Mathematical Statistics
Stockholm University

Ancillarity and conditional inference for ML and interval estimates in some classical genetic linkage trials

Rolf Sundberg

Research Report 2001:3

ISSN 1650-0377

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.matematik.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2001:3**,
http://www.matematik.su.se/matstat

Ancillarity and conditional inference for ML and interval estimates in some classical genetic linkage trials

Rolf Sundberg*

February 2001

Abstract

The main object of study here is a classical example of linkage analysis, in which there are two separately but not jointly ancillary statistics, which are mutually exchangeable. In such cases it is not obvious how or even if the statistical inference about the parameter of interest (here the recombination probability) should be a conditional inference. We consider various precision measures, viz. the observed and the expected (Fisher) information quantities, and various conditional expected values in between, and we compare their ability to quantify the precision of the parameter estimate, as well as to quantify the confidence to be attached to interval estimates. The general conclusion drawn is that there is not much to be gained but much to be risked by conditional inference in this example.

Keywords: confidence, precision, recombination probability, relevance.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@matematik.su.se. Financial support from the Swedish Natural Science Research Council is gratefully acknowledged.

1 Introduction

Classical crossing trials in genetics, carried out in order to study the linkage of different genes, that is to study the probability that the genes are inherited together from the same of the two crossed gametes, typically give rise to 2×2 or larger tables and a multinomial type statistical model with one or a few unknown parameters. Row or column sums may turn out to be ancillary statistics, in the (weak) sense of simply having distributions not depending on the parameters. This is the case in the examples to be discussed, and in the second, main example there are even two exchangeable statistics which are separetely but not jointly ancillary. The conditionality principle tells us to make the inference conditional on ancillary statistics, in order to make the inference more relevant to the data obtained. Is a conditional inference given such an ancillary statistic more relevant in these cases? Are there other statistics on which it could pay to condition, in terms of increased relevance?

We will study conditional inference concerning the precision of maximum likelihood (ML) point estimates and the degree of confidence of the corresponding large sample interval estimates, based on various conditionally expected information statistics. In this connection the observed information for the actual sample can be regarded as an extremal case.

1.1 Linkage analysis

The aim of the statistical analyses to be discussed below is to get information about the distance between the different loci (positions) for two genes on a chromosome, and we study a population of individuals with two different gene forms (alleles) in each of these places, A or a at one locus, B or b at the other one. The chromosomes appear in pairs, inherited one from each parent of the individual, so each locus is represented by a pair of genes. This pair specifies the genotype. That is, at one locus we have one of the three genotypes AA, Aa and aa, and analogously one of BB, Bb and bb at the other locus. When such specification is necessary, we write for example AB—ab to indicate that A and B are on one chromosome and a and b on the other. At reproduction, a new chromosome (a gamete) is formed from the two original ones and inherited. At each locus, both genes have the same probability 1/2 to be selected. However, these Bernoulli trials ("coin tosses") are not independent between loci, because whole segments of the gamete are taken together in the formation of the new one. This is called *linkage*. The probability that the genes come from different gametes is called the *recombination probability* and will be denoted p, with $q = 1 - p$ for its complement. For simplicity we assume that the recombination probability is the same in both male and female gametogenesis. Knowledge about p carries information about the distance between the two loci. If the loci are close together, p should be close to zero, whereas if they are at a large distance, p should be close to 1/2. Most models for recombination imply that p increases monotonically with increasing distance. Values of $p > 1/2$ are less natural, or even "inadmissible on biological grounds" as stated by Sham (1998).

In a classical crossing trial, the recombination probability p can be estimated from observed frequencies of different offspring phenotypes (distinguishable appearances). In two genetically different such situations we will discuss the precision of the ML estimate and the confidence to be attached to a conventional ML-based interval estimate.

1.2 Prediction of precision and confidence

Here we prepare for later sections, in particular Section 3, by briefly discussing how precision and confidence should be attached to point estimates and interval estimates, respectively. We will argue that precision and confidence are quantities to be predicted, rather than estimated, and we will see that this point of view admits a quantification of the concept of relevance, that can be used to compare various conditional inferences, for example. A more extensive discussion is found in Sundberg (1996).

Let θ be a parameter of interest, which could be the recombination probability p or some function of it, and let $\hat{\theta}$ denote an estimator of θ, e.g. the MLE. We will restrict ourselves to the special case of one-dimensional parameters here, since this is sufficient for the examples to be discussed (i.e. no nuisance parameters involved). Conventionally, assuming $\hat{\theta}$ unbiased or approximately unbiased, the precision of $\hat{\theta}$ is expressed by its variance, $\mathrm{Var}(\hat{\theta})$, with $\hat{\theta}$ inserted for the unknown θ in this expression. Generally, approximate estimator variances for the MLE are obtained by using the inverse of the observed or the expected (Fisher) information derived from the likelihood of the data. If there is an ancillary statistic, i.e. vaguely a statistic whose distribution does not depend on θ, the conditionality principle tells us to use the conditional Fisher information, given the ancillary. However, even if we adhere to this principle, the vagueness of the concept of ancillarity carries over to the principle. There are more or less demanding versions of ancillarity, there are counter-examples where conditioning leads to absurd consequences, and there are cases where an ancillary statistic would be desirable but does not exist, but where there are other statistics that may be conditioned on. In the latter case particularly, but also more generally, when does it pay to make the inference conditional?

More specifically, if we use the observed information or a conditional or unconditional expected information to construct a variance estimator (by inversion), which of the possible variance estimators is the best one? Until possible parameters have been replaced by estimates, observed and conditionally expected information quantities all have the same expected value, identical with the unconditional Fisher information. But since they are based on more or less conditioning they will not be equally relevant to the data at hand. At the same time, inversion of the information statistic to yield a variance is a nonlinear transformation, so results about expected values and other moments for information statistics do not immediately carry over to variance estimators. When these variance estimators next are used as standard errors in normality-based (asymptotic) confidence intervals, even more nonlinearities come in. Which interval yields the most reliable or relevant confidence statement?

