Estimating Omissions from Searches

Anthony J. Webster* and Richard Kemp

United Kingdom Atomic Energy Authority,

Culham Science Centre, Abingdon, Oxon, OX14 3DB.

Published in:

The American Statistician,
Volume 67, Issue 2, pages 82-89, 2013.

Abstract

The mark-recapture method was devised by Petersen in 1896 to estimate the number of fish migrating into the Limfjord, and independently by Lincoln in 1930 to estimate waterfowl abundance. The technique can be applied to any search for a finite number of items by two or more people or agents, allowing the number of searched-for items to be estimated. This ubiquitous problem appears in fields from ecology and epidemiology, through to mathematics, social sciences, and computing. Here we exactly calculate the moments of the hypergeometric distribution associated with this long-standing problem, confirming that widely used estimates conjectured in 1951 are often too small. Our Bayesian approach highlights how different search strategies will modify the estimates. The estimates are applied to several examples. For some published applications substantial errors are found to result from using the Chapman or Lincoln-Petersen estimates.

Keywords: capture-recapture; hypergeometric distribution; Lincoln-Petersen; Mark-recapture; PRISMA; systematic reviews

*email: anthony.webster@ccfe.ac.uk

1 Introduction

If a finite set is searched by two or more people it is possible to estimate how many of the searched-for items have been missed. The simple Lincoln-Petersen estimate was independently developed by Petersen (1896) to estimate fish numbers migrating between the German sea and the Limfjord, and by Lincoln (1930) to estimate waterfowl abundance. The technique has rapidly grown in popularity since a more rigorous treatment by Chapman (1951), especially in the context of ecological census techniques (Seber 1982, Sutherland 2006) and epidemiology (Hook and Regal 1995). Our interest arose from the technique's application to assess the accuracy of a literature search. In 1938 such a literature search led to the re-discovery of Alexander Fleming's papers on penicillin (Masters 1946, Lax 2004), and penicillin's subsequent development. Today literature searches are a valued method for identifying and appraising evidence, particularly in evidence-based healthcare (Sackett et al. 1996). Reviews often search thousands of papers, and standardised guidelines have developed for reporting search terms and the databases used (Liberati et al. 2009, Higgins & Green 2011). Common practice involves an electronic search to retrieve hundreds or even thousands of potentially relevant articles, that are subsequently searched by the authors for pertinent material. Inevitably, even if multiple authors search the database, human error may cause some papers to be erroneously missed at this stage, leading to a less comprehensive review (Edwards et al. 2002). The Lincoln-Petersen estimator has previously been used to assess the completeness of medical databases (Spoor et al. 1996, Bennett et al. 2004, Poorolajal et al. 2010), and to provide "stopping rules" to help determine when searches are complete (Kastner et al. 2009, Booth 2010); surprisingly, standard practice does not include an estimate for the number of papers unintentionally omitted by a search.

Here we derive some simple but rigorous results for estimating the number of items missed from a search, including exact expressions for the average, standard deviation, and skewness. They correct a widely used conjecture from Chapman's 1951 paper and a subsequent widely used approximation for the variance. Despite their extensive use (Seber 1982, Hook and Regal 1995, Sutherland 2006), we confirm the suggestion (García-Pelayo 2006) that previous conjectured and approximated estimates can be inaccurate for many cases of interest, including assessing the

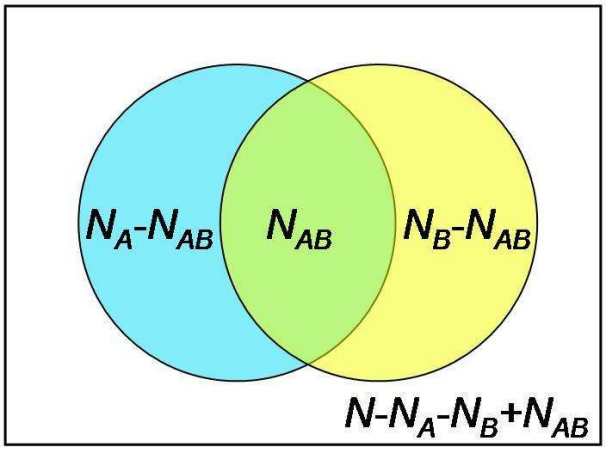

Figure 1: The total number of papers found (N_f) equals the number found by A (N_A), plus the number found by B (N_B), minus the number found by both A and B (N_{AB}), that have been counted twice.

accuracy of literature searches.

