Remodeling Musicological Icebergs: Analyzing the Historical Question of Musical Loss

William Watson

More than ten years ago now, Michael Cuthbert proposed a radical idea. Commenting on the then-widespread belief that the surviving corpus of written music from early fifteenth-century Italy constituted a "small, possibly unrepresentative sliver of the original written repertory," the metaphorical "tip of the iceberg," Cuthbert instead proposed that "between half and three-quarters of the written music of late-Trecento and early-Quattrocento Italy is available for study" [1, pg. 39, 59]. This surprising suggestion emerged out of Cuthbert's consideration of a wide range of indicators, including evidence from literary sources, concordances in surviving indexes from otherwise-lost manuscripts, and, most importantly for present purposes, statistical models. Cuthbert's blending of statistical models with historical analysis made his argument compelling in a way unusual within the musicological literature, and the piece made quite a splash in the realm of medieval musicology. Especially when seen in the context of estimates of medieval manuscript survival, such as the 6% figure given by Eltjo Buringh for fifteenth-century Europe, Cuthbert's estimates paint a surprisingly rosy picture of how much music survives from centuries long past [2, pg. 261].

However, there remains an unresolved tension at the heart of Cuthbert's article. The statistical models that Cuthbert draws upon all deal with the problem of estimating an unknown population size given a corpus of random samples taken from that population with replacement—the original repertoire is treated as the population and extant manuscripts stand in for the samples.[1] And yet, these models are extremely different from each other. One of them, used by Bradley Efron and Ronald Thisted to estimate the number of words that Shakespeare knew but never used in his writings, has obvious shortcomings pointed out by Cuthbert, and I will not discuss it here. Rather, the focus of this article is on the other two: a probabilistic method that Cuthbert himself proposed, and an analytic method proposed by Shahar Boneh, Arnon Boneh, and R. J. Caron (hence "BBC," following Cuthbert's practice) [1, 3]. Cuthbert makes no claims about the relative merit of these methods, and even uses the relative similarity of their results to assuage any anxieties his readers may have about trusting a statistical model designed by a musicologist. I claim not only that Cuthbert's probabilistic method is demonstrably better than BBC's at producing reasonable estimates of population size, but that BBC's method is in fact extremely ill-suited for this task—a task for which it was never really designed.

In what follows, I first describe Cuthbert's and BBC's methods of estimation in some detail, paying especially close attention to the role played by assumptions of randomness and

[1] The aptness of this metaphor, particularly as randomness comes into play, will be discussed further below.

by equivalences of expected values. I then use a series of statistical thought-experiments to show that Cuthbert's method outperforms BBC's both in terms of accuracy and theoretical soundness. Finally, I extend Cuthbert's method so that it produces distributions of estimates for a population size rather than single estimates, and then apply it to study the survival of several subrepertoires of fifteenth-century European vernacular polyphony, as represented by the contents of the Digital Index of Late Medieval Song (DILMS). The argument relies on a collection of Python scripts, all of which are accessible at `https://www.github.com/wcwatson/modeling_icebergs`. I encourage the reader to experiment with these scripts (and with the supplementary Jupyter Notebook that walks through the thought-experiment described in section II.3) in order to get a feeling for how the statistical models work.

I Two Models of Loss

In this section I describe two very different ways of approaching the same question: given a corpus of random samples taken with replacement from a population of unknown size, how can we estimate that size? In the ensuing discussion, parameters are defined in the following manner. The true population size (i.e., the original number of songs) is N. The number of distinct entities observed (i.e., the extant number of songs) is n, and the number of entities observed k times (i.e., the number of songs with k extant copies) is n_k, including n_0, which is the number of unobserved entities in the population. The proportion of the population observed is $p = \frac{n}{N}$. The number of samples (i.e., the number of extant manuscripts) is y, and the size of the k^{th} sample (i.e., the number of songs in the k^{th} manuscript) is x_k.

I.1 Cuthbert

Cuthbert's method of estimation is a two-stage process that relies on nothing more than probabilistic reasoning. In the first stage, we momentarily assume that the samples are independent, random, and select from the entire population with replacement (i.e., that the probability of a song appearing in a manuscript is more-or-less independent of the particular song and manuscript in question, and that every manuscript potentially contains items from the entirety of the once-extant written repertoire).[2] Under this assumption, the probability that a given entity will appear in the k^{th} sample is $\frac{x_k}{N}$, and the probability that it does not appear in any sample (i.e., that a given song does not appear in any manuscript and, thus, is no longer extant) is merely the product of the probabilities that it does not appear in each sample: $\mathbb{P}(\textit{unobserved}) = \prod_{k=1}^{y}(1 - \frac{x_k}{N}) = \dfrac{\prod_{k=1}^{y} N - x_k}{N^y}$. Moreover, since under this assumption each of the N entities in the population has an identical likelihood of not appearing in any sample (i.e., all songs are equally likely to have been lost), the expected number of unobserved entities is $E(n_0) = N \cdot \mathbb{P}(\textit{unobvserved}) = \dfrac{\prod_{k=1}^{y} N - x_k}{N^{y-1}}$. And since, by definition, $n_0 = N - n$, an estimate for N can be generated by solving for it in the

[2] In a musical context, we might say that the second condition amounts to a claim that the repertoire is homogeneous and bounded.

following equation (all parameters other than N are known).[3]

$$\frac{\prod_{k=1}^{y} N - x_k}{N^{y-1}} - (N - n) = 0 \tag{1}$$

This expression is essentially a ratio of polynomials in N of approximate degree y. The value of y can easily be quite large in the situations under discussion here, but since the value of N is assumed to be a positive integer greater than n but less than some reasonably large upper bound, it is relatively expedient to find an approximate solution to this equation by computational brute force—analytic precision is not needed here. My implementation of this process can be seen in `cuthbert.py`, included in the GitHub repository described above.[4] The value of N that emerges at the end of this first stage is our initial estimate of the total size of the written repertoire.

