Conditional, posterior and fiducial sampling

Bo Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim

http://www.math.ntnu.no/~bo/
bo@math.ntnu.no


Joint work with Gunnar Taraldsen.
Contents of Talk

- Conditional sampling given sufficient statistics
  - an algorithm
  - a caution
- Conditional sampling in more generality
  - sufficiency
  - Bayesian posterior
  - fiducial distribution
- When is a fiducial distribution a Bayes posterior?
- Generalized fiducial distributions
- Briefly on improper distributions
- Concluding remarks
Sufficient statistics

Let \((X, T)\) be a pair of random vectors with joint distribution indexed by \(\theta\). Typically,

- \(X = (X_1, \ldots, X_n)\) is a sample
- \(T = T(X_1, \ldots, X_n)\) is a statistic.

Assume that \(T\) is **sufficient** for \(\theta\) w.r.t. \(X\), meaning that:

- The conditional distribution of \(X\) given \(T = t\) does not depend on \(\theta\).
- or equivalently: Conditional expectations \(E\{\phi(X) \mid T = t\}\) do not depend on \(\theta\).

Applications:

- Construction of optimal estimators and tests,
- Nuisance parameter elimination,
- Goodness-of-fit testing.

Bo Lindqvist
Conditional, posterior and fiducial sampling
The data generating function*

Let $X$ be a sample from a statistical model with parameter $\theta$.

Let $T = T(X)$ be a statistic.

Assume that $X$ and $T$ for a given parameter value $\theta$ can be represented as

\[
X = \chi(U, \theta) \\
T = \tau(U, \theta)
\]

where $U$ has a completely known distribution with density $f(u)$.

An obvious application of this is to simulate $X$ and $T$.

*This construction is well known in the literature, with different names (functional model, fiducial model, structural model)*
\[ \mathbf{X} = (X_1, \ldots, X_n) \text{ are i.i.d. from } \text{Exp}(\theta): \quad f_{\mathbf{X}}(x; \theta) = \theta e^{-\theta x} \]

\[ T = \sum_{i=1}^{n} X_i \text{ is sufficient for } \theta \]

\[ \mathbf{U} = (U_1, \ldots, U_n) \text{ are i.i.d. } \text{Exp}(1) \]

\[ \chi(\mathbf{U}, \theta) = (U_1/\theta, \ldots, U_n/\theta) \sim^{\theta} \mathbf{X} \]

\[ \tau(\mathbf{U}, \theta) = \sum_{i=1}^{n} U_i/\theta \sim^{\theta} T \]
To sample \( X = (X_1, X_2, \ldots, X_n) \) conditional on \( T = \sum X_i = t \).

Recall:
\[
\chi(U, \theta) = (U_1/\theta, \ldots, U_n/\theta) \sim X,
\]
\[
\tau(U, \theta) = \sum_{i=1}^{n} U_i/\theta \sim T.
\]

**Algorithm (Engen and Lillegård, 1997):**

- Draw \( U = (U_1, \ldots, U_n) \) i.i.d. Exp(1).
- Adjust \( \theta \) so that \( \tau(U, \theta) = t \), i.e., let \( \theta = \hat{\theta}(U, t) = \sum U_i/t \).
- Use sample

\[
X_t(U) = \chi(U, \hat{\theta}(U, t)) = \left( \frac{tU_1}{\sum_{i=1}^{n} U_i}, \ldots, \frac{tU_n}{\sum_{i=1}^{n} U_i} \right)
\]
Algorithm 1: Conditional sampling of $X$ given $T = t$

The algorithm can more generally be described as follows:

Recall: $\chi(U, \theta) \sim \theta X$, $\tau(U, \theta) \sim \theta T$

ALGORITHM 1

- Generate $U$ from the known density $f(u)$
- Solve $\tau(U, \theta) = t$ for $\theta$ to get $\hat{\theta}(U, t)$
- Return $X_t(U) = \chi\{U, \hat{\theta}(U, t)\}$
Proof of Algorithm 1

Let $\phi$ be any function. For all $\theta$ and $t$ we can formally write:

\[
E\left\{ \phi(X) | T = t \right\} = E[\phi\{\chi(U, \theta)\} | \tau(U, \theta) = t] \\
= E[\phi\{\chi(U, \theta)\} | \hat{\theta}(U, t) = \theta] \\
= E(\phi[\chi\{U, \hat{\theta}(U, t)\}] | \hat{\theta}(U, t) = \theta) \\
= E(\phi[\chi\{U, \hat{\theta}(U, t)\}]) \equiv E\{\phi(X_t(U))\}
\]

