The Sampling Problem

Peter Occil

The Sampling Problem

This version of the document is dated 2023-07-22.

Peter Occil

This page is about a mathematical problem of **sampling a probability distribution with unknown parameters**. This problem can be described as sampling from a new distribution using an endless stream of random variates from an incompletely known distribution.

Suppose there is an endless stream of numbers, each generated at random and independently from each other, and as many numbers can be sampled from the stream as desired. Let $(X_0, X_1, X_2, X_3, ...)$ be that endless stream, and call the numbers *input values*.

Let `InDist` be the probability distribution of these input values, and let λ be an unknown parameter that determines the distribution `InDist`, such as its expected value (or mean or "long-run average"). Suppose the problem is to **produce a random variate with a distribution `OutDist` that depends on the unknown parameter λ**. Then, of the algorithms in the section "**Sampling Distributions Using Incomplete Information**[1]":

- In **Algorithm 1** (Jacob and Thiery 2015)[2], `InDist` is arbitrary but must have a known minimum and maximum, λ is the expected value of `InDist`, and `OutDist` is non-negative and has an expected value of $f(\lambda)$.
- In **Algorithm 2** (Duvignau 2015)[3], `InDist` is a fair die with an unknown number of faces, λ is the number of faces, and `OutDist` is a specific distribution that depends on the number of faces.
- In **Algorithm 3** (Lee et al. 2014)[4], `InDist` is arbitrary, λ is the expected value of `InDist`, and `OutDist` is non-negative and has an expected value equal to the mean of $f(X)$, where X is an input value taken.
- In **Algorithm 4** (Jacob and Thiery 2015)[5], `InDist` is arbitrary but must have a known minimum, λ is the expected value of `InDist`, and `OutDist` is non-negative and has an expected value of $f(\lambda)$.
- In **Algorithm 5** (Akahira et al. 1992)[6], `InDist` is Bernoulli, λ is the expected value of `InDist`, and `OutDist` has an expected value of $f(\lambda)$.
- In the **Bernoulli factory problem**[7] (a problem of turning biased coins to biased coins), `InDist` is Bernoulli, λ is the expected value of `InDist`, and `OutDist` is Bernoulli with an expected value of $f(\lambda)$.

In all cases given above, each input value is independent of everything else.

There are numerous other cases of interest that are not covered in the algorithms above. An example is the

[1] https://peteroupc.github.io/randmisc.md#Sampling_Distributions_Using_Incomplete_Information

[2] Jacob, P.E., Thiery, A.H., "On nonnegative unbiased estimators", Ann. Statist., Volume 43, Number 2 (2015), 769-784.

[3] Duvignau, R., "Maintenance et simulation de graphes aléatoires dynamiques", Doctoral dissertation, Université de Bordeaux, 2015.

[4] Lee, A., Doucet, A. and Łatuszyński, K., 2014. "**Perfect simulation using atomic regeneration with application to Sequential Monte Carlo**", arXiv:1407.5770v1 [stat.CO]. https://arxiv.org/abs/1407.5770v1

[5] Jacob, P.E., Thiery, A.H., "On nonnegative unbiased estimators", Ann. Statist., Volume 43, Number 2 (2015), 769-784.

[6] AKAHIRA, Masafumi, Kei TAKEUCHI, and Ken-ichi KOIKE. "Unbiased estimation in sequential binomial sampling", Rep. Stat. Appl. Res., JUSE 39 1-13, 1992.

[7] https://peteroupc.github.io/bernoulli.html

case of **Algorithm 5** except `InDist` is any discrete distribution, not just Bernoulli. [8] An interesting topic is to answer the following: In which cases (and for which functions f) can the problem be solved…

- …when the number of input values taken is random (and may depend on already taken inputs), but is finite with probability 1 (a *sequential unbiased* estimator)?[9]
- …when only a fixed number n of input values can be taken (a *fixed-size unbiased* estimator)?
- …using an algorithm that produces outputs whose expected value *approaches* $f(\lambda)$ as more input values are taken (an *asymptotically unbiased* estimator)?

The answers to these questions will depend on—

- the allowed distributions for `InDist`,
- the allowed distributions for `OutDist`,
- which parameter λ is unknown,
- whether the inputs are independent, and
- whether outside randomness is allowed (that is, whether the estimator is *randomized*).

An additional question is to find lower bounds on the input/output ratio that an algorithm can achieve as the number of inputs taken increases (e.g., Nacu and Peres (2005, Question 2)[10]).

My interest on the problem is in the existence and construction of simple-to-implement algorithms that solve the *sampling problem* given here. In addition, the cases that most interest me are when—

- λ is `InDist`'s expected value, and
- `OutDist` has an expected value of $f(\mathbb{E}[X])$ or $\mathbb{E}[f(X)]$, where X is an input value taken,

with or without other conditions.

1 Results

It should be noted that many special cases of the sampling problem have been studied and resolved in academic papers and books.

The problem here is one of bringing all these results together in one place.