In the second stage of Cuthbert's method, the initial estimate of N is cross-validated in order to put pressure on the assumption that independence and randomness structure the corpus of samples. It is vital to note that the independence and randomness of the individual samples themselves is not really being tested here, and it is possible that emphatically non-random samples may fare relatively well.[5] Rather, the relevant question here is the degree to which the distribution of entities among the corpus of samples, *as a whole*, approximates the distribution that would be expected of truly random samples.

In order to cross-validate the initial estimate of N, we simulate a population of that size that includes every observed entity along with $n_0 = N - n$ "dummy" entities, each representing an unobserved entity assumed to be in the original population (i.e., the dummies represent those songs assumed to have been lost to the whims of history). We then randomly select a number of the original samples as a "validation set," and note how many entities out of the n observed would and would not have still been observed if those samples had not been collected (i.e., if those manuscripts had not survived). We then construct a "simulated set" of samples by iterating through the validation set and taking an identical number of truly random samples from the simulated population (including the dummy entities). Each of the samples in the simulated set will be uniquely paired with a sample in the validation set and will be identical to it in size. We then count how many "new" entities are in the simulated set, from the perspective of the corpus of known samples *not* in the validation set. If the assumption that samples are structured by independence and randomness holds, then the number of new entities in the simulated set should be approximately equal to the number of new entities in the validation set. A significant under-estimate would be an indication that the assumption may not hold. Cuthbert notes that, since the results of this experiment are highly dependent on the selection of a particular validation set, it should ideally be performed many times to mitigate this dependence, although he was not able to

[3]There is a typographical error in [1] at the point that equation 1 appears; a sign of equality is substituted for a sign of subtraction.

[4]My implementation is not particularly efficient at present, but there are several possibilities for improving its performance in the future.

[5]Indeed, in the musical context of fifteenth-century polyphonic songs it would be ludicrous to do make this assumption. The extant corpus of manuscripts evinces copious evidence of careful and deliberate compilation, for instance in the curated collection of songs copied into Oxford 213, or in the opening dozen songs of the Wolfenbüttel Chansonnier whose initials form the dedicatory acrostic "A Etjene Petjt."

do this himself due to the construction of his database. Some combination of the estimate of N from the first stage and the results of the cross-validation experiments performed in this second stage constitutes the results of Cuthbert's method. I extend this method beyond what Cuthbert himself proposed in section III below, but for now let us turn to the other method under discussion here.

I.2 Boneh, Boneh, and Caron (BBC)

The second method under consideration here, due to BBC, is completely different in its motivation and execution. It foregoes the naïve probabilistic analysis in Cuthbert's method for a more sophisticated mathematical armature. BBC motivate their discussion by invoking a multinomial distribution, which describes the outcome of sampling N objects with replacement and with probabilities $p_1, \ldots, p_N$. They then observe that this is related in the limit to a situation in which there are N independent Poisson processes with parameters $\lambda_1, \ldots, \lambda_N$ (recall that a Poisson process S with parameter λ_S is a discretely valued random variable on the positive real numbers such that $\mathbb{P}(S(t) = k) = \frac{(\lambda_S t)^k}{k!} e^{-t\lambda_S}$). The relation is fairly transparent: if we track these Poisson processes in the interval $[0, 1]$ and count how many of them occur once, how many occur twice, etc., then this is identical to generating values for $\{n_1, n_2, \ldots n_m\}$, where m is the maximum number of times that any individual Poisson process is detected (the largest number of extant copies of any song).

In order to use this information to estimate the total number of Poisson processes, N, it is useful to define the auxiliary function $D(t)$ to be the number of processes detected in the interval $(1, t + 1]$ that were not first detected in the interval $[0, 1]$, and the function $\Psi(t) = E(D(t))$, which BBC call "the prediction function."[6] $\Psi(t)$ has several attractive mathematical properties, most pertinent of which for this discussion are that it has infinite order alternating copositivity (that is, its k^{th} derivative takes positive values on the positive half-line for all odd k and negative values for all even k) and that it is bounded, which conspire to mean that it has an asymptotic limit as t increases without bound. Thus, computing this limit is tantamount to generating an estimate for n_0 (we might imagine stumbling across an arbitrarily large number of new manuscripts containing songs from our assumedly finite and homogeneous repertoire).

BBC show that this limit may be estimated with a relatively simple two-part process. First, we calculate a biased estimate, $\hat{\Psi}(\infty)$, using a simple sum of exponentials:

$$\hat{\Psi}(\infty) = \sum_{k=1}^{m} n_k e^{-k} \tag{2}$$

Then, we obtain an unbiased estimate, $\hat{n}_0$, by numerically solving the equation:

$$\hat{n}_0 \left(1 - e^{\frac{-n_1}{\hat{n}_0}}\right) = \hat{\Psi}(\infty) \tag{3}$$

BBC also give some algorithmic details for how this equation may be solved numerically in a relatively efficient manner; my implementation thereof may be found within `bbc.py`,

[6] BBC define $D(t)$ slightly differently, but it is all a matter of horizontal displacement and there is no meaningful distinction to be made between our definitions.

included in the GitHub repository described above. A seemingly small element of BBC's discussion that will become important below is their invocation of a result of Robbins's [4], according to which $E(n_1) = E(\sum_{S \in \mathbf{U}} \lambda_S)$, where $\mathbf{U}$ is the set of Poisson processes that go unobserved in the interval $[0, 1]$. As we shall see, this interacts with some of the assumptions made in Cuthbert's method in some interesting ways.

