To be submitted to *Bernoulli*

Sufficiency and conditional Monte Carlo

BO H. LINDQVIST and GUNNAR TARALDSEN*

Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. E-mail: bo@math.ntnu.no

**SINTEF Information and Communication Technology (ICT) O.S.Bragstads plass, N-7465 Trondheim, Norway. E-mail: gunnar.taraldsen@sintef.no*

Engen & Lillegård (1997) presented a most interesting approach for doing Monte Carlo simulations conditioned on a sufficient statistic. It turns out that one of their main claims is incorrect due to the ignorance of a Borel paradox when conditioning on a zero set. Lindqvist & Taraldsen (2004) present ways to mend the problem and suggest how the modified claims can be proven. Measure theoretic proofs of these claims are given here. The modified claims are also presented in the form of three simulation algorithms. The formulas obtained are similar in form to the ones presented by Trotter & Tukey (1956) who introduced the term *conditional Monte Carlo*, but the methods are otherwise rather different.

Keywords: conditional distribution; Monte Carlo simulation; nuisance parameters; point estimation; sufficiency

1 Introduction

The concept of sufficiency is due to Fisher (1920) according to Savage (1976, p.453), Rao (1992, p.42), and Lehmann & Casella (1998, p.143). It is a part of the foundations of statistics through the sufficiency principle, and it is practically important with applications for example in construction of optimal estimators and nuisance parameter elimination (Halmos & Savage 1949, Welsh 1996, E.L.Lehmann 1997, Lehmann & Casella 1998). The required conditioning may however be difficult to implement in practical problems. The traditional conditional Monte Carlo method of Trotter & Tukey (1956) is one possible approach, but this is also sometimes difficult to implement. Section 2 below presents alternative related approaches in the form of algorithms. In contrast to the article by Trotter & Tukey (1956) the proofs behind these algorithms rely on the assumed sufficiency. The above background is first explained in some more detail.

1.1 Rao-Blackwellization

Of the many possible definitions of sufficiency the following is convenient in the context here: A statistic T is *sufficient* for a parameter θ compared to a statistic X if for all ϕ there exist a ψ such that $E^\theta(\phi(X)\,|\,T=t)=\psi(t)$. Definition 2 in Section 4 gives a precise definition. The crucial point is that the conditional expectation is not a function of the parameter θ. This definition of sufficiency is closely related to the definition given by Blackwell (1953, p.266) and is more general than the conventional since it is not assumed that T is a function of X.

An important consequence of this definition is that the conditional expectation of an estimator with respect to a sufficient statistic is again an estimator. It is actually a better or equally good estimator: The bias is unchanged, and the variance is less than or equal to the original variance. More generally, since Jensen's inequality holds for conditional expectations, the expectation of any convex loss function is decreased by the conditional expectation. This improved estimator is the Rao-Blackwellization of the original estimator, and the original proofs generalize to the setting here (Rao 1945, Blackwell 1947, Lehmann & Casella 1998).

Manuscript version of 4th October 2004

Lehmann & Scheffe (1950) introduced the notion of completeness. A statistic T is *complete* for a parameter θ if $E^\theta\{\psi(T)\} = 0$ for all θ implies $P^\theta(\psi \neq 0) = 0$ for all θ. Completeness implies that the Rao-Blackwellization of an unbiased estimator is the unique best estimator in the class of unbiased estimators From this it is not surprising that a complete sufficient statistic is automatically *minimal* in the sense that the generated σ-field is contained in the σ-field of any other sufficient statistic (Lehmann & Casella 1998, p.42). There exist different definitions of completeness, for instance given by consideration of different classes of functions ψ (Mattner 1993), and the concept is also important for non-parametric statistics (Bell, Blackwell & Breiman 1960).

The problem of existence of an unbiased estimator for a given estimand is treated nicely by Halmos (1946), and more recent references are given by Lehmann & Casella (1998, p.111). Unbiased estimation is however, even if possible, not always a reasonable approach as demonstrated by simple examples (Rao 1973, p.333). The main point, however, remains true quite generally: Given any reasonable estimator, such as empirical mean $\bar{X}$, variance S^2, or distribution F_n, or some slightly biased estimator, the estimator may be improved by conditioning on a sufficient statistic, and the methods presented here give alternative routes for the actual calculation.

1.2 Nuisance parameter elimination

Conditioning on a sufficient statistic for a nuisance parameter is a convenient and natural method for the elimination of the nuisance (Fraser 1956, Basu 1977, Reid 1995). In good cases, typically if the sufficient statistic is complete and a component of a complete sufficient statistic for the model under consideration, the resulting inference is optimal if the method of inference for the conditional model is optimal (E.L.Lehmann 1997, p.147). There are also cases where optimal inference is too much to hope for, such as in goodness-of-fit tests with a large class of alternative distributions. As a concrete example in Section 5 it is explained how exact p-values in a Kolmogorov-Smirnov type test for the gamma distribution can be obtained by conditioning on a sufficient statistic. The idea is simply to replace the more traditional estimate of the distribution under the null hypothesis via estimation of the parameters by the optimal unbiased estimate of the distribution at a finite number of points. The distribution of the corresponding unconditional test statistic may depend on the parameters, but exact p-values can be obtained from consideration of a corresponding conditional test. A related Kolmogorov-Smirnov type test is described by Kumar & Pathak (1977), but the above observation regarding exact p-values is perhaps a novelty here. This idea may be transferred to other tests, which depend on the use of a given distribution function, such as the χ^2 test. A related, but more restricted approach, is to use Monte Carlo simulation to estimate the power and the p-value of *any* given test statistic (E.L.Lehmann 1997, p.151). Again, by considering the conditional test, it is possible to arrive at exact p-values for the resulting unconditional test. The actual calculations can be done with the sufficient conditional Monte Carlo method.

1.3 Conditional Monte Carlo

The most straightforward way of computing a conditional expectation $E^\theta(\phi(X)\,|\,T = t)$ would be to compute for a fixed parameter value θ_0. If the distribution of T given θ_0 dominates the family of distribution of T, then this gives a reduction to the problem of calculating conditional expectations in a purely probabilistic setting without parameters.

For this case Trotter & Tukey (1956) introduced the term *conditional Monte Carlo* (Hammersley 1956, Wendel 1957, Dubi & Horowitz 1979, Granovsky 1981)(Ripley 1987, p.136) (Evans & Swartz 2000, p.224). The idea is to determine a weight $w_t(X)$ and a modified sample $X_t = \chi(X,t)$ such that $E(\phi(X)\,|\,\tau(X) = t) = E\{\phi(X_t)w_t(X)\}$ for any function ϕ, where it is assumed that $T = \tau(X)$. This is an important simplification since a conditional expectation is replaced by an ordinary expectation which is suited for Monte Carlo computation. A given sample $x_1, \ldots, x_N$

can be used to get weighted conditional samples for all values of t. The modified sample fulfills the condition $\tau(X_t) = t$, and is obtained through adjustment of an artificially introduced parameter.

The traditional conditional Monte Carlo can be explained in terms of importance sampling and a change of variables as explained by Dubi & Horowitz (1979), or more directly connected to the original Trotter-Tukey article in group-theoretic language involving Haar measure and a homogeneous condition as explained by Wendel (1957). The first approach is more general and most straightforward and is the survivor in textbooks (Ripley 1987, Evans & Swartz 2000). Both approaches are related to the results here and involves in particular a similar adjustment of a parameter and an arbitrary distribution for this parameter. The skullduggery related to the introduction of an arbitrary new measure is most nicely explained by Trotter & Tukey (1956), and will not be repeated here.

In addition to all of the above ingredients the sufficient conditional Monte Carlo relies on the additional assumption of sufficiency, and this assumption is fulfilled in many interesting cases. Sufficiency is indeed the main new ingredient introduced here in relation to conditional Monte Carlo. This ingredient was also the starting point in the article by Engen & Lillegård (1997), but their main result is unfortunately not correct (Lindqvist, Taraldsen, Lillegård & Engen 2003). Lindqvist & Taraldsen (2004) sketch correct arguments based on Bayes theorem which lead to new versions of the original claims. Here proofs based on measure theory replace these arguments.

1.4 Outline

The plan of the paper is as follows. Section 2 presents three Algorithms for the simulation of samples from the conditional distribution given the sufficient statistic. Simple examples illustrate the Algorithms. The suggested methods for computation of conditional expectations are based on a choice of an arbitrary σ-finite measure on the parameter space. This leads to the need for extending the definition of conditional expectation to the case of σ-finite measures, and this is done in Section 3. This material could be of independent interest in a Bayesian setting. Section 4 contains statements and proofs of the main theoretical results. Section 5 relies on the presented Algorithms for the analysis of a pressure vessel example involving the gamma distribution.

