Conditional statistical inference and quantification of relevance

Rolf Sundberg*, Stockholm University

December 29, 2000

Abstract

We argue that it can be fruitful to take a predictive view on notions as the precision of a point estimator and the confidence of an interval estimator in frequentist inference. This predictive approach has implications for conditional inference, because it immediately allows a quantification of the concept of relevance for conditional inference. Conditioning on an ancillary statistic makes inference more relevant in this sense, provided that the ancillary is a precision index. Not all ancillary statistics satisfy this demand. We discuss the problem of choice between alternative ancillary statistics. The approach also has implications for the best choice of variance estimator, taking account of correlations with the squared error of estimation itself. The theory is illustrated by numerous examples, many of which are classical.
Keywords: ancillarity, confidence, precision index, precision of estimate, predictive approach, variance estimators.

1 Introduction

In this paper we will discuss frequency-based statistical inference about a model parameter in a statistical model. Quoting Cox (1958), we state that the aim of statistical inference about a model parameter θ is to find out "what we can learn from the data that we have". When all probability statements about the parameter are interpreted as long-run frequencies under hypothetical repetition, we should make sure that the long run matches the given data in relevant aspects. This is often accomplished by restricting the hypothetical repetitions to a more relevant set than the whole sample space. With the notion of ancillarity, one (not quite precise) specification of this argument is expressed in the conditionality principle, that inference about θ should be conditional on any ancillary statistic, that is the conclusions should be drawn as if the ancillary statistic had been fixed at its observed value, se e.g. Cox & Hinkley (1974, Sec. 2.3). This is a way of making the inference *relevant* to the particular set of data (Cox & Hinkley, 1974, Sec. 2.4, Barndorff-Nielsen & Cox, 1994, Sec. 2.4). On the other hand, it is well-known that we cannot argue for conditioning using standard optimization criteria like minimum variance of estimators,

*Research financed by the Swedish Natural Science Research Council

minimum length of confidence intervals, or maximum power of significance tests, see e.g. Cox & Hinkley, Sec. 4.4. Therefore it has appeared in literature as if we can only argue for conditioning by reference to intuition and common sense (and well-behaved examples), as expressed by McCullagh (1992):

> "My own assessment of the case is that, even though the evidence is largely circumstantial, the argument for conditioning is extremely compelling. The fact that the theory is rooted in examples is simultaneously its strength and its weakness."

The motivation behind the present paper was a desire to understand better the concept of relevance, the case for conditioning, and the role of ancillarity. It will be argued that it can be fruitful to regard precision and confidence as quantities to be predicted. This *predictive approach* immediately opens for a natural *quantification* of the concept of relevance (Sec. 2.1). When the predictive principle is applied to situations with ancillary statistics it tells us to make the inference conditional, provided that the ancillary statistic is a *precision index*, that only affects the precision, in a sense to be specified in Sec. 2.2.

Consider for example the MLE $\hat{\theta}$, which is the same whether we condition on an ancillary or not, and perhaps for simplicity assume $\hat{\theta}$ unbiased. The basic problems of statistical inference are to attach to the observed value of $\hat{\theta}$ with maximum relevance a standard error, a confidence interval, and perhaps also a significance level (p-value). We will firstly and mostly discuss precision of point estimators, Sec. 2. This is not to pretend that confidence statements should be less important than standard errors (which do not have the same invariance properties), but variance estimation seems more primitive and basic, and confidence results follow mostly by analogy, Sec. 3. We will discuss properties required from ancillary statistics to make them suitable for conditioning (Sec. 2.2), and the problem of choice between alternative ancillary statistics, when such exist (Sec. 2.5). Variances, conditional or not, must usually be estimated. In Sec. 2.3 aspects of variance estimation are found, taking into account possible correlation between a variance estimator and the actual quadratic error. An extension of the quantitative criterion for relevance from quadratic error to arbitrary loss functions is given in Sec. 2.4.

Another argument for constructing conditional procedures is the wish to obtain similarity by eliminating nuisance parameters. In principle this argument applies to a different class of situations than the one considered in the present paper, even though there is a considerable overlap. Example 15 below illustrates such aspects.

In order not to shade the simple main theme of the paper we will avoid mathematical formalities and intricacies. We do not want here to get involved in the mathematical complications of exact higher moment calculations, cf. Sundberg (1994), or of higher order asymptotics, cf. Reid (1995) for a review and Lindsay & Li (1997) for recent results.

2 Precision of point estimators

2.1 The predictive principle

Suppose we have a statistical model with a parameter θ of interest. Let $\hat{\theta}$ be an unbiased estimator of θ. Traditional theory and practice of statistical inference constructs a precision value (standard error) to an observed $\hat{\theta}$ in two steps:

- As a measure of the squared error of $\hat{\theta}$, use its expected value $v = E[(\hat{\theta} - \theta)^2]$
- Typically this variance of $\hat{\theta}$ depends on θ and/or on nuisance parameters, $v = v_{\theta,\psi}$, so construct an estimate V for it, for example by inserting parameter estimates in the variance v. This yields a (squared) *standard error* of $\hat{\theta}$.

Optimization is typically done separately for these two steps; thus in the second step we will be told to estimate the variance v as precisely as possible. This is not a compelling argumentation, however, because when doing so the aim of v has been forgotten, to represent our knowledge about the squared error of $\hat{\theta}$, $(\hat{\theta} - \theta)^2$. Different variance estimators can be more or less correlated with this squared error (see Sec. 2.3). This can be taken into account by selecting the variance estimator according to its properties as a *predictor* of the actual (random) squared error $(\hat{\theta} - \theta)^2$. If we look for an (observable) statistic that can be used to represent this squared error, we are in fact trying to predict it, since it is random. This is the background for the following principle.

Predictive principle: A squared standard error should aim at predicting best possible the squared error $(\hat{\theta} - \theta)^2$.

This principle is immediately quantifiable: Calculate the expected value of some measure of size of the prediction error $(\hat{\theta} - \theta)^2 - V$, where V is a statistic to be used as a predictor for $(\hat{\theta} - \theta)^2$, for example an estimator of v. If we use a quadratic measure of size of this prediction error we are led to consider the quantity

$$M_V = E[\{(\hat{\theta} - \theta)^2 - V\}^2]. \tag{1}$$

We will use this quantity, and it will be called the *MSE of PSE* for V, the mean squared error of the predicted squared error. Alternatively (but not equivalently) we could have used some other measure of the error than the squared error, for example the relative squared error, or the absolute value of the error, cf. Section 2.4 where we replace $(\hat{\theta} - \theta)^2$ by a general loss function $L(\hat{\theta}, \theta)$. Version (1) is often the only one allowing explicit theoretic calculations, but for simulation studies we may as well use a different version.

Sandved (1968) formulated this criterion for an arbitrary loss function, even though she called it estimation of the loss function. Her paper seems to have been largely forgotten,

but Goutis & Casella (1995) called attention to it. The present author (Sundberg, 1994) proposed and used the measure (1) to compare variance estimators in sample survey inference. The criterion has also been used by Lindsay & Li (1997) to show that the inverse of the observed Fisher information is the asymptotically optimal choice of predictor.

In Sandved's Theorem 1 she showed that if the parametric model has a complete sufficient statistic T, there is at most one statistic V, function of T, that is unbiased, i.e. has expected value v, so if there is one, it is the unique best unbiased. Here is a simple illustration of this result.