In a previous manuscript (Sundberg, 1996) we have argued in detail that variance estimators V should not be regarded as estimators of $\mathrm{Var}_\theta(\hat{\theta}) = E_\theta\{(\hat{\theta}-\theta)^2\}$, but as predictors of the actual quadratic error, $(\hat{\theta}-\theta)^2$. The question which is the best one among some such predictors V may be evaluated by comparing for example their expected squared prediction errors $E_\theta[\{V-(\hat{\theta}-\theta)^2\}^2]$. To allow theoretical calculations such a quadratic measure is the most suitable one. For example, it is not difficult to prove that if $\hat{\theta}$ is conditionally unbiased, given an ancillary statistic U, and if $V = v(U)$ is the true conditional variance of $\hat{\theta}$, then

$$E_\theta[\{V-(\hat{\theta}-\theta)^2\}^2] = \mathrm{Var}[(\hat{\theta}-\theta)^2] - \mathrm{Var}[v(U)]. \tag{1}$$

This result shows the advantage of conditioning on an ancillary statistic in simple cases. More generally, if V is an unbiased estimator of $E_\theta[(\hat{\theta}-\theta)^2]$, of course it holds that

$$E_\theta[\{V-(\hat{\theta}-\theta)^2\}^2] = \mathrm{Var}[(\hat{\theta}-\theta)^2] - 2\,\mathrm{Cov}[(\hat{\theta}-\theta)^2,\, V] + \mathrm{Var}(V). \tag{2}$$

Hence, for judging an unbiased variance estimator V, not only its variance, but also its covariance with the actual quadratic error, is important. Even more generally, a bias component can also play an important role. The measure proposed on the left hand side of (1) and (2) above will be called the *mean squared error of predicted squared error*, abbreviated the MSE of PSE. However, in principle we could think of alternative measures, and for simulation studies any other function than the quadratic could be used as easily.

Analogously, a basis for comparison of interval estimators is formed by

1. Fixing a reference interval construction for θ, for example $|\hat{\theta}-\theta|/se(\hat{\theta}) \leq 1.96$, where $se(\hat{\theta})$ is the marginal standard deviation of $\hat{\theta}$, either as a function of θ or else with $\hat{\theta}$ inserted;

2. Allowing the attached confidence $1-\alpha$ to be a random variable, $1-\hat{\alpha}$, different for different interval estimators (for example, one of them might depend on an ancillary whereas another one does not);

3. Compare them as predictors of the indicator for coverage/non-coverage.

Let ξ be an indicator that is 0 if the interval covers θ, else 1. The criterion to judge a confidence predictor $1-\hat{\alpha}$ will be $E_\theta\{(\hat{\alpha}-\xi)^2\}$, the *mean squared error of predicted confidence*, abbreviated the MSE of PC. Like in (2), if $E_\theta(\xi) = E_\theta(\hat{\alpha})$ we may write the MSE of PC on the form

$$E_\theta\{(\hat{\alpha}-\xi)^2\} = \mathrm{Var}(\xi) - 2\,\mathrm{Cov}[\xi,\, \hat{\alpha}] + \mathrm{Var}(\hat{\alpha}), \tag{3}$$

where the first term on the right hand side can be expressed even more explicitely as a binomial variance, $\alpha(1-\alpha)$ if $E(\xi) = \alpha$. Effects of the correlation between ξ and $\hat{\alpha}$ will be illustrated in Sec. 3.

Thus, as measures of relevance we will use the MSE of the predicted squared error of the point estimator and the MSE of the predicted confidence of the interval estimator.

Note that, like for example the Wald statistic, the MSE of PSE comparisons are not invariant under nonlinear transformations of the parameter. Hence, for use of the MSE of PSE measure it must be decided what is the parameter of primary interest. For confidence there is also a lack of invariance, in the sense that comparison results to some extent will depend on the choice of reference interval. Note that the reference interval need not be taken as the best construction, to be used in practice in the future. For example, the reference interval might be unconditionally constructed while the comparisons are indicating that inference should be conditional on some statistic a. Then it would probably be more natural to response to the results of the comparisons by changing to a construction that is conditional, too.

2 A first example

The main role of this first example is to be genetically an introduction to the next situation, which will be statistically less trivial. Suppose a so called double heterozygote AB—ab is crossed with itself, for example by self-fertilization, to yield a progeny of size n. The following three tables show the possible outcomes, notations for the corresponding frequencies, and the corresponding probabilities for these outcomes. The probability table corresponds to Table 67, §57.1, in Fisher's "Statistical Methods for Research Workers", where both this situation and the next one are discussed, and illustrated with data from pea crossing trials, carried out by the Swedish geneticists Tedin & Tedin (Fisher, 1990).

	BB	Bb	bb
AA	AB—AB	AB—Ab	Ab—Ab
Aa	AB—aB	AB—ab Ab—aB	Ab—ab
aa	aB—aB	aB—ab	ab—ab

	BB	Bb	bb
AA	n_{11}	n_{12}	n_{13}
Aa	n_{21}	n_{221} n_{222}	n_{23}
aa	n_{31}	n_{32}	n_{33}

- $\frac{1}{4}$

q^2	$2pq$	p^2
$2pq$	$2q^2$ $2p^2$	$2pq$
p^2	$2pq$	q^2

Note that we should distinguish the outcomes AB—ab and Ab—aB, which both have one of each allele but represent no recombination and double recombination, respectively.

In this example we assume that all these possible outcomes are identifiable. for example by subsequent self-fertilization trials. The multinomial likelihood for the data is then given by

$$L(p) \propto p^{n_{21}+2n_{31}+n_{12}+2n_{222}+n_{32}+2n_{13}+n_{23}} \times (1-p)^{2n_{11}+n_{21}+n_{12}+2n_{221}+n_{23}+n_{33}} \qquad (4)$$

Note that the two exponents sum to $2n$, that is twice the total progeny. The likelihood is proportional to that of a binomial, Bin($2n$,p), with

$$T = n_{21} + 2n_{31} + n_{12} + 2n_{222} + n_{32} + 2n_{13} + n_{23},$$

as sufficient statistic. It follows that the ML estimator is $\hat{p} = T/2n$. It is easily seen that $\hat{p} = T/2n$ is unbiased.

Since the sufficient statistic is one-dimensional, there is no room for any ancillary statistic in the strict sense, that is for a statistic with distribution free of p which is also a component of the minimal sufficient statistic. However, there are plenty of statistics which are ancillary in the weak sense of only satisfying the first requirement, to have distributions which do not depend on the parameter. As an example, consider the set of column sums (or equivalently row sums), $U = \{n_1, n_2, n_3\}$, which is trinomial($2n$; 1/4, 1/2, 1/4). Could

it be that U carries any (conditional) information about the precision of $\hat{p}$? Intuitively we perhaps do not expect this. Let us confirm that such an expectation is correct.