2 Sufficient conditional Monte Carlo

The intended purpose of this Section is to present simplified versions of the main results in a form which is well suited for the implementation of the methods for practical calculations. The simulation methods are presented in the form of algorithms adapting the format used by Ripley (1987, Chapter 3). A statistic is in particular a function of the parameter θ and a statistic U with a known distribution. Throughout this Section the statistic $T = \tau(U, \theta)$ is assumed to be sufficient for the parameter θ compared to the statistic $X = \chi(U, \theta)$. Precise assumptions and results are presented in Section 4.

The most convenient case for simulation is when Algorithm 1 can be used. A sufficient and necessary condition for Algorithm 1 to give samples X_t from the conditional distribution of X given $T = t$ is that X_t is independent of T as explained in Theorem 1 in Section 4. This can in principle be tested in each case by simulation, but is not a recommended approach. Theorem 1 give sufficient and necessary conditions related to the Basu theorem (Basu 1955, Basu 1958, Lehmann & Casella 1998). It was claimed by Engen & Lillegård (1997, p.237) that sufficiency and a unique solution for θ of the equation $\tau(U, \theta) = t$ implies that Algorithm 1 gives samples from the conditional distribution. Unfortunately this is wrong as shown by an example by Lindqvist et al. (2003).

ALGORITHM 1

Let $X = \chi(U,\theta)$ and $T = \tau(U,\theta)$.

1. Generate U.
2. Solve $\tau(U,\theta) = t$ for θ. The solution is $\hat{\theta}(U,t)$.
3. Return $X_t = \chi\{U,\hat{\theta}(U,t)\}$.

Theorem 2 proves that a sufficient additional assumption is given by a certain pivotal structure. The basic condition is that $\tau(u,\theta)$ depends on u only through a function $r(u)$, where the value of $r(u)$ can be uniquely recovered from the equation $\tau(u,\theta) = t$ for given θ and t. This means that there is a function $\tilde{\tau}$ such that $\tau(u,\theta) = \tilde{\tau}(r(u),\theta)$ for all (u,θ), and a function $\tilde{v}$ such that $\tilde{\tau}(r(u),\theta) = t$ implies $r(u) = \tilde{v}(\theta,t)$. In this case $\tilde{v}(\theta,T)$ is a pivotal quantity in the classical meaning, and is moreover invertible according to the definition by Tukey (1957). The basic idea of the proof is that both $\tau(u,\theta)$ and $\hat{\theta}(u,t)$ are in one-to-one correspondence with $r(u)$ as explained in Lindqvist & Taraldsen (2004). This is stated and proved more precisely in Section 4. Here this will be explained by a familiar example. The point is mainly to illustrate Algorithm 1 and the conditions of Theorem 1 and Theorem 2.

Example 1 (Uniform distribution) Let $Y_1,\dots,Y_n$ be independent samples from the uniform distribution on $(0,\theta)$. The statistic $T = \max_i Y_i$ is complete and sufficient for the parameter θ compared to the statistic $X = Y_1$.

Let $U_1,\dots,U_n$ be independent samples from the uniform distribution on $(0,1)$. For the purpose here it may be assumed that $X = \theta U_1$ and $T = \theta \max_i U_i$.

With notation as in Algorithm 1 this gives $\chi(u,\theta) = \theta u_1$, $\tau(u,\theta) = \theta \max_i u_i$, $\hat{\theta}(u,t) = t/(\max_i u_i)$, and

$$X_t = \chi\{U,\hat{\theta}(U,t)\} = t\frac{U_1}{\max_i U_i} \tag{1}$$

The statistic T is sufficient for θ compared to the ancillary statistic $X_t = tY_1/\max_i Y_i$ as required in Theorem 1. The statistic X_t has then the conditional distribution of X given $T = t$.

The pivotal condition is also satisfied in this case with $r(U) = \max_i U_i$. The pivotal quantity is given by $\tilde{v}(\theta,T) = T/\theta$. Since θ is a scale parameter for T it follows also directly that T/θ is pivotal. Statistical inference regarding θ can be based on the pivotal quantity (Casella & Berger 1990, p.405).

□

A more common case is that there is a unique solution $\hat{\theta}(U,t)$, but the remaining conditions of Theorem 1 are not satisfied. Algorithm 2 is applicable for this case. Algorithm 2 is similar to Algorithm 1, but sampling from the distribution of U is replaced by weighted sampling in the spirit of Trotter & Tukey (1956), where in both cases the weight depends on the choice of the distribution of the variable Θ. If $\tau(u,\theta) = t$ is uniquely solvable for θ, then X_t given by Algorithm 2 is a sample from the conditional distribution of X given $T = t$. A more precise statement is found in Theorem 4. Algorithm 2 is also a consequence of equation (9) of Lindqvist & Taraldsen (2004).

In many cases it is not necessary to actually sample V as described in Algorithm 2. If, as in the numerical example in Section 5, the problem is to compute many conditional expectation for a fixed condition $T = t$, then this can be done with a weighted sample: The pair $\{\chi\{U,\hat{\theta}(U,t)\}, w_t(U)\}$ is

ALGORITHM 2

Let $X = \chi(U,\theta)$, $T = \tau(U,\theta)$, and let $t \mapsto w_t(u)$ be the density of $\tau(u,\Theta)$.

1. Generate V from a density proportional to w_t times the density of U.
2. Solve $\tau(V,\theta) = t$ for θ. The solution is $\hat{\theta}(V,t)$.
3. Return $X_t = \chi(V,\hat{\theta}(V,t))$.

a weighted sample from the conditional distribution in the terminology of Trotter & Tukey (1956). Algorithm 2 is illustrated by the example given in Section 5

The calculation of an expectation with respect to a weight function is a thoroughly studied problem, and many improved methods exist (Ripley 1987, Evans & Swartz 2000). Samples from the conditional distribution may be obtained by rejection sampling provided an envelope for the function $w_t(u)$ times the density of U is available. Alternatively, one may find it more convenient to use the ratio of uniforms method (Ripley 1987), Markov Chain Monte Carlo methods (Tierney 1994) or the SIR-algorithm of Rubin (Tanner 1996).

The general sufficient conditional Monte Carlo method is given by Algorithm 3 and includes cases where the assumption of a unique solution $\hat{\theta}(U,t)$ is dropped. A more precise statement

ALGORITHM 3

Let $X = \chi(U,\theta)$, $T = \tau(U,\theta)$, and let $t \mapsto w_t(u)$ be the density of $\tau(u,\Theta)$.

1. Generate V from a density proportional to w_t times the density of U.
2. Generate Θ_t from the conditional distribution of Θ given $\tau(v,\Theta) = t$, where $V = v$ from the above.
3. Return $X_t = \chi(V,\Theta_t)$.

of conditions that imply that X_t has the conditional distribution is found in Theorem 4. Generally speaking Step 2 is comparable with the initial conditional problem, but the two preceding Algorithms should have convinced the reader that this switch from U to Θ sometimes leads to simplifications. An example with several roots demonstrates this further.

Example 2 (Non-unique solution for θ) Let

$$f(x,\theta) = \begin{cases} \frac{\theta}{a^\theta + b^\theta}\, x^{\theta-1} & \text{if } 0 < x < a \\ \frac{\theta}{a^\theta + b^\theta}\, (x-a)^{\theta-1} & \text{if } a < x < a+b \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $a, b > 0$ are given constants. If $X_1, \ldots, X_n$ are independent samples from this distribution, then

$$T = \sum_{X_i < a} \log X_i + \sum_{X_i > a} \log(X_i - a) \tag{3}$$

is sufficient for the parameter θ compared to $X = (X_1, \ldots, X_n)$.

For given values of θ the statistic X can be simulated from the density $f(x,\theta)$ by ordinary inversion. This gives

$$X_i = \chi_i(U,\theta) = \begin{cases} \{U_i(a^\theta + b^\theta)\}^{1/\theta} & \text{if } U_i < a^\theta/(a^\theta + b^\theta) \\ (U_i(a^\theta + b^\theta) - a^\theta)^{1/\theta} + a & \text{if } U_i > a^\theta/(a^\theta + b^\theta) \end{cases} \tag{4}$$

where $U_1, \ldots, U_n$ are independent samples from the uniform distribution on $(0,1)$.