Example 1. The precision of a normal mean.
Suppose we have a fixed size n sample $(y_1, \ldots, y_n)$ from a normal distribution $N(\theta, \sigma^2)$, with unknown parameters, and let θ be estimated by the sample mean $\hat{\theta} = \bar{y}$. Since $T = (\bar{y}, s^2)$ is sufficient and complete, with s^2 unbiased for σ^2, Sandved's result implies that s^2/n is the one and only unbiased predictor of $(\hat{\theta} - \theta)^2$ that is a function of T. □

Applications like this one ensure that a shift of paradigm to the predictive approach would not bring us away from what we have conventionally learnt to do in standard situations. However, we have less standard situations in mind here, and the next example illustrates the type of results to come. There is certainly a consensus that we should condition on an ancillary random sample size for the sake of relevance, so qualitatively Example 2 will does not bring much new; the novel feature is that we have a quantitative criterion (1) of the predictive principle establishing that conditioning improves relevance.

Example 2. A random size sample without nuisance parameters.
Suppose a random sample of random size $N > 0$ has unknown mean θ but known variance σ_0^2. To estimate θ we use the usual sample average $\bar{y}$, and we all know that conditionally on the sample size, $\mathrm{var}(\bar{y}|N) = \sigma_0^2/N$. By using this conditional variance as predictor of the quadratic error, instead of the constant total variance $\mathrm{var}(\bar{y}) = \sigma_0^2\, E(1/N)$, the MSE of PSE is reduced by the quantity $\sigma_0^4 \,\mathrm{var}(1/N)$, from

$$\mathrm{var}\{(\hat{\theta} - \theta)^2\},$$

the total variance of the squared error, to

$$\mathrm{var}\{(\hat{\theta} - \theta)^2\} - \sigma_0^4 \,\mathrm{var}(1/N).$$

□

The result of Example 2 is a direct consequence of Proposition 1 below. As before $\hat{\theta}$ is assumed to be an unbiased estimator of θ, and even more, conditionally unbiased given a

statistic U. The unbiasedness is not only imposed for mathematical convenience, but also represents a necessity that there must not be serious (conditional) systematic error in $\hat{\theta}$, see further Secion 2.2.

Proposition 1 *Let U be a statistic such that*

$$\begin{aligned} E(\hat{\theta}|U) &= \theta, \\ var(\hat{\theta}|U) &= v(U)\,, \end{aligned}$$

where $V = v(U)$ is a statistic to be used as predictor in (1). Then V has MSE of PSE

$$M_V = var\{(\hat{\theta} - \theta)^2\} - var(V). \tag{2}$$

Proof: Elementary, but since several analogous results will follow, the proof is written out this first time. Adding and subtracting the expected value $v = E\{v(U)\}$ within (1) we find

$$E[\{(\hat{\theta} - \theta)^2 - V\}^2] = \mathrm{var}\{(\hat{\theta} - \theta)^2\} - 2\,\mathrm{cov}\{(\hat{\theta} - \theta)^2, V\} + \mathrm{var}\{V\}.$$

Here the covariance equals the last variance, since $E\{(\hat{\theta} - \theta)^2|U\} = V$, and hence the desired result follows. □

Note that formula (2) *per se* does not require V to be a statistic (i.e. parameter-free). We will later refer to Prop. 1 even when $\mathrm{var}(\hat{\theta}|U)$ involves parameters, and in Sec. 2.3 we consider the modifications necessary when V is an estimator of $\mathrm{var}(\hat{\theta}|U)$.

Note also that the first term of (2) is independent of the choice of V, whereas the second term quantifies the possible *gain from conditioning* on U. The proposition induces a partial ordering of all possible U, as expressed in the corollary below.

Corollary 1 *According to the predictive principle, the finer the partitioning induced by U, the more relevant it is to measure the precision conditional on the observed value of U, for U such that $\hat{\theta}$ is conditionally unbiased (and $var(\hat{\theta}|U)$ is a statistic).*

Example 2 was a simple situation where Proposition 1 applies. Here is a modified example of clear practical importance.

Example 3. Precision under post stratification.
Suppose a finite population is divided in H strata of known stratum sizes N_h, $\sum N_h = N$, with unknown mean values $\overline{Y}_h$ but strata variances S_h^2 regarded as known from previous studies. Of interest are the strata mean values and the population mean value $\overline{Y} = \sum(N_h/N)\overline{Y}_h$. After taking a simple random sample of fixed size n from the population, this sample is post stratified, yielding a set $\mathbf{n}$ of random strata sample sizes n_h, and observed strata means $\bar{y}_h$. We assume the expected n_h-values so large that the probability for any n_h being zero is negligible. The natural post stratified estimator of the population

mean is $\widehat{Y} = \sum(N_h/N)\bar{y}_h$. Conventionally its precision would be expressed by the formula for its total sampling variance, v say, but Holt & Smith (1979) argued for conditioning on the ancillary n_h-values and for using instead the usual variance under pre-stratification,

$$v(\mathbf{n}) = \sum(N_h/N)^2(1 - f_h)S_h^2/n_h,$$

Since $\widehat{Y}$ is conditionally unbiased, $v = E\{v(\mathbf{n})\}$, and Proposition 1 applies with $V = v(\mathbf{n})$. Hence the argumentation by Holt & Smith is reinforced by the quantification of relevance given above. An analogous argument can be given in the case of estimation of a particular stratum mean, and approximate explicit formulae for the MSE of PSE can easily be derived. The modifications needed when S_h^2 are estimated follow as in Sec. 2.3.

In less simple sample survey situations the choice of precision formula must typically be discussed in the setting of a superpopulation model, cf. Sundberg (1994). □

2.2 Ancillarity and precision indices

Consider a statistic U that is *ancillary* for θ, in the sense of having a distribution that does not involve the parameter θ. Furthermore let us assume that $\hat{\theta}$ is conditionally unbiased, given U, as in Proposition 1. Then the predictive principle supports the *conditionality principle*, by implying that we should measure the precision conditionally on U.

The demand only that an ancillary statistic have a distribution not involving the parameter θ is deliberately somewhat vague and less restrictive than for example the definition in Barndorff-Nielsen & Cox (1994) or Cox & Hinkley (1974). They avoid some of the counter-examples to the conditionality principle by demanding that the ancillary statistic be part of the minimal sufficient statistic, in other words, that we first reduce data to a minimal sufficient statistic before we look for an ancillary statistic. For a comprehensive account of various ancillarity definitions, see for example Barndorff-Nielsen (1978, Ch. 4). In the present approach, based on criterion 1, does not explicitly involve the likelihood; Our use of the notion of ancillarity includes for example S-ancillarity, where the distribution of the ancillary statistic is allowed to depend on a nuisance parameter, variation independent of the parameter of interest (Barndorff-Nielsen & Cox, 1994, Sec. 2.5). It might appear a defect that we do not explicitly demand variation independence here, but instead we assumed conditional unbiasedness of $\hat{\theta}$. This is a different but related assumption, see Lemma 1 below. Essentially *conditional unbiasedness* is an assumption on U to be a *precision index*, that affects only the precision of $\hat{\theta}$, not its location. This seems to conform with what Fisher (1935, p. 48) had in mind when he wrote

> "...ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it."

That the distribution of U does not involve θ is not a sufficient condition for an unbiased estimator $\hat{\theta}$ to be also conditionally unbiased, given U, not even for the ML estimator. The third of Basu's (1964) examples of anomalous ancillary statistics provides a simple but striking counter-example to such dreams. There are more of his examples in which the conditional unbiasedness is violated, but they are not as simple as this one:

Example 4. Basu's third example, lacking conditional unbiasedness.
Let Y be a single observation from a uniform distribution on the unit length interval $[\theta, \theta + 1]$, where the parameter θ is any real number. There is no unique ML estimator of θ since the likelihood is constant on an interval. One unbiased estimator is $\hat{\theta} = [Y]$, the integer part of Y, and the fractional part $U = Y - [Y]$ is ancillary, having a uniform distribution on the unit interval. For simplicity of writing, suppose $0 < \theta < 1$. Then $\hat{\theta}$ has a two-point distribution on 0 and 1, with probabilities $(1 - \theta)$ and θ, respectively. It follows that $\hat{\theta}$ is unbiased. However, given U, $\hat{\theta}$ has a one-point distribution on either 0 or 1, so $\hat{\theta}$ is far from conditionally unbiased for any value of U. □

Basu's example shows that conditional unbiasedness or something else is required. Another simple such example is discussed in Sec. 3 of Helland (1995). This example can be taken to represent classical randomization-based sampling inference in general.