Conditionally on the column sums U, we are seen to have three multinomials with essentially identical probability vectors. If we just merge n_{12} and n_{32} in the middle column, they will be exactly the same, namely trinomial with probabilities p^2, $2pq$, and q^2. That we can merge n_{12} and n_{32} is because they represent the same probability, $2pq$. Now, if we have three samples of fixed sample sizes from one and the same multinomial (trinomial) distribution, we certainly lose no information about the multinomial probabilities by merging the three samples, and only the total sample size is relevant for precision, in this case. We could in fact go one step further to find that the experiment can be represented by a Bin($2n$,p) distribution, but this is not necessary for us to be able to draw the following conclusion: Conditioning on the weak ancillary U in this example does not contribute to the relevance of the statistical inference about p. (This does not say, of course, that this conditioning could not be useful in connection with model checking.)

Alternatively, we could have calculated the conditional variance of $\hat{p}$, given U, and would then have found that $\mathrm{Var}(\hat{p}|U)$ depends on U only through the fixed sum $2n = n_1 + n_2 + n_3$. Hence the conditional variance is independent of the conditioning ancillary statistic, so no precision could be gained/lost by conditioning.

Remark: It is not necessary that the recombination frequency is the same in male and female gametogenesis. With two different parameters, p_1 and p_2, instead of a common p, we just replace p^2 by p_1p_2, $2pq$ by $p_1q_2+p_2q_1$, and q^2 by q_1q_2. The argumentation above remains essentially the same, but with a two-dimensional minimal sufficient statistic and a two-dimensional parameter, identifiable except for a symmetry ambiguity between sexes, $q_1q_2 = q_2q_1$. ■

Remark: If instead of crossing AB—ab with itself, Ab—aB is crossed with itself, the only difference is a matter of notations, of course. Just change B to b and vice versa in all tables, or p to q and vice versa in the probability table. However, when A and B are dominant over a and b, the question is not only notational, see the next section. ■

Remark: Another modification of the previous example is when A and a are not distinguishable unless there is at least one B present at the other locus. Also such a situation is discussed by Fisher (1990, Statist. Methods §57.1), describing the pea crossing trials mentioned above. In this case the bb column above is condensed into a single bb outcome. Since the probability for this outcome is independent of p, it yields no information about linkage, so the column can simply be neglected. However, the argumentation above works analogously with just two columns instead of all three. ■

3 The main example

In this example we start out from the situation of the previous section, but additionally assume that alleles A and B are dominant over a and b, respectively. That allele A is dominant means that as soon as A is present in any of the two chromosomes we will observe one and the same phenotype, that is an observable characteristic of the individual. The alternative phenotype in this trial requires that allele a is present in both chromosomes. The data in this section are taken to be the frequencies of the phenotypes observable. As a consequence, we cannot distinguish individuals in columns 1 and 2 of Table 1, nor

observations in rows 1 and 2. The resulting aggregated data form a 2×2 table with cell probabilities obtained by summing over the corresponding cell probabilities in the original genotype table. The resulting table of probabilities is the following, where the marginal row and column probabilities are also shown, and where θ denotes the q^2 of the previous table:

	BB or Bb	bb	
AA or Aa	$\frac{1}{4}(2+\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{3}{4}$
aa	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$	

Here θ can take any value, $0 \leq \theta \leq 1$, but only $\theta \geq 1/4$ is genetically reasonable.

Remark: If male and female recombination frequencies are not assumed equal, θ should be interpreted as $q_1 q_2$, cf. Sec. 2. If Ab—aB is self-fertilized instead of AB—ab, $\theta = p^2$ (or $p_1 p_2$). Then, the genetically natural values are $\theta \leq 1/4$. However, starting from Ab—aB would lead to a relatively inefficient design. This can be seen by calculating and comparing the Fisher information quantities in the two cases, see Fisher (1990, Experim. Design §71). ■

The situation and the model appears in both old and recent texts on statistical genetics, as exemplified by Fisher (1990, Statist. Methods §57.1-2 & Experim. Design §71), Mather (1938) and Lange (1997, p. 30). The model has also appeared frequently in the general statistical literature in order to illustrate computational aspects of ML estimation, even though an explicit solution does exist, see (5) below. Rao (1952, Sec. 4c.2, & 1973, Sec. 5g) used the model to illustrate the Newton–Raphson and the Fisher scoring methods. Dempster, Laird & Rubin (1977), Tanner (1991, Sec. 4), and McLachlan & Krishnan (1997, Ex. 1.1) all have had it as their introductory example on the EM algorithm.

Let the observed 2×2 table of frequencies for a progeny of size n, with its row and column sums and total, be represented by

n_{11}	n_{12}	r
n_{21}	n_{22}	$n-r$
s	$n-s$	n

The basic likelihood properties of the model are as follows. The likelihood is

$$L(\theta) \propto (2+\theta)^{n_{11}} \times (1-\theta)^{n_{12}+n_{21}} \times \theta^{n_{22}} .$$

It is evident that the minimal sufficient statistic is of dimension 2, consisting of any pair of the three statistics n_{11}, $n_{12}+n_{21}$, n_{22}, since the sum n of the table is fixed. We have a curved exponential family, index (2,1). Since the minimal sufficient statistic is of dimension 2 we have reason to look for an ancillary component of it. Unfortunately, an

ancillary in this strict sense does not exist (according to definition in Sec. 2). However, the row and column sums of the probability table immediately tell that both the row totals, r and $n-r$, and the column totals, s and $n-s$, have marginal distributions independent of θ, so each of the marginals is ancillary in the weak sense. Hence, we may pose the question whether it pays to condition on one of them. But then, on rows or columns? We return to this paradoxical question shortly.

The score function is

$$D \log L(\theta) = \frac{n_{11}}{2+\theta} - \frac{n_{12}+n_{21}}{1-\theta} + \frac{n_{22}}{\theta}.$$

Setting this to zero yields the MLE in regular cases, when the likelihood has a maximum in the interior of the interval (0, 1). The MLE $\hat{\theta}$ is then the solution of a quadratic equation,

$$\hat{\theta} = b + \sqrt{b^2 + c}, \tag{5}$$

where $b = (n_{11} - 2(n_{12}+n_{21}) - n_{22})/2n$ and $c = 2n_{22}/n$. In the special case $n_{22} = 0, n_{11} < 2n/3$ the likelihood maximum is attained at the boundary point $\theta = 0$, and in the special case of no off-diagonal observations, $n_{12} + n_{21} = 0$, the likelihood is maximized at $\theta = 1$.