Now consider the equation $\tau(u,\theta) = t$. This equation has a finite but varying number of solutions for θ, depending on the values of u and t. Let $\Gamma(u,t) = \{\theta \,|\, \tau(u,\theta) = t\}$. As an illustration, when $n = 2, a = 3, b = 1$ and $u = (0.5, 0.9)$, there are two solutions for θ if $t \leq 0.2$, one solution if $t \geq 1.9$ and no solution if $0.2 \leq t \leq 1.9$ (approximate values). For example, $t = 0$ gives the solutions $\theta_1 = 0.925$, $\theta_2 = 1.801$.

The density $t \mapsto w_t(u)$ of $\tau(u,\Theta)$ with respect to Lebesgue measure is

$$w_t(u) = \sum_{\theta \in \Gamma(u,t)} \frac{g(\theta)}{|\det \partial_\theta \tau(u,\theta)|} \tag{5}$$

where g is the density of Θ. The conditional distribution of Θ given $\tau(u,\Theta) = t$ is concentrated on $\Gamma(u,t)$ and given by the discrete distribution

$$\frac{g(\theta)}{|\det \partial_\theta \tau(u,\theta)| w_t(u)}, \quad \theta \in \Gamma(u,t) \tag{6}$$

□

In the preceding example a difficult conditioning on $\tau(U,\theta)$ was effectively replaced by a more tractable conditioning on $\tau(u,\Theta)$. It is more tractable since the conditional distribution of Θ is discrete. For discrete distributions the solution set $\Gamma(u,t) = \{\theta \,|\, \tau(u,\theta) = t\}$ is typically a continuum, and Algorithm 3 may prove useful also in this case. Depending on the problem, other kinds of simplifications may be possible.

3 Conditional expectations for σ-finite measures

The suggested method for computation of conditional expectations is based on a choice of a σ-finite measure on the parameter space. This leads to the need for extending the definition of conditional expectation to the case of σ-finite measures, and corresponds to the use of improper priors in Bayesian analysis (Jeffreys 1946, Schervish 1995, Berger 1985). Two traditional alternative ways to make the concept of improper priors mathematically precise is given by: (i) Allowing infinite probability measures (Hartigan 1983). (ii) Allowing finitely additive probability measures (Heath & Sudderth 1989). The approach here is a special case of alternative (i). It is surprising that a presentation of the following basic definitions and results in the literature is difficult to find. The present section provides this foundation briefly, and is an extension of the corresponding theory for finite measures presented by Halmos & Savage (1949).

If $T : \mathcal{X} \to \mathcal{Y}$ is a measurable function and μ is a measure on $\mathcal{X}$, then the distribution μ_T of T is the measure on $\mathcal{Y}$ defined by $\mu_T(A) := \mu(T \in A)$. The notation $(T \in A) := \{x \,|\, T(x) \in A\}$ is used as an alternative to the more standard notation $T^{-1}(A)$. The set of locally μ-integrable functions $L_{1,\mathrm{loc}}(\mu)$ is defined as the equivalence class of functions $X : \mathcal{X} \to \mathbb{R}$ which are μ-integrable on sets with finite μ-measure. In the following other standard concepts from measure theory (Rudin 1987) and probability and statistics (Schervish 1995) will be used without further comments.

Definition 1 *Let $T : \mathcal{X} \to \mathcal{Y}$ be a measurable function. Let μ be a measure on $\mathcal{X}$ and assume that μ_T is σ-finite. The conditional expectation $\mu(X \,|\, T = t)$ of a locally μ-integrable function X is a measurable function of t such that*

$$\mu(Xg(T)) = \int \mu(X \,|\, T = t) g(t) \, \mu_T(dt) \tag{7}$$

for all indicator functions $g \in L_{1,loc}(\mu_T)$.

The measure determined by $g \mapsto \mu(Xg(T))$ is absolutely continuous with respect to μ_T, and the Radon-Nikodym theorem ensures existence and uniqueness of the conditional expectation identified as the density with respect to μ_T. The conditional expectation $\mu(X|T=t)$ defined in this way is determined as an element of $L_{1,\mathrm{loc}}(\mu_T)$. The different functions $\mu(X|T=t)$ satisfying (7) are called versions of the conditional expectations.

The assumption of σ-finite distribution for T ensures that the family of indicator functions of sets with finite μ_T-measure is non-trivial and furthermore makes the Radon-Nikodym theorem available (Rudin 1987, p.121-124). The assumption means in particular that T cannot be a constant unless the measure μ is finite. Equation (7) could be used as a definition in this special case where $\mu(X) = \infty$ also, but the resulting conditional expectation would then be a completely arbitrary measurable function since $g = 0$ is the only available indicator function in $L_{1,\mathrm{loc}}(\mu_T)$. In the converse direction it can be observed that σ-finiteness of μ follows from the σ-finiteness of μ_T.

The modern definition of conditional expectation $\mu(X\,|\,\mathcal{F})$ (Schervish 1995, p.616, B.23) with respect to a σ-field $\mathcal{F}$ in $\mathcal{X}$ may be defined in a similar manner by the replacement of the above $g(t)$'s with indicator functions for the sets in $\mathcal{F}$, replacing μ_T with μ, and the demand that $\mu(X\,|\,\mathcal{F})$ is measurable with respect to $\mathcal{F}$. Existence of the conditional expectation follows from the Radon-Nikodym theorem in this case also. With additional assumptions the two definitions of the conditional expectation are connected by consideration of the σ-field $\mathcal{F}$ generated by T (Schervish 1995, p.616, B.24). In the context here it is convenient to follow Halmos & Savage (1949) and take Definition 1 as the starting point.

The existence and uniqueness of the conditional expectation in Definition 1 can be interpreted as a form of Bayes theorem. In a Bayesian setting a model is often formulated in terms of a prior distribution for a parameter Θ and a conditional distribution for the observation X given Θ. This gives the joint distribution $\mu_{(X,\Theta)}$ of (X,Θ). If the observation X has a σ-finite distribution μ_X in the model, then Definition 1 gives the conditional expectation of $\phi(\Theta) \in L_{1,\mathrm{loc}}(\mu)$ given $X = x$ on which Bayesian inference can be based.

A salient feature of the conditional expectation is that it is normalized even when the initial distribution has infinite mass.

Proposition 1 *Let $T : \mathcal{X} \to \mathcal{Y}$ be a measurable function and let μ be a measure on $\mathcal{X}$. If the distribution μ_T is σ-finite, then*

$$\mu(1\,|\,T=t) = 1 \tag{8}$$

Proof. If g is the indicator function of A with $\mu_T(A) < \infty$, then

$$\mu(1\,g(T)) = \int g(T(x))\,\mu(dx) = \mu\{x\,|\,T(x) \in A\} = \mu_T(A) = \int 1\,g(t)\,\mu_T(dt) \tag{9}$$

which proves the claim. □

Proposition 1 is a generalization of the same observation for finite measures (Halmos & Savage 1949, p.230). Hartigan (1983, p.24-30) allows infinite conditional probabilities, but in the statement of Bayes theorem a σ-finiteness condition is used and the resulting conditional probability is normalized.

Many results from the theory of conditional expectation generalize verbatim to the more general case of σ-finite measures. Three results are essential in the following and are listed in the following proposition.

Proposition 2 *Let $T : \mathcal{X} \to \mathcal{Y}$ be a measurable function. Let μ be a measure on $\mathcal{X}$ and assume that μ_T is σ-finite. Let $X \in L_{1,loc}(\mu)$.*

If h is one-to-one and measurable, then

$$\mu(X|T=t) = \mu(X|h(T)=h(t)) \tag{10}$$

If $\mathcal{X}$ is a Borel space, then there exist a regular conditional distribution $\mu(\cdot|T=t)$. Furthermore, if $f : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is measurable and $f(Z,T) \in L_{1,loc}(\mu)$, then

$$\mu(f(Z,T)|T=t) = \int f(Z(x),t)\,\mu(dx|T=t) \tag{11}$$

Proof. The first claim is left to the reader. The existence of a regular conditional distribution follows from the proof given by Schervish (1995, p.618). The last claim follows from the proof of a corresponding result of Bahadur & Bickel (1968). □

The class of Borel spaces is large and includes in particular every complete separable metric space. A regular conditional distribution $A \mapsto \mu(A|T=t)$ is a measure on $\mathcal{X}$ for each t where $\mu(1_A|T=t) = \mu(A|T=t)$. The right hand side of (11) can be considered to define a version of $\mu(f(Z,t)|T=t)$, so (11) can be viewed as a result on substitution in conditional expectations.

