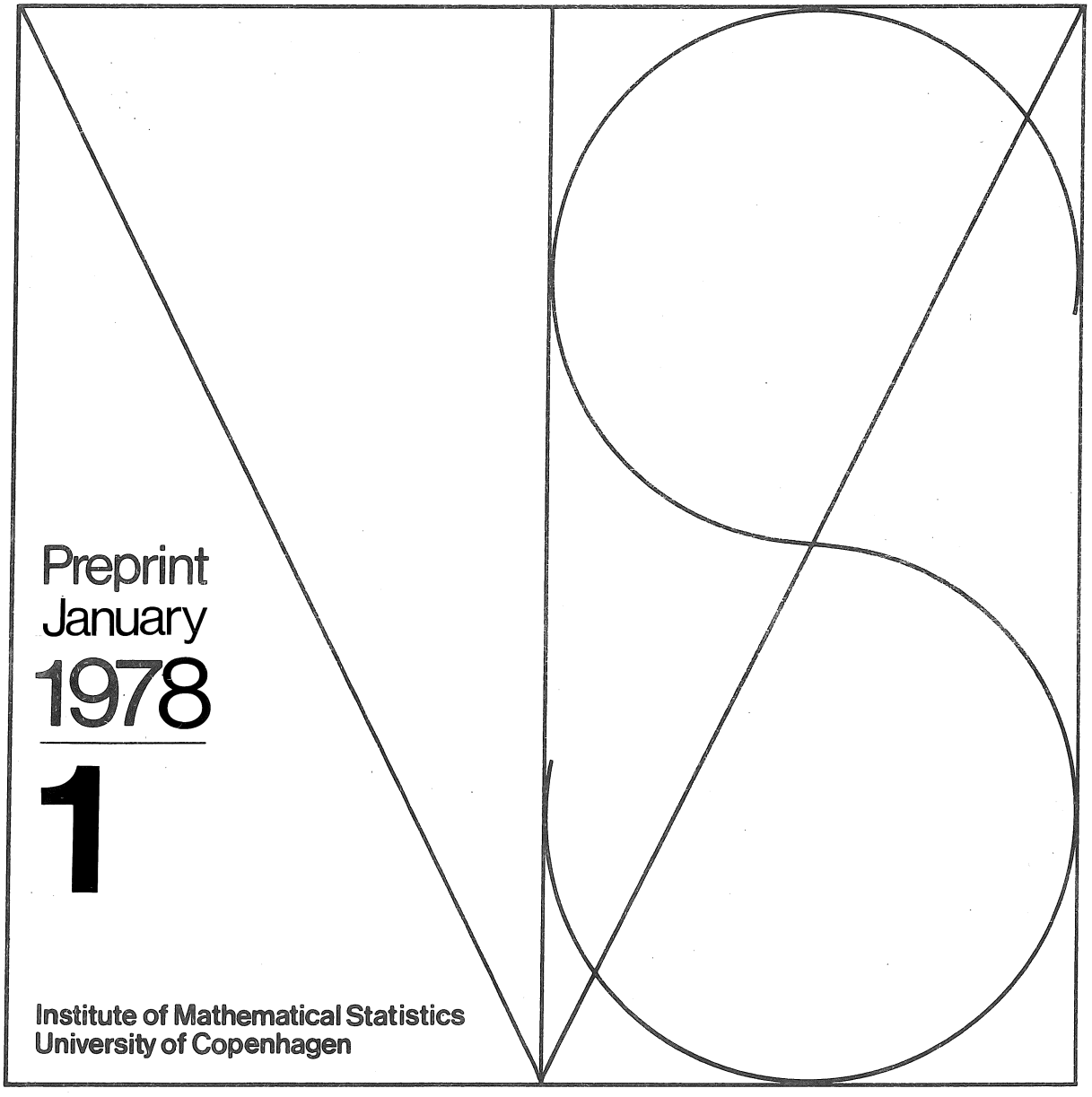
Tue Tjur

Statistical Inference under the Likelihood Principle



Preprint
January
1978

1

Institute of Mathematical Statistics
University of Copenhagen

Tue Tjur

STATISTICAL INFERENCE UNDER THE LIKELIHOOD PRINCIPLE

Preprint 1978 No. 1

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

January 1978

Abstract:

A short review of the arguments in favour of the likelihood principle is given. It is noticed that the likelihood principle in itself is not in conflict with inference from marginal or conditional experiments. An approach to statistical inference, based on the likelihood principle, is outlined. This approach includes inference from marginal or conditional likelihood functions and significance testing.

Key words:

Statistical inference
The likelihood principle
Partial likelihood
Stopping rule paradoxes

1. General remarks on the foundations of statistics.

In 1962, Allan Birnbaum proved a result which has had a very crucial influence on the discussion of the foundations of statistics since then. What he proved was, essentially, that a consistent theory of statistical inference , based on the idea of a parametric model, leads via some very natural principles to the likelihood principle.

The likelihood principle states that all information about an unknown parameter is contained in the likelihood function. Distributions, sample sizes etc. should only affect the inference through the likelihood function. For a review of the most important examples and "paradoxes" for and against the likelihood principle, see Basu (1974).

The likelihood principle fits nicely into the Bayesian approach, since the likelihood function is simply the density of the posterior distribution with respect to the prior distribution. This makes Birnbaum's theorem a very good argument for the Bayesian way of doing statistics. But many statisticians reject Bayesian inference , because they don't want their statements about unknown quantities to depend on more or less arbitrary prior distributions. The Bayesian answer to the problem of objective inference is, that all the statistician can do is to report the likelihood function (or -almost equivalently- a catalogue of posterior distributions corresponding to a reasonably broad class of priors). But many statisticians feel that it should be possible to summarize data in a more accessible form.

Non-Bayesian likelihood approaches to statistical inference (see Edwards (1972) , Kalbfleisch and Sprott (1970)) have had very limited success. The problem is that it is far from obvious how to draw conclusions from a likelihood function. In the presence of nuisance parameters, it is almost impossible. The value of the likelihood function at a point can not be taken as a naive measure of the "degree of belief" one should have in the hypothesis that this is the true value of the parameter. This is most

convincingly demonstrated by the case of a single observation from a normal distribution with unknown parameters, where the likelihood function suggests "infinite belief" in the hypothesis $\sigma^2 = 0$.