Minus the second derivative of $\log L$ will be denoted J and is given by

$$J(\theta) = -D^2 \log L(\theta) = \frac{n_{11}}{(2+\theta)^2} + \frac{n_{12}+n_{21}}{(1-\theta)^2} + \frac{n_{22}}{\theta^2}.$$

The value $J(\hat{\theta})$, calculated in a regular MLE, is the observed information in the data. The expected (Fisher) information $I(\theta)$ is the expected value of J, or equivalently the variance of the score function, that is

$$I(\theta) = -E\, D^2 \log L(\theta) = \frac{n}{4}\left(\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta}\right) = n\frac{0.5+\theta}{\theta(2+\theta)(1-\theta)}.$$

The inverses of these information quantities J and I (with $\hat{\theta}$ inserted for θ) can be used to approximate the variance of the MLE. A plot of $I(\theta)^{-1}$ against θ, found as Figure 1a, gives an impression of how the estimation precision depends on θ. The Fisher information for $q = 1 - p$ is found as Figure 1b. For general arguments supporting the observed information in large sample situations, see Lindsay & Li (1997). Considering that row totals (or column totals) are ancillary statistics, a further alternative is to condition on such an ancillary and use the inverse of

$$I_r(\theta) = E\{J(\theta)|r\} = \frac{n\theta + 2(n-r)}{\theta(2+\theta)(1-\theta)},$$

(or $I_s(\theta)$, analogously defined). But how could we choose between I_r and I_s? By symmetry they are equivalent, until we have seen their values. Could we condition on both of them jointly, and use the information $I_{rs}(\theta) = E\{J(\theta)|r,s\}$? We are then not conditioning on an ancillary statistic, but on the other hand we are still in a case in between the two defendable statistics I and J, only closer to J. As another alternative, what happens if one tries to "cheat nature "by conditioning on the one that yields the highest information value?

It may be argued (Fisher, 1990, Scientific Inference, Ch. VI) that the unconditional Fisher information $I(\theta)$ represents the total information in data, and if we do not recover this information by conditioning, we should replace I by the information in the marginal distribution for $\hat{\theta}$. This distribution does not have an explicit form, however, and the idea will not be pursued here.

An analogous discussion to the one on point estimation can be given for interval estimates and the degree of confidence to be attached to them. Also in this case we may compare unconditional and various conditional procedures.

We will present the results of a study which was based on complete enumeration of all possible outcomes, for the two sample sizes $n = 25$ and $n = 50$.

- Comparisons of MSE of PSE functions of θ for four or five different information quantities, their inverses regarded as variance estimators, or more precisely here, as different predictors of the quadratic estimation error.

- The corresponding MSE of PSE comparison under parametrization with the recombination probability itself, $q = \sqrt{\theta}$ or $p = 1 - q$ instead of θ.

- Comparison of true and predicted coverage probabilities.

- Comparison of the MSE of PC measure for relevance of confidence statements.

The different variance estimators and confidence predictors are constructed from the following information statistics, the variance estimators simply as the inverses of the corresponding information statistics.

1. Standard Fisher information, $I(\hat{\theta})$

2. Conditional information, given row (or column) sums, $I_r(\hat{\theta})$

3. Maximal conditional information, given row or column sums, $max\{I_r(\hat{\theta}), I_s(\hat{\theta})\}$

4. Conditional information, given row and column sums jointly, $I_{rs}(\hat{\theta})$

5. Observed information $J(\hat{\theta})$

As reference interval, for the comparisons of coverage probabilities and of the MSE of PC measure of relevance of confidence statements for the five different choices above, is taken the nominal 95% interval based on inversion of an assumed (approximate) marginal normality of $(\hat{\theta} - \theta)/I(\theta)^{1/2}$. Slightly extending the plausible region $\theta \geq 0.25$ we will restrict attention to the interval $\theta > 0.2$. Five diagrams are shown for each of the two n-values, representing MSE of PSE comparisons for θ and for p or q as the parameters of interest, MSE of PC for confidence, and mean confidence and coverage probabilities for the reference interval estimate, and correlations between predicted confidence and coverage.

We first consider Figures 2 and 3, which show MSE of PSE comparisons with θ or p (q) as the parameter of interest. Qualitatively all four diagrams in Figures 2 and 3 are similar. The forms seen in Figures 2 and 3 should differ, depending on the parameterization. The crude form for the dependence on θ in Figure 2, when $1/I(\hat{\theta})$ is used as variance estimator (solid curve) is obtained by thinking of $\hat{\theta}$ as approximately N(θ, $1/I(\theta)$), which from (1)

would yield an MSE of PSE for $I(\theta)$ of size $2(1/I(\theta))^2$. Compare Figure 1a, which shows the form of $1/I(\theta)$. Analogously for p (or q), Figure 1b shows $1/I(p)$ when the crude form of the MSE of PSE should follow $2(1/I(p))^2$.

The dotted curves in Figures 2 and 3 represent conditioning on the ancillary row sum r (or s, exchangeable with r). Comparing with the solid curves, it is seen that this conditioning yields a very tiny gain, if any. In some diagrams it is difficult to see the dotted curve because it falls so close to the solid one. It turned out that the curves for conditioning on $max\{I_r(\hat{\theta}), I_s(\hat{\theta})\}$ also differed that little from $I(\hat{\theta})$, and these curves are therefore not even shown in any of the diagrams.

Could the ancillary r be regarded as a precision index? For a sample size such as $n = 25$ or more, the conditional bias $E(\hat{\theta}|r) - \theta$ was found to be quite small for all r, within the interval $\theta > 0.2$. Hence, with good approximation r can be accepted as a precision index, albeit one with a negligible influence on precision.

For confidence the MSE of PC curves shown in Figure 4 (and interpreted below) are also extremely close for $I(\theta)$ (solid) and $I_r(\theta)$ (dotted). The practical overall conclusion is that it neither pays nor costs to condition on r (or s). Thus, the problem of choice between the ancillary statistics r and s is superficial. Not even conditioning on the 'best' of r and s had any noticeable effect (not included in the diagrams shown). The same conclusions were drawn for $n = 25$ as for $n = 50$.

Returning to Figures 2 and 3, it is seen that use of $I_{rs}(\hat{\theta})$ or $J(\hat{\theta})$ instead of $I(\hat{\theta})$ or $I_r(\hat{\theta})$ makes a substantial difference. Over part of the parameter interval the variance estimators $1/I_{rs}(\hat{\theta})$ or $1/J(\hat{\theta})$ both have a reduced MSE of PSE, but in other parts it is increased (Figure 2). The relationships with p or q as parameter of interest (Figure 3) are qualitatively similar but considerably distorted, in comparison.