The following are examples of results for this problem. The estimators are allowed to be randomized (to use outside randomness) unless specified otherwise.

- Suppose `InDist` takes an unknown finite number n of values with unknown probabilities ($n \geq 1$), λ is n, and `OutDist` has an expected value of λ.
  - No sequential non-randomized unbiased estimator exists, even if n is known to have a maximum of 2 or greater (Christman and Nayak 1994)[11]. [12]
  - Not aware of conditions for sequential randomized unbiased estimators.
  - Not aware of conditions for fixed-size unbiased estimators.
  - Not aware of conditions for asymptotically unbiased estimators.

[8] Singh (1964, "Existence of unbiased estimates", Sankhyā A 26) claimed that an estimation algorithm with expected value $f(\lambda)$ exists for a more general class of `InDist` distributions than the Bernoulli distribution, as long as there are polynomials that converge pointwise to f, and Bhandari and Bose (1990, "Existence of unbiased estimates in sequential binomial experiments", Sankhyā A 52) claimed necessary conditions for those algorithms. However, Akahira et al. (1992) questioned the claims of both papers, and the latter paper underwent a correction, which I haven't seen (Sankhyā A 55, 1993).

[9] An algorithm that takes a finite number of inputs with probability 1 is also known as a *closed sampling plan* in papers and books about sequential estimation.

[10] Nacu, Şerban, and Yuval Peres. "**Fast simulation of new coins from old**", The Annals of Applied Probability 15, no. 1A (2005): 93-115. https://projecteuclid.org/euclid.aoap/1106922322

[11] Christman, M.C., Nayak, T.K., "**Sequential unbiased estimation of the number of classes in a population**", Statistica Sinica 4(1), 1994. https://www.jstor.org/stable/24305291

[12] Christman and Nayak (1994) did not study the case when the estimator can use outside randomness or the case when n is known to have a *minimum* of 2 or greater. Duvignau (2015) studied a closely related problem.

- Suppose `InDist` is a fair die with an unknown number of faces (1 or greater), λ is the number of faces, and `OutDist` has an expected value of $f(\lambda)$.
  - If there is no maximum sample size, a sequential unbiased estimator exists for every f (Christman and Nayak 1994)[13].
  - If f is unbounded (including when $f = \lambda$), there is no fixed-size non-randomized unbiased estimator that is based only on the sample size and the number of unique items sampled (Christman and Nayak 1994)[14].
  - Not aware of conditions for more general fixed-size unbiased estimators.
  - Not aware of conditions for asymptotically unbiased estimators.
- Suppose `InDist` takes on numbers from a finite set; λ is the expected value of `InDist`; and `OutDist` has an expected value of $f(\lambda)$.
  - A fixed-size nonrandomized unbiased estimator exists only if f is a polynomial in homogeneous form of degree n or less, where n is the number of inputs taken (Lehmann (1983, for coin flips)[15], Paninski (2003, proof of Proposition 8, more generally)[16]).
  - The existence of randomized sequential unbiased estimators is claimed by Singh (1964)[17]. But see Akahira et al. (1992)[18].
  - Not aware of conditions for sequential nonrandomized unbiased estimators.
  - Not aware of conditions for fixed-size randomized unbiased estimators.
  - Not aware of conditions for asymptotically unbiased estimators.
- Suppose `InDist` has a finite mean, λ is the expected value of `InDist`, and `OutDist` is nonnegative and has an expected value of $f(\lambda)$.
  - There is no sequential unbiased estimator (and thus no fixed-size unbiased estimator) (Jacob and Thiery 2015)[19].
  - Not aware of conditions for asymptotically unbiased estimators.
- Suppose `InDist` has a finite mean and is supported on $[a, \infty)$, λ is the expected value of `InDist`, and `OutDist` is nonnegative and has an expected value of $f(\lambda)$.
  - A sequential unbiased estimator exists only if f is nowhere decreasing (Jacob and Thiery 2015)[20].[21]
  - Not aware of conditions for fixed-size unbiased estimators.
  - Not aware of conditions for asymptotically unbiased estimators.
- Suppose `InDist` has a finite mean and is supported on $(\infty, a]$, λ is the expected value of `InDist`, and `OutDist` is nonnegative and has an expected value of $f(\lambda)$.
  - A sequential unbiased estimator exists only if f is nowhere increasing (Jacob and Thiery 2015)[22].
  - Not aware of conditions for fixed-size unbiased estimators.
  - Not aware of conditions for asymptotically unbiased estimators.
- Suppose `InDist` is Bernoulli (a "biased coin"), λ is the expected value of `InDist`, and `OutDist` is Bernoulli with an expected value of $f(\lambda)$.
  - Let D be the set of allowed values for λ. Thus, D is either the closed unit interval or a subset thereof.
  - A sequential unbiased estimator exists if and only if f is everywhere 0, everywhere 1, or continuous

[13]Christman, M.C., Nayak, T.K., "**Sequential unbiased estimation of the number of classes in a population**", Statistica Sinica 4(1), 1994. https://www.jstor.org/stable/24305291

[14]Christman, M.C., Nayak, T.K., "**Sequential unbiased estimation of the number of classes in a population**", Statistica Sinica 4(1), 1994. https://www.jstor.org/stable/24305291

[15]Lehmann, E.L., *Theory of Point Estimation*, 1983.