The key equality is

\[
E[\phi\{\chi(U, \theta)\} | \tau(U, \theta) = t] = E[\phi\{\chi(U, \theta)\} | \hat{\theta}(U, t) = \theta]
\]

**CAUTION:**

- A possible ’Borel paradox’ if conditioning events have probability 0 – this might cause trouble here unless ....
- the two events can be described by the same function of $U$ (same $\sigma$-algebra) – see the pivotal condition next
**The pivotal condition:** Assume that \( \tau(u, \theta) \) depends on \( u \) only through a function \( r(u) \), where we have unique representation \( r(u) = v(\theta, t) \).

(Then \( v(\theta, T) \) is a **pivot** in the classical sense.)

*Exponential example:*

\[
\tau(u, \theta) = \sum_{i=1}^{n} \frac{u_i}{\theta} \\
r(u) = \sum_{i=1}^{n} u_i \\
\text{pivot is } \theta T
\]

Algorithm 1 may (will?) fail when the pivotal condition does not hold, even if \( \hat{\theta}(U, t) \) is uniquely given.

**Examples:** truncated exponential distribution; gamma-distribution; inverse Gaussian distribution.
Counter example: Truncated exponential

\(X = (X_1, \ldots, X_n)\) are i.i.d. on \([0, 1]\) with density

\[
f(x, \theta) = \begin{cases} 
\frac{\theta e^{\theta x}}{e^{\theta} - 1} & \text{if } \theta \neq 0 \\
1 & \text{if } \theta = 0
\end{cases}
\quad \text{for } 0 \leq x \leq 1, \quad -\infty < \theta < \infty
\]

\(T = \sum_{i=1}^{n} X_i\) is sufficient for \(\theta\).

Conditional distribution of \(X\) given \(T = t\) is that of \(n\) independent uniform \([0, 1]\) random variables given their sum.

Let \(n = 2\). Then \(X_1\) given \(X_1 + X_2 = 1\) is \(U[0, 1]\), while Algorithm 1 would give that \(X_1\) given \(X_1 + X_2 = 1\) is distributed as

\[
\log \frac{1 - U_1}{U_2} / \log \frac{1 - U_1}{U_2} + \log \frac{1 - U_2}{U_1}; \quad U_1, U_2 \sim \text{iid} U[0, 1].
\]
Counter example: Densities

\[ \{ X_1 | X_1 + X_2 = 1 \} \sim U[0, 1] \]

Algorithm 1:

\[ \frac{\log \frac{1-U_1}{U_2}}{\log \frac{1-U_1}{U_2} + \log \frac{1-U_2}{U_1}} ; \quad U_1, U_2 \sim \text{iid } U[0, 1]. \]
In the non-pivotal case we need to extend the approach, which also opens for new applications.

Suppose $X, T$ are random vectors and we want to sample from the conditional distribution

$$\{X \mid T = t\}$$

**General problem:** $f(x \mid t) = f(x, t)/f(t)$ is often not tractable since $(X, T)$ may not have a joint density.

**Possible trick:** If $X = (X_1, \ldots, X_n)$, $T = (T_1(X), \ldots, T_k(X))$, we may compute density $f(x_1, \ldots, x_{n-k}, t_1, \ldots, t_k)$.

**Our approach will instead use auxiliary random variables, following an idea going back at least to Trotter and Tukey (1956).**
Introduce new variables \((X^*, T^*)\), possibly improper, so that
\[
\{X^* \mid T^* = t\} \sim \{X \mid T = t\}
\]
but \((X^*, T^*) \not\sim (X, T)\).

We will do this by letting \(X^*, T^*\) be given through \((U, \Theta)\) with
\[
X^* = \chi(U, \Theta)
\]
\[
T^* = \tau(U, \Theta).
\]

where \(U, \Theta\) are suitable random quantities.
The general setup

Let

\[ X^* = \chi(U, \Theta) \]
\[ T^* = \tau(U, \Theta) \]

where

- \( U \) has a known proper distribution with density \( f(u) \)
- \( \Theta \) has a possibly improper and \( \sigma \)-finite distribution with density \( \pi(\theta) \)

Let the joint density of \((U, \Theta)\) be \( f(u)\pi(\theta) \) (so that \( U, \Theta \) are “independent”). We will use this joint density to sample from \( \{(U, \Theta) | \tau(U, \Theta) = t\} \).