This has forced most non-Bayesian statisticians to the conclusion that the likelihood principle is wrong, and -consequently- that a formalization based solely on parametric models is not possible. It has been suggested that other structures on the sample space and the parameter space should be taken into account (see Barnard's contribution to the discussion of Birnbaums 1962-paper). For example, it has been argued that certain invariance properties of a statistical model should allow for a more concrete sort of inference (fiducial and structural inference, see Fisher (1956) and Fraser (1968)). It has been argued that order structures should support the relevance of tail probabilities (see Barndorff-Nielsen (1973) and Cox (1977)). It has been argued that the presence of a "canonical" underlying measure should support comparison of values of a density at different points of the sample space (see Martin-Löf (1974) and Barndorff-Nielsen (1976)). And it has very often been argued, that the presence of a "repetitive structure" (that is a natural way of extending the experiment, for example by letting the number of observations tend to infinity) should support the use of inference procedures with desirable asymptotic properties. Some of these arguments have a very strong intuitive appeal. Unfortunately, it is hard to see how an additional structure can affect the very simple arguments given by Birnbaum.

My intention with this paper is to defend the likelihood principle, and to indicate that it <u>is</u> possible to make statistical inference under the likelihood principle. It seems to be generally believed that most classical inference procedures (with maximum likelihood estimation as the obvious exception) are inconsistent with the likelihood principle. As we shall see, this is not quite true. Inference from conditional or marginal experiments can

be motivated on likelihood grounds, and even significance testing has its place in the likelihood hierachy.

2. The likelihood principle.

In the following, we are dealing with parametric models $(X , (\pi_\vartheta \mid \vartheta \in \Theta))$. Here, X is the sample space, and we have a family π_ϑ, $\vartheta \in \Theta$, of probability measures on X.

It should perhaps be emphasized that our starting point is that we believe in the model -or, at least, we decide to act as if we did. Questions of robustness etc. are sometimes put forward in the discussion of the foundations, but that seems to be a quite different matter.

By an experiment (or, more precisely, a performed experiment) we mean a parametric model together with an element (the observation) $x \in X$. The idea is that we think of x as the observed value of a stochastic variable with distribution π_ϑ, and we want to make conclusions about the unknown value of ϑ (for example, to estimate ϑ or to test a hypothesis about ϑ).

If we are to develop a consistent theory of statistical inference, based on the concept of a parametric model, then there are certain principles which can hardly be avoided. Birnbaum stressed two of them:

The sufficiency principle: Let $t: X \to Y$ be sufficient for the family (π_ϑ). Then the reduced experiment $t(x) \in (Y , (t(\pi_\vartheta) \mid \vartheta \in \Theta))$ contains exactly the same information about ϑ as did the original experiment $x \in (X , (\pi_\vartheta \mid \vartheta \in \Theta))$.

The ancillarity principle (or the conditionality principle): Let $s: X \to Z$ be ancillary (i.e. the distribution of $s(x)$ is independent of ϑ). Then, the conditional experiment $x \in (X , (\pi_\vartheta^{s(x)} \mid \vartheta \in \Theta))$ contains exactly the same

information about ϑ as did the original experiment $x \in (X , (\pi_\vartheta \mid \vartheta \in \Theta))$. (Here, $\pi_\vartheta^{s(x)}$ denotes the conditional distribution of x, given $s(x)$).

Notice that by information we mean qualitative information. The idea is that if two experiments contain the same information, then our conclusions from those experiments should be the same.

The two principles are very similar, and they are motivated by the same intuitive argument: Suppose that our experiment can be simulated by two consecutive experiments, where the second experiment may depend on the outcome of the first. And suppose that one of these two experiments is completely irrelevant (i.e. the distribution of its outcome is independent of ϑ). Then, inference should be based on the other part only.

We shall not give a more detailed motivation here. See Birnbaum (1962) or Basu (1974) for more detailed arguments.

In the following we shall assume that the probability measures π_ϑ have densities p_ϑ with respect to some underlying measure. We shall not discuss regularity conditions, but readers who prefer to be quite sure about what is going on may assume that X is discrete (or even finite) and that the underlying measure is counting measure. After all, models with continuous state space are idealizations, approximating the case of discrete state space, and general principles of inference for the discrete case ought to be valid in the continuous case also.

Birnbaum proved that the above two principles imply the following:

The likelihood principle: Let $x \in (X , (\pi_\vartheta \mid \vartheta \in \Theta))$ and $x' \in (X' , (\pi'_\vartheta \mid \vartheta \in \Theta))$ be two experiments with the same parameter space. Suppose that the two likelihood functions are proportional, i.e. there exists a constant $c > 0$ such that $p_\vartheta(x) = c \cdot p'_\vartheta(x')$ for all $\vartheta \in \Theta$. Then the two experiments contain exactly the same information about ϑ.

It should perhaps be mentioned that the likelihood principle follows from weaker assumptions than those given by Birnbaum, see Basu (1974).

Example 1. Let ϑ denote the unknown probability of an event. Consider the following two experiments:

(1) (fixed sample size) 100 independent repetitions are performed. The event occurs 15 times. The remaining 85 times it does not occur.

(2) (inverse sampling) Independent repetitions are continued until 15 successes have been observed. The 15th success occurs at the 100th repetition.

The two experiments have the same likelihood function (namely $\vartheta^{15}(1-\vartheta)^{85}$), and so -according to the likelihood principle- our conclusions about ϑ should be the same after the two experiments. But confidence intervals for ϑ and significance tests for specific values of ϑ are different in the two experiments. Even though the differences are small, this indicates that the classical concepts of confidence intervals and significance testing are in conflict with the likelihood principle.

Birnbaum's theorem is, I think, a very convincing argument for the likelihood principle. But other arguments exist. Three of them are given here:

(1) The concept of likelihood is fundamental: The likelihood function gives the probability of what happened as a function of the unknown parameter, and what more can we ask for ? Many classical inference procedures rely on considerations of what might have happened. This argument was put forward by Fisher (see e.g. Fisher (1956)), and it seems to be the only argument behind Edwards' acceptance of the likelihood function as the central tool in inference.

