\text{cov}(i_A^B - 2i_A, i_A^B)$. A sufficient condition for this to vanish is that i_A^B be free of A . The conditions quoted are in turn sufficient for this.

Proof of Theorem 1. The functional equation (4.1) together with the requirement of continuity leads to the family of solutions

$$D^2(A) = \alpha U(A)^\beta R(A)^\gamma \quad \text{or} \quad \gamma U(A)^\beta + \alpha R(A)^\beta$$

for arbitrary real α, β, γ . The first inequality required by validity leads to the rejection of the first solution else $D^2(C) = 0 = D^2(S)$. It further implies that $\gamma > 0$ in the second solution so we may take it to equal 1. The second inequality implies that $\alpha > 0$ while the requirement that D^2 increase in both its arguments implies that $\beta > 0$.

Proof of Theorem 2. If Δ^2 is an absolute measure of merit then it defines a relative measure of merit

$$d^2(B, A) = \frac{\Delta^2(B)}{\Delta^2(A)} = \frac{D^2(B)}{D^2(A)}$$

and by Theorem 1, Δ^2 must have the same form as D^2 given there, but rescaled by some constant. Since $D^2(C) = i^\beta$ the appropriate constant is $i^{-\beta}$, i.e.

$$\Delta^2(A) = \frac{U(A)^\beta + \alpha R(A)^\beta}{i^\beta}.$$

Next let A_1 be defined on experiment 1 with $\Delta^2(A_1) = 1$ and C_2 a constant defined on experiment 2. Further let $\ell = i^{A_1}/i_1$ be the proportion of information lost in conditioning on A_1 and $r = i_1/(i_1 + i_2)$ be the information in experiment 1 as a proportion of the total. Then taking $\Delta^2(A_1) = 1$ and applying Lemma 2, a little algebra leads to

$$\Delta^2(A_1, C_2) = (1 - r\ell)^\beta + r^\beta(1 - (1 - \ell)^\beta),$$

which necessarily equals 1 if and only if $\beta = 1$.

Lemma 3. Let g, h, k be twice differentiable functions from \mathbb{R} to \mathbb{R} , and let $\{F_n(\cdot)\}$ be a sequence of twice differentiable functions satisfying

$$F_n(c) = g(c) + \frac{h(c)}{\sqrt{n}} + \frac{k(c)}{n} + O(n^{-3/2}).$$

Suppose that g is maximised at c_0 , that $g'(c_0) = 0$ and $g''(c_0) < 0$. Then $F_n(\cdot)$ is maximised at c_n where

$$c_n = c_0 - \frac{h'(c_0)}{g''(c_0)\sqrt{n}} + O(n^{-1}),$$

and

$$F(c_n) = F(c_0) - \frac{[h'(c_0)]^2}{2n|g''(c_0)|} + O(n^{-3/2}).$$

The main point to note about the lemma is that c_n is obtained by differentiating F_n only down to the $O(n^{-1/2})$ term, but the increase in the value of the

objective function F_n is only $O(n^{-1})$. Thus, given an expansion of an objective function F_n down to the $O(n^{-1/2})$ term the improvement is not apparent to this order.

References

- BASU, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A* **26**, 3–16.
- BARNARD, G.A. (1945). A new test for 2×2 tables. *Nature* **156**, 177.
- BARNDORFF-NIELSEN, O. & COX, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. Roy. Statist. Soc. Ser. B* **41**, 279–312.
- BERKSON, J. (1978). In dispraise of the exact test. *J. Statist. Plann. Inference* **2**, 27–42.
- COX, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.
- (1971). On the choice between alternative ancillary statistics. *J. Roy. Statist. Soc. Ser. B* **33**, 251–255.
- (1980). Local ancillarity. *Biometrika* **67**, 279–286.
- (1988). Some aspects of conditional and asymptotic inference: a review. *Sankhyā Ser. A* **50**, 314–337.
- DAVISON, A. (1988). Approximate conditional inference in generalised linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 445–461.
- FISHER, R.A. (1934). Two new properties of the mathematical likelihood. *Proc. Roy. Soc. London Ser. A* **144**, 285–307.
- (1948). Conclusions fiduciaire. *Ann. Inst. H. Poincaré Probab. Statist.* **10**, 191–213.
- (1959). *Statistical Methods and Scientific Inference*, 2nd edn. Oliver & Boyd, London.
- GODAMBE, V.P. (1984). On ancillarity and Fisher information in the presence of nuisance parameters. *Biometrika* **71**, 626–629.
- LEHMANN, E.L. & SCHEFFÉ, H.O. (1950). Completeness, unbiasedness and similar regions: Part 1. *Sankhyā* **10**, 305–339.
- LINDSAY, B. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503–512.
- LLOYD, C.J. (1988). Some issues arising from the analysis of 2×2 contingency tables. *Austral. J. Statist.* **30**, 35–46.
- (1991). Asymptotic expansions of the information in a mean. *Statist. Probab. Lett.* **11**, 133–137.
- YATES, F. (1984). Tests of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. Ser. A* **147**, 426–463.