4 Sufficiency and Monte Carlo conditioning

Let $\{P^\theta\}$ be a family of σ-finite measures on the measurable space Ω. The corresponding linear functionals are in this case denoted by E^θ. A statistic T is a measurable function from Ω into a measurable space Ω_T. The statistic T is ancillary if its distribution does not depend on θ: $P_T^\theta = P_T$.

An alternative definition for a statistic will also be allowed here. Let P be a σ-finite distribution on the measurable space Ω. An ancillary statistic U is a measurable function from Ω into a measurable space Ω_U. A statistic $T = \tau(U,\theta)$ is specified by a measurable function τ and an ancillary statistic U. It follows that if the distribution P_T^θ does not depend on θ then T can be identified with an ancillary statistic.

In the first definition the measurable space Ω and the family of measures $\{P^\theta\}$ is fixed, and every statistic is based on this underlying structure. In the alternative definition the measure space (Ω, P) is fixed, and every statistic is based on this underlying structure. In both cases the result is a family of distributions $\{P_T^\theta\}$ for each statistic T. The alternative definition is convenient in the context here since it has as a consequence that the family of distributions is measurable: $\theta \mapsto P_T^\theta(A) = E(\tau(U,\theta) \in A)$ is measurable. This claim follows from the Fubini theorem since P is σ-finite (Rudin 1987, p.164).

The following two results also show a technical advantage of the alternative definition of a statistic. The first result is a generalization of the Radon-Nikodym theorem to measurable families of measures.

Proposition 3 *Let $\pi(d\theta)$ be a σ-finite measure. If $\{\nu^\theta\}$ is a measurable family of measures dominated by a σ-finite measure γ, then there exists a measurable f such that $\nu^\theta(dx) = f(x,\theta)\gamma(dx)$ for almost all θ. More generally, if $\nu^\theta \ll \mu^\theta$ for almost all θ, where $\{\mu^\theta\}$ is a measurable family of σ-finite measures, then there exist a measurable f such that $\nu^\theta(dx) = f(x,\theta)\,\mu^\theta(dx)$ for almost all θ.*

Proof. Define the measure ν by

$$\int h(x,\theta)\,\nu(dx,d\theta) = \int h(x,\theta)\,\mu^\theta(dx)\pi(d\theta) \tag{12}$$

and likewise for μ. The Radon-Nikodym theorem and $\nu \ll \mu$ gives a measurable f such that $\nu(dx,d\theta) = f(x,\theta)\,\mu(dx,d\theta)$. Substitution of functions on the form $h(x)g(\theta)$ in (12) shows that $\nu^\theta(dx) = f(x,\theta)\,\mu^\theta(dx)$ holds for almost all θ with respect to π. □

The following result ensures joint measurability of a conditional expectation as a function of the parameter and the statistic it is conditioned on.

Proposition 4 *Let $X = \chi(U,\theta)$ and $T = \tau(U,\theta)$ be statistics with measurable χ, τ and an ancillary statistic U. Let $\pi(d\theta)$ be a σ-finite measure. Assume that P_T^θ is σ-finite for almost all θ, and that $X \in L_{1,loc}(P)$ for almost all θ. Then, for almost all θ, there exists a version of the conditonal expectation $E^\theta(X\,|\,T=t)$ such that $(\theta,t) \mapsto E^\theta(X\,|\,T=t)$ is measurable.*

Proof. Define the measure ν^θ by $\nu^\theta(g) = E^\theta(Xg(T))$. The conditional expectation $E^\theta(X\,|\,T=t)$ is the density of ν^θ with respect to P_T^θ and Proposition 3 gives the claim. □

It can be remarked that in a Bayesian setting, where $P^\theta(dx\,|\,T=t) = P(dx\,|\,T=t,\Theta=\theta)$, the previous joint measurability is automatically fulfilled.

Halmos & Savage (1949) introduced the concept of dominated families of measures in their proof of the factorization theorem for sufficient statistics. The following result is needed below.

Proposition 5 *A dominated family of σ-finite measures has an equivalent countable subfamily, and there exists a σ-finite measure equivalent to the family.*

Proof. This follows from the results of Halmos & Savage (1949, p.232–233) since a σ-finite measure is equivalent to a finite measure. □

The concept of sufficiency is one of the cornerstones of statistics. Halmos & Savage (1949, p.239–241) give a good explanation for what a sufficient statistic is sufficient for doing. This explanation is relevant also for the definition given here. There exist several definitions, and the following is convenient for the purposes here.

Definition 2 *A statistic T is sufficient for a parameter θ compared to a statistic X if to each $f \in \bigcap_\theta L_{1,loc}(P_X^\theta)$ there exists a measurable g such that $E^\theta(f(X)\,|\,T=t) = g(t)$ for all θ.*

This means that g is a version of the conditional expectation for each θ, and the notation $E(f(X)\,|\,T=t) = g(t)$ denotes this particular version of the conditional expectation. Definition 2 is similar to the general sufficiency concept of Blackwell (1953), and includes the more standard concept where it is assumed in addition that T is a function of X (Halmos & Savage 1949) (E.L.Lehmann 1997, p.19) (Schervish 1995, p.84). If Ω_X is a Borel space, then T is sufficient for a parameter θ compared to X if the regular conditional distribution of X given T does not depend on θ. The weaker concept of pairwise sufficiency (Halmos & Savage 1949) can be generalized in a similar manner.

The Basu theorem (Basu 1955, Basu 1958, Lehmann & Casella 1998), including its proof, holds also with the definition of a statistic given above. The statement and proof is included here since the result will be shown to be closely linked to Algorithm 1.

Proposition 6 *If T is a complete sufficient statistic for the parameter θ compared to an ancillary statistic A, then T and A are independent.*

Proof. Let ϕ be bounded and measurable. From sufficiency compared to A define $\psi(t) = E^\theta(\phi(A)\,|\,T=t)$. It will be proven that $\psi(t) = E^\theta\phi(A)$. The calculation $E^\theta\phi(A) = E^\theta\{E^\theta(\phi(A)\,|\,T)\} = E^\theta\{\psi(T)\}$ gives $E^\theta\{\psi(T) - E^\theta\phi(A)\}$ for all θ. Completeness gives the claim $\psi(t) = E^\theta\phi(A)$, since

$\psi(T) - E^\theta\phi(A)$ is a function of T only. This holds for the first term due to sufficiency and for the last term since A is ancillary. □

As in the conventional setting with normalized distributions a complete sufficient statistic also gives the uniformly minimum variance unbiased estimator by conditioning of an unbiased estimator. The proof is unchanged and is not repeated here.

The next result corresponds to Algorithm 1 in the case of a complete sufficient statistic. The Theorem gives sufficient and necessary conditions such that the resulting X_t is distributed like X given $T = t$. This result seems to be new and is in particular not contained in the related paper by Lindqvist & Taraldsen (2004).

Theorem 1 *Assume $T = \tau(U,\theta)$ to be a complete sufficient statistic for the parameter θ compared to $X = \chi(U,\theta)$ with measurable χ, τ and an ancillary statistic U where Ω_U is a Borel space. Assume there is a unique solution $\hat\theta(u,t)$ of the equation $\tau(u,\theta) = t$ for each pair (u,t).*

The variable $X_t = \chi(U,\hat\theta(U,t))$ is distributed like X given $T = t$ if and only if T is a sufficient statistic for the parameter θ compared to X_t.

Proof. Let ϕ be bounded and measurable. The identity $X = \chi[U,\hat\theta\{U,\tau(U,\theta)\}]$ and equation (11) in Proposition 2 gives

$$E(\phi(X)\,|\,T = t) = E^\theta(\phi(X_t)\,|\,T = t) \tag{13}$$

since Ω_U is a Borel space.

Assume first that T is a sufficient statistic compared to X_t for the parameter θ. Proposition 6 gives independence of X_t and T, so $E^\theta(\phi(X_t)\,|\,T) = E^\theta\{\phi(X_t)\}$. The conditioning on the right hand side of equation (13) can hence be removed and X_t has the conditional distribution.

Assume next that the variable $X_t = \chi(U,\hat\theta(U,t))$ is distributed like X given $T = t$. Equation (13) gives independence of X_t and T. It must be proven that $E^\theta(\phi(X_t)\,|\,T = s) = \psi_t(s)$ with no dependence on θ. The independence gives a trivial s dependence, and the distribution assumption gives $E^\theta(\phi(X_t)\,|\,T = s) = E(\phi(X)\,|\,T = t)$ which proves sufficiency. □

The next result also corresponds to Algorithm 1, but in this case completeness is not assumed. The conditions are more constructive and imply that the resulting X_t is distributed like X given $T = t$. The Theorem corresponds to the pivotal case as explained by Lindqvist & Taraldsen (2004).