(2) Bayesian statistics obeys the likelihood principle: Most statisticians agree that if we have a known prior distribution (for example if ϑ can be regarded as a member of a well known population) then the Bayesian way of making inference is the correct one. Now, it is not always obvious whether we have a prior or not. We may have a prior estimated from a very small number of ϑ's etc. It may be acceptable that our methods of inference should be different, according to whether we have a prior or not. But it would be surprising and suspicious if information about sample space etc. -which is irrelevant under any prior distribution- should suddenly become important in case we are not able to specify a prior.

(3) The sample space is not always well defined: We shall illustrate this by a sequential experiment where the stopping rule is only vaguely defined. It is well known that the likelihood function is independent of the stopping rule (cfr. example 1), and very often this is considered an argument against the likelihood principle. But it may as well be turned the other way around: Consider the situation of example 1. Suppose we start out making independent repetitions without much idea about when to stop. Our intention may be to prove a theory according to which $\vartheta = 1/2$. When observing the 15th success in experiment no. 100, we realize that this is more than enough, and since 100 is such a nice number we decide to stop here. What is the stopping rule ? And do we really need to know it ? After all, there is no doubt about what we have observed; the stopping rule is merely a part of our private opinion about how we managed to observe it. The likelihood function is known, and the data seem to contain very relevant information about the hypothesis $\vartheta = 1/2$.

3. Estimation.

In this and the following section, I shall give a very short outline of a likelihood approach to statistical

inference. I have tried to make a distinction between the two main types of inference: Estimation and hypothesis testing. However, this does not mean that it is possible to skip this section.

Our starting point can be described as follows: We believe in the likelihood principle, and we believe in the Bayesian argument whenever we are willing to state our prior knowledge (or lack of knowledge) in terms of a probability distribution. But unfortunately we are not willing to do that very often. In most situations our prior knowledge is better described as a vague feeling of what a reasonable prior distribution would be like.

However, there are situations where such a vague feeling is all we need. This is illustrated by example 3 below. But before that we shall give a very simple example, where a direct, non-Bayesian estimation procedure seems to be acceptable:

Example 2. Suppose that the parameter space consists of two points only, say 1 and 2 . Birnbaum treated this case separately in a paper from 1961. We have two probability densities p_1 and p_2 , and we observe an x with distribution either p_1 or p_2 . Suppose that we have no prior knowledge indicating that one of the two hypotheses should be more likely than the other. What can we say about the true value of ϑ ? We proceed as follows: Consider the values $p_1(x)$ and $p_2(x)$ of the densities at the observed point x. These numbers are (proportional to) the probabilities of the event observed under the two hypotheses. If they are approximately equal, we can obviously say nothing of interest. If one (for example $p_2(x)$) is considerably larger than the other, the corresponding ϑ -value is considered the most likely, and the likelihood ratio (in this case $p_2(x)/p_1(x)$) is regarded as a measure of "confidence" in this statement about ϑ .

This procedure is very acceptable from a Bayesian point of view: In case of a symmetric prior, the likelihood

quotient is simply the ratio between the two probabilities of the posterior distribution. Even a Neyman-Pearson-like argument supports the procedure: It follows from the Neyman-Pearson lemma, that the maximum likelihood estimator in this case minimizes the sum of the two error probabilities π_1(estimating ϑ to be 2) and π_2(estimating ϑ to be 1).

Example 3. Suppose that Θ is the real axis (or an interval) and that we have obtained a likelihood function of shape something like fig. 1 . It is assumed that our prior

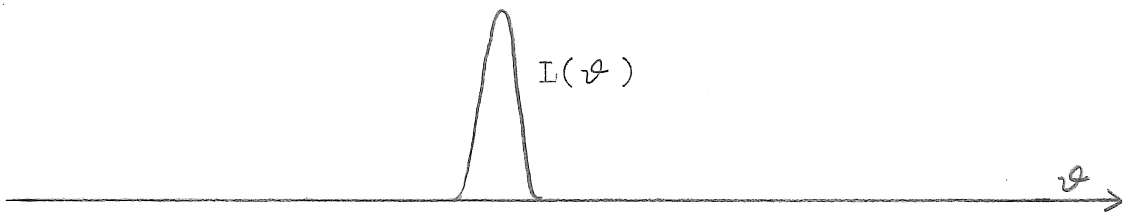


fig. 1

knowledge about ϑ is almost empty, compared to the accuracy indicated by the likelihood function. Typically, the likelihood function may be that of a very precise measurement (cfr. Savage (1960)) of an only vaguely known physical constant ϑ . Our problem is to estimate ϑ . We proceed as follows: Suppose we were to express our prior knowledge in terms of a prior distribution. This distribution would obviously have to be very flat, and in the small interval of interest (just around the peak of the likelihood function) the prior distribution would be approximately proportional to Lebesgue measure. This means that the posterior distribution is approximately equal to the distribution given by the density const.·L with respect to Lebesgue measure. Thus, the posterior distribution is almost independent of the prior, as long as only "reasonable" priors are considered. In this case the use of Bayesian estimates and Bayesian confidence statements is suggested.

Many statisticians reject this method, simply because

it is Bayesian. But it can hardly be denied that it is possible to make inference from a likelihood function in this way.

The above argument can be applied to any parameter space, not only the real axis. But it should be noticed that the "robustness" of the likelihood function in case of many parameters is very often surprisingly small. A seemingly small change of the prior distribution may change completely the distribution (prior as well as posterior) of certain parametric functions.

Example 4. Now, consider a situation where the likelihood function is less robust, in the sense that different prior distributions give rise to rather different posterior distributions. As a more concrete example, suppose we obtain a likelihood function like that of fig. 2,

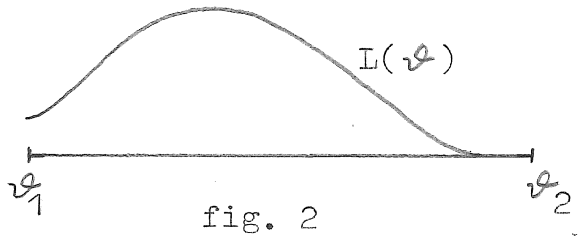


fig. 2

Θ being the bounded interval $[\vartheta_1, \vartheta_2]$. My personal opinion is that all one can do in such a situation is to ask questions of the type "What would the posterior distribution be like if the prior was ... ". The likelihood function should be judged on the way it transforms prior distributions into posterior distributions. In the present situation, it may be possible to end up with something like an upper confidence bound ϑ_0 for ϑ, while a point estimate of ϑ would be of little use. After all, if we are not able to say very much about ϑ in advance, and if the experiment is rather useless, why should we be able to say very much after the experiment ? Classical statistical methods (like the notion of a confidence interval) suggest that it is possible to be very specific in this

situation. So much the worse for those methods.