Example 5. Helland's sampling example
A coin flip decides whether Y is an observation from $N(\theta + \alpha, 1)$ or from $N(\theta - \alpha, 1)$, where θ is the parameter of interest and α is a nuisance parameter. In other words, a sample of size 1 from a population of size 2, with measurement errors, and the population mean to be estimated. The result of the coin flip is ancillary, even in the strict sense of being part of the minimal sufficient statistic. In this case a conditional inference provides no information about θ, whereas inference in the total model does, with $\hat{\theta} = Y$. Helland comments on the difficulties to formulate a general principle telling why we should condition in Cox's classical example of two measuring instruments of different precision, but not in this example (or in Basu's examples). Based on criterion 1 we can again resolve the problem by noting that Y is an unbiased estimator of θ that is not conditionally unbiased, so the coin flip does not serve as a precision index. □

Conditional unbiasedness should of course not be regarded as an absolute condition, but we must be able to rule out statistics U yielding serious systematic conditional errors of $\hat{\theta}$, like in Examples 4 and 5. The MSE of PSE criterion can do this, and the following proposition shows how the MSE of PSE is affected by conditional bias, if we use the conditional variance $v(U)$ to predict $(\hat{\theta} - \theta)^2$ in spite of the bias.

Proposition 2 *Let $b(U) = E(\hat{\theta}|U) - \theta$ and $v(U) = var(\hat{\theta}|U)$ be the conditional bias and conditional variance of $\hat{\theta}$, respectively, and suppose $V = v(U)$ is used to predict the squared error $(\hat{\theta} - \theta)^2$. Then V has MSE of PSE*

$$M_V = var\{(\hat{\theta} - \theta)^2\} - var\{V\} + [E\{b^2(U)\}]^2 - 2\, cov\{b^2(U), V\}.$$

Proof: Elementary. □

Example 4. Basu's third example, cont'd.
In this example, with its unconditionally unbiased but conditionally extremely biased $\hat{\theta}$, it can be quantified by Proposition 2 that it is (of course) much worse to use the conditional variance $v(U) \equiv 0$ to measure precision than the unconditional variance $\theta(1-\theta)$. The latter has $M_V = \theta(1-\theta)(1-2\theta)^2$. With $v(U) = 0$ instead, the MSE of PSE is always higher, an increase by addition of the squared expected squared bias $\theta^2(1-\theta)^2$. We conclude from Proposition 2 that the predictive criterion tells us *not* to make the inference conditional, in full agreement with common sense, whereas a blind conditioning on the ancillary statistic according to a general conditionality principle would lead totally wrong. □

In some situations, like Example 4, we have reason to reconsider the choice of $\hat{\theta}$. In other situations it might be reasonable to neglect the conditional bias, because it has a small influence. We will return to this discussion later.

One immediate conclusion from Proposition 2 is that if $b(U) =$ constant, be it zero or not, we gain in MSE of PSE by replacing the unconditional variance by the conditional one as predictor. As we have seen, not all ancillary statistics are such, but when the ancillary U can be imagined as generated from a distribution in a complete family, unbiasedness actually implies conditional unbiasedness. We formulate this positive result in the following Lemma, also found in slightly less generality as Theorem 3 (a) in Sandved (1968):

Lemma 1 *Suppose the statistic U has a distribution that can be embedded in a complete family $g(u;\nu)$ of distributions, with ν and θ variation independent and $E(\hat{\theta})$ independent of ν. Then $E(\hat{\theta}|U)$ does not depend on U, that is*

$$E(\hat{\theta}|U) = E(\hat{\theta})$$

for each (θ, ν). In particular, if $\hat{\theta}$ is unbiased, $E(\hat{\theta}) = \theta$, it is also conditionally unbiased.

Proof: Just the definition of completeness. □

Note that the family of distributions of U need not have been given in advance, it is sufficient that we can embed the actual distribution of U in such a family. However, the practical usefulness of the lemma appears to be rather limited, simply because properties of estimators are typically analysed more easily conditionally than in the whole model.

Example 2, cont'd. A random size sample without nuisance parameters.
Whether the distribution of the random sample size N is known or unknown, it is easy to imagine this distribution embedded in a family of distributions on the positive integers large enough for completeness to hold. Since $\hat{\theta}$ is unbiased, the lemma implies that it is also conditionally unbiased. However, we already knew this from the construction of $\hat{\theta}$. □

In all problematic examples of ancillary statistics, such as Basu's third example and Helland's example above, the conditional unbiasedness fails seriously. In Helland's example we can easily embed U in a complete family, the Bernoulli distribution family, but the Bernoulli parameter would go into $E(\hat{\theta})$ and violate the assumption that $E(\hat{\theta})$ must not be affected. Basu (1964) realized that the problems in his examples appeared because U was intrinsically defined within the experiment, and proposed that we must distinguish between *real* (performable) and *conceptual* (non-performable) conditional experiments; only the former type is naturally consistent with the conditionality principle. Kalbfleisch (1975) tried to formalize this idea by his notions of *experimental* and *mathematical* ancillaries: An experimental ancillary represents the first stage of a two-stage experiment, called an *experimental mixture*. What is and what is not an experimental mixture is to Kalbfleisch dependent on the experiment actually performed. Buehler (1982) imposed other restrictions, essentially demanding that the family of distributions of the MLE be conditionally a location family in some parametrization, not dependent on U.

In our much more operational setting, based on the MSE of PSE criterion and referring to a specific estimator to be used, conditional unbiasedness (or more generally constant conditional bias) given an ancillary will guarantee that we can gain by conditioning on the ancillary. This property will rarely hold unless Lemma 1 can be applied, that is unless the distribution of U can be embedded in a complete family. This condition is not only more operational but also slightly weaker than the demands of Basu and Kalbfleisch. For example they would hardly accept as real or experimental configuration statistics in location–scale families in general.

2.3 Variance estimators

Typically $\text{var}(\hat{\theta}|U)$ will also depend on the parameters, (θ, ψ), where ψ is a nuisance parameter. We must distinguish between the theoretical conditional variance and a statistic V used as an estimator for $\text{var}(\hat{\theta}|U)$. Next proposition tells the MSE of PSE for a conditionally unbiased V, and shows that in a choice between variance estimators the most precise one is not necessarily the best one, because alternatives can be more or less correlated with the actual squared error. Examples follow.

Proposition 3 *Let U and V be statistics such that*

$$\begin{aligned} E(\hat{\theta}|U) &= \theta, \\ E(V|U) &= var(\hat{\theta}|U). \end{aligned}$$

Then the predictor V has MSE of PSE

$$M_V = var\{(\hat{\theta}-\theta)^2\} - var\{var(\hat{\theta}|U)\} + E\{var(V|U)\} - 2\,E[cov\{(\hat{\theta}-\theta)^2, V|U\}].$$

Here the first term is independent of how V and U are chosen, the second term represents the ideal gain from conditioning on U (when $var(\hat{\theta}|U)$ is known), the third term represents the cost paid for estimating the unknown $var(\hat{\theta}|U)$, and the last term quantifies the gain or loss when V is conditionally correlated with the actual squared error.

Proof: Elementary. □

It is helpful to see explicitly the simple version for an empty U:

Corollary 2 *Let $\hat{\theta}$ be an unbiased estimator of θ and V an unbiased estimator of $var(\hat{\theta})$. Then V regarded as predictor has MSE of PSE*

$$M_V = var\{(\hat{\theta}-\theta)^2\} + var(V) - 2\,cov\{(\hat{\theta}-\theta)^2, V\}.$$

Remark: An example from sample survey inference, where the best choice of V (least MSE of PSE) depends crucially on the correlation between V and $(\hat{\theta}-\theta)^2$, is analysed in Sundberg (1994), cf. also Example 7 below.