*

For two methods that purport to answer similar questions, they are strikingly different. The one intuitively transparent but computationally unwieldy, the other more difficult to grasp but substantially more mathematically elegant. However, they do have one thing in common: their propensity to traffic in seemingly strange assumptions when it comes to the scenario of songs in manuscripts that motivated this discussion. Cuthbert's method, for instance, assumes that all entities have the same probability of being observed in any sample. While this may be reasonable in certain circumstances, it is somewhat difficult to believe that all songs have the same probability of being copied into any manuscript. Even discounting differences of geographic and chronological origin, we might wonder about the possibility of a "popularity multiplier"—whether songs that are already relatively widely copied are more likely to be recopied than new repertoire is to be copied for the first time.

BBC's method, on the other hand, explicitly does not assume that all entities are equally likely to be observed (the probability that an entity S will be observed is instead a function of λ_S). And yet, it still seems to smuggle in a few assumptions about the *distribution* of those probabilities, which is to say the distribution of $\{\lambda_S\}$. This is where their invocation of Robbins's result flagged above becomes important. Stare at that equality long enough, and it becomes difficult to convince yourself that there is not some funny business going on under the hood here. Granted, the equality places no real restrictions on the size of $\mathbf{U}$, but realistically a constraint on the expected value of a sum of positive values will place certain restrictions on the *likely* estimates of the number of those values, which is to say the *likely* estimates of n_0. Unfortunately, I have yet to find a way to state this constraint in formal mathematical terms, partially because the theoretical situation that BBC present encompasses a far wider range of conditions than we should realistically expect for any actual situation in which their method could be applied. Indeed, the significance of the limitations placed on the distribution of $\{\lambda_S\}$ should become clearer in the course of the thought-experiments undertaken below.

II Evaluation

In this section I evaluate the relative merits of Cuthbert's and BBC's methods through a series of three statistical thought-experiments. Two of these deal with edge cases—specifically, the case where all sampled entities are unique, and the case where all samples are identical. The third uses random simulation to study each method's behavior in a variety of circumstances. However, before turning to these thought-experiments I would like to draw attention to two broader, "theoretical" points.

First, while Cuthbert's method of simulation takes as its input—as its data—information about the distribution of observations into samples (songs into manuscripts), BBC's method

takes information about the distribution of entities into observations (inscriptions into songs). In other words, the first method appears to be about the relation of sample to population while the second appears to be about the relation of entity to sample.[7] Of course, each method is in some sense about entity, sample, and population all at once, but the distinction is best appreciated by considering the information that each method ignores. Cuthbert's method would be oblivious to the difference between an entity observed 10 times and an entity observed once; it would only register whether the combined 11 observations are spread among 11 samples or 10. On the other hand, BBC's method would be oblivious to the difference between a sample with 100 observations and a sample with 10; it would only register how many times each of the observations comprising those samples had been detected in total. Given this, it is relatively easy to make a case that Cuthbert's method is more intuitively motivated for the task of estimating population size. It asks what kind of population (what kind of collection of entities) would be likely to produce the collection of samples at hand. BBC's method, on the other hand, seems relatively uninterested in contemplating the population at all, even to the point of almost seeming to assume the characteristics that it seeks to estimate. Whether or not this is truly circular, it should at least make us raise a skeptical eyebrow.

The second point comes down to two words: "cross validation." While both methods make similar (though not identical) assumptions about the role that randomness plays in the constitution of samples and observations, the obvious way of testing these assumptions—the method of cross-validation described in section I.2 above—relies on the kinds of information that Cuthbert's method uses. While a similar exercise could be carried out using a BBC-derived estimate of the overall repertoire size, it would raise a serious question as to why the relations of entity and observation used to generate that estimate should be different from the relations of sample and population used to test it. With Cuthbert's method there is no such conflict. Moreover, I suspect that attempts to generate analogous methods of cross validation that relied on the information used to generate BBC's estimate would be unsatisfactory, due in part to the results of the thought-experiments to which I now turn.

II.1 In a world where every song is an *unicum*...

For the first thought-experiment, let us consider the limit case where each observation is of a distinct entity (in musical terms, where each song is an *unicum*). For present purposes, let us assume that all samples contain ten observations—that $x_k = 10\ \forall k \in \{1, 2, \ldots, y\}$, and thus that $n = 10y$. While BBC's method encounters no problems in this scenario, Cuthbert's method actually breaks—it cannot produce a finite estimate of the population size. To help it along, we can fudge the scenario when running Cuthbert's method and say that two observations are of the same entity, so that $n = 10y - 1$. Because the relations between y, n, and n_1 are all known, both methods' estimates of N can be written as implicit equations dependent only on y. That is, because of the constraints of this experiment, the estimates of the population size are entirely dependent on the number of samples taken (the number of manuscripts extant). It is thus a simple matter to produce a series of estimates

[7] In musical terms, we might cast this as the difference between circulation (*qua* the distribution of songs among manuscripts) and ontology (*qua* the grouping of inscriptions into songs).

y	Cuthbert ($n = 10y - 1$)			BBC ($n = 10y$)		
	$\hat{n}_0$	$\hat{N}$	$\hat{p}$	$\hat{n}_0$	$\hat{N}$	$\hat{p}$
5	941	990	0.04949	20	70	0.71428
10	4374	4473	0.02213	40	140	0.71428
15	10308	10457	0.01424	60	210	0.71428
20	18741	18940	0.01050	80	280	0.71428
25	24650	24899	0.01000	100	350	0.71428
30	29600	29899	0.01000	120	420	0.71428
40	39500	39899	0.01000	160	560	0.71428
50	49400	49899	0.01000	201	701	0.71326
75	74150	74899	0.01000	301	1051	0.71360
100	98900	99899	0.01000	401	1401	0.71377
150	148400	149899	0.01000	602	2102	0.71360
200	197000	199899	0.01000	802	2802	0.71377

Table 1: Estimates when every entity is observed once

such as is given in table 1.