Until now, the methods suggested have not differed much from those recommended by the Bayesian school, see in particular Savage (1960) and Dickey (1976). I believe that the Bayesian approach is essentially the correct one for estimation in the absence of nuisance parameters. In the presence of nuisance parameters, the Bayesian solution is more difficult to swallow.

Example 5. Let $x_1, y_1, \ldots, x_n, y_n$ be independent, normally distributed with parameters

$$E\, x_i = E\, y_i = \xi_i$$

$$\mathrm{var}(x_i) = \mathrm{var}(y_i) = \sigma^2 .$$

We want to estimate the variance. Likelihood inference from this experiment is complicated by the many nuisance parameters. Bayesian elimination of the nuisance parameters from the likelihood function is possible if the ξ_i's belong to some welldefined population, but if this is not the case, the situation is rather hopeless. It can be described as follows: We have made an experiment in order to determine σ^2, but we don't know which experiment (it depends on the nuisance parameters). Data do not tell us very much about which experiment we have made, and our conclusions about σ^2 depend very much on this. The situation is impossible, and all we can do is to start looking for an experiment with a more approachable likelihood function.

Now, we notice that the experiment described above has a very nice subexperiment. Suppose, namely, that we observe only $x_1 - y_1, \ldots, x_n - y_n$. These variables are stochastically independent, normally distributed with mean 0 and variance $2\sigma^2$, and for n large we shall find ourselves in a situation like that described in example 3. So why not base the estimation of σ^2 on this subexperiment ; pretending that the remaining part of the data was never

observed ? I think that even orthodox Bayesians would tend to prefer this solution (or, at least: They would tend to wish that they had only observed the differences).

For some reason, it seems to be generally agreed that this method is inconsistent with the likelihood principle (see, however, Barnards comment in the discussion of Kalbfleisch and Sprott (1970) and the book of Edwards (1972)). Obviously, the method involves arguments which are not based on the likelihood function of the total experiment. However, the method can be justified on likelihood grounds as follows:

Suppose that, before the experiment is carried out, we realize that the total experiment is useless, since the likelihood function will contain nuisance parameters. Having realized this, we notice that a certain marginal experiment has more desirable properties, and so we decide to carry out that marginal experiment. Unfortunately, the only way of doing this is to carry out the total experiment, followed by a data reduction. But this should not force us to believe that the total experiment is the relevant one. The likelihood principle tells us something about what to do when an experiment has been carried out, but it does not tell us which experiment to consider. In particular, it does not forbid us to consider a marginal experiment.

This justification of the marginal experiment as the basis for inference is based on the rather unrealistic assumption that the statistician is present before the experiment. In practice, this is usually not the case, but our common sense tells us that this should not affect the argument. All the statistician has to do is to "act as if data had not been reported yet ". The decision to reduce to a marginal experiment should be based on the description of the total experiment, but (at least in principle) it should not depend on the data.

It may be argued that the reduction to a subexperiment is somewhat against the spirit of the likelihood principle.

But this criticism of the marginal likelihood method subsumes another principle, according to which a bigger experiment should always be preferred to a smaller one. Or, in terms of likelihoods: The relevant likelihood function is always that of the biggest possible experiment involving the parameter of interest. I think that principle is wrong.

Birnbaum was aware of this problem in 1962. In an answer to a question posed by Kempthorne, he suggested that the outcome of an experiment based on a randomized design should be analyzed as if the (random) allocation of treatments to plots had not been observed. This suggestion seems to solve the controversy between randomization and the ancillarity principle.

Even if reduction to a subexperiment is in formal agreement with the likelihood principle, it is obviously not a thing one should do whenever it makes inference easier. The extreme way of obtaining easy inference is to reduce to the empty experiment, where not only the nuisance parameters, but also the parameter of interest has disappeared from the likelihood function. I insist that this behaviour would not be in conflict with the likelihood principle, but it might violate some other principles which we have not discussed here. However, the likelihood principle in itself indicates the criteria according to which one should choose the best subexperiment: An experiment should be judged on the likelihood functions it can produce. Two criteria present themselves immediately:

(1) The likelihood function should be such that likelihood inference is possible (i.e. our conclusions should not depend too much on more or less arbitrary prior distributions).

(2) The experiment should be as informative (i.e. as big) as possible among those satisfying (1).

These criteria are very vague, and I don't think it is possible to give simple rules as to which experiment one

should prefer. Even in the Neyman-Scott example (example 5) it is not a trivial matter to decide whether one should look at the total experiment or at the differences only. It depends on our knowledge about the nuisance parameters. However, our conclusion does not depend on a detailed discussion of criteria for optimal choice of an experiment. The point is that a subexperiment may be preferable in some situations, and in such situations we are free to base our inference on the likelihood function of that subexperiment.

Conditional and partial likelihoods. Until now, we have argued as if our subexperiment had to be a marginal experiment. But obviously, by the same argument we may prefer a conditional experiment, or a marginal experiment derived from a conditional experiment. That is, if $t(x)$ and $s(x)$ are statistics such that the observation of $t(x)$, given $s(x)$, is considered more relevant than the observation of the outcome of the total experiment, then we are free to base our inference on the corresponding "marginal-conditional" likelihood function.

Even a more general kind of "subexperiments" may be considered: Suppose that we make a finite sequence of experiments. After any experiment, we are free to decide whether or not to go on to the next experiment, and the next experiment may depend on the outcome of the previous ones. Moreover, some of these experiments may be marginal experiments. Thus, the final likelihood function -the so called partial likelihood function (Cox (1975))- is a product of marginal likelihood functions, each of which is derived from a conditional experiment, given the previous ones. Such a likelihood function can not always be derived as the likelihood function of a conditional or marginal (or "conditional-marginal") experiment. We shall not give a more precise definition of partial likelihood functions here. The following example illustrates the idea in a special case:

Example 6. Suppose we observe k generations of some animal population under conditions where the usual branching process model is applicable. However, we are not interested in the offspring distribution, but only in the probability p that a newborn individual dies immediately after birth from a certain disease. The obvious estimate of p is the number x of deaths of this kind, divided by the total number n of births observed. Intuitively we feel that, according to accuracy, this estimate can be regarded as if the total number of births was fixed, in which case we would have the simple situation of estimating in a binomial distribution. This can be motivated as follows: Let n_i denote the total number of individuals born into the i^{th} generation, and let x_i denote the number of those who die from the disease immediately after birth. First we observe x_1, given n_1, which gives us the likelihood function $p^{x_1}(1-p)^{n_1-x_1}$. Then we observe a lot of things which we are not interested in, namely the growth and parental behaviour of the first generation. All this information is disregarded. We are now in the position to make a new relevant observation, namely x_2, given n_2. The likelihood function $p^{x_2}(1-p)^{n_2-x_2}$ is multiplied by the likelihood function of the previous experiment. Continuing in this manner, we end up with the partial likelihood function $p^{x}(1-p)^{n-x}$.