Example 6. A random size normal sample with nuisance parameter.
Let a random sample of random size N be taken from $N(\theta, \sigma^2)$, and let the parameters be estimated by the usual sample mean $\bar{y}$ and the sample variance s^2, respectively, both conditionally unbiased. The usual estimate of the conditional variance $\mathrm{var}(\bar{y}|N) = \sigma^2/N$ is $V = s^2/N$ and since $\mathrm{var}(s^2|N) = 2\,\sigma^4/(N-1)$, the corollary yields the MSE of PSE

$$M(s^2/N) = \mathrm{var}\{(\hat{\theta}-\theta)^2\} - \sigma^4\,\mathrm{var}(1/N) + 2\sigma^4\,E[1/\{N^2(N-1)\}].$$

In comparison with Example 1, the last term here is new, representing the unavoidable cost for having to estimate the nuisance parameter σ^2 in $\mathrm{var}(\hat{\theta}|U)$. The covariance term of Proposition 3 does not appear since s^2 and $\bar{y}$ are conditionally independent. Note also that the distribution of N need not be known for the result to hold. □

Example 7. Inference about the variance of a normal distribution.
Now let the variance be the parameter θ of interest in a normal distribution. Suppose that we have a sample of size n from $N(\mu, \theta)$, with $\hat{\theta} = s^2$. As well-known,

$$\mathrm{var}(\hat{\theta}) = 2\,\theta^2/(n-1),$$

an unbiased estimator of which is $V = 2s^4/(n+1)$. In this case V and $\hat{\theta}$ are correlated, and the result of Corollary 2 (with the three terms in the same order) reads

$$M_V = \mathrm{var}\{(\hat{\theta}-\theta)^2\} + 32\,\theta^4(n+2)/\{(n-1)^3(n+1)\} - 96\,\theta^4/(n-1)^3.$$

Note the remarkable fact that the covariance term (the last term) dominates the preceding term (cost for estimating $\mathrm{var}(\hat{\theta})$). In other words, V is so highly positively correlated with the squared error $(\hat{\theta}-\theta)^2$ that we gain by using the estimate V instead of its expected value $E(V) = \mathrm{var}(\hat{\theta})$! □

Example 8. Simple linear regression.
Let (X_i, Y_i), $i = 1, \ldots, n$, be a sequence of n independent bivariate normally distributed pairs of random variables, and write

$$\begin{aligned} E(Y|X) &= \alpha + \beta X, \\ \mathrm{var}(Y|X) &= \sigma^2\,. \end{aligned}$$

Thus the conditional distribution of Y has three parameters. The mean and the variance τ^2 of the marginal distribution of X may be regarded as known or fully unknown. In the former case $U = \{X_i\}$ is ancillary in the simple sense, in the latter case it is S-ancillary for inference about α, β and σ. Let us consider inference about β based on the usual least squares estimator $\hat{\beta} = s_{xy}/s_{xx}$, which is conditionally unbiased with conditional variance σ^2/s_{xx} (s_{xx} is the sum of squares). A conditionally unbiased estimator of this variance is $V = \hat{\sigma}^2/s_{xx}$, with the usual $\hat{\sigma}^2$. Application of Proposition 3 yields the MSE of PSE

$$\begin{aligned} M_V &= \mathrm{var}\{(\hat{\beta}-\beta)^2\} - \sigma^4\,\mathrm{var}(1/s_{xx}) + 2\,\sigma^4\,E(1/s_{xx}^2)\,/\,(n-2) \qquad (3) \\ &= \mathrm{var}\{(\hat{\beta}-\beta)^2\} - 2\,(\sigma/\tau)^4/\{(n-3)^2(n-5)\} + 2\,(\sigma/\tau)^4/\{(n-2)(n-3)(n-5)\}. \end{aligned}$$

The near-equality for large n of the estimation cost and the gain from conditioning can be understood from the form of V as proportional to the ratio $\hat{\sigma}^2/\hat{\tau}^2$ of two mutually independent variance estimators with almost the same degrees of freedom, and the latter representing U.

If we knew τ^2, could we gain by using $V_\tau = \hat{\sigma}^2/\tau^2$ instead of $V = \hat{\sigma}^2/s_{xx}$? Corollary 2 yields the MSE of PSE for V_τ as

$$\mathrm{var}\{(\hat{\beta}-\beta)^2\} + 2\,(\sigma/\tau)^4/\{n^2(n-2)\},$$

corresponding to the first and last terms in (3). The missing gain from conditioning term confirms that V is preferable to V_τ. □

2.4 Extension to general loss functions

The choice to measure the error of $\hat{\theta}$ by its squared deviation from θ, $(\hat{\theta}-\theta)^2$, is to some extent arbitrary. We could choose any loss function $L(\hat{\theta},\theta)$ and formulate the predictive principle and subsequent criteria and results in terms of L, following Sandved (1968). Our 'standard error' should then be judged according to how well it predicts $L(\hat{\theta},\theta)$, and we will still use a quadratic measure of the prediction error. In particular, the following analogue of Proposition 1 is easily demonstrated.

Proposition 4 *Let $L = L(\hat{\theta},\theta)$ be a loss function such that its second order moment exists, and let U be such that $E\{L(\hat{\theta},\theta)|U\} = l(U)$ is a statistic that might be used to predict $L(\hat{\theta},\theta)$. Then,*

$$E[\{L(\hat{\theta},\theta) - l(U)\}^2] = var\{L(\hat{\theta},\theta)\} - var[l(U)]. \tag{4}$$

Hence, according to the predictive principle for loss function L it is more relevant to measure the precision conditional on the observed value of U (unless $l(U) = constant$).

Proof: Immediate. □

Proposition 1 is retained for quadratic loss L and $l(U) = \mathrm{var}(\hat{\theta}|U)$, when additionally $E(\hat{\theta}|U) = \theta$. Proposition 4 shows that in principle the previously established relevance in conditioning on ancillary information is not dependent on a quadratic loss. In particular, Proposition 4 can be applied in some non-regular situations where the second moment of the quadratic loss does not exist. The Cauchy location model with $n = 2$ provides a relatively simple such example:

Example 9. Cauchy models.
In the Cauchy location model, with density

$$p(y;\theta) = \frac{1}{\pi}\,\frac{1}{1+(y-\theta)^2}\,,$$

the configuration U of an iid sample is ancillary. With a sample size $n = 2$ explicit calculations are possible. The MLE is then

$$\hat{\theta} = \bar{Y}\,.$$

This estimator does not have finite mean and variance, since $\bar{Y}$ has the same Cauchy distribution as the individual Y_i. Conditional moments exist, however. The configuration can be represented by

$$U = (Y_1 - Y_2)/2\,,$$

which is also Cauchy. The pair $(\hat{\theta}, U)$ is minimal sufficient and the distribution of U does not depend on θ. For given U, the conditional density of $\hat{\theta}$ is

$$p(\hat{\theta}|U=u) = \frac{2}{\pi} \frac{1+u^2}{\{1+(\hat{\theta}-\theta+u)^2\}\{1+(\hat{\theta}-\theta-u)^2\}},$$

so in particular $\hat{\theta}$ is conditionally unbiased for each fixed u, Also the conditional variance $v(U)$ exists, albeit not its expected value over the distribution of U. Since necessary moments thus do not exist we cannot apply Proposition 1 to motivate that inference should be conditional. However, Proposition 4 is applicable as soon as the loss function has a second moment. Thus we conclude even in this non-regular example that we should make the inference conditional on the value of U. In the Cauchy location–scale model (with $n \geq 3$), the problems are more intricate, however, see the continued discussion in the next section. □

2.5 Choice between ancillary statistics

Consider an estimator $\hat{\theta}$ in a model, e.g. the ML estimator. The MSE of PSE measure (1) can be used to find among a number of potential error predictors the one which is most relevant for the inference about $\hat{\theta}$. Particular candidates are based on the conditional variance given a conditioning statistic U, $v(U) = \mathrm{var}(\hat{\theta}|U)$, cf. Proposition 1.

