2: Settings With No Model Uncertainty

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

Short Course (Day 4)
University of Reading (UK)

© David Draper (all rights reserved)
You’ll recall that optimal model specification consists of conditioning only, and exhaustively, on propositions rendered true by the context of the problem and the design of the data-gathering process.

In Day 2 (Lecture Notes, Part 2) we looked at optimal prior distribution specification; what about sampling distributions?

Optimal sampling-distribution specification. Sometimes the sampling distribution is uniquely specified by problem context.

These cases are of two kinds: based on theoretical definition-matching or exchangeability.

Case 1: Theoretical Definition-Matching

Example 6. In random sampling from a finite population with dichotomous outcomes, if you can actually achieve the theoretical goal of sampling at random either with or without replacement, then (by definition) you have no uncertainty about the resulting sampling distribution: binomial with replacement, hypergeometric without replacement.
Example 7. Consider estimating the number \(0 < N < \infty\) of individuals in a finite population (such as \(\mathcal{P} = \{\text{the deer living on the UCSC campus as of 1 July 2013}\}\)).

One popular method for performing this estimation is capture-recapture sampling; the simplest version of this approach proceeds as follows.

In stage I, a random sample of \(m_0\) individuals is taken, and all of these individuals are tagged and released; then, a short time later, in stage II a second independent random sample of \(n_1\) individuals is taken, and the number \(m_1\) of these \(n_1\) individuals who were previously tagged is noted.

If You can actually achieve the theoretical goals of simple random sampling (SRS: at random without replacement) in stage I and IID sampling (at random with replacement) in stage II, then (by definition) the conditional sampling distribution for \(m_1\) given \(N\) is \((m_1|N \sim \text{Binomial}(n_1, \frac{m_0}{N})\).
Example 8. You’re watching a counting process unfold in time, looking for the occurrences of specific events; if this process satisfies the following three basic assumptions, then the sampling distribution for the number \( N(t) \) of events occurring in \([0, t]\) is (by definition) Poisson(\( \lambda t \)):

- \( P[N(t) = 1|\mathcal{B}] = \lambda t + o(t); \)
- \( P[N(t) = 2|\mathcal{B}] = o(t); \)
- The numbers of events in disjoint time intervals are independent.

Example 9. You’re watching a counting process unfold in time, keeping track of the elapsed times \( T_1, T_2, \ldots \) between events; if this process satisfies the three basic assumptions above, then the sampling distribution for the \( T_i \) is (by definition) IID exponential with mean \( \frac{1}{\lambda} \).

Example 9, continued. If the scientific context of the problem ensures that the \( T_i \) are memoryless — i.e., if \( P(T_i > s + t|T_i > t\mathcal{B}) = P(T_i > s|\mathcal{B}) \) for all \( s, t \geq 0 \) — then again
(by definition) the sampling distribution for the $T_i$ is IID exponential.

**Example 10.** Paleobotanists estimate the moments in the remote past when a given species first arose and then became extinct by taking cylindrical, vertical core samples well below the earth’s surface and looking for the first and last occurrences of the species in the fossil record, measured in meters above the unknown point $A$ at which the species first emerged.

Let $y_{ij} (j = 1, \ldots, J)$ denote the distance above $A$ at which fossil $j$ is found in core sample $i \in (1, \ldots, I)$.

Under the scientifically reasonable assumption that these fossil records are found at random points along the core sample (this would be part of $B$), then You again have no sampling-distribution uncertainty: by definition $(y_{ij} \mid A, B, B) \overset{\text{iid}}{\sim} \text{Uniform} (A, B)$, where $B$ is the unknown point at which the species went extinct.

**Example 11.** The astronomer John Herschel (1850) was interested in characterizing the two-dimensional probability distribution of errors in measuring the position of a star.
Let $x$ and $y$ be the errors in the east-west and north-south directions, respectively; Herschel wanted the joint sampling distribution $p(x, y | B)$.

He took the following two statements as axioms, based on his astronomical intuition:

$(A_1)$ Errors in orthogonal directions should be independent, i.e.,
$$p(x, y | B) = p(x | B) p(y | B).$$

An equivalent expression for $p(x, y | B)$ is obtainable by transforming to polar coordinates:
$$p(x, y | B) = f(r, \theta | B).$$

$(A_2)$ In this new coordinate system, the probability density of the errors should be the same no matter at what angle the telescope is pointed; i.e., $f$ should not depend on $\theta$, i.e.,
$$f(r, \theta | B) = f(r | B).$$

He then showed that under these two axioms the only possible sampling distribution has $x$ and $y$ as independently Normal with mean 0 and the same SD $\sigma$.

James Clerk Maxwell (1860) used the same argument 10 years later.
to characterize the unique three-dimensional sampling distribution of velocities of molecules in a gas.

**Case 2: Exchangeability**

**Example 3 (Day 2, Lecture Notes Part 2, continued.** We’ve already seen an example in which exchangeability led to a unique sampling distribution: the binary mortality indicators $y_i$ for the heart attack patients in calendar 2014.

Recall that de Finetti’s Representation Theorem for binary outcomes said informally that if Your uncertainty about binary $(y_1, y_2, \ldots)$ is exchangeable, then the only logically-internally-consistent inferential model (prior + sampling distribution) is

$$
(\theta | B) \sim p(\theta | B) \\
(y_i | \theta B) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta),
$$

where $\theta$ is both the marginal death probability $P(y_i = 1 | \theta B)$ for patient $i$ and the limiting (population) mean of $(y_1, y_2, \ldots)$. 