For confidence prediction, Figure 4 shows the MSE of PC measure for the four predictors. The variability in the solid line shows essentially how the variance of the binomial ξ varies with θ. Added to this is the variance of the predictor, and minus twice the covariance, see formula (**??**). Over the central part of the interval the predictor variances are in fact so small that the covariances do not matter, either. Again, as for the MSE of PSE, the dotted curve follows the solid curve, so it does neither pay nor cost in relevance to condition on the ancillary r. For the more strongly conditional predictors, based on I_{rs} and J, and relatively large values of θ, the predictor variances are high, and this is not compensated for by the covariance terms, but magnified, because the correlation between ξ and these two predictors are negative for high θ, as illustrated in Figure 6. In the left part of the region (around $\theta = 0.25$), these correlations are mostly positive, however, and more or less compensate for the predictor variances. The exception is for the observed information $J(\theta)$ when $n = 25$, where the upper diagram of Figure 4 shows the dashed curve is clearly above the others in the leftmost part.

Does not the possible bias in predicted confidence matter for these methods? Figure 5 shows the actual coverage probability for the interval estimate ($E(\xi)$; solid curve) and the mean values of the predicted confidences. The fluctuations in coverage probability reflect the discrete character of data. It might appear as if these mean value curves should explain the relationships in MSE of PC, but in fact the 'squared bias' contributions to the MSE of PC are mostly quite small in comparison with the variance contributions. Therefore much of the rugged features seen in Figure 5 are smoothed out in Figure 4.

Conclusions: Our study of this classical example has shown that relevance is not increased by making the inference conditional on row or column sums, so this in a sense resolves the ancillarity paradox, at least from the practical point of view. Nor does it appear adviceable to base inference on the observed information, because even though it has a somewhat increased relevance in some parameter regions, it is much worse in other regions. Thus, this seems to be an example where the asymptotic advantage of the observed information (Lindsay & Li, 1997) does not carry over immediately to the finite sample situations, essentially because the asymptotics of Lindsay & Li is not uniform in the parameter. The example also asks for a statistic that is better approximated by the normal than are $\hat{\theta}$ or $\hat{p}$, but that is outside the scope of the present paper.

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B* **39**, 1–38.

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

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis.* Springer-Verlag.

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

Mather, K. (1932). *The Measurement of Linkage in Heredity.* Methuen & Co, London.

McLachlan, G.J. & Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley.

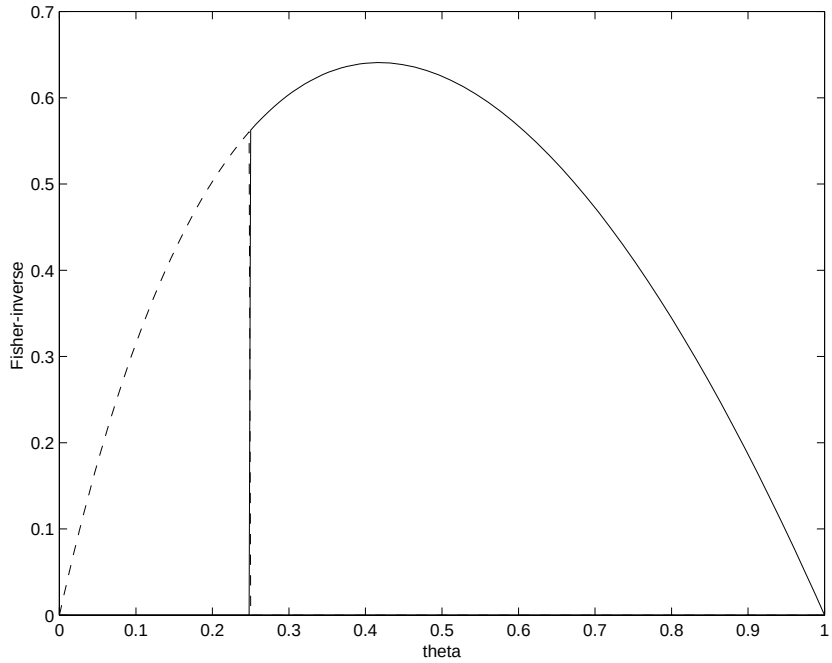
Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research.* John Wiley & Sons, New York.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications, 2nd ed.* John Wiley & Sons, New York.

Sham, P. (1998). *Statistics in Human Genetics.* Arnold, London.

Sundberg, R. (1996). Conditional statistical inference and quantification of relevance. Manuscript, revised version 2000.

Tanner, M.A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods.* Springer-Verlag (Lecture notes in statistics)



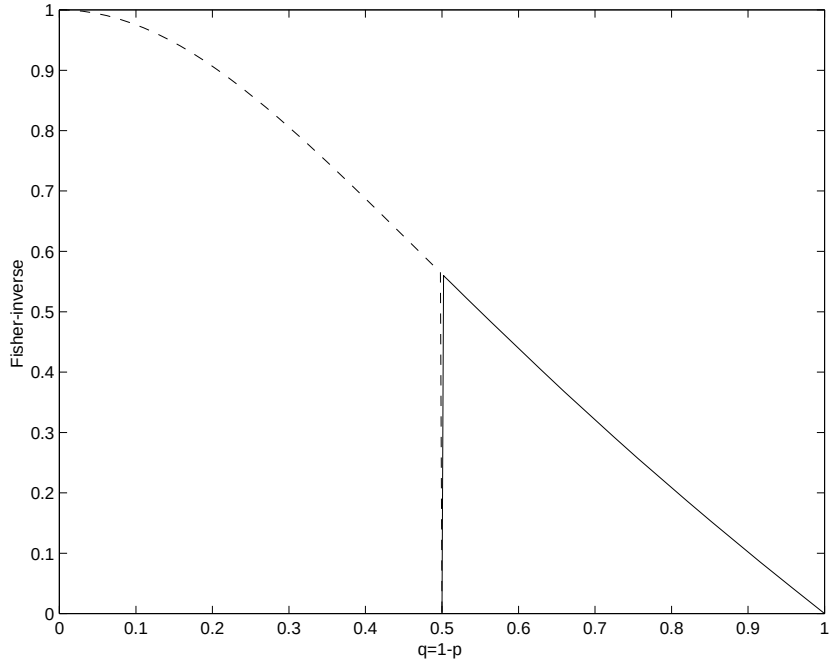


Figure 1. Inverse of Fisher information under two different parametrizations, $I(\theta)^{-1}$ and $I(q)^{-1} = I(p)^{-1}$ of θ and q or p, respectively, where $q = 1 - p = \sqrt{\theta}$.