Theorem 2 *Assume $T = \tau(U,\theta)$ to be a sufficient statistic for the parameter θ compared to $X = \chi(U,\theta)$ with measurable χ, τ and an ancillary statistic U where Ω_U is a Borel space. Assume in addition that the three following conditions are satisfied.*

- **Unique solution:** *There is a unique solution $\hat\theta(u,t)$ of the equation $\tau(u,\theta) = t$ for each pair (u,t).*
- **The pivotal condition:** *There exist measurable functions r and $\tilde\tau$ with $\tau(u,\theta) = \tilde\tau(r(u),\theta)$, and such that there is a unique solution $\tilde v(\theta,t)$ of the equation $\tilde\tau(v,\theta) = t$ for each pair (θ,t).*
- **Dominating measures:** *There exist two σ-finite measures ν,γ which dominate respectively the family of distribution of $\tau(U,\theta)$ and the family of distribution of $\hat\theta(U,t)$.*

The variable $\tilde v(\theta,T)$ is pivotal. For ν-almost all t the variable $X_t = \chi(U,\hat\theta(U,t))$ is independent of $\{r(U),\hat\theta(U,t)\}$ and X_t is distributed like X given $T = t$.

Proof. The variable $\tilde{v}(\theta,T) = \tilde{v}(\theta,\tau(U,\theta)) = r(U)$ is pivotal. Because of Proposition 5 it can be assumed that ν and γ are equivalent to the families they dominate. Let $\psi(t) = E(\phi(X)\,|\,T=t)$ where ϕ is bounded and measurable. From the pivotal assumption there is a one-to-one and measurable correspondence between $r(U)$ and $\tilde{\tau}(r(U),\theta)$ for each θ. Proposition 2 gives

$$\psi(t) = E(\phi(X)\,|\,\tilde{\tau}(r(U),\theta)=t) = E\{\phi(X)\,|\,r(U)=\tilde{v}(\theta,t)\} \tag{14}$$

for all θ and all $t \in A$ where $\nu(A^c)=0$. Proposition 4 gives joint measurability of the right hand side with respect to (θ,t). The assumptions give that $\hat{\theta}(u,t)$ depends on u only through $r(u)$, so let $\hat{\theta}(u,t) = \tilde{\theta}(r(u),t)$. Consider the calculation

$$E\{\phi(X)\,|\,r(U)=\tilde{v}(\theta,t)\} = E(\phi(X)\,|\,\tilde{\theta}(r(U),t)=\theta) = E(\phi(X_t)\,|\,\hat{\theta}(U,t)=\theta) \tag{15}$$

which holds for all t and all $\theta \in B$ where $\gamma(B^c)=0$. The first equality follows since for fixed t, $\tilde{\theta}(r(U),t)$ is a one-to-one and measurable function of $r(U)$. The second equality follows from $\hat{\theta}(u,t) = \tilde{\theta}(r(u),t)$ and substitution which is allowed since Ω_U is a Borel space. Equation (14) and equation (15) combine to give

$$\psi(t) = E(\phi(X_t)\,|\,\hat{\theta}(U,t)=\theta) = E\{\phi(X_t)\} \tag{16}$$

where the first equality holds for $(t,\theta) \in A \times B$, and the second equality holds since there is no dependence on θ in ψ. This proves thatX_t has the conditional distribution.

The independence result also follows from equation (16) and the given one-one correspondence between $\hat{\theta}(U,t)$ and $r(U)$. □

The variable $\tilde{v}(\theta,T)$ is a pivotal quantity in the classical meaning, and is moreover invertible according to the definition by Tukey (1957).

The assumptions in Theorem 2 are fulfilled in many interesting and important cases, including in particular the multinormal (Lindqvist & Taraldsen 2004). The following result is essential for the derivation of similar results under more relaxed conditions. It is a more precise and general version of equation (2) derived by Lindqvist & Taraldsen (2004).

Theorem 3 *Assume $T = \tau(U,\theta)$ to be a sufficient statistic for the parameter θ compared to $X = \chi(U,\theta)$ with measurable χ, τ and an ancillary statistic U. Let $\mu = P_U \otimes \pi$, where π is a σ-finite distribution on the set of parameters such that the family of distributions $\{P_T^\theta\}$ is dominated by μ_τ. If $\phi \in L_{1,loc}(\mu_\chi)$ and μ_τ is σ-finite, then $\mu(\phi(\chi)\,|\,\tau=t)$ is a version of $E(\phi(X)\,|\,T=t)$*

Proof. Let $\psi_1(t) = E(\phi(X)|\,T=t)$ and $\psi_2(t) = \mu(\phi(\chi)\,|\,\tau=t)$. The claim follows by proving $\mu\{\psi_1(\tau) \neq \psi_2(\tau)\} = 0$. Recall that by sufficiency $E(\phi(X)\,|\,T=t)$ is a version of $E(\phi\{\chi(U,\theta)\}\,|\,\tau(U,\theta)=t)$ for all θ. Let $g \in L_{1,\mathrm{loc}}(\mu_\tau)$ be an indicator function. The claim follows from

$$\begin{aligned}\mu\{g(\tau)\psi_2(\tau)\} = \mu\{g(\tau)\phi(\chi)\} &= \int E[g\{\tau(U,\theta)\}\phi\{\chi(U,\theta)\}]\,\pi(d\theta) \\ &= \int E[g\{\tau(U,\theta)\}\psi_1\{\tau(U,\theta)\}]\,\pi(d\theta) = \mu\{g(\tau)\psi_1(\tau)\}\end{aligned} \tag{17}$$

□

The point of the domination assumption is most easily illustrated by an example where $X = (X_1,\ldots,X_n)$ is a vector of n independent variables uniformly distributed on $(0,\theta)$, with $\theta > 0$. Then $T = \max X_i$ is sufficient compared to X. Now for a fixed value θ_0 the support of T is $[0,\theta_0]$. Thus conditional expectations of functions $\phi(X)$ given $T=t$ under P^{θ_0} can be assigned

an arbitrary value for $t > \theta_0$. The conditional expectation $\mu[\phi(\chi) \,|\, \tau = t]$ with $\pi = \delta_{\theta_0}$ can not be used as representative for $E[\phi(X) \,|\, T = t]$ in this case.

The next theorem gives the precise version of Algorithm 3, of which Algorithm 2 is a special case. Sufficient conditions are sufficiency as in Theorem 3 and existence of a distribution π for Θ such that the distributions of both $\tau(u,\Theta)$ and $\tau(U,\theta)$ are dominated by the distribution of $\tau(U,\Theta)$. The next Theorem gives assumptions that are slightly more general.

Theorem 4 *Assume $T = \tau(U,\theta)$ to be a sufficient statistic for the parameter θ compared to $X = \chi(U,\theta)$ with measurable χ, τ and an ancillary statistic U. Let $\mu = P_U \otimes \pi$, where π is a σ-finite distribution such that μ_τ is σ-finite and such that the family of distributions $\{P_T^\theta\}$ is dominated by μ_τ. Define $\Theta(\theta) = \theta$. Assume that the family of distribution of $\tau(u,\Theta)$ for almost all u is dominated by a σ-finite measure ν. Let $t \mapsto w_t(u)$ be the density of $\tau(u,\Theta)$ with respect to ν. Let $z_t(u) = \pi(\phi(\chi(u,\Theta)) \,|\, \tau(u,\Theta) = t)$ where $\phi \in L_{1,loc}(\mu_\chi)$. The functions $w_t(u)$ and $z_t(u)$ can then be chosen jointly measurable in (u,t), and*

$$E(\phi(X) \,|\, T = t) = \frac{E(w_t(U) z_t(U))}{E(w_t(U))} \tag{18}$$

Proof. The joint measurability of $w_t(u)$ and $z_t(u)$ follows from Proposition 3 and Proposition 4. A short calculation gives

$$\mu_\tau(dt) = \{E w_t(U)\} \nu(dt) \tag{19}$$

Let $g \in L_{1,\mathrm{loc}}(\mu_\tau)$ be an indicator function. The definition of $z_t(u)$ as a conditional expectation and the definition of $w_t(u)$ as a density give

$$\int \phi\{\chi(u,\theta)\} g\{\tau(u,\theta)\} \, \pi(d\theta) = \int z_t(u) g(t) w_t(u) \nu(dt) \tag{20}$$

Integrating both sides of (20) with respect to P_U and using Fubini's theorem give

$$\mu\{\phi(\chi) g(\tau)\} = \int \frac{E\{w_t(U) z_t(U)\}}{E w_t(U)} g(t) \, \mu_\tau(dt) \tag{21}$$

Equation (21) and Theorem 3 prove that (18) holds. □

From the proof it follows that the assumed existence of a dominating ν is equivalent with the assumption that the distribution of $\tau(u,\Theta)$ is dominated by the distribution of τ for almost all u. The measure ν can then be taken to be μ_τ. In this case $E w_t(U) = 1$ so $E(\phi(X) \,|\, T = t) = E(w_t z_t)$. Note also that $0 < E w_t(U) < \infty$ for almost all t, because of (19) and the assumption of μ_τ beeing $\sigma-$finite.