It is not obvious that it makes sense to multiply these likelihood functions together as if they were likelihoods of independent experiments. However, the procedure can be justified by a Bayesian argument: Suppose we start out with a prior distribution for the parameter of interest (p, in the example). Then, after the first experiment, we end up with a posterior distribution. This distribution represents our present belief, and the outcome of the first experiment can now be regarded as a given constant. Now we are free to carry out any new experiment, for example a subexperiment of some experiment which is determined by the outcome of the previous experiment. Again, the prior distribution (i.e. the posterior distribution from the first

experiment) is multiplied by the likelihood function to form a new posterior distribution. Continuing in this way, we obviously end up with a posterior distribution which is the product of the first prior distribution and the partial likelihood function. Thus, as long as our inference from the partial likelihood function is essentially Bayesian, the partial likelihood function has exactly the same operational meaning as a proper likelihood function. It should be noticed, however, that a partial likelihood function is not a likelihood function of an experiment. It is not obvious to me how such a partial likelihood function can be ascribed any meaning in a non-Bayesian framework.

Reduction of an experiment for computational convenience. Our reason for preferring subexperiments in the previous examples was to avoid nuisance parameters. However, other reasons may exist.

Example 7. Let $x_1, \ldots, x_n$ be independent, identically distributed with density $p(x+\vartheta)$ with respect to Lebesgue measure on the real line, and suppose we want to estimate ϑ. The correct way of doing this is of course to compute the likelihood function

$$L(\vartheta) = p(x_1+\vartheta)\ldots p(x_n+\vartheta)$$

and make inference as indicated in examples 3 and 4. However, this may be computationally very tedious. An easy way of escaping this problem is the following: Suppose we have a statistic $t(x_1, \ldots, x_n)$ (for example the median or the mean of the empirical distribution) which is known to be approximately normally distributed with mean ϑ and variance close to the Cramér-Rao lower bound. It is intuitively obvious, then, that t contains almost all the information we are interested in. So why not act as if we had only observed t? It makes inference much easier, because the likelihood function is (approximately) known.

It may seem surprising, that such methods have a place

in likelihood inference. But, again, the likelihood principle in itself is not violated by this method. All we can say against the method is that it obviously wastes some information, namely the information contained in the conditional distribution of the sample given t. But if this information is found to be so small that it is not worth the trouble to use the correct likelihood function, then our marginal inference procedure seems to be perfectly well motivated.

4. Hypothesis testing.

An important criticism against the likelihood principle is that it is impossible to make anything like a significance test, based on the likelihood function of the total experiment. In this section, we shall see that significance testing can be regarded as likelihood inference from a subexperiment. But first, we shall discuss how to handle a likelihood function in a testing situation.

There is one (and essentially only one) testing situation where a pure likelihood approach is possible:

Simple hypothesis against simple alternative. Suppose that the parameter space consists of two points 1 and 2, corresponding to the densities p_1 and p_2. From a single observation x we want to test the hypothesis $\vartheta = 1$. That is, we want to know if data are decisively against the hypothesis $\vartheta = 1$ in favour of $\vartheta = 2$. The obvious thing to do is the following: Consider the likelihood quotient $Q = p_2(x)/p_1(x)$. If this number is very large, it means that the probability of the event observed is very much larger under the alternative hypothesis $\vartheta = 2$. Thus, if we have to make a decision we shall obviously reject the hypothesis $\vartheta = 1$ for Q greater than some number c, for example $c = 20$ or $c = 100$. This procedure has a very desirable property, which may also be of help when we are to choose c: The size of the test (i.e. the probability of rejecting the null hypothesis

when it is true) is smaller than $1/c$. Moreover, by the Neyman-Pearson lemma, this test is the most powerful among tests of the same size (i.e. it minimizes π_2(accepting $\vartheta = 1$) for fixed π_1(rejecting $\vartheta = 1$)).

This approach was suggested by Dempster (1974), and the following approach to the case of simple hypothesis against composite alternative is also very similar to Dempster's:

Simple hypothesis against a composite alternative. For simplicity, the parameter space is assumed to be the real axis (or an interval). We want to test the hypothesis $\vartheta = \vartheta_0$ against the alternative $\vartheta \neq \vartheta_0$.

This situation can be reduced to the situation of a simple hypothesis against a simple alternative, if we are willing to specify the alternative $\vartheta \neq \vartheta_0$ in terms of an "alternative prior". By an alternative prior we simply mean a probability distribution α on the set $\Theta \setminus \{\vartheta_0\}$. Having specified this, the test is carried out as a test of the simple hypothesis $\vartheta = \vartheta_0$ against the simple hypothesis "ϑ has distribution α". The test statistic Q can be written as

$$Q = \int L(\vartheta)\, \alpha(d\vartheta) \,/\, L(\vartheta_0) ,$$

where L is the likelihood function.

The obvious interpretation of α is that α specifies our "prior knowledge, given that $\vartheta \neq \vartheta_0$". The procedure is essentially Bayesian, since the conclusion of the test is very similar to the conclusion we would obtain if we started up with a prior distribution of mixed type, namely a convex combination of α and the Dirach measure at ϑ_0. However, the applicability of the test is not restricted to situations where such a prior knowledge can be specified. The alternative prior α specifies the alternative, but as to the test itself, this only means that α specifies the conditions under which the test is optimal. For example, if we want a test which is powerful against small deviations from ϑ_0, we should take care

that α has a substantial part of its mass close to ϑ_0. And if we suspect certain alternative values of ϑ, α should have a substantial part of its mass at (or close to) those points. For statisticians of the Neyman-Pearson school, the best interpretation of α is probably the following: The alternative prior α represents a criterion for maximization of the power function. The test we end up with is optimal in the sense that the integral of the power function with respect to α is maximal among tests of the same size.