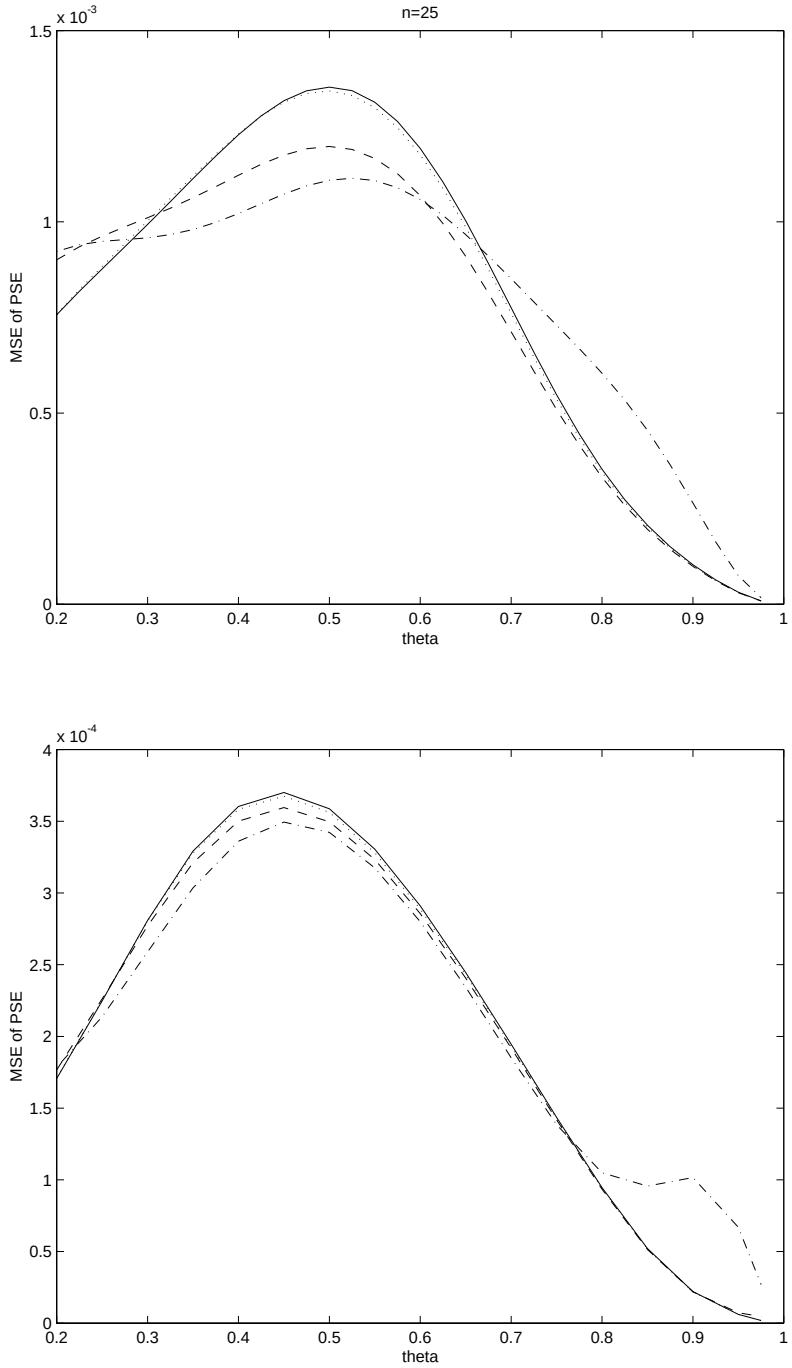


Figure 2. MSE of PSE comparisons with $\theta = (1-p)^2$ as parameter of interest.
Sample sizes $n = 25$ (upper) and $n = 50$ (lower).
Solid curve: $V = 1/I(\hat{\theta})$
Dotted curve: $V = 1/I_r(\hat{\theta})$
Dash-dotted curve: $V = 1/I_{rs}(\hat{\theta})$
Dashed curve: $V = 1/J(\hat{\theta})$