The next result applies typically to cases where T has a discrete distribution for all θ. The proposition is a corollary of Theorem 3, but an elementary direct proof is given here. The result is essentially the main result of Engen & Lillegård (1997), but the presented proof is different.

Theorem 5 *Assume $T = \tau(U,\theta)$ to be a sufficient statistic for the parameter θ compared to $X = \chi(U,\theta)$ with measurable χ, τ and an ancillary statistic U. Let π be a σ-finite measure, let ϕ be bounded and measurable and define*

$$y_t(u) = \int \phi(\chi(u,\theta)) \cdot [\tau(u,\theta) = t] \, \pi(d\theta), \quad w_t(u) = \int [\tau(u,\theta) = t] \, \pi(d\theta) \tag{22}$$

Then for all t with $0 < E(w_t(U)) < \infty$

$$E(\phi(X) \,|\, T = t) = \frac{E(y_t(U))}{E(w_t(U))} \tag{23}$$

Proof. The composed function $\phi(\chi)$ is not more general than χ alone, so in the proof it may be assumed that $\phi(x) = x$. By considering only θ's such that $P^\theta(T = t) > 0$ it follows that

$$E(X \,|\, T = t) = \frac{E(X[T = t])}{P^\theta(T = t)} = \frac{\int E(X[T = t])\, \pi(d\theta)}{\int E^\theta(T = t)\, \pi(d\theta)} \tag{24}$$

The above integration is valid since the sufficiency gives that the fraction is independent of θ. The conclusion follows from the Fubini theorem which allows us to change the order of integration since P_U and π are σ-finite. □

In some cases, there is a function $\tilde{x}_t(u)$ such that $\chi(u,\theta) = \tilde{x}_t(u)$ for all θ solving $\tau(u,\theta) = t$ (Engen & Lillegård 1997). This is more general than the assumptions defining X_t previously. In this more general case the $y_t(U)$ of Theorem 5 factors as $y_t(U) = \phi(\tilde{x}_t(U))w_t(U)$. Moreover, if $\tilde{x}_t(U)$ and $w_t(U)$ are independent, then $\tilde{x}_t(U)$ is distributed like the conditional distribution of X given $T = t$. A similar argument holds for Theorem 4 and gives a slight generalization of Algorithm 2.

5 A failure time example

Consider the following failure times (in hours)

274	28.5	1.7	20.8	871	363	1311	1661	236	828
458	290	54.9	175	1787	970	0.75	1278	776	126

(25)

of $n = 20$ similarly constructed pressure vessels subjected to constant pressure (Keating, Glaser & Ketchum 1990, Wong 1992, Welsh 1996). The sample empirical mean and variance are respectively $\bar{x} = 575.53$ and $s^2 = 3.3021{\cdot}10^5$. It will be assumed that the failure times are independent samples from the gamma density

$$f_X(x) = \{\beta^\alpha \Gamma(\alpha)\}^{-1}\, x^{\alpha-1} e^{-x/\beta}, \;\; \text{shape } \alpha > 0, \text{scale } \beta > 0 \tag{26}$$

With this assumption $\bar{X}$ is the best unbiased estimator for the mean $\mu = EX_i = \alpha\beta$, and it coincides with the maximum likelihood estimator. The empirical variance S^2 is however not an optimal estimator for the variance $\sigma^2 = \operatorname{Var} X_i = \alpha\beta^2$, but the best unbiased estimator is obtained by conditioning on a sufficient statistics. The actual calculation of this best unbiased estimator is not quite straightforward, and the authors have not been able to find references where this calculation is done neither numerically nor analytically. The method of sufficient conditional Monte Carlo permits a direct calculation which is also valid for small samples. In the following the best unbiased estimator for the variance will be calculated, and compared with the maximum likelihood estimator.

5.1 Estimation

The maximum-likelihood estimates of shape and scale are well known (Welsh 1996, p.190), and leads to the estimate $\hat{\sigma}^2 = \bar{x}^2/\hat{\alpha}$ of the variance, where $\hat{\alpha}$ solves $w := \tilde{x}/\bar{x} = \hat{\alpha}^{-1} \cdot \exp\psi(\hat{\alpha})$, $\tilde{x}$ is the geometrical mean, and $\psi(\alpha) = \partial_\alpha \log\Gamma(\alpha)$ is the digamma function. It should be observed that the maximum likelihood estimator of the shape parameter is in a one-one correspondence with the statistic $W := \tilde{X}/\bar{X}$. The exact distribution of W is known, and the distribution is sometimes referred to as the Bartlett distribution (Keating et al. 1990). In the following W will be referred to as the Bartlett statistic. For the given pressure vessel data $w = \tilde{x}/\bar{x} = 0.3422$ and $\hat{\sigma}^2 = 5.719{\cdot}10^5$, which is somewhat different from the empirical variance. For later reference note also the estimates $\hat{\alpha} = 0.5792$ and $\hat{\beta} = 993.7$.

Next, consider the calculation by the sufficient conditional Monte Carlo method. The statistic $(\bar{X},\tilde{X})$ is complete and sufficient. This is then also true for the statistic $T=(\bar{X},\tilde{X}/\bar{X})$, which has better properties in some respects. The Bartlett statistic $\tilde{X}/\bar{X}$ has no β dependence, and is hence ancillary with respect to β. The Basu theorem gives independence of $\{\bar{X},\tilde{X}/\bar{X}\}$, since $\bar{X}$ is complete and sufficient for β. The range of the statistic T is hence $(0,\infty)\times(0,1)$.

The best unbiased estimate of the variance is given by $\hat{\sigma}_n^2=E(S^2\,|\,\bar{X}=\bar{x},\tilde{X}/\bar{X}=\tilde{x}/\bar{x})$. A change of scale argument gives the decomposition $\hat{\sigma}_n^2=\bar{x}^2/\hat{\alpha}_n$, where $1/\hat{\alpha}_n=E(S^2\,|\,\bar{X}=1,\tilde{X}/\bar{X}=\tilde{x}/\bar{x})$. This estimator has the same dependence on $\bar{x}$ as the maximum-likelihood estimator, but $\hat{\alpha}_n$ depends on the sample size n in addition to the dependence on $\tilde{x}/\bar{x}$, and is hence more complicated than $\hat{\alpha}$. The aim here is to estimate the variance, but it can be noted that natural estimates of shape and scale follow from the one-one correspondence with (μ,σ). In this context it can be remarked further that from unbiasedness and the above independence $E(1/\hat{\alpha}_n(W))=1/(\alpha+1/n)$.

Let $\chi_i(u,\theta)=\beta F^{-1}(u_i;\alpha)$, where F is the CDF of the gamma distribution with scale $\beta=1$, and $\theta=(\alpha,\beta)$. Define further $\tau_1=\bar{\chi}$, and $\tau_2=\tilde{\chi}/\bar{\chi}$. If $U_1,\ldots,U_n\sim\mathbf{U}(0,1)$ are independent, then the inversion method gives $(\chi(U,\theta),\tau(U,\theta))\sim(X,T)$. This is then a model of the sufficient conditional Monte Carlo type. Define the function $\check{\theta}(u,t)$ as the solution of $\tau(u,\check{\theta})=t$. This is two real equations with two real unknowns, but the components are given by the solution of the single equation $\tilde{\chi}(u,\check{\alpha})/\bar{\chi}(u,\check{\alpha})=t_2$, and $\check{\beta}=t_1/\bar{\chi}(u,\check{\alpha})$. Here the convention $\chi(u,\alpha,1)=\chi(u,\alpha)$ is used. From this $\check{\chi}_i(u,t):=\chi_i(u,\check{\theta}(u,t))=t_1\chi_i(u,\check{\alpha}(u,t_2))/\bar{\chi}(u,\check{\alpha}(u,t_2))$. The density of $\tau(u,\Theta)$ is given by $w(t,u)=f_\Theta(\check{\alpha},\check{\beta})/[\bar{\chi}(u,\check{\alpha})\partial_\alpha(\tilde{\chi}/\bar{\chi})\,|_{\alpha=\check{\alpha}}]$.