The differences between the two methods' estimates are striking. Cuthbert's method yields estimates of p that are never greater than 0.14 and that rapidly drop to 0.01 (the minimum value allowed by my implementation of this algorithm) as y increases. On the other hand, BBC's method produces estimates of this value that all range between 0.71 and 0.72, and the estimates of n_0 track $4y$ almost exactly. Thus, it seems that there is a lower bound on the estimates of p that BBC's method will produce; it will never suggest that we have sampled anything less than 70% of any population. I would (cautiously) suggest that this lower bound may be related to the assumptions about the distribution of $\{\lambda_S\}$ smuggled into BBC's estimates via the chained operations of "reasoning by maximum likelihood" mentioned above.

While table 1 alone should be worrying enough to tip the scales in favor of Cuthbert's method, perhaps an even more important consideration is that Cuthbert's method fails for the "pure form" of this experiment. After all, that's exactly what it should do. To believe that we would be in any position to speak of the possible size of a population, each of whose entities we have observed at most once, would be to court absurdity. In the end, the strange honesty that Cuthbert's method forces us to adopt about the limited accessibility of the knowledge it helps to produce turns out to be one of its great strengths.

II.2 In a world where all books are identical...

The premise of the second experiment is in many ways the opposite of the first: the contents of every single sample are identical. Let us again proceed with the convenient and decimal-friendly set-up of samples from the first thought-experiment, except that this time, given the new constraints, n always equals 10. Perhaps it is no surprise to learn that this time, while Cuthbert's method encounters no problems, BBC's method breaks for $y > 1$; the fact that $n_1 = 0$ makes the second stage of its calculation (the bias correction) impossible to execute. And so here too we can fudge our scenario. Let us say that, for BBC's calculations,

y	Cuthbert ($n = 10$)			BBC ($n = 11$)		
	$\hat{n}_0$	$\hat{N}$	$\hat{p}$	$\hat{n}_0$	$\hat{N}$	$\hat{p}$
5	0	10	1.00	0.5046	11.5046	0.95613
10	0	10	1.00	0.4014	11.4014	0.96479
15	0	10	1.00	0.4008	11.4008	0.96484
20	0	10	1.00	0.4008	11.4008	0.96485
25	0	10	1.00	0.4008	11.4008	0.96485
30	0	10	1.00	0.4008	11.4008	0.96485
40	0	10	1.00	0.4008	11.4008	0.96485
50	0	10	1.00	0.4008	11.4008	0.96485
75	0	10	1.00	0.4008	11.4008	0.96485
100	0	10	1.00	0.4008	11.4008	0.96485
150	0	10	1.00	0.4008	11.4008	0.96485
200	0	10	1.00	0.4008	11.4008	0.96485

Table 2: Estimates when every sample is identical

one of the samples has 11 observations, one of which is not observed in any other sample, so that $n_1 = 1$ and $n = 11$. As before, both estimates can be written strictly in terms of y. The relevant equation for Cuthbert's method is $\frac{(N-10)^y}{N^{y-1}} = N - 10$, which only has one viable solution independent of y: $N = 10$.[8] That is, the estimate of the population size is independent of the number of samples taken. By contrast, and as may be seen in table 2, the estimate of the total repertoire size produced by BBC's method never quite dips down to 11, although it does start off quite close and rapidly approaches what seems to be an asymptote around 11.4.

Whereas the failure of Cuthbert's method in the first thought-experiment was desirable, the failure of BBC's method in this scenario must be seen as a detriment. If each one of a collection of samples contains exactly the same 10 entities, the only samples that could be said to belong to the population in question are, tautologically, precisely those that only contain these 10 entities.[9] There would be no call to speculate about the existence of unknown entities; indeed, to do so might even be irresponsible. The only responsible claim is not that we are in no position to say what the limits of the population might be, but rather that, to the best of our knowledge, the population consists of exactly those 10 entities that show up in every sample. Thus, here too Cuthbert's method actually encourages greater honesty about what statistical knowledge really tells us.

II.3 In a world made of silicon and electrons...

For the third thought-experiment, let us now suppose that there exists a population of 10,000 entities. Given a collection of samples and asked to use them to estimate the true population

[8]Technically, if y is odd then $N = 5$ is also a "solution" to this equation, but this is obviously nonsense—N cannot be less than n.

[9]This is not to say that it is impossible to continuously generate identical random samples from a larger population; rather, it is to say that this technicality of non-zero probability does not negate the underlying methodological point.