[16]Paninski, Liam. "Estimation of Entropy and Mutual Information." Neural Computation 15 (2003): 1191-1253.

[17]R. Singh, "Existence of unbiased estimates", Sankhyā A 26, 1964.

[18]AKAHIRA, Masafumi, Kei TAKEUCHI, and Ken-ichi KOIKE. "Unbiased estimation in sequential binomial sampling", Rep. Stat. Appl. Res., JUSE 39 1-13, 1992.

[19]Jacob, P.E., Thiery, A.H., "On nonnegative unbiased estimators", Ann. Statist., Volume 43, Number 2 (2015), 769-784.

[20]Jacob, P.E., Thiery, A.H., "On nonnegative unbiased estimators", Ann. Statist., Volume 43, Number 2 (2015), 769-784.

[21]In addition, Jacob and Thiery (2015) conjecture that this estimator exists if and only if f is writable as $f(\lambda) = c_0(\lambda - a)^0 + c_1(\lambda - a)^1 + \ldots$, where $c_0, c_1, \ldots$ are all nonnegative. In that case, they showed that the random number of inputs need not depend on inputs already taken.

[22]Jacob, P.E., Thiery, A.H., "On nonnegative unbiased estimators", Ann. Statist., Volume 43, Number 2 (2015), 769-784.

and polynomially bounded on D (Keane and O'Brien 1994)[23].

  - The estimator can be nonrandomized whenever D contains neither 0 nor 1 (Keane and O'Brien 1994)[24]. See **this section**[25] for results when λ is allowed to be 0 or 1 (the coin can show heads every time or tails every time).

- A fixed-size unbiased estimator exists if and only if f is writable as a polynomial of degree n with $n + 1$ Bernstein coefficients in the closed unit interval, where n is the number of inputs taken (Goyal and Sigman 2012)[26].
- Perhaps it is true that an asymptotically unbiased estimator exists if and only if there are polynomials $p_1, p_2, ...$ that converge pointwise to f on D (that is, for each λ in D, $p_n(\lambda)$ approaches $f(\lambda)$ as n increases), and the polynomials' Bernstein coefficients lie in the closed unit interval (see also Singh (1964)[27]).

There are also two other results on the existence of fixed-size and asymptotically unbiased estimators, but they are relatively hard to translate to this problem in a simple way: Liu and Brown (1993)[28], Hirano and Porter (2012)[29]. Other results include Gajek (1995)[30] (which has a result on building unbiased estimators from asymptotically unbiased ones), Rychlik (1995)[31].

In a result closely related to the sampling problem, given a stream of independent random variates each distributed as φ with probability λ or as Q otherwise (where φ and Q are probability distributions, φ and λ are known, and Q is unknown), there is no way in general to generate a variate distributed as Q, even if values from Q and φ must come from the same set of numbers [32].

1.1 Question

For any case of the sampling problem, suppose the number of input values taken is random. If the number of inputs is allowed to depend on previously taken inputs, do more sequential unbiased estimators exist than otherwise?

2 Notes

[23]Keane, M. S., and O'Brien, G. L., "A Bernoulli factory", *ACM Transactions on Modeling and Computer Simulation* 4(2), 1994.

[24]Keane, M. S., and O'Brien, G. L., "A Bernoulli factory", *ACM Transactions on Modeling and Computer Simulation* 4(2), 1994.

[25]https://peteroupc.github.io/bernsupp.html#Which_functions_don_t_require_outside_randomness_to_simulate

[26]Goyal, V. and Sigman, K., 2012. On simulating a class of Bernstein polynomials. ACM Transactions on Modeling and Computer Simulation (TOMACS), 22(2), pp.1-5.

[27]R. Singh, "Existence of unbiased estimates", Sankhyā A 26, 1964.

[28]Liu., R.C., Brown, L.D., "Nonexistence of informative unbiased estimators in singular problems", Annals of Statistics 21(1), 1993.

[29]Hirano, Keisuke, and Jack R. Porter. "Impossibility results for nondifferentiable functionals." Econometrica 80, no. 4 (2012): 1769-1790.

[30]Gajek, L. (1995). Note on unbiased estimability of the larger of two mean values. Applicationes Mathematicae, 23(2), 239-245.

[31]Rychlik, Tomasz. "A class of unbiased kernel estimates of a probability density function." Applicationes Mathematicae 22, no. 4 (1995): 485-497.

[32]Henderson, S.G., Glynn, P.W., "Nonexistence of a class of variate generation schemes", *Operations Research Letters* 31 (2003). It is also believed that the paper's Theorem 2 remains true even if Q must be a polynomial.