Obviously, the conclusion of the test varies with α, and this is the main argument against the method. However, this variation depends very much on the actual likelihood function. We shall illustrate this by three more specific examples:

Example 8. Suppose we have a likelihood function like that of fig. 3. In this case, the conclusion of the test

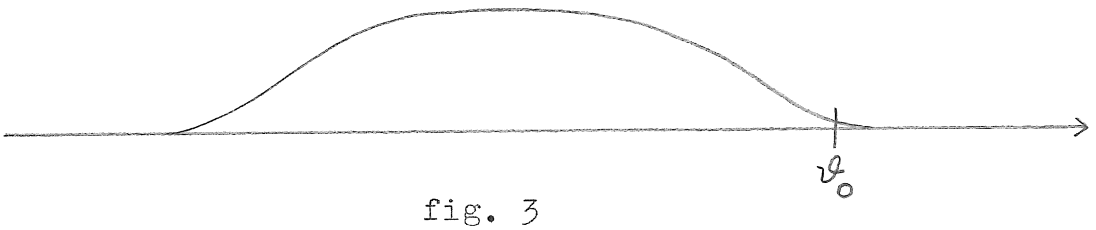


fig. 3

depends strongly on α. In order to apply the method indicated above we must be willing to state rather explicitely what we want to test against (the alternative prior). It should be noticed, however, that large values of the test statistic Q always indicate that the null hypothesis is false. For example, $Q > 100$ means that the alternative hypothesis "ϑ has distribution α" makes data at least 100 times more probable than does the null hypothesis, and even if α was specified rather arbitrarily (before the experiment), this is a decisive argument against the null hypothesis. For $\vartheta = \vartheta_0$, the probability of obtaining

$Q > 100$ is smaller than $1/100$.

Example 9. Now, consider a likelihood function like that of fig. 4. Obviously, we have to accept the null hypothesis

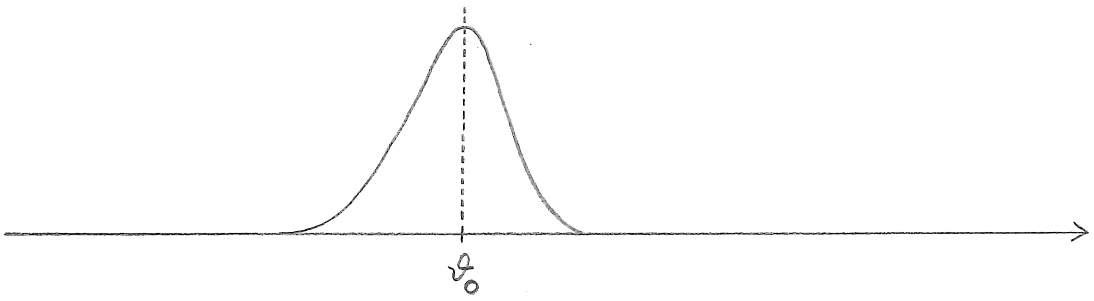


fig. 4

in this situation: For any alternative prior α (even for α concentrated at a single point) we obtain $Q \leq 1$.

More generally, suppose that $\max L$ is not much greater than $L(\vartheta_0)$, for example that

$$L(\vartheta) / L(\vartheta_0) \leq 5 \qquad \text{for all } \vartheta.$$

Then we have $Q \leq 5$ for any α, and so we would accept the null hypothesis for any α. In this sense, the likelihood function allows for objective inference.

Example 10. This time we assume a likelihood function as indicated by fig. 5. More precisely, let us say that the

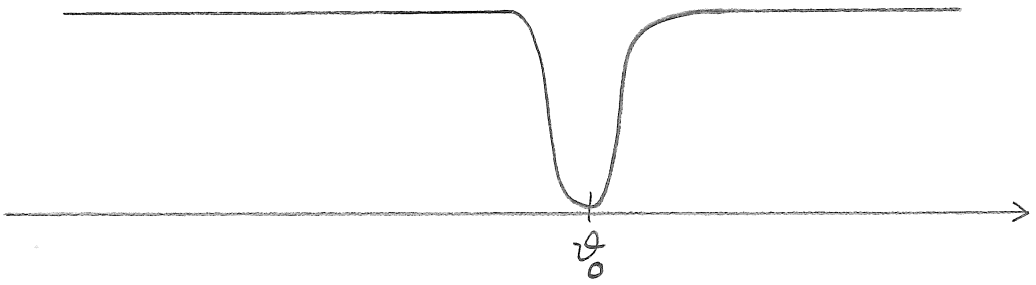


fig. 5

likelihood function has its values close to 1, except on a small neighbourhood of ϑ_0, and that the value at ϑ_0 is .01. Then, for any α which is reasonably flat

(or just with only little of its mass contained in the small neighbourhood of ϑ_0) we have approximately $Q = 100$. Loosely speaking, this means that for any reasonable alternative prior (even for one-point alternatives not too close to ϑ_0) the hypothesis $\vartheta \neq \vartheta_0$ makes the observed data about 100 times more probable than does the null hypothesis $\vartheta = \vartheta_0$, and by this argument we reject the null hypothesis. Hence, in this situation it is not necessary to specify the alternative prior in full detail, if we are willing to accept that alternative values of ϑ close to ϑ_0 can not be detected by the test. In this sense, we can reject the null hypothesis on objective grounds.