$x_k, k \in \{1, 2, \ldots y\}$	y	Cuthbert		BBC	
		$\mu(\hat{N})$	$\sigma^2(\hat{N})$	$\mu(\hat{N})$	$\sigma^2(\hat{N})$
10	20	13774	36744260.00	278	4.77
	50	9916	5159771.00	680	25.53
	100	9967	1259836.00	1323	67.90
	150	10283	866699.40	1934	218.49
	250	9871	228336.40	3037	445.92
25	20	12255	25274670.00	683	31.55
	50	10245	1062180.00	1635	131.45
	100	9988	370856.30	3044	664.80
	150	10016	80882.41	4266	588.44
	250	9966	21631.68	6216	891.54
50	20	10467	3563189.00	1328	118.67
	50	9955	162751.10	3047	288.26
	100	10072	63474.99	5333	1230.79
	150	9985	22979.22	7005	1667.65
	250	10006	8807.46	9163	2389.19
100	20	9946	678820.20	2512	530.82
	50	9968	50028.42	5330	995.47
	100	9975	12429.22	8256	2017.11
	150	9985	2893.01	9820	1255.36
	250	9990	1109.48	11028	1299.44

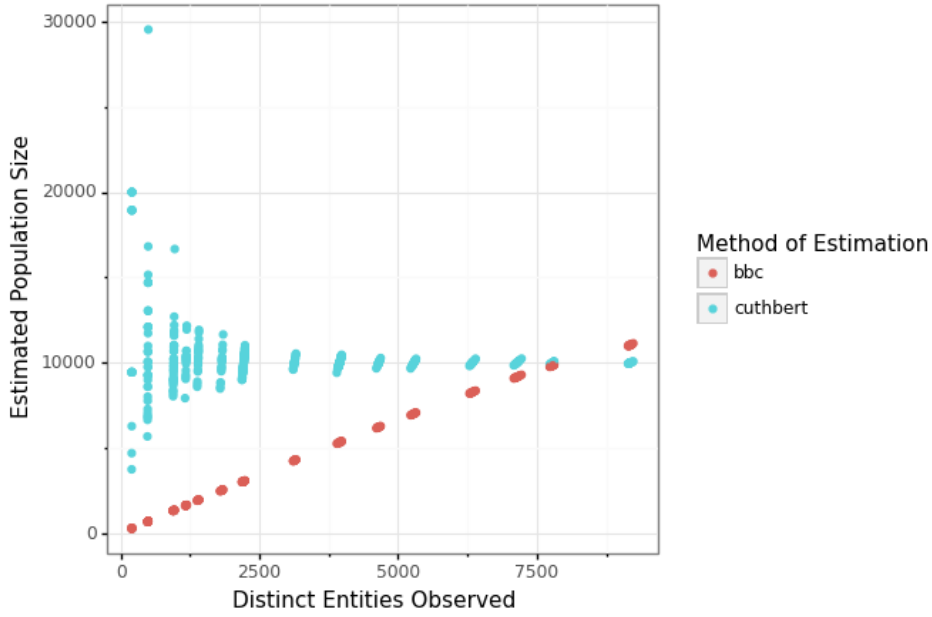
Table 3: Simulations given a variety of sample sizes and numbers

size, how should we expect each method to perform? This is a thorny problem to tackle analytically, but it can be simulated relatively easily; `random_simulation.py`, included as part of the GitHub repository described above, exposes a function that yields both methods' estimates.[10] In order to responsibly assess each method's performance, we should evaluate their performance on a variety of sample sizes and with a variety of numbers of samples. For present purposes I have run 20 random simulations using 20 starting conditions: 20, 50, 100, 150, and 250 uniformly sized samples each containing 10, 25, 50, or 100 entities each. Figure 1 plots the number of distinct entities observed (i.e., the value of n) against the two methods' estimated population sizes (i.e., the value of $\hat{N}$) for each simulation, and table 3 gives the mean and variance of $\hat{N}$ for each set of conditions (each pairing of y and x_k).

To put it bluntly, BBC's method does not emerge from this experiment looking particularly good. Not only does Cuthbert's method consistently produce more accurate estimates of the true population size, but it becomes more accurate the more information it has—the variance of the estimates it yields decreases as the number of entities observed (and thus the proportion of entities observed) increases. In other words, it has all the markings of a well-behaved statistical method. BBC's method, on the other hand, produces generally inaccurate estimates whose variance seems actually to *increase* as a function of the number

[10]The analysis undertaken here is reproduced in the Jupyter Notebook, `sim_sandbox.ipynb`, also included in the GitHub repository described above.

Figure 1: Plot of number of distinct entities observed against estimated population size



of entities observed. Figure 1 paints an even starker picture; while the estimate produced by Cuthbert's method cluster around the true population size, the estimates produced by BBC's method seem to increase linearly with the number of distinct entities observed. Indeed, a linear regression on these simulated BBC estimates yielded an R^2 of 0.9966, and the resultant model suggests that estimates produced by BBC's estimates will tend to yield an estimate of p in the neighborhood of 0.78. The main takeaway here is that BBC's method is over-confident and inaccurate, whereas Cuthbert's method is generally accurate and reasonably forthright about the uncertainties in which it traffics.

*

Perhaps my opinions were already apparent in the theoretical discussion that opened this section, but they should now be perfectly perspicuous. Not only does Cuthbert's method arguably stand on more solid theoretical ground than the alternative, but it also outperforms BBC's method in each of the three statistical thought-experiments presented here. Thus, for the remainder of the present article I take Cuthbert's method as my preferred method.

III Fifteenth-Century Vernacular Polyphony

In this section I extend Cuthbert's method to make it more transparently reflexive, and then apply this extended version to estimate survival rates for several subcorpora of fifteenth-century vernacular polyphony. More specifically, the extension I propose involves adapting cross-validation to produce distributions of "corrected estimates" for N, as a means of clarifying the effects of non-random sampling from a population. After all, as mentioned above, individual fifteenth-century song manuscripts are not perfectly random samples of the entire repertoire of fifteenth-century literate polyphony. My proposed extension helps to pin

down the degree to which the actual collection of possibly non-random samples that we do have (here, the corpus of manuscripts) behaves, as a whole, in ways that we would expect a collection of truly random samples to behave.

