The Sampling Problem

Peter Occil

The Sampling Problem

This version of the document is dated 2023-07-15.

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 finite with probability 1 (a *sequential unbiased* estimator)?
- …when only a fixed number n of input values can be taken (a fixed-sample-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.

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)[9]).

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.

- 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 unbiased estimator exists only if f is a polynomial of degree n or less, where n is the number of inputs taken (Lehmann (1983, for coin flips)[10], Paninski (2003, proof of Proposition 8, more generally)[11]).
  - The existence of sequential unbiased estimators is claimed by Singh «1964|R. Singh, "Existence of unbiased estimates", Sankhyā A 26, 1964.». But see Akahira et al. (1992)[12].
- Suppose `InDist` is Bernoulli, 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 and polynomially bounded on D (Keane and O'Brien 1994)[13].
  - Then a fixed-size unbiased estimator exists if and only if f is a polynomial of degree n with $n+1$ Bernstein coefficients in the closed unit interval, where n is the number of inputs taken «Goyal

[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]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

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

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

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

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

and Sigman 2012|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.».

- 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, the $p_n(\lambda)$ approaches $f(\lambda)$ as n increases), and the polynomials' Bernstein coefficients lie in the closed unit interval (see also Singh «1964|R. Singh, "Existence of unbiased estimates", Sankhyā A 26, 1964.»).

2 Notes