Before giving the algorithm, we look at possible applications.
Suppose $T$ is sufficient for $\theta$ and we want to sample from 
$\{X|T = t\}$ (which does not depend on $\theta$).

This is equivalent to sampling from 
$$\{\chi(U, \theta)|\tau(U, \theta) = t\}$$

for any fixed $\theta$ (where Algorithm 1 adjusted $\theta$ according to $U$).

In the generally valid method we instead give $\Theta$ a distribution, “independent” of $U$, and it can then be shown that 
$$\{\chi(U, \Theta)|\tau(U, \Theta) = t\} \sim \{X|T = t\}$$

(This is the approach of L&T 2005).
Recall data generating functions: \( X = \chi(U, \theta), \ T = \tau(U, \theta). \)

The *prior distribution* for \( \theta \) is represented by a random quantity \( \Theta \) with density \( \pi(\theta) \), which in applications often is *improper*.

We want the posterior distribution for \( \Theta \) given data.

If we assume that \( T \) is a sufficient statistic, then this is given as

\[
\{ \Theta | \tau(U, \Theta) = t \}
\]

and is hence covered by the general algorithm, which samples from

\[
(U, \Theta) | \tau(U, \Theta) = t.
\]
Data generating equation: \( X = \chi(U, \theta), \ T = \tau(U, \theta) \).

Suppose \( X = x \) and \( T = t \) are observed.

Suppose \( \tau(U, \theta) = t \) can be uniquely solved for \( \theta \) by \( \hat{\theta}(U, t) \).

Then the **fiducial distribution** of \( \theta \) is the distribution of \( \hat{\theta}(U, t) \).

If \( \theta \) is one-dimensional (and \( \hat{\theta}(U, t) \) is unique), then this distribution is a **confidence distribution** in the sense of Schweder and Hjort (2016), giving rise to exact confidence intervals of all sizes.

This is also consistent with the **fiducial distribution of Fisher**.
The conditional sampling algorithm

Recall that we want to sample from \( \{(U, \Theta) | \tau(U, \Theta) = t\} \).

**General algorithm:**

1. Sample \( u \) from \( \{U | \tau(U, \Theta) = t\} \)
2. Sample \( \theta \) from \( \{\Theta | U = u, \tau(U, \Theta) = t\} \)

If the equation \( \tau(u, \theta) = t \) has a unique solution for \( \theta \), then step 2 is trivial; put \( \theta = \hat{\theta}(u, t) \), so algorithm is

1. Sample \( u \) from \( \{U | \tau(U, \Theta) = t\} \)
2. Solve \( \tau(u, \theta) = t \) to get \( \theta = \hat{\theta}(u, t) \),

*In the latter case, all we need is the distribution \( \{U | \tau(U, \Theta) = t\} \).*
Recall: \((U, \Theta)\) has density \(f(u)\pi(\theta)\).

By transformation \((U, \Theta) \rightarrow (U, \tau(U, \Theta))\)
the joint density of \((U, \tau(U, \Theta))\) is

\[
h(u, t) = f(u)w(t, u)
\]

where \(t \mapsto w(t, u)\) is the density of \(\tau(u, \Theta)\) for fixed \(u\).

Thus:

Conditional density \(\{U|\tau(U, \Theta) = t\} \propto f(u)w(t, u)\).

Practice:

Choose a suitable density \(\pi(\theta)\) for \(\Theta\), calculate \(w(t, u)\), draw \(u\)
from \(f(u)w(t, u)\).
Special cases

ALGORITHM 2 (sampling given **sufficient** statistic, \( T = t \))

- Generate \( U \) from the density \( \propto f(u)w(t, u) \) (*)
- Solve \( \tau(U, \theta) = t \) for \( \theta \) to get \( \hat{\theta}(U, t) \)
- Return \( X_t(U) = \chi\{U, \hat{\theta}(U, t)\} \)

ALGORITHM 3 (sampling from **posterior** with prior \( \pi(\theta) \))

- Generate \( U \) from the density \( \propto f(u)w(t, u) \) (*)
- Return \( \theta_t(U) = \hat{\theta}(U, t) \}

ALGORITHM 4 (sampling from **fiducial** distribution for \( \theta \))

- Generate \( U \) from the density \( f(u) \)
- Return \( \theta_t(U) = \hat{\theta}(U, t) \}