The conditioning statistic U need not be an ancillary statistic, even though it frequently will be. The problem of choice of an ancillary U has been particularly discussed in the literature, since the conditionality principle is formulated for ancillary statistics. If U is a function of another ancillary U_1 that is a precision index, it follows from Corollary 1 that U_1 is preferable to U, at least when we neglect the possible problems caused by parameter estimation in the conditional variances (cf. Proposition 3). The serious problems appear when there is no maximal ancillary statistic, but two or more statistics are individually but not jointly ancillary. That such anomalous situations exist was first pointed out by Basu (1964). Example 12 below shows that they can also occur is statistical practice. As a resolution of the problem of choice between ancillary statistics for conditioning, Cox (1971) proposed that the variance of the conditional Fisher information should be used as criterion, a higher such variance indicating a higher potential to distinguish between outcomes of more or less information about the parameter, and in this way best representing the relevance. Related discussion can be found in Fraser (1973), Becker & Gordon (1983), Lloyd (1992), McCullagh (1992).

From a principle point of view the predictive approach makes the problem much simpler and makes the anomaly less disturbing, since the role of ancillarity is diminished. According to Proposition 1 (assuming its requirement of conditional unbiasedness satisfied) the

predictive principle says that among precision indices we should select that particular conditional variance $v(U) = \text{var}(\hat{\theta}|U)$ which has the largest variance, $\text{var}\{v(U)\}$. According to the Cox (1971) criterion we should select the U that maximizes $\text{var}\{i(\theta|U)\}$, where $i(\theta|U)$ is the conditional expected Fisher information. In regular cases the conditional information will be exactly or approximately equal to the inverse of the conditional variance of the ML estimator, so the Cox criterion may then be written approximately as 'maximize $\text{var}\{1/\text{var}(\hat{\theta}|U)\}$'. This is not identically the same criterion as in Proposition 1, since the conditional variance has now been inverted. However, both criteria will typically yield the *same ordering*. Specifically, a propagation of errors calculation gives

$$\text{var}\{1/\text{var}(\hat{\theta}|U)\} \approx \text{var}\{\text{var}(\hat{\theta}|U)\}/E\{\text{var}(\hat{\theta}|U)\}^4 = \text{var}\{\text{var}(\hat{\theta}|)U\}/\text{var}(\hat{\theta})^4 \,.$$

The equality to the right holds provided $\hat{\theta}$ is conditionally unbiased.

In the predictive approach we can also easily incorporate the effects of estimated parameters in the variances, as shown in Proposition 3 above. To the contrary, it is certainly not evident how Cox's criterion should be regarded and modified in the light of parameter uncertainty.

The discussion above assumed that each U was a precision index, so that (conditional) unbiasedness was satisfied. Otherwise neither variance nor Fisher information would be of relevance on its own. Going now to known examples of non-uniqueness of ancillary statistics (i.e. lack of maximal ancillary), it turns out that in none of them is this requirement satisfied! We will discuss some examples here, but must leave open the question whether there exist situations without a maximal ancillary in which the competing ancillaries are proper precision indices.

Example 10. Multinomial 2×2 *table with separately but not jointly ancillary marginals.* Consider n observations multinomially distributed in a 2×2 table with cell and marginal probabilities as shown. The parameter θ is of course restricted to the interval [-1,1].

$(2+\theta)/6$	$(2-\theta)/6$	2/3
$(1-\theta)/6$	$(1+\theta)/6$	1/3
1/2	1/2	

From the marginal probabilities it is seen that both row sums and column sums are ancillary — they are not jointly so, however. This is the example used by Cox (1971) to illustrate the use of his criterion, and from which he concluded that we should condition on rows rather than on columns. It has been much discussed in the literature ever since it appeared in Basu (1964) in a slightly different version. The question whether the ancillary statistics are precision indices has typically been neglected, however. Strictly speaking they are not. Therefore this discussion is not valid for small samples, when bias is not negligible, with the paper by Barnard & Sprott (1971) on transformation invariance and likelihood shape aspects as an exception.

A numerical study of relevance for sample sizes up to $n = 50$ has been described in some detail in Sundberg (1996). In this study conditional and unconditional variances and inverses of expected and observed informations were compared as virtual predictors (that is with true θ instead of $\hat{\theta}$). Except for very small samples, conditioning on row sums was found uniformly (in θ) more relevant than conditioning on column sums, in agreement with Cox's criterion, and also more relevant than not conditioning at all. On the other hand, for parts of the parameter space, widening with increasing n, it was even more relevant to condition on row and column sums jointly, even though they are not jointly ancillary, or to use the inverse of the observed information. For some set of θ-values relatively close to 1, the inverse observed information was much worse, however.

A general conclusion drawn from this observation is that the asymptotic behaviour (as $n \to \infty$) of the inverse observed information demonstrated by Lindsay & Li (1997) need not be uniform in θ, so the asymptotic result need not carry over to finite n. □

Example 11. Bivariate normal with only correlation unknown.
Suppose we have a sample $\{x_i, y_i\}_1^n$ from $N(0, 0, 1, 1, \theta)$, that is a bivariate normal with correlation coefficient θ, whose marginals are standard normal, $N(0, 1)$. This is Example 1 of Basu (1964). The minimal sufficient statistic can be represented by $t = (s_{xy}, s_{xx} + s_{yy})$, where $s_{xy} = \sum x_i y_i / n$ and s_{xx} and s_{yy} are defined analogously. The MLE $\hat{\theta} = \hat{\theta}(t)$ is the solution of a third degree equation, so it cannot be explicitly studied. There is no ancillary statistic that is a function of t. If we allow ancillary statistics which are not functions of t, we could choose any of the two equivalent statistics s_{xx} and s_{yy}, but then the new problem appears that, for symmetry reasons, there is no criterion that helps choosing between them. This was the dilemma pointed out by Basu (1964). For situations when there is no exact ancillary function of t, much intricate work has been devoted to the construction of approximately (asymptotically, locally) ancillary statistics by second order asymptotics, (Efron & Hinkley, 1978, Barndorff-Nielsen, 1980, Cox, 1980, Hinkley, 1980, McCullagh, 1984, 1987, Severini 1993, Barndorff-Nielsen & Cox, 1994). If we look for approximate ancillary statistics in this case, the statistic $s_{xx} + s_{yy}$ is not ancillary since s_{xx} and s_{yy} are correlated to an extent that depends on θ, but it is second order locally ancillary at the parameter point $\theta = 0$ (since its expected value is constant, $= 2$, and its variance is locally quadratic at $\theta = 0$, being $4(1 + \theta^2)/n$).

For precision estimation local ancillaries are more problematic than for hypothesis testing, when the hypothesis specifies a particular parameter value of interest. Also, the adequacy for moderately sized samples is unclear. In Sundberg (1996) was demonstrated, by referring to a simulation study with moderate n and to asymptotics, including Cox's criterion without the ancillarity requirement, that conditional precision given the approxi-

mate ancillary $s_{xx}+s_{yy}$ was more relevant than given either of the exact ancillary statistics s_{xx} or s_{yy}. □

Example 12. A 2 × 2 table from genetic linkage analysis, with separately but not jointly ancillary margins.