First, we must generate an "uncorrected estimate" of N (the result of the first stage of Cuthbert's method described in section I.1), which I call $\hat{N}_0$. Following this, we perform a number of cross-validation experiments (the second stage of Cuthbert's method as described in section I.1), recording for each experiment the true number of "newly lost" entities in the validation set, n_{lost}, and the modeled number of "newly lost" entities in the simulated set, $\hat{n}_{lost}$. The precise number of cross-validation experiments and their parameters (the proportion of the population held for validation, etc.) are not particularly important, although a sufficient number of experiments should be performed so that a reasonably robust sense of their distribution can be obtained. The "error factor" associated with cross-validation experiment i, ε_i, is then either the extent by which the true number exceeds the simulated number relative to the smaller quantity, or simply 1 if the simulated number is greater than the true number: $\varepsilon_i = 1 + \max\left(\frac{n_{lost}-\hat{n}_{lost}}{\hat{n}_{lost}}, 0\right)$. The reason for ensuring that $\varepsilon_i \geq 1$ is to make sure that cross-validation only increases our estimates of N (or equivalently, that it only decreases our estimates of p), which is to say that cross-validation should only ever inspire greater caution about claims of omniscience and should never encourage them. The "corrected estimate" for cross-validation experiment i, $\hat{N}_i$, is then:

$$\hat{N}_i = n + \varepsilon_i\left(\hat{N}_0 - n\right) \tag{4}$$

The resultant distribution of $\{\hat{N}_i\}$ across all cross-validation experiments (i.e., all values of i) can then indicate something about the stability of $\hat{N}$ for a given population—which is to say, the sensitivity of this estimate to non-randomness in the samples. More specifically, while the definition of the error factors ensures that most (if not all) distributions of $\{\hat{N}_i\}$ will exhibit some leftward skew, the relative severity of this skew can still indicate whether the estimate is comparatively stable or unstable.

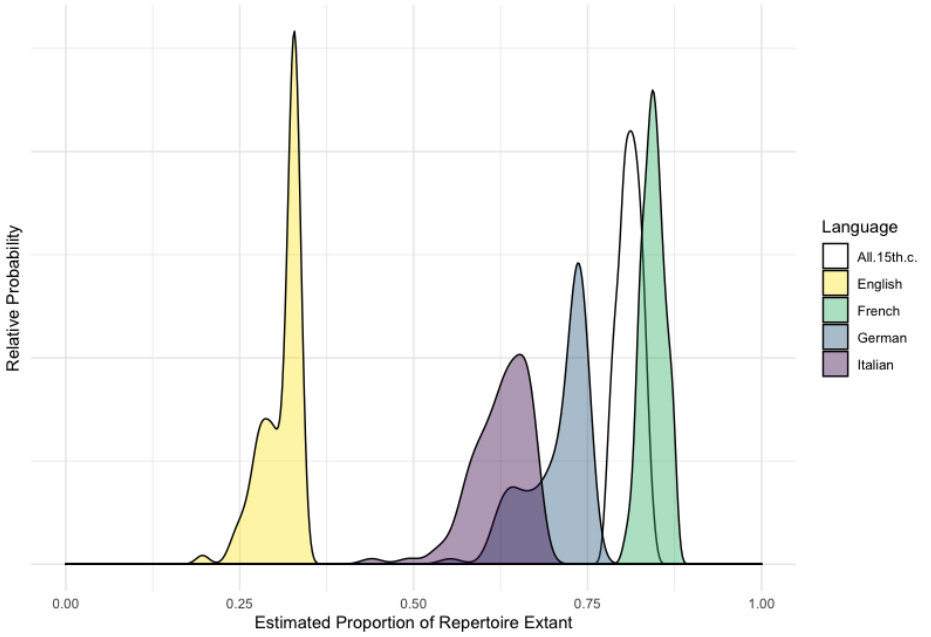
To see how this can work in practice, let us turn to fifteenth-century vernacular polyphony, or rather to the information about that corpus of music encoded into the Digital Index of Late Medieval Song (DILMS).[11] This database aspires to index every extant copy of every song written down in fifteenth-century Europe, and presently indexes approximately 5800 copies of 2200 songs. Since DILMS includes locational information about each of these copies (which is to say, information about the manuscripts into which they are copied), a simple query can yield all of the information needed to apply my extension of Cuthbert's method to this repertoire (that is, the value n and the set of values $\{x_k\}, k \in \{1, 2, \ldots, y\}$). Moreover, since DILMS collates a variety of information about the copies it indexes, including the language(s) in which the text accompanying the music is written and date ranges in which each song manuscript may have been produced, it is also possible to split this musical corpus into a number of non-disjoint subcorpora delimited by language and/or chronology.[12] With this

[11] These calculations were performed on the contents of DILMS as of March 9, 2020.

[12] See [5, Ch. 2] for a full description of DILMS, including a longer explanation of why these subcorpora are not disjoint—e.g., why a "song" may appear in both the "Francophone" and "Anglophone" repertoires, or even in both chronologically differentiated "Francophone" repertoires.

information in hand, I was able to generate distributions containing 100 corrected estimates of the original sizes of each of these subcorpora. Table 4 gives five-number summaries (the minimum, first quartile, median, third quartile, and maximum estimates) of the distributions of these corrected estimates for the original "population" size of several linguistically differentiated bodies of music, as well as those of two chronologically differentiated subsegments of Francophone song. Note that, since subscripts refer to the error measurements, and since $\hat{p}$ shrinks as ε and $\hat{N}$ grow, $\hat{p}_{max}$ will actually be the *smallest* value of p in a given distribution. In addition to the table, figure 2 shows a density plot of $\hat{p}$ for the linguistically differentiated subcorpora.