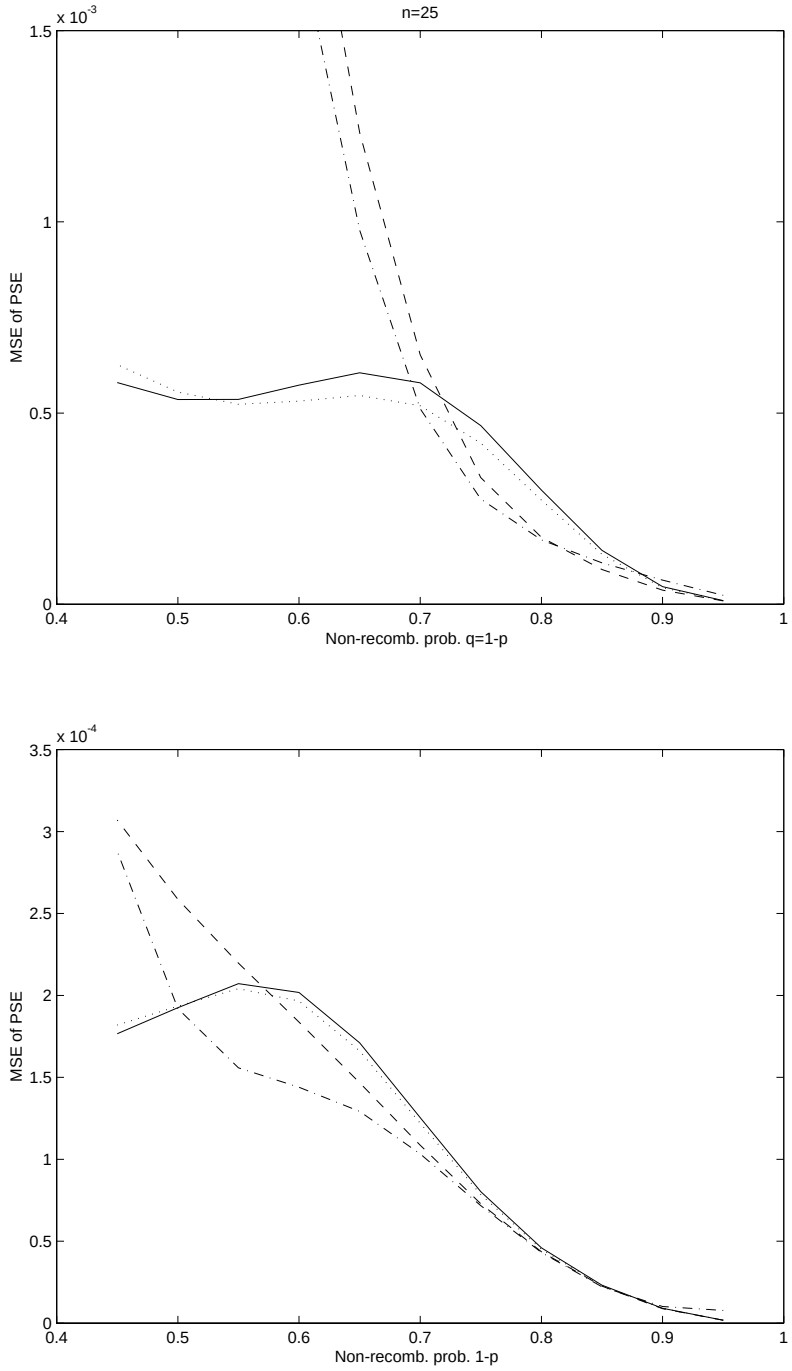


Figure 3. MSE of PSE comparisons with p (or $q = 1 - p$) as parameter of interest, sample sizes $n = 25$ (upper) and $n = 50$ (lower).
Solid curve: $V = 1/I(\hat{p})$
Dotted curve: $V = 1/I_r(\hat{p})$
Dash-dotted curve: $V = 1/I_{rs}(\hat{p})$
Dashed curve: $V = 1/J(\hat{p})$

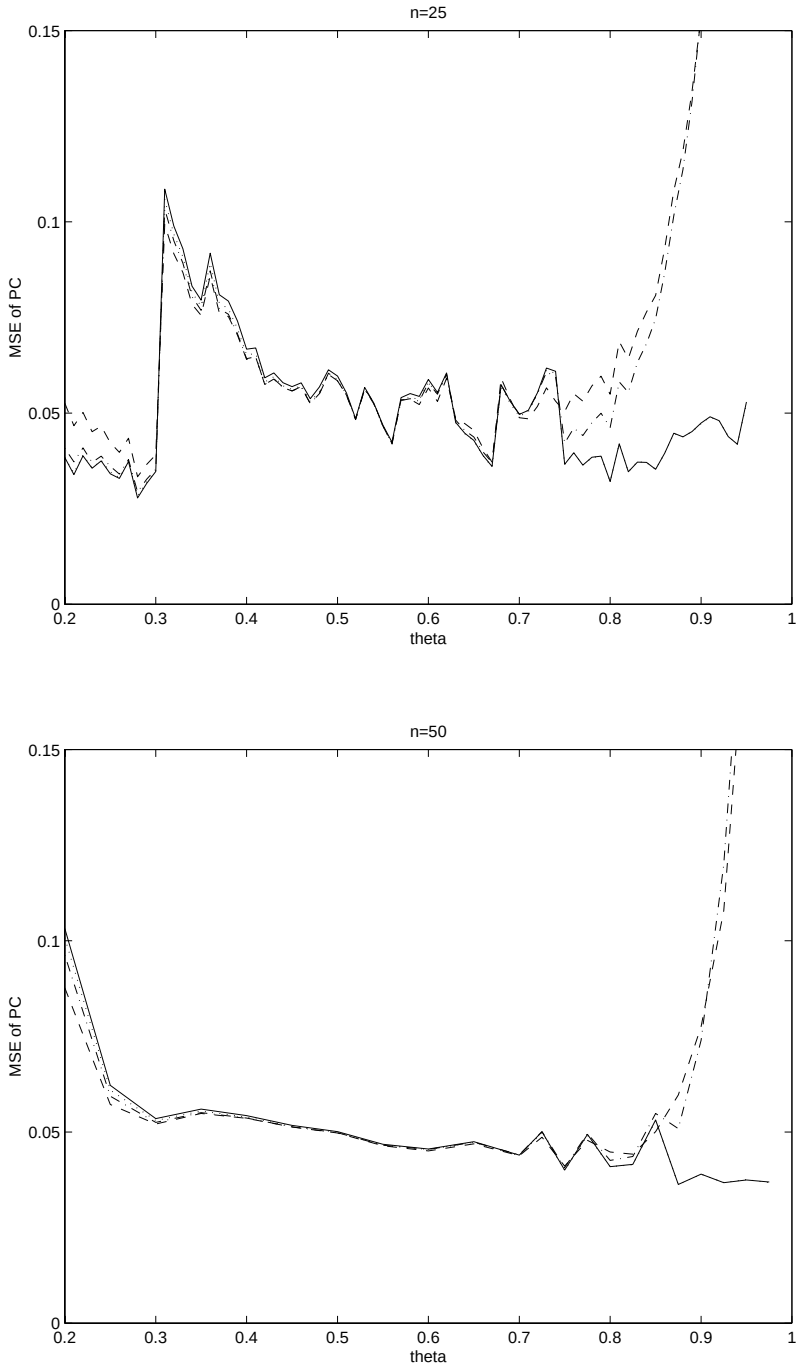


Figure 4. MSE of PC comparisons of confidence relevance, with $\theta = (1-p)^2$ as parameter of interest. Sample sizes $n = 25$ (upper) and $n = 50$ (lower).
Solid curve: $\hat{\alpha}$ as if $\hat{\theta} \sim \mathrm{N}(\theta,\, 1/I(\hat{\theta}))$
Dotted curve: $\hat{\alpha}$ as if $\hat{\theta} \sim \mathrm{N}(\theta,\, 1/I_r(\hat{\theta}))$, given r
Dash-dotted curve: $\hat{\alpha}$ as if $\hat{\theta} \sim \mathrm{N}(\theta,\, 1/I_{rs}(\hat{\theta}))$, given (r, s)
Dashed curve: $\hat{\alpha}$ as if $\hat{\theta} \sim \mathrm{N}(\theta,\, 1/J(\hat{\theta}))$