The authors have not been able to prove that $\alpha\mapsto\tilde{\chi}/\bar{\chi}$ is strictly increasing. The conclusion in the general case is the representation

$$1/\hat{\alpha}_n=\frac{E\sum_{\alpha\in\Gamma}\frac{n}{n-1}\overline{(\chi/\bar{\chi}-1)^2}f_\Theta/[\bar{\chi}\partial_\alpha(\tilde{\chi}/\bar{\chi})]}{E\sum_{\alpha\in\Gamma}f_\Theta/[\bar{\chi}\partial_\alpha(\tilde{\chi}/\bar{\chi})]},\tag{27}$$

where $\Gamma=\Gamma(U,t_2)=\{\alpha\,|\,\tilde{\chi}(U,\alpha)/\bar{\chi}(U,\alpha)=t_2\}$. The simple indicator choice $f_\Theta(\alpha,\beta)=[\alpha>\alpha_{\min}]$ is one way to avoid numerical problems for small α. The value $\alpha_{\min}=0.01$ is used in the simulations, and the result is that in some rare cases this gives $\Gamma=\emptyset$, but not so frequently that this caused any problems. A possible improvement is to multiply this with the Jeffreys' prior $\sqrt{\alpha\psi'(\alpha)-1}/\beta$, and in the simulations this seemed to give an improvement.

Numerical routines in S-Plus, MATLAB with the Statistics tool-box, and Fortran 95 with the IMSL library have been implemented in order to evaluate the integrands in the two $n=20$ dimensional integrals in equation (27). One reason for the implementation on three independent platforms was initial numerical problems with the function $\chi(u,\alpha)$ for small α. In MATLAB the function gaminv() gives erroneous answers, and in the IMSL library the corresponding gamin() function sometimes returns without convergence. The solution was a modification of the MATLAB function, and a corresponding implementation in Fortran. The end result was three independent implementations, and it is reassuring that they produce consistent answers. The underlying pseudo-random generators are also different on the three platforms.

The integrals are approximated by the mean values of the integrands evaluated at $u(1),\ldots,u(m)$ obtained from the pseudo-random generator. Numerical experiments indicate that two digit accuracy is obtained with $m=10000$, and the estimate is then $\hat{\sigma}_{20}^2=5.5e5$. This is quite close to the maximum likelihood estimate, and indicates that $n=20$ is a "large sample" for this specific case, and hence that the maximum likelihood estimator is close to optimal.

It is not clear from the previous whether the UMVU estimate $\hat{\sigma}_{20}^2$ is an improvement relatively to the MLE estimate $\hat{\sigma}^2$. It is not the purpose here to settle this question, but to get an indication simulations with $m=100000$ Monte Carlo samples of size $n=20$ from the MATLAB pseudo-random generator *gamrnd*() with $\alpha=0.6$ and $\beta=1000$ have been carried out. For each

Table 1: Empirical mean and relative root-mean-square error of estimates of the variance in a **Gamma**$(0.6, 1000)$ distribution from 100000 pseudo-random samples of size 20 and 3. The exact variance is 6e5.

Estimator	Sample size	Mean	Relative root-mean-square error
Empirical variance	20	5.99e5	77.7%
Maximum likelihood	20	6.21e5	69.3%
Algorithm 2 (UMVU)	20	5.99e5	65.5%
Algorithm 1	20	6.02e5	65.9%
Empirical variance	3	5.99e5	208%
Maximum likelihood	3	6.53e5	269%
Algorithm 2 (UMVU)	3	6.00e5	200%
Algorithm 1	3	6.06e5	197%

Monte Carlo sample the MLE estimate, the empirical variance, and estimates corresponding to Algorithm 1-2 have been computed. The empirical mean and empirical relative root-mean-square error of the estimators indicate that $\hat{\sigma}^2_{20}$ is to be preferred, but it is hardly distinguishable from the result of Algorithm 1. The MLE estimator is biased and tends to give a too large answer, which is comparable with the pressure vessel estimates.

The extreme small sample case $n = 3$ has also been considered, and then the MLE is clearly not a good choice. More surprising is perhaps the relatively good result for the empirical variance: It is superior to the MLE estimator, and almost as good as the UMVU estimator. The reason is most likely that the reduction of the data through the sufficient statistic is only from 3 to 2 numbers in this case. The estimator from Algorithm 1 seems to be biased, and indicates that X_t from Algorithm 1 does not have the conditional distribution of X given $T = t$ in this case. In terms of root-mean-square error the estimator from Algorithm 1 is slightly better than the UMVU estimator from Algorithm 2 in this case.

The results are summarized in Table 1. In order to do these simulations the estimators have been sampled at several w values, and afterwards interpolation have been used to evaluate the estimators. The corresponding estimators of the shape are shown in Figure 1.

5.2 Goodness of fit

In previous studies of the present pressure vessel data set the authors have considered goodness of fit tests for testing exponentiality $\alpha = 1$ against the gamma alternative with $\alpha < 1$ (Keating et al. 1990, Wong 1992, Welsh 1996). A particular conclusion is rejection of exponentiality at the 1% level for a test based on W (Keating et al. 1990). It is possible to repeat this with the uniformly most powerful unbiased test for this case (E.L.Lehmann 1997, p.147), and the actual calculations may be done with the sufficient conditional Monte Carlo method. A related problem is considered here: Is the pressure vessel data from a gamma distribution?

A Kolmogorov-Smirnov type test will be used, where the null hypothesis H_0 is that the density f is a gamma density. It is assumed that the data are independent from a common density f. The most common approach would be to estimate the parameters to obtain a specific distribution which can be used in a Kolmogorov-Smirnov test. Exact p-values and critical values can be obtained by simulation. Instead of an estimate of the distribution function through the parameters the uniformly minimum variance unbiased estimate of the distribution function at the given failure times will be used. To obtain exact p-values a conditional test given the sufficient statistic for H_0 will be considered. This approach is quite general and it is possible and much simpler to use this

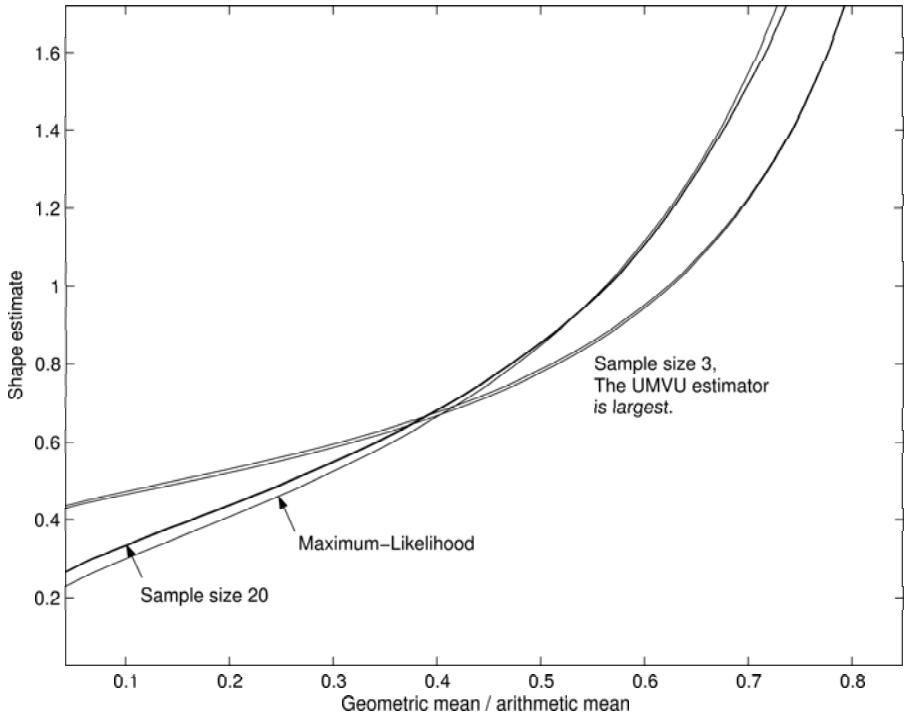

Figure 1: Maximum-likelihood estimator of shape and the estimator $\hat{\alpha}_n$ from equation (27) from samples of size $n = 3$ and $n = 20$ from a gamma distribution. The numerical distance between the two estimators is small for these two sample sizes.

in cases where Algorithm 1 is available, including in particular the exponential and the normal family. It is definitely of practical importance to consider the power of the resulting tests and compare it with alternative tests, but this will not be considered here. The purpose here is to demonstrate how the sufficient conditional Monte Carlo method can be used to calculate an exact p-value in a nontrivial case.