Figure 2: Density plot of $\hat{p}$ across linguistically differentiated corpora of 15^{th}-century song



Even taking the most cautious of these estimates, the surviving corpus of manuscripts would appear to transmit more than three-quarters of the fifteenth-century song repertoire alluded to by the contents of DILMS. What is more, the relative symmetry and slenderness of the void curve in figure 2 suggest that these estimates of the total repertoire's size are relatively robust—they are unlikely to be seriously impacted by the discovery of new materials. Especially in comparison with Buringh's single-digit-percentage estimates of manuscript survival, things would seem to be not nearly as bleak as they might. However, it is important to remember the assumptions that stand behind these estimates, and the limitations of interpretation that necessarily ensue. In particular, these estimates can only aspire to account for now-lost musical materials that were sufficiently similar to what presently survives that we might be able to imagine what they were like. There can be no accounting for songs that would completely upend our conception of the fifteenth-century song repertoire. And yet there can be no doubt that such music existed, and moreover that manuscripts containing such music existed. The early fifteenth-century Cypriot-French manuscript Turin J.II.9, which has no known concordances with any other manuscript and whose contents are stylistically quite unlike almost any other music that we know from the period, reminds us of

Subcorpus	n	ε_{max}		ε_{Q3}		ε_{med}		ε_{Q1}		ε_{min}	
		$\hat{N}_{max}$	$\hat{p}_{max}$	$\hat{N}_{Q3}$	$\hat{p}_{Q3}$	$\hat{N}_{med}$	$\hat{p}_{med}$	$\hat{N}_{Q1}$	$\hat{p}_{Q1}$	$\hat{N}_{min}$	$\hat{p}_{min}$
All	2185	2807	(0.7784)	2731	(0.8001)	2694	(0.8111)	2655	(0.823)	2590	(0.8436)
Francophone	1347	1674	(0.8047)	1618	(0.8325)	1597	(0.8435)	1575	(0.8552)	1544	(0.8724)
Fr. pre-1460	*611*	*1271*	*(0.4807)*	*1069*	*(0.5716)*	*1047*	*(0.5836)*	*1012*	*(0.6038)*	*1012*	*(0.6038)*
Fr. post-1460	*940*	*1100*	*(0.8545)*	*1057*	*(0.8893)*	*1049*	*(0.8961)*	*1043*	*(0.9012)*	*1026*	*(0.9162)*
Germanophone	329	595	(0.5529)	486	(0.677)	450	(0.7311)	445	(0.7393)	445	(0.7393)
Italophone	215	489	(0.4397)	358	(0.6006)	339	(0.6342)	326	(0.6595)	321	(0.6698)
Anglophone	47	239	(0.1967)	160	(0.2938)	143	(0.3287)	143	(0.3287)	143	(0.3287)

Table 4: Summaries of distributions of estimates of survival rates for several segments of 15^{th}-century vernacular song

that . This being the case, it is highly unlikely that 77–84% of the literate fifteenth-century song repertoire actually survives; the real number is certainly much lower. This raises a rather pointed question: what, then, is the use of these estimates at all?

While there are many possible answers to this question, I have found it useful to shift focus away from the individual distributions of estimates and toward relative differences between the distributions. For it is within those differences that we can start to tease historical and musicological meaning out of these numbers. For a start, consider the striking difference between the values of $\hat{p}$ for the Francophone repertoire ($\hat{p}_{med} = 0.8435$) and for every other linguistically determined subcorpus ($\hat{p}_{med} = 0.7311, 0.6342$, and 0.3287 for German-, Italian-, and English-texted repertoire, respectively). Whatever relation obtains between these estimates and the "real" values of p, they still strongly suggest that Francophone repertoire has been significantly more likely to survive than texted repertoire in any other language. This then easily lends itself to historical interpretation. Perhaps the outsized survival of Francophone repertoire is explicable with recourse to a combination of the generally high status of French within courtly and commercial circles across Western Europe and the outsized importance of Francophone regions for manuscript production during this period [6]? But then again, each of these linguistic buckets also invites its own contemplation, and it could be a mistake to explain away the lower survival rates of the other segments as a mere result of their being "not French." For instance, the lower survival rate of Italophone songs might be best attributable to a comparatively large emphasis on non-literate or para-literate forms of music-making in the Italian peninsula during this period [7, 8]. And the frankly abysmal survival rates for English-texted music seem likely to be due to some combination of the oft-lamented paucity of surviving music manuscripts from fifteenth-century Britain and the relatively high prestige of French among those members of Anglophone society likely to be responsible for commissioning a book of songs in the first place [9, 10]. As for the German-language subcorpus, the structure of DILMS and the nature of the selection criteria ensure both that it excludes a rich tradition of producing Latin sacred contrafacts of French secular songs throughout German-speaking regions during this period, and that it conceals within itself an equally rich tradition of German-language contrafaction (both sacred and secular) of music from all over Western and Central Europe. That is, none of these linguistic categories is so cleanly separable from broader musical practices and trends as might be computationally or statistically desirable, and yet there is still meaning to be found here.