This result can be summarized as follows:

For binary observables $y_i$, exchangeability $+$ ______ $\rightarrow$ unique Bernoulli sampling distribution, where in this case no additional assumptions are needed to fill in the blank.

This gives rise immediately to questions like the following: what’s needed in the blank to make this statement true?

For non-negative integer observables $y_i$,

exchangeability $+$ ______ $\rightarrow \left\{ \begin{array}{l} (\lambda | B) \sim p(\lambda | B) \\ (y_i | \lambda, B) \overset{\text{iid}}{\sim} \text{Poisson}(\lambda) \end{array} \right\}$. \(2\)

Many people have worked on de-Finetti-style Representation Theorems of this type; here’s an example.

**Example 12.** To get the Poisson result above, the following assumption has to fill in the blank:

the conditional distribution $(y_1, \ldots, y_n | s_n, B)$, where $s_n = \sum_{i=1}^n y_i$ is a minimal sufficient statistic in the Poisson($\lambda$) sampling model, is Multinomial on \{\text{n-tuples of non-negative integers with sum } s_n\} with Uniform probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$. 
Here are two more examples of this basic idea.

**Example 13.** For continuous observables $y_i$ on $(0, \infty)$,

\[
\text{exchangeability} + \quad \rightarrow \quad \left\{ \begin{array}{l}
(\eta | \mathcal{B}) \sim p(\eta | \mathcal{B}) \\
(y_i | \eta \mathcal{B}) \sim_{\text{IID}} \text{Exponential}(\eta)
\end{array} \right.,
\]

where ______ is the following:

the conditional distribution $(y_1, \ldots, y_n | s_n \mathcal{B})$, where $s_n = \sum_{i=1}^{n} y_i$ is a minimal sufficient statistic in the $\text{Exponential}(\eta)$ sampling model, is Uniform on the simplex $\{(y_1, \ldots, y_n) : y_i \geq 0 \text{ with } \sum_{i=1}^{n} y_i = s_n\}$.

**Example 14.** For continuous observables $y_i$ on $(-\infty, \infty)$,

\[
\text{exchangeability} + \quad \rightarrow \quad \left\{ \begin{array}{l}
(\sigma | \mathcal{B}) \sim p(\sigma | \mathcal{B}) \\
(y_i | \sigma \mathcal{B}) \sim_{\text{IID}} N(0, \sigma^2)
\end{array} \right.,
\]

where ______ is the following:

the conditional distribution $(y_1, \ldots, y_n | t_n \mathcal{B})$, where $t_n = \sqrt{\sum_{i=1}^{n} y_i^2}$ is a minimal sufficient statistic in the $N(0, \sigma^2)$ sampling model,
is **uniform** on the \((n - 1)\)-dimensional sphere of radius \(t_n\) in \(\mathbb{R}^n\) (this condition is equivalent to the **joint distribution** \((y_1, \ldots, y_n|\mathcal{B})\) being rotationally symmetric).

[short course web page: Singpurwallah (2006), pages 45–57, gives a comprehensive catalog of all known sampling-distribution-via-exchangeability results]

You can see that all of these findings have a **common pattern**:

1. **You have to be prepared** to assume the ________ condition, which is of the form \(\{\text{the conditional distribution of the data vector, given a minimal sufficient statistic in the desired sampling model, is uniform on some space}\}\), and

2. **You will rarely work** on a problem in which that condition is automatically rendered true by the problem context.

This makes the Bernoulli result look like the only useful one arising from exchangeability considerations, but de Finetti (1937) himself proved one more **Representation Theorem** that’s even more important and potentially useful than the Bernoulli case:
Bayesian Nonparametric Methods

**de Finetti’s Representation Theorem for Continuous Outcomes.**

You observe \((y_1, \ldots, y_n)\), with the \(y_i\) conceptually continuous in \(\mathbb{R}\);
Your uncertainty about the \(y_i\) is exchangeable.

If You’re prepared to extend Your judgment of exchangeability from
\((y_1, \ldots, y_n)\) to \((y_1, y_2, \ldots)\), then — letting \(F\) denote the empirical
cumulative distribution function (CDF) of the \((y_1, y_2, \ldots)\) values —
the only logically-internally-consistent inferential model based on the
observables is

\[
(F | B) \sim p(F | B) \quad (5)
\]

\[
(y_i | F, B) \overset{\text{IID}}{\sim} F.
\]

(Note that de Finetti’s Representation Theorem for binary
outcomes is a special case of this result.)

This new theorem requires You to place a scientifically-meaningful
prior distribution on the space \(\mathcal{F}\) of all CDFs on \(\mathbb{R}\), which de Finetti
didn’t have the slightest idea how to do in 1937.)
Putting priors on functions (rather than scalars, vectors or matrices) is the subject addressed by Bayesian nonparametric methods; this is an issue we’ll talk more about in Part 3 of the Lecture Notes.

One more example in which both the prior and the sampling distribution arise directly from problem context, i.e., in which optimal Bayesian model specification is possible:

[short course web page: Lecture Notes Part 2A (Bayesian Qualitative-Quantitative Inference)]