A classical crossing trial in genetic linkage analysis combines features of Examples 10 and 11, having exchangeable ancillary statistics like in Example 11. Individuals having alleles A and B at two loci on a chromosome and a and b on the other one in the chromosome pair are cross-fertilized. If A and B are dominant over a and b respectively, the corresponding four different offspring phenotypes have multinomial probilities according to the following table, with θ being the single parameter of interest, $\theta = (1-p)^2$ in terms of the recombination probability p. Independence between loci corresponds to $p = 1/2$, that is $\theta = 1/4$, and only values $1/4 \leq \theta \leq 1$ are genetically reasonable.

$(2+\theta)/4$	$(1-\theta)/4$	$3/4$
$(1-\theta)/4$	$\theta/4$	$1/4$
$3/4$	$1/4$	

This model has been used in genetics since the 1920's and is for example discussed in two of Fisher's books (Fisher, 1990, Statistical Methods 57.1-2 & Experimental Design 71). Later it appears to have become a favourite example in introductions to the EM algorithm, despite the fact that the MLE can be explicitly written down. However, it appears not to have been discussed from the point of view of ancillarity and conditional inference, in contrast to the related artificial Examples 10 and 11 above.

The model is a curved exponential family, index (2,1). Minimal sufficient statistic is any pair of components of the three statistics n_{11}, $n_{12}+n_{21}$, and n_{22}, with obvious notations n_{ij}, and with $\sum n_{ij} = n$ fixed. There is no ancillary component of it. However, the row sums, $r = \bar{n}_{1\cdot}$ (and $n-r$), or the column sums, $s = \bar{n}_{\cdot 1}$ (and $n-s$), are ancillary in the (weaker) sense of having distributions free from the parameter, as obvious from the margins of the probability table. Hence, we may pose the question whether it pays to condition on one of them. But then, on rows or columns? For symmetry reasons, Cox's criterion does not help us select. If we select for example r, is r a precision index?

Complete enumeration studies for some n-values up to $n = 50$ were carried out to compare the MSE of PSE for different information statistics, as functions of the true θ. Consider first the conditional $V = i(\hat{\theta}|r)^{-1}$ and the unconditional $V = i(\hat{\theta})^{-1}$, with the MLE $\hat{\theta}$ inserted. At least for $n \geq 25$ the conditional bias of $\hat{\theta}$ was negligible, so r could well be regarded as a precision index. However, the MSE of PSE showed only a very tiny gain from conditioning on r, irrespective of θ. From a practical point of view, it could be concluded that the single margin r (or s) was essentially irrelevant for precision. This in

a sense resolved the ambiguity of choice between r and s.

The asymptotic result by Lindsay & Li (1997) speaks for the inverse of the observed information, $V = j(\hat{\theta})^{-1}$ as an alternative, and in between we could also imagine a conditioning on $\{r, s\}$ jointly, $V = i(\hat{\theta}|r, s)^{-1}$, even though this statistic admittedly is not ancillary and is numerically inconvenient. It turned out that these two statistics had a clearly reduced MSE of PSE over a large central part of the parameter space, but in other parts they showed the opposite behaviour. Hence none of them could be advocated as uniformly more relevant than the ordinary Fisher information for these finite samples.

Going over from θ to p as parameter of interest, the MSE of PSE curves changed, of course, but qualitatively the results and conclusions remained the same. □

The previous examples concerned situations with two ancillary statistics to choose between. Lloyd (1992) in his Example 1 has a variation with one exact and one approximate ancillary, where only the latter is informative. If an MSE of PSE type criterion is followed, the more informative conditioning variable will be selected.

Example 9. Cauchy models, cont'd.
Another problem appears in the Cauchy location-scale family (McCullagh, 1992). If Y is Cauchy, so is also $1/(Y + c)$ for any c, but the ancillary configuration statistics for all these Cauchy families are different, and not jointly ancillary. McCullagh notes that for symmetry reasons Cox's criterion does not help us choose between them. However, as soon as a specific θ is declared to be the parameter of interest, the predictive principle should (at least in principle) be able to resolve the problem, because different ancillary statistics will yield different conditional distributions of $\hat{\theta}$, and therefore different MSE of PSE values. In practice, the numerical problems of carrying out such comparisons might be substantial, however. □

3 Interval confidence and test significance

To a point estimate we attach a measure of its error in the long run. To an interval estimate we attach a confidence number between 0 and 1 telling how well the interval construction succeeds in covering the true parameter value in the long run (the probability that the interval covers the true parameter value). Conventionally this interval is chosen such that the confidence level has a constant prespecified value. For discussion of ancillary statistics and relevance of conditional inference it is natural to allow non-constant confidence levels, and the confidence level is then naturally regarded from the predictive point of view, in analogy with the predictive principle for the error of a point estimator. We predict whether

the interval covers the parameter or not, and a constant confidence level will then be the best constant predictor in the sense of minimizing among constants a mean squared error criterion, see below. However, in some situations better predictors exist.

Suppose data represented by $(\hat{\theta}, \hat{\psi}, U)$ are observed, where $\hat{\theta}$ and $\hat{\psi}$ are estimates of the parameter of interest, θ, and a nuisance parameter ψ, respectively, and U is an ancillary or other potential conditioning variate. In contrast to Section 2, we need no longer assume that θ is scalar. Let $R(\hat{\theta}, \hat{\psi}, U)$ be a confidence region for θ, and let

$$\xi = \begin{cases} 0 & \text{if } \theta \in R(\hat{\theta}, \hat{\psi}, U) \\ 1 & \text{otherwise} \end{cases}$$

If $R(\hat{\theta}, \hat{\psi}, U)$ is a $1-\alpha$ confidence procedure in the total model, then

$$P(\xi = 0) = 1 - \alpha \text{ for any } \theta.$$

As an alternative consider now confidence statements conditional on U for the same region $R(\hat{\theta}, \hat{\psi}, U)$. Suppose we can attach a confidence level $1 - \alpha(U)$ to $R(\hat{\theta}, \hat{\psi}, U)$ for given U, that is suppose there is a function $\alpha(U)$ independent of θ such that

$$P(\xi = 0|U) = 1 - \alpha(U) \text{ for any } \theta. \tag{5}$$

The crucial requirement (5) is essentially the same as Property P5 in Buehler's (1982) list of properties which can (but need not) be possessed by an ancillary statistic U.

That U be ancillary is neither necessary or sufficient for (5) to hold, of course. Also, (5) trivially implies $E\{\alpha(U)\} = \alpha$, be U ancillary or not. This has been stressed for example in the reflective paper by Dawid (1991).

To answer whether a conditional confidence statement is more relevant than an unconditional one we use the predictive principle as basis and consider α or $\alpha(U)$ as predictor of ξ. To judge a predictor $\alpha(U)$ we will use the MSE of predicted confidence, MSE of PC, and calculate

$$E[\{\xi - \alpha(U)\}^2]. \tag{6}$$

This MSE type measure is chosen for convenience and is analogous with the MSE of PSE studied in Section 2. From different perspectives it has been used in several papers from Robinson (1979) onwards, see Goutis & Casella (1995, Sec. 4). Among constant predictors of ξ, α yields the minimium value $\alpha(1-\alpha)$ of the measure (6). The next proposition shows that a variable $\alpha(U)$ satisfying (5) is a more efficient predictor, i.e. is more relevant. The result is essentially a special case of Proposition 4.

Proposition 5 *Under the set-up formulated above, with condition (5) satisfied, we have*

$$E[\{\xi - \alpha(U)\}^2] = var(\xi) - var\{\alpha(U)\} = \alpha(1-\alpha) - var\{\alpha(U)\},$$

so $\alpha(U)$ is more relevant than a constant α, and the finer the partitioning induced by U, the more relevant is $\alpha(U)$.