But perhaps these readings of the distributions of $\hat{p}$ leave you unsatisfied. Perhaps you would object that latent uncertainties about the "unknown unknowns" make this kind of historical interpretation little better than idle speculation. This is not a wholly unreasonable position. Even so, these numbers can still tell us something about how we, as twenty-first-century musicologists (or statisticians) encounter the fifteenth-century song repertoire, and about what the structure of those encounters is like. With a plethora of editions, facsimiles, recordings, and writings available online and in many hundreds of libraries, a very different sense of "fifteenth-century song" is available to us than would have been available at most other points in history, including the fifteenth century itself. The kind of intimate knowledge we have of disparate repertoires and manuscripts would have been unthinkable to the people we study. But our knowledge nonetheless remains structured by the patterns of *apparent* survival at which the plotted distributions of $\hat{p}$ gesture. Even with all of those resources at our fingertips, the Francophone and Anglophone fifteenth-century repertoires feel qualitatively

different from each other, and have managed to attract very different kinds of research. Seeing the vastly different values of $\hat{p}$ for those repertoires gives us a way to understand why, for instance, scholarship on fifteenth-century Anglophone song has been so consistently obsessed with hunting for "hidden" specimens of English song in the form of contrafacts. That is, even if you don't believe that these numbers tell us anything about historical realities, it is still hard to deny that they tell us something about the enterprise of history itself.

For another example of this duality, let's turn to the two chronologically differentiated segments of Francophone song (pre- and post-1460). While the distributions of $\hat{p}$ for these repertoires are quite different from each other, the corresponding distributions of $\hat{N}$ are quite similar. The two $\hat{N}_{med}$ estimates are almost exactly the same, at 1047 for pre-1460 and 1049 for post-1460, even though there are far fewer surviving members of the former segment than there are of the latter (611 vs. 940). Historically, this might suggest both that the production of written polyphonic song proceeded at a relatively constant rate throughout the fifteenth century (with perhaps a slight uptick to account for the earlier segment?s larger catchment period), and that the propensity for literate polyphonic song to survive into the present day increased significantly as the century drew to a close (p_{med} is only 0.5836 for the earlier segment, whereas it jumps to 0.8961 for the later segment).[13] That said, since the distribution of $\hat{p}$ is wider for the earlier segment than it is for the later segment, any estimate of how much French-texted repertoire survives from earlier in the century must be less secure than a similar estimate for later in the century. In terms of our engagement with the music, this time we are confronted with a confounding equality. A relatively constant long-term rate of musical production is not something generally associated with fifteenth-century musicking—imagine Binchois and Agricola discussed in terms of long-term constancy! And yet, the estimates presented here might cause us to question why that is, to begin to imagine how we might begin to approach the history of fifteenth-century song with a more egalitarian perspective.

*

Thought-experiments may have a certain artificial elegance, but real data from the real world always throw up substantial challenges to statistical inference. In discussing the results of my experiments with DILMS, I have tried to be attentive to the very real limitations of what the methods of estimation under discussion here can really tell us. And as with a great deal of statistically grounded research, the sticking points coalesce around the interstices between the phenomena under investigation, their remediation into computationally tractable objects, the assumptions of the statistical methods being employed, and the results that those methods yield. Only in paying close attention to how those chains of mediations function can we honestly articulate what our estimates of $\hat{p}$ or $\hat{N}$ actually mean in any given situation. Said another way, numbers are only useful insofar as you understand what they are counting.

[13] Here "production" would refer to the actual fashioning of material musical inscriptions, rather than the abstract composition of songs.

IV Conclusion

Musicological study of the European Middle Ages will never be completely free of anxieties about the question of lost music. But as Cuthbert first suggested a decade ago, the careful application of statistical methods can help us shift our perspective on this question and others like it, revealing our eyes to be half-open rather than half-closed. In particular, the extension of Cuthbert's probabilistic method of estimation that I propose here allows us to generate distributions of estimates for the original size of musical repertoires that are now only partially extant. In the realm of fifteenth-century vernacular polyphony more specifically, it allows us to estimate that, at least when it comes to the music that we might be able to imagine based on what we have, upwards of three-quarters of what was once written down still survives. But just as importantly, careful attention to the instability of population estimates across various segments of that repertoire yields a way of thinking through inequities in its historical production and modern reception alike. Cuthbert may have shown us a way to assuage our anxieties about the size of hidden masses of lost music—the submerged portion of the looming music-historical iceberg—but when it comes to the analytic and historical possibilities opened up by his methods, we have barely begun to scratch the surface. There will assuredly be many more musicological icebergs to survey and to contemplate in the years to come.

References

[1] Cuthbert, Michael Scott. 2009. "Tipping the Iceberg: Missing Italian Polyphony from the Age of Schism," *Musica Disciplina* 54: 39–74.

[2] Buringh, Eltjo. 2011. *Medieval Manuscript Production in the Latin West: Explorations with a Global Database* (Leiden: Brill).

[3] Boneh, Shahar, Arnon Boneh, and R. J. Caron. 1998. "Estimating the Prediction Function and the Number of Unseen Species in Sampling with Replacement," *Journal of the American Statistical Association* 93: 372–79.

[4] Robbins, H.E. 1968. "Estimating the Total Probability of the Unobserved Outcomes of a Random Experiment," *The Annals of Mathematical Statistics* 39: 256–57.

[5] Watson, William. 2020. "Circulating Song from the Century before Print," PhD diss., Yale University.

[6] Fallows, David. 1989. "French as a Courtly Language in Fifteenth-Century Italy: The Musical Evidence," *Renaissance Studies* 3: 429–41.

[7] Elmi, Elizabeth. 2019. "Singing Lyric among Local Aristocratic Networks in the Aragonese-Ruled Kingdom of Naples: Aesthetic and Political Meaning in the Written Records of an Oral Practice," PhD diss., Indiana University.

[8] Rodgers, Mark. 2018. "Renaissance Formalisms in the Cultural Archive of Tonality," PhD diss., Yale University.

[9] Colton, Lisa. 2016. *Angel Song: Medieval English Song in History* (London: Routledge).

[10] Fallows, David. 1977. "English Song Repertoires of the Mid-Fifteenth Century," *Proceedings of the Royal Musical Association* 103: 61–79.