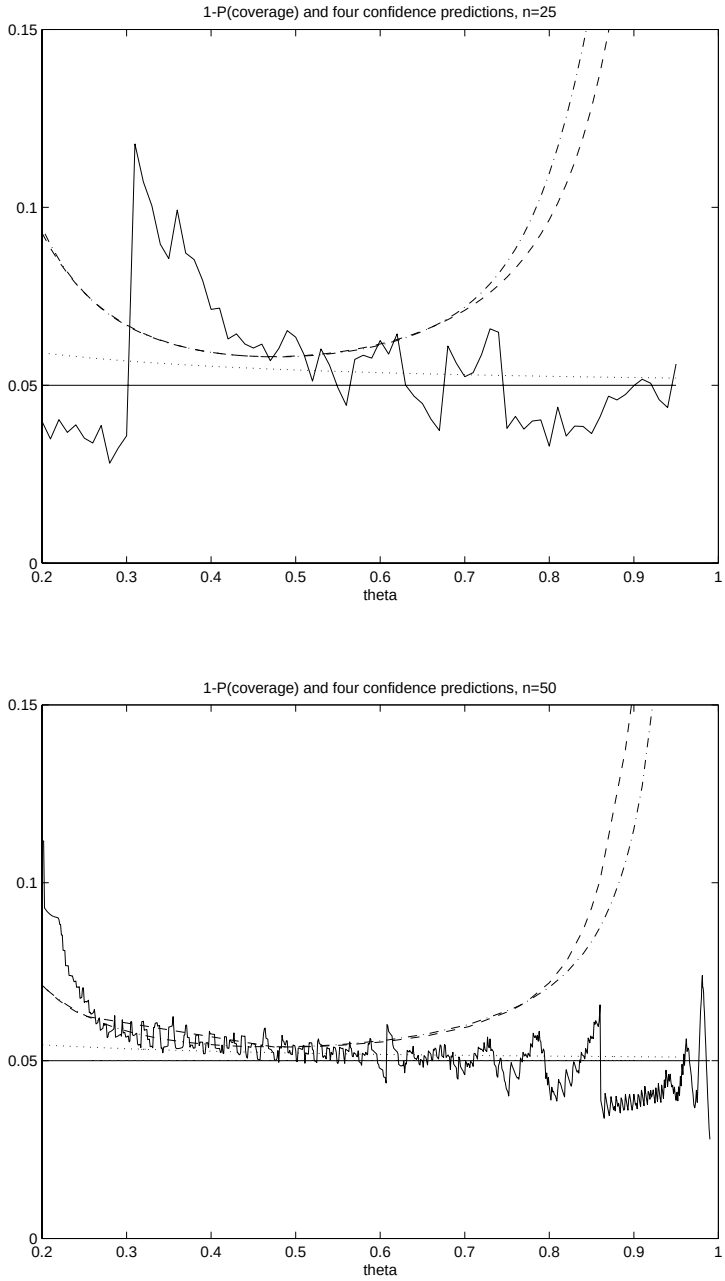


Figure 5. Coverage probability and four confidence predictions (complementary values shown) for an interval estimate of nominal 95% confidence, as functions of θ.
Sample sizes $n = 25$ (upper) and $n = 50$ (lower).
Solid curve (rugged): Actual non-coverage probability for the interval estimate, i.e. $E_\theta(\xi)$
Solid horizontal line: Nominal value of predicted confidence by construction, $\alpha = 5\%$
Dotted curve: Mean value of predicted confidence based on $I_r(\theta)$.
Dash-dotted curve: Mean value of predicted confidence based on $I_{rs}(\theta)$.
Dashed curve: Mean value of predicted confidence based on $J(\theta)$.

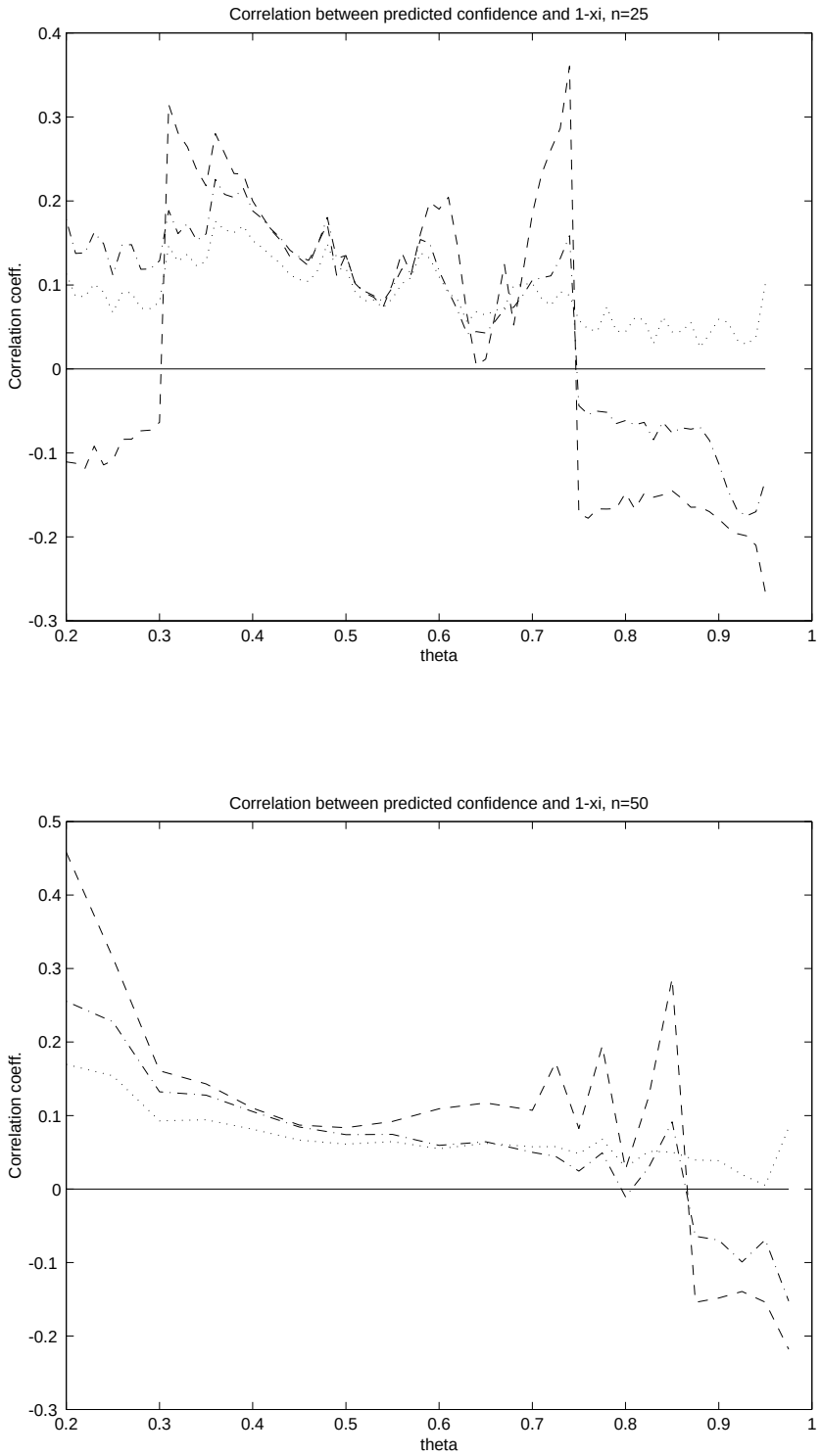


Figure 6. Correlation between predicted confidence and coverage $1 - \xi$. Sample sizes $n = 25$ (upper) and $n = 50$ (lower).