\( (*) \ t \mapsto w(t, u) \) is density of \( \tau(u, \Theta) \) for fixed \( u \).
From Algorithms 3 and 4 we can make the following observation:

*If $\Theta$ has a distribution $\pi(\theta)$ such that the density $w(t, u)$ of $\tau(u, \Theta)$ does not depend on $u$, then the fiducial distribution is a Bayes posterior distribution.*

A sufficient condition for this is that the pivotal condition holds.
The exponential examples

The (full) exponential distribution:

We need the density of $\tau(u, \Theta) = \sum_{i=1}^{n} u_i / \Theta$ when $u$ is fixed:

$$w(t, u) = \frac{\sum u_i}{t^2} \cdot \pi \left( \frac{\sum u_i}{t} \right)$$

The choice $\pi(\theta) = 1/\theta$ gives $w(t, u) = 1/t$.

Since this does not depend on $u$, Algorithm 1 and 2 are the same, and posterior and fiducial distributions are the same by Algorithms 3 and 4 for this prior distribution for $\theta$.

The truncated exponential distribution:

There is no prior $\pi(\theta)$ for which $w(t, u)$ does not depend on $u$, so the fiducial (which still is given by the $\hat{\theta}(U, t)$, and is a confidence distribution) is not a posterior distribution for any prior $\pi(\theta)$.
Lindley (1958): (One-dimensional $\theta$). The distribution of $\hat{\theta}(U, t)$ is a posterior distribution for some (possibly improper) prior distribution for $\theta$ if and only if $T$, or a transformation of it, has a location distribution.

Exponential example: The example is covered by Lindley’s result since $\log X$ has a location distribution.

Fraser (1961): (Multi-dimensional $\theta$). The distribution of $\hat{\theta}(U, t)$ is a posterior distribution if the sample space and parameter sets are transformation groups, and the distributions are given by means of density functions with respect to right Haar measure.
Consider the data generating equation $X = \chi(U, \theta)$, and suppose $X = x$ has been observed.

The main idea for deriving the fiducial distribution is to generate $U$ and then solve for $\theta$ to get realizations from the fiducial.

But what if there is no solution, or if there are several solutions?

If there is no solution, then this $U$ cannot be the origin of the observed data, so it is discarded. Then draw a new $U$, etc... This means that one conditions on the $U$ for which there is a solution.

If there are more than one solution $\theta$, one of them is selected according to some random rule.

The probability of an acceptable $U$ may be 0, so: accept $U$ for which $x$ is in an $\epsilon$-neighborhood of $\chi(U, \theta)$. 
Bayes’ formula: \[ \pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\eta)\pi(\eta)d\eta} \]
gives proper posteriors even if \( \pi(\theta) \) is improper, if denominator is finite.

But in what sense can we do valid probability calculations with improper distributions?

Introduce, as in Kolmogorov’s theory, a sample space \( \Omega \) and a measure \( \Pr \) on it.

Random quantities \( X : \Omega \to \Omega_X \) will then have

\[ \Pr_X(\Omega_X) = \Pr(\Omega) \]

and if we let the random parameter \( \Theta \) be defined on \( \Omega \) as well, and let it have infinite mass, then

\[ \Pr(\Omega) = \Pr_{\Theta}(\Omega_{\Theta}) = \infty \]

and hence \( \Omega \) and any random quantity has infinite mass.
Conditional distributions are proper...fortunately

Conditional distributions

\[ \Pr_{\Theta}^X(\cdot) = \Pr_{\Theta}(\cdot | X = x) \]

can be defined in a way so that they still are proper probability distributions, using the same Radon-Nikodym approach that is used for ordinary conditional probabilities.

But - requirement is that we only condition on random quantities \( X \) which are \( \sigma \)-finite, i.e., there is a countable partition, \( \Omega_X = \bigcup_i E_i \), with all \( \Pr_X(E_i) < \infty \).

**Consequences:**

- IMPROPER POSTERIORS ARE NOT ALLOWED
- MARGINALIZATION PARADOXES ARE AVOIDED
Concluding remarks

- A general approach for conditional sampling has been considered, unifying Bayesian, Frequentistic and Fiducial methods.

  "... but here is a safe prediction for the 21st century: ... I believe there is a good chance that ... something like fiducial inference will play an important role ... Maybe Fishers biggest blunder will become a big hit in the 21st century!"

- So, perhaps Bayesian, Frequentistic and Fiducial (BFF) people may be ... “Best Friends Forever”