Proof: Elementary, and quite analogous with Proposition 1. □

A classical simple and drastic example for the application of Proposition 5 concerns the conditional confidence level of an unconditional construction for the location parameter of a fixed length uniform distribution. The sample range is a conditioning ancillary statistic. For recent literature, see Barndorff-Nielsen & Cox (1994, Ex. 2.17) and Goutis & Casella (1995, Ex. 4) . We will not go into details here.

If we have a confidence interval construction of total confidence level $1-\alpha$, for which Proposition 5 states that a variable confidence level $1-\alpha(U)$ is more relevant, we may of course respond by changing the construction to be conditionally of constant level $1-\alpha$. This is what we should do if we want to specify the confidence level to be α already before we have observed U. Examples 13 and 14 below illustrate this principle.

Example 13. A normal sample confidence interval.
Suppose we have a sample from $N(\theta, 1)$, and suppose we do not utilize the known variance in the confidence interval construction for θ, but form a conventional t-interval around the sample mean, based on the sample variance s^2. The conditional confidence level given s will exist and depend on s, being $\alpha(s) = \Phi(t_{1-\alpha/2}\, s) - \Phi(t_{\alpha/2}\, s)$, where t_α denotes the α quantile of the t distribution in question. Proposition 5 states that if we use this confidence interval construction in the given model, and believe in the model, we should better tell the conditional level $1-\alpha(s)$ than the total level $1-\alpha$. If we next realize we want a conditionally constant level, we modify the width of each conditional interval to obtain this desired level. We should of course not be surprised to find that the new interval is now simply the standard interval for known variance. □

As far as significance tests concern parameters and can be represented as inverse confidence procedures the results on conditioning will be applicable to testing as well. For other types of hypothesis testing, for example goodness-of-fit tests, the present ideas seem not applicable, because test procedures per se do not have quantities to be predicted. Our treatment of significance tests goes through their relationship with confidence regions and confidence levels. Suppose we have constructed a size α test of the hypothesis $\theta = \theta_0$ by rejecting the hypothesis if θ_0 is outside the confidence region $R(\hat{\theta}, \hat{\psi}, U)$. Let ξ be defined as above, and $\alpha(U)$ the conditional size of the test, given U. Then, in the same way as for confidence levels Proposition 5 follows for tests of size α:

Corollary 3 *Confidence regions and rejection regions for which we can split the sample space according to the values of a statistic U and obtain conditional confidence levels*

(5) and corresponding conditional tests, respectively, should be performed as conditional procedures. Thus, if we want to state a fixed confidence level $1-\alpha$ *or test size* α*, the region should be constructed to be conditionally* $1-\alpha$ *and* α*, respectively.*

Example 14. Two measuring instruments.
This much quoted example appears in slightly different versions in Cox (1958) and Cox & Hinkley (1974). Two measuring instruments are known to have high and low precision, respectively. One instrument is selected by coin tossing and a measurement is obtained for testing whether the mean value is some θ. The most powerful size α test approximately rejects as a size 2α test when we have chosen the precise instrument, and as a size ≈ 0 test in the other case. Analogously, the confidence procedure of total confidence level $1-\alpha$ that yields intervals of constant length or smallest expected length will have a very high conditional confidence when the precise instrument is used, but confidence close to $1-2\alpha$ when the imprecise instrument is used. According to Corollary 3 we should instead construct the most powerful conditional size α test, or the equivalent best conditional level $1-\alpha$ confidence interval procedure. This is in full agreement with the result of the qualitative reasoning about relevance found in the references cited. □

Example 15. Test for association in a 2×2 *table.*
Suppose a random sample of individuals from a large population are sorted in a 2×2 table according to two binary criteria, and that we want to test for independence between the two criteria, or more generally for a specified degree of association. Typically the marginal probabilities are unknown nuisance parameters, but they can be eliminated by conditioning on the margins U (which under independence are jointly sufficient for the marginal probabilities). This leads to Fisher's hypergeometric exact test, but even its asymptotic χ^2 version may be regarded from the conditional point of view, as stressed by Yates (1984). The margins are not ancillary in the simple sense, but only in a generalized meaning, see Barndorff-Nielsen & Cox (1994, Ex. 2.22), Zhu & Reid (1994), or Yates (1984): "margins provide virtually no information on the existence of association". This has influenced arguments against conditional testing, see for example the review given by Yates (1984) and the two papers by Upton (1982, 1992), who converted in the mean time from the unconditional to the conditional standpoint.

The predictive approach yields some support to using a conditional test. The hypothesis of independence may be parametrized by the log odds ratio as parameter θ of interest. Suppose we consider a test for $\theta=0$ that has a conditional interpretation given U, i.e. that satisfies condition (5) for some size $\alpha(U)$. We may extend this test to a conditional test of the same size $\alpha(U)$ for any $\theta=\theta_0$ (randomization can cope with the discreteness), and

regard the equivalent confidence region interpretation. Proposition 6 states that for sake of relevance the conditional confidence or significance level should be stated. This is the support given by the predictive approach. On the other hand we cannot use the predictive approach to argue against another test that does not allow a conditional interpretation. This corresponds to the fact that the predictive approach does not tell us what estimator $\hat{\theta}$ to use for θ, only about how its precision should be quantified. □

A frequent complication for the application of Proposition 5 is that the degree of confidence/significance to be attached is often known only approximately or asymptotically, based on a normal or χ^2 approximation of a pivotal or test statistic. This will add squared bias contributions and additional variance and covariance terms to the expression for the MSE of PC in Proposition 5, like in Proposition 3 above, cf. next example (Ex. 12 cont'd). For example, it might be better to use a known and accurate fixed α than an inaccurately known $\alpha(U)$. However, often when U is an ancillary statistic only affecting the precision, like the random size examples above, $\alpha(U)$ will be known more accurately than α.

Example 12. A 2×2 *table from genetic linkage analysis, cont'd.*
This example was discussed in Section 2.5 from the point of view of variance estimators and their MSE of PSE functions. Here we give some results on confidence prediction from the same enumeration studies, $n = 25$ and $n = 50$. For more details, see Sundberg (2000).

An interval estimator was constructed by simply approximating the distribution of $\hat{\theta}$ by a $N(\theta, i(\theta)^{-1})$. Comparisons were carried out for confidence predictors based on normal distributions with inverse variance $i(\theta)$, $i(\theta|r)$, $i(\theta|r,s)$, or $j(\theta)$. In a central part of the parameter space the choice of variance had not much influence on either coverage probabilities or MSE of PC. Even outside the central region, $i(\theta|r)$ agreed almost perfectly with $i(\theta)$. For small θ there was some advantage in using $i(\theta|r,s)$ or $j(\theta)$, but for large θ on the other hand, they behaved much worse in coverage and MSE of PC. These effects could be well explained from the magnitudes and signs of the variances and covariances of the predicted confidence $\alpha(U)$ and the indicator variable ξ. The bias had generally little influence. Hence the conclusion is analogous with that for the MSE of PSE:

- It does not pay to base confidence statements on $i(\theta|r)$ (or $i(\theta|s)$)
- In some parameter interval, $i(\theta|r,s)$ or $j(\theta)$ can yield an increased relevance, but in other intervals the opposite holds. Hence, their usage cannot be generally recommended over $i(\theta)$, either.

This means again that the asymptotic advantage of the observed information (Lindsay & Li, 1997) does not carry over immediately to the finite sample situations. The example

also asks for a statistic that is better approximated by the normal than $\hat{\theta}$ itself.