Thus it seems that likelihood functions like those of fig. 4 and fig. 5 are preferable to the (more typical) likelihood function of fig. 3. The reader may find that the function of fig. 5 does not look very much like a likelihood function. But the whole point of this section is, that if we want to be able to reject a hypothesis $\vartheta = \vartheta_0$ on objective grounds, then we have to make experiments that can produce such a likelihood function. This is where significance testing comes in. We shall illustrate the idea by a more concrete example:

Example 11. Let ϑ denote the value of some physical constant, and suppose that somebody has put forward the hypothesis $\vartheta = \vartheta_0$. We do not believe in this hypothesis. How can we prove that it is false (or, at least, very unlikely to be true) ? To our disposal we have a very accurate instrument for measuring quantities like ϑ. Whatever we do should include such a measurement. The question is how to formalize the inference procedure. We shall consider two solutions to this problem:

(1) We specify the alternative $\vartheta \neq \vartheta_0$ in terms of an alternative prior α. Then we carry out the measurement. If the test statistic $Q = \int L(\vartheta)\,\alpha(d\vartheta) \,/\, L(\vartheta_0)$ is very large (which is what

we expect), for example $Q \geq 1000$, then we can argue as follows: Even if our specification of the alternative is rather arbitrary, it makes data at least 1000 times more probable than does the hypothesis $\vartheta = \vartheta_0$. In case $\vartheta = \vartheta_0$, this would happen with probability less than 1/1000.

This argument requires that α is specified before the experiment. Arguments of the type "for any reasonable prior... " may be possible in some situations, but a reasonable prior can not be defined as a sufficiently flat prior. The more flat we make the alternative prior, the smaller becomes the test statistic Q. Whatever x is, the hypothesis $\vartheta = \vartheta_0$ will always be accepted for α flat enough. This means that the "proof " of falsity of the null hypothesis depends entirely on our choice of α. This is not very desirable, since it reduces a scientific inference procedure to a sort of gambling. In order to reject the null hypothesis, we must be able to prove (by witnesses or whatever) that α was chosen in advance.

In order to avoid these difficulties, we may prefer to base the inference on a subexperiment:

(2) The measurement is carried out, but we reduce to the summary statistic

$$y = \begin{cases} 0 & \text{if } |x| > 3\sigma \\ 1 & \text{if } |x| \leq 3\sigma \end{cases}$$

where σ denotes the (known) standard deviation of our measuring instrument. In case we obtain $y = 0$ (which is what we expect), we have a marginal likelihood function like that of fig. 5 (namely the power function of a u-test), and the null hypothesis is rejected by the argument given in example 10.

Hence, it seems that the approach (2) allows for a somewhat more convincing "proof " of the falsity of the

null hypothesis. The reduction to a binary subexperiment is very analogous to the reductions made in section 3 in order to avoid nuisance parameters. From the point of view of hypothesis testing, the specific value of the parameter (in case the null hypothesis is false) is a nuisance parameter, and the reduction to the binary experiment is made in order to obtain a likelihood function which is approximately independent of that nuisance parameter.

Significance testing. The above example indicates the role of significance testing in likelihood inference: A significance test is a subexperiment with two outcomes, 0 (or "reject") and 1 (or "accept"). The likelihood function corresponding to the outcome "reject" is called the power function, and the likelihood function corresponding to the outcome "accept" is then one minus the power function. As in any other experiment, the outcome of a significance test should be judged on its likelihood function only. The advantage of a significance test -as compared to the total experiment- is that the likelihood function allows for a more objective sort of conclusions (in case of test of a simple hypothesis, it is subsumed that the two possible likelihood functions should be of shape like those of fig. 4 and 5).

It should be emphasized, that from this point of view a significance test is not a way of making inference from the total experiment. A significance test is an experiment. This solves the "stopping rule paradox":

Example 12. Let $x_1, x_2, \ldots$ denote the decrease of blood pressure obtained by a certain treatment for patients $1, 2, \ldots$. A doctor wants to prove that the treatment has an effect -positive or negative. To this end, he continues sampling until significance is obtained by an ordinary t-test on the one per cent level. It is well known that this is going to happen sooner or later, even if the treatment has no effect at all. But the doctor argues as follows: "According to the likelihood principle, the final experiment is equivalent to an experiment with a

fixed stopping time. In that experiment, we would reject the null hypothesis by a t-test " .

This "paradox " has very often been considered an important argument against the likelihood principle. But from our point of view, the doctor's argument is not a correct application of the likelihood principle. It is correct that the total experiment is equivalent to an imagined experiment of fixed sample size. This means that our situation, when we try to make inference from the sequential experiment, is exactly as it would be if we tried to make inference from the total imagined experiment. But this does not include a t-test, since a t-test makes inference from a subexperiment of the imagined experiment.

Thus, as long as we intend to base our inference on the likelihood function of the total experiment, the stopping rule plays no role. We can sample freely, until we decide to stop for one or another reason. But as soon as we want to consider a subexperiment -like a t-test or a sequential test- we must specify the stopping rule in advance, because the likelihood function of the subexperiment does depend on the stopping rule.

Test of composite hypotheses. In test of a simple hypothesis, it is very often a matter of taste whether one should make inference directly from the total experiment, or from a significance test. We may have an idea about the order of magnitude of ϑ , and that imposes a bound on the flatness of the alternative prior. In that case, statements of the type "for any reasonable alternative prior ... " may be acceptable.

When testing a composite hypothesis (against a composite alternative) the situation is much more complicated, and this is, perhaps, where significance testing really comes into its own right. Let $\Theta_0 \subset \Theta$ denote a subset of the parameter space, and suppose that we want to test the hypothesis $\vartheta \in \Theta_0$. Typically, Θ will be a subset of a Euclidean space, and Θ_0 will be a manifold of lower

dimension. In order to reduce to the case of simple hypothesis against simple alternative, we have to specify the alternative as well as the hypothesis itself in terms of "prior" distributions. The test statistic becomes

$$Q = \int L(\vartheta)\,\alpha(d\vartheta) \Big/ \int L(\vartheta)\,\beta(d\vartheta) \quad ,$$

where α specifies the alternative and β specifies the hypothesis to be tested. The conclusion of such a test depends in a very complicated manner on α and β. It is not true -as it was in case of a simple hypothesis- that the event $Q \geq c$ has probability $\leq 1/c$ under the null hypothesis. This is only true if our specification of the null hypothesis is correct (whatever that means in practice).

Obviously, the first thing one should look for in this situation is a nice subexperiment, parametrized by a function $\varphi = \varphi(\vartheta)$, such that our composite hypothesis is reduced to a simple hypothesis $\varphi = \varphi_0$ (where $\Theta_0 = \{\vartheta \mid \varphi(\vartheta) = \varphi_0\}$). But even if such a reduction is not possible, we may still be able to make a significance test:

Let $t: X \to \{0,1\}$ be a transformation such that

(1) $\pi_\vartheta\{ t(x) = 0 \}$ is small (for example $\leq .01$) for any $\vartheta \in \Theta_0$.