In order to formulate the test some functions are introduced. They will generally depend on $t = (\bar{x}, w)$, but this dependence will not always be made explicit. The conditional Kolmogorov-Smirnov distance $r = r(x)$ is a function of $x = (x_1, \ldots, x_n)$, and actually a function of the ordered tuple $(x_{(1)}, \ldots, x_{(n)})$. It is defined by

$$r(x) = \max_i r_i(x), \quad r_i(x) = \max\{\left|F_t(x_{(i)}) - i/n\right|, \left|F_t(x_{(i)}) - (i-1)/n\right|\} \tag{28}$$

where $F_t(x) = P(X_1 \leq x \,|\, T = t)$. A calculation of $r(x)$ can be done by $n = 20$ conditional expectations as in the previous case of the variance. The conditional probability

$$p(x) = P(r(X) \geq r(x) \,|\, T = t(x)) \tag{29}$$

gives the exact p-value. In this case, as in many similar cases, the p-value has the additional interpretation as a normalized test statistic. The rejection region $\{x \,|\, p(x) \leq \alpha\}$ is a precise level α test, meaning that $P(p(X) \leq \alpha) = \alpha$ given H_0. Given H_0 the statistic $p(X)$ is distributed uniformly on $(0, 1)$.

The actual calculation of the p-value in the example involves a huge number of conditional expectations. The Kolmogorov-Smirnov distance $r(x)$ is determined by $n = 20$ conditional expectations. To calculate the conditional expectation in the expression for $p(x)$ the integrand must be evaluated many times, for instance $m = 10000$ times. For each of these evaluations $n = 20$ conditional expectations must be calculated in order to calculate $r(X)$.

The computational cost could seem to be prohibitive, but all of these conditional expectations are conditioned on $T = t$. This means that a large weighted sample can be generated from $X \mid T = t$ once, and then this single weighted sample can be used for the calculation of all of the above conditional expectations. Implementation of this gives the resulting p-value, and in this case $p(x) = 0.43$. Numerical experiments indicate that this answer has the indicated two digit accuracy with $m = 10000$. The gamma assumption is in particular not rejected at the 1% level.

It is tempting to say that the p-value 0.43 is rather high, but this is only if the p-value is considered as a measure on how well the gamma distribution fits the data. As mentioned above the p-value is uniformly distributed on $(0,1)$ given the gamma assumption, and $p = 0.43$ should then not be a surprising result.

References

Bahadur, R. & Bickel, P. (1968), 'Substitution in conditional expectation', *Ann.Math.Statist.* **39**(2), 377–378.

Basu, D. (1955), 'On statistics independent of a complete sufficient statistic', *Sankhya* **15**, 377–380.

Basu, D. (1958), 'On statistics independent of a sufficient statistic', *Sankhya* **20**, 223–226.

Basu, D. (1977), 'On the elimination of nuisance parameters', *Journal of the American Statistical Association* **72**, 355–366.

Bell, C., Blackwell, D. & Breiman, L. (1960), 'On the completeness of order statistics', *Ann. Math. Statist.* **31**, 794–797.

Berger, J. (1985), *Statistical decision theory and Bayesian analysis*, Springer (second edition).

Blackwell, D. (1947), 'Conditional Expectation and Unbiased Sequential Estimation', *Ann.Math.Statist.* **18**, 105–110.

Blackwell, D. (1953), 'Equivalent comparisons of experiments', *Ann.Math.Statist.* **24**, 265–272.

Casella, G. & Berger, R. (1990), *Statistical inference*, Duxbury.

Dubi, A. & Horowitz, Y. (1979), 'The interpretation of conditional Monte Carlo as a form of importance sampling', *SIAM J.APPL.MATH*.

E.L.Lehmann (1997), *Testing statistical hypotheses*, Springer (second edition).

Engen, S. & Lillegård, M. (1997), 'Stochastic simulations conditioned on sufficient statistics', *Biometrika* **84**(1), 235–240.

Evans, M. & Swartz, T. (2000), *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford.

Fisher, R. (1920), 'A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error', *Monthly Notices Roy. Astronom. Soc.* **80**, 758–770.

Fraser, D. (1956), 'Sufficient statistics with nuisance parameters', *Ann. Math. Statist.* **27**, 838–842.

Granovsky, B. (1981), 'Optimal formulae of the conditional Monte Carlo', *SIAM J.Alg.Disc.Meth.* **2**, 289–294.

Halmos, P. (1946), 'The Theory of Unbiased Estimation', *Ann. Math. Statist.* **17**, 34–43.

Halmos, P. & Savage, L. (1949), 'Application of the Radon-Nikodym theorem to the theory of sufficient statistics', *Ann. Math. Statist.* **20**, 225–241.

Hammersley, J. (1956), 'Conditional Monte Carlo', *J.Assoc.Comp.Mach.* **3**, 73–76.

Hartigan, J. (1983), *Bayes theory*, Springer.

Heath, D. & Sudderth, W. (1989), 'Coherent inference from improper priors and from finitely additive priors', *Ann. Math. Statist.* **17**, 907–919.

Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation problems', *Proc.Roy.Soc. A* **186**, 453–461.

Keating, J., Glaser, R. & Ketchum, N. (1990), 'Testing Hypotheses About the Shape Parameter of a Gamma Distribution', *Technometrics* **32**, 67–82.

Kumar, A. & Pathak, P. (1977), 'Sufficiency and Tests of Goodness of Fit', *Scand. J. Statist.* **4**.

Lehmann, E. & Casella, G. (1998), *Theory of Point Estimation*, Springer.

Lehmann, E. & Scheffe, H. (1950), 'Completeness, similar regions and unbiased estimation', *Sankhya* **10**, 305–340.

Lindqvist, B. & Taraldsen, G. (2004), 'Monte Carlo conditioning on a sufficient statistic', *Biometrika* **to appear**.

Lindqvist, B., Taraldsen, G., Lillegård, M. & Engen, S. (2003), 'A counter example to a claim on stochastic simulations', *Biometrika* **90**(2), 489–490.

Mattner, L. (1993), 'Some incomplete but boundedly complete location families', *Ann. Math. Statist.* **21**, 2158–2162.

Rao, C. (1945), 'Information and accuracy attainable in the estimation of statistical parameters', *Bull.Cal.Math.Soc.* **37**, 81–91.

Rao, C. (1973), *Linear Statistical Inference and Its Applications*, Wiley (second edition).

Rao, C. (1992), 'R.A. Fisher: The Founder of Modern Statistics', *Statistical Science* **7**(1), 34–48.

Reid, N. (1995), 'The Roles of Conditioning in Inference', *Statistical Science* **10**(2), 138–157.

Ripley, B. (1987), *Stochastic simulation*, Wiley.

Rudin, W. (1987), *Real and Complex Analysis*, third edn, McGraw-Hill, USA.

Savage, L. (1976), 'On Rereading R.A. Fisher', *Ann. Statist.* **4**(3), 441–500.

Schervish, M. (1995), *Theory of Statistics*, Springer.

Tanner, M. (1996), *Tools for Statistical Inference*, third edn, Springer.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions (with discussion)', *Ann.Statist.* **22**, 1701–1762.

Trotter, H. & Tukey, J. (1956), 'Conditional Monte Carlo for normal samples', *Symposium on Monte Carlo Methods.* H.A. Meyer, Ed. Wiley, New York pp. 64–79.

Tukey, J. (1957), 'Some examples with fiducial relevance', *Ann. Math. Statist.* **28**, 687–695.

Welsh, A. (1996), *Aspects of Statistical Inference*, Wiley.

Wendel, J. (1957), 'Groups and conditional Monte Carlo', *Ann. Math. Statist.* **28**, 1048–1052.

Wong, A. (1992), 'Inferences on the Shape Parameter of a Gamma Distribution: A Conditional Approach', *Technometrics* **34**, 348–351.