4 Concluding discussion

In this paper we have argued that precision (variance) of a point estimator and confidence of an interval estimator represent quantities to be predicted. This predictive approach could be regarded as a new paradigm for frequency-based statistical inference. However, it must be stressed that the principle does not tell what estimator or interval (region) to use, only how we should look at its precision/confidence. Also, the principle does not imply that we should change standard methodology. Its most important aspect might be that it admits a natural MSE type quantification of the concept of relevance, which has implications for when precision/confidence should be stated conditionally. Ancillary statistics often qualify for this, but not necessarily (they need also be precision indices). The classical ancillarity paradoxes become less paradoxical in the light of the pragmatic relevance criterion. On the other hand non-ancillary statistics may qualify, depending on the situation and on the particular estimator or interval construction in question. In connection with ML estimators the observed Fisher information is an obvious candidate, because it may be taken as approximately ancillary, see Efron & Hinkley (1978) and Lindsay & Li (1997) for asymptotically valid results supporting the observed Fisher information over the expected. However, the exact calculations in Examples 10 and 12 indicated that the asymptotics were not uniform, which means that the asymptotic results do not necessarily carry over to finite samples.

The theory given might seem largely unrelated with the concepts of likelihood and invariance, and in a strict sense it is. Standard errors and interval estimates as concepts are not directly connected with likelihoods and they are often not constructed to be invariant under nonlinear reparametrizations (e.g. standard-error-based confidence intervals). However, there is no conflict between the predictive approach and desires to use procedures derived from likelihoods (MLE, Fisher information, likelihood-based intervals) or to use invariant procedures. Specifically, note the role of likelihood for suggesting statistics to condition on. Cox's criterion would not have been less likelihood-based if it had been formulated in terms of the inverse conditional information instead of the conditional information itself. For precision indices it would then have been even more like the MSE of PSE criterion, but much less tractable for theoretical calculations than it is now. More important, it would remain to connect Cox's criterion with a demand for the ancillary to be a precision index, and then we would be even closer to the predictive approach. The deviation from likelihood and invariance comes when we specify the MSE of PSE or some similar quantitative measure for assessing relevance. The main motivation for the

quadratic form of the measures used above is that it makes them relatively easy to deal with theoretically. For enumeration or simulation studies other measures could be used, according to taste.

It must be admitted that many of the examples given in the paper are artificial, aimed at being either simple or classics. Several of the classics have turned out to require extensive studies for a more complete understanding. The same is true for many realistic situations of practical interest, like the genetics example, Example 12. Another type of situations where an extensive numerical study sometimes could contribute further to the question of conditional inference, are those with approximate ancillary statistics. An example is the three-parametric generalized extreme value distribution, studied by Dixon *et al.* (1998) and applied to sea-levels and insurance claims data. But in each such case one must ask whether the calculations could be worth-while.

Some examples which first looked fruitful for a predictive study of conditional inference have turned out to be dominated by the bias aspect. As an example, in an exponential life test study with censoring, should precision be measured conditionally, given the total time on test? It turned out that bias in coverage probabilities was the factor of main importance (Sundberg, 2001).

For some sample survey estimators of population characteristics, proposed variance estimators exist in abundance in the literature. The MSE of PSE criterion makes it possible to rank them, see Sundberg (1994) for a study of variance estimators for use with the ratio estimator. The result of that paper was that some of the variance estimators could be ruled out as unsatisfactory, and previous rankings based only on the variances of the variance estimators could be revised. Other situations which could be analysed in the same way include the regression estimator, quite analogous to the ratio estimator, and estimators used in subsampling designs.

References

Barnard, G.A. and Sprott, D.A. (1971) A note on Basu's examples of anomalous ancillary statistics (with discussion). In *Foundations of Statistical Inference* (eds V.P. Godambe and D.A. Sprott), 163–176, Toronto: Holt, Rinehart and Winston of Canada.

Barndorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory.* Chichester: Wiley.

Barndorff-Nielsen, O. (1980) Conditionality resolutions. *Biometrika*, **67**, 293–310.

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

Basu, D. (1964) Recovery of ancillary information. *Sankhyā A*, **26**, 3–16.

Becker, N. and Gordon, I. (1983) On Cox's criterion for discriminating between alternative ancillary statistics. *Int. Statist. Rev.*, **51**, 89–92.

Buehler, R.J. (1982) Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Assoc.*, **77**, 581–594.

Cox, D.R. (1958) Some problems connected with statistical inference. *Ann. Math. Statist.*, **29**, 357–372.

Cox, D.R. (1971) The choice between alternative ancillary statistics. *J. Roy. Statist. Soc. Ser. B*, **33**, 251–255.

Cox, D.R. (1980) Local ancillarity. *Biometrika*, **67**, 279–286.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics.* London: Chapman & Hall.

Dawid, A.P. (1991) Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. Ser. B*, **53**, 79–109.

Dixon, M.J., Ledford, A.W. & Marriott, P.K. (1998) Finite sample inference for extreme value distributions. Manuscript.

Efron, B. and Hinkley, D.V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with Discussion). *Biometrika*, **65**, 457–487.

Fisher, R.A. (1935) The logic of inductive inference. *J. Roy. Statist. Soc.*, **48**, 39–54.

Fisher, R.A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference.* Re-issue, Oxford University Press, Oxford.

Fraser, D.A.S. (1985) The elusive ancillary. In *Multivariate Statistical Inference* (eds

D.G. Kabe and R.P. Gupta), 41–48. Amsterdam: North Holland.

Goutis, C. and Casella, G. (1995) Frequentist post-data inference. *Int. Statist. Rev.*, **63**, 325–344.

Helland, I.S. (1995) Simple counterexamples against the conditionality principle. *Amer. Statist.*, **49**, 351–356.

Hinkley, D.V. (1980) Likelihood as approximate pivotal distribution. *Biometrika*, **67**, 287–292.

Holt, D. &, Smith, T.M.F. (1979) Post stratification. *J. Roy. Statist. Soc. Ser. A*, **142**, 33–46.

Kalbfleisch, J.D. (1975) Sufficiency and conditionality. *Biometrika*, **62**, 251–259.

Lindsay, B.G. & Li, B. (1997) On second-order optimality of the observed Fisher information. *Ann. Statist.*, **25**, 2172–2199.

Lloyd, C. (1992) Effective conditioning. *Austral. J. Statist.*, **34**, 241–260.

McCullagh, P. (1984) Local sufficiency. *Biometrika*, **71**, 233–244.

McCullagh, P. (1987) *Tensor Methods in Statistics.*. London: Chapman & Hall.

McCullagh, P. (1992) Conditional inference and Cauchy models. *Biometrika*, **79**, 247–259.

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science*, **10**, 138–157.

Robinson, G.K. (1979) Conditional properties of statistical procedures. *Ann. Statist.*, **7**, 742–755.

Sandved, E. (1968). Ancillary statistics and estimation of the loss in estimation problems. *Ann. Math. Statist.*, **39**, 1756–1758.

Severini, T.A. (1993) Local ancillarity in the presence of a nuisance parameter. *Biometrika*, **80**, 305–320.

Sundberg, R. (1994). Precision estimation in sample survey inference: A criterion for choice between variance estimators. *Biometrika*, **81**, 157–172.

Sundberg, R. (1996). Conditional statistical inference and quantification of relevance. Res. report, Mathem. statistics, Stockholm Univ.

Sundberg, R. (2000). Ancillarity and conditional inference for ML and interval estimates in classical genetic linkage trials. Res. report, Mathem. statistics, Stockholm Univ.

Sundberg, R. (2001). Comparison of confidence procedures for type I censored exponential lifetimes. To appear in *Lifetime Data Analysis*

Upton, G.J.G. (1982) A comparison of alternative tests for the 2×2 comparative trial. *J. Roy. Statist. Soc. Ser. A*, **145**, 86–105.

Upton, G.J.G. (1992) Fisher's exact test. *J. Roy. Statist. Soc. Ser. A*, **155**, 395–402.

Yates, F. (1984) Tests of significance for 2×2 contingency tables (with Discussion). *J. Roy. Statist. Soc. Ser. A*, **147**, 426–463.

Zhu, Y. & Reid, N. (1994) Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Canad. J. Statist.*, **22**, 111–123.