(2) $\pi_\vartheta\{ t(x) = 0 \}$ is close to 1 for $\vartheta \notin \Theta_0$, except for ϑ very close to Θ_0.

Inference from this subexperiment is easy: In case we observe $t(x) = 1$, we obviously have to accept the null hypothesis, since all values of the likelihood function $L(\vartheta) = \pi_\vartheta\{ t(x) = 1 \}$ for $\vartheta \in \Theta_0$ are very close to $\sup L$. For $t(x) = 0$, we can argue as follows: For any β specifying the null hypothesis and for any alternative prior α with only little of its mass close to Θ_0, Q is very large (for example ≥ 100). Thus, we reject

the null hypothesis $\vartheta \in \Theta_0$.

Again, the "approximate objectivity" of the conclusion is obtained by reduction to a subexperiment. The point is that the binary subexperiment $y = t(x)$ contains very little "nuisance" information. From the point of view of hypothesis testing, a likelihood function completely free of nuisance information would be of the form

$$L(\vartheta) = \begin{cases} a & \text{for } \vartheta \in \Theta_0 \\ b & \text{for } \vartheta \notin \Theta_0 \end{cases} .$$

Usually, a likelihood function of this form can not be obtained (most likelihood functions are continuous), but the likelihood functions of a significance test are approximately of this form.

In the choice of significance test, all sorts of classical considerations may be of relevance. Power considerations are very important, because the power function determines the two possible likelihood functions. Obviously, a uniformely most powerful test should be preferred when it exists. If not so, the notion of an unbiased and locally most powerful test may serve as a guideline in the choice of test. Also invariance properties of tests should be taken into account.

A final remark: It has been assumed in this section that a significance test has to be a marginal binary experiment. But obviously, by the same arguments we may prefer to base inference on a binary statistic in a conditional experiment. Thus, also conditional significance testing comes out as a special sort of likelihood inference.

5. Conclusions.

In practice, the statistical analysis of a set of data (given the model) involves consideration of many different subexperiments. Many different conclusions are made about the parameters, and each of these conclusions is based on the subexperiment which is considered the most relevant for

this specific purpose. The thing, that makes the likelihood principle seem so very naive from a practical point of view, is the idea that this whole process of inference can be replaced by some considerations based on a single function, the likelihood function of the total experiment. But, as we have seen, this idea is not a necessary consequence of the likelihood principle. The likelihood principle does not reduce statistical inference to a triviality. Inference from a likelihood function is not easy, and to this comes the problem of choosing the best subexperiment for each of the many kinds of inference we want to make. It seems that very little can be said in general about the last problem. All we can say is that a subexperiment should be judged on the likelihood functions it can produce (or more precisely: on the distribution of its likelihood function). Reduction to a subexperiment is a way of avoiding nuisance information, for example nuisance parameters for which we have no prior distribution.

Thus, the conclusion of this paper is that the "likelihood paradox" (i.e. the inconsistency of classical inference methods with the likelihood principle) disappears when we introduce a careful distinction between an experiment and any of its subexperiments. We have to give up the idea that an experiment (in our technical use of this word) contains the information of any of its subexperiments.

References.

Barndorff-Nielsen, O. (1973): Exponential Families and Conditioning. Sc. D. thesis, University of Copenhagen.

Barndorff-Nielsen, O. (1976): Plausibility Inference. JRSS B 38 103-131.

Basu, D. (1964): Recovery of Ancillary Information. Sankhya A 26 3-16.

Basu, D. (1974): Statistical Information and Likelihood. Proceedings of conference on foundational questions in statistical inference, Aarhus May 7-12, 1973. Inst. of Math., University of Aarhus. Also published in Sankhya A 37 1-71.

Birnbaum, A. (1961): On the Foundations of Statistical Inference; Binary Experiments. Ann. Math. Stat. 32 414-35.

Birnbaum, A. (1962): On the Foundations of Statistical Inference. JASA 57 269-326.

Cox, D.R. (1975): Partial Likelihood. Biometrika 62, 2 269-276.

Cox, D.R. (1977): The Role of Significance Tests. Scand. J. Stat. 4 49-70.

Dempster, A.P. (1974): The Direct Use of Likelihood for Significance Testing. Proceedings of conference on foundational questions in statistical inference, Aarhus May 7-12, 1973. Inst. of Math., University of Aarhus.

Dickey, J.M. (1976): Approximate Posterior Distributions. JASA 71 680-689.

Edwards, A.W.F. (1972): Likelihood. Cambridge University Press.

Fisher, R.A. (1956): Statistical Methods and Scientific Inference. Hafner.

Fraser, D.A.S. (1968): The Structure of Inference. Wiley.

Kalbfleisch, J.D. and Sprott, D.A. (1970): Application of Likelihood Methods to Models Involving Large Numbers of Parameters. JRSS B 32 175-208.

Martin-Löf, P. (1974): Exact Tests, Confidence Regions and Estimates. Proceedings of conference on foundational questions in statistical inference, Aarhus May 7-12, 1973. Inst. of Math., University of Aarhus.

Savage, L.J. (1960): The Foundations of Statistical Inference. London, Methuen.

PREPRINTS 1977

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Asmussen, Søren & Keiding, Niels: Martingale Central Limit Theorems and Asymptotic Estimation Theory for Multitype Branching Processes.

No. 2 Jacobsen, Martin: Stochastic Processes with Stationary Increments in Time and Space.

No. 3 Johansen, Søren: Product Integrals and Markov Processes.

No. 4 Keiding, Niels & Lauritzen, Steffen L. : Maximum likelihood estimation of the offspring mean in a simple branching process.

No. 5 Hering, Heinrich: Multitype Branching Diffusions.

No. 6 Aalen, Odd & Johansen, Søren: An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations.

No. 7 Johansen, Søren: The Product Limit Estimator as Maximum Likelihood Estimator.

No. 8 Aalen, Odd & Keiding, Niels & Thormann, Jens: Interaction Between Life History Events.

No. 9 Asmussen, Søren & Kurtz, Thomas G.: Necessary and Sufficient Conditions for Complete Convergence in the Law of Large Numbers.

No. 10 Dion, Jean-Pierre & Keiding, Niels: Statistical Inference in Branching Processes.

PREPRINTS 1978

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Tjur, Tue: Statistical Inference under the Likelihood Principle.