The problem is as follows. Authors A and B each separately search a given set of references for relevant articles. (It is assumed that after agreement by both authors, papers that are included are definitely relevant.) The result is that N_A and N_B articles are found by authors A and B respectively with N_{AB} of those found by both authors. If we assume all papers are equally likely to be found, then a simple estimate can be made as follows. Taking N as the total number of papers searched for, and taking probabilities p_A, p_B, and p_{AB} for A, B, and both (A and B) finding N_A, N_B, and N_{AB} papers respectively, then we can estimate p_A, p_B, and p_{AB}, from

$$\begin{aligned} p_A &\approx \tfrac{N_A}{N} \\ p_B &\approx \tfrac{N_B}{N} \\ p_{AB} &\approx \tfrac{N_{AB}}{N} \end{aligned} \tag{1}$$

Because the probability p_{AB} of a paper being found by both authors is $p_{AB} = p_A \times p_B$, we can combine and solve (1) for N, giving an estimate for N as

$$N \approx \frac{N_A N_B}{N_{AB}} \tag{2}$$

The number of papers missed, X, is then estimated to be $X = N - N_f$, where $N_f = N_A +$

$N_B - N_{AB}$ is the total number of different papers found by both authors (figure 1), finding after a little algebra,

$$X \approx \frac{(N_A - N_{AB})(N_B - N_{AB})}{N_{AB}} \tag{3}$$

Equations (2) and (3) are often reasonable estimates if the numbers involved are large. However these estimates are clearly misleading if $N_{AB} = N_A$, N_B, or is zero: for the former cases because there can be papers that both authors have missed (although the estimate suggests not); and for the latter case because an infinite estimate is inconsistent with searching a finite set. More importantly, there is no indication for the accuracy of the estimate, so used in isolation it is impossible to know whether it is reasonable or not. Improved estimates are given later by (19), (20), (23), and (24); the need for them and their derivation is explained in the following sections. The key assumption underlying all of these estimates is that all items are equally likely to be found. As is discussed at the end of Section 3, when this assumption is true or a reasonable approximation, then the estimates can be used.

The paper proceeds as follows. Section 2 uses a Bayesian approach to allow a rigorous mathematical derivation of the probability density function for the number of items missed. Section 3 considers the calculation of its moments. "Exact estimates", refer to exactly calculated moments of the distribution. "Approximate estimates", refer to approximations for the moments, usually found by expanding about the distribution's maximum. Consequently approximated averages are often close to the "most probable" estimate, where the distribution is a maximum. Section 4 comments on the effects of different assumptions on the final answer, and finds explicit prior assumptions for which Chapman's estimate is exactly the most probable estimate. The main result of this paper is to show that the moments can be calculated exactly, subsequently finding that Chapman's extensively used estimate can sometimes be misleading. A recently published example discussed in Section 3 emphasises this.

Throughout the paper we refer to two search procedures. In the example above, both authors searched for all the papers (N) and compared the number found by both (N_{AB}) to estimate $N \approx N_A N_B / N_{AB}$. An alternative approach is for A and B to search for a *predetermined* number of items N_A and N_B respectively, stopping when that number is found, and again using the

number N_{AB} found by both to estimate $N \approx N_A N_B / N_{AB}$. Whereas the former approach is more sensible for a literature search, the latter approach allows a comparatively small sample of animals to provide an estimate for their abundance. Mathematically the difference can be important. If a fixed number of items N_A are searched for, then other than the requirement that $N_A \leq N$, N_A is *independent* of N. In contrast, if all items are searched for then the probability of A finding N_A items is *dependent* on N. Equivalent remarks apply to B. Section 2 uses Bayes theorem to rigorously formulate the problem for both search procedures. Section 3 notes that provided that a large number of items are found, then the moments of both problems are closely related, and the moments of one can be used to closely approximate the moments of the other. The consequences of different search procedures are discussed further in Section 4. Section 5 summarises the paper's conclusions.

2 Bayesian formulation

The shortcomings with (2) and (3) arise from the estimates of $p_A \simeq N_A/N$, $p_B \simeq N_B/N$, and $p_{AB} \simeq N_{AB}/N$. They improve with increasing values of N_A, N_B, and N_{AB}, but are nonetheless estimates. Specifically, if we know the probability p_A of author A finding any given paper (we continue to assume all papers are equally difficult to find), and if we also knew the total number of papers N that the author is searching for, then the probability of author A finding N_A papers is given by the binomial distribution,

$$P(N_A|N, p_A) = \left(\begin{array}{c} N \\ N_A \end{array} \right) p_A^{N_A} \left(1 - p_A\right)^{N - N_A} \tag{4}$$

The expected number of papers to be found is then $\langle N_A \rangle \equiv \sum_{N_A=0}^{N} N_A P(N_A|N, p_A) = p_A N$ (*e.g.* Stirzaker (1994)). Therefore provided $N_A \simeq \langle N_A \rangle$, as on *average* it will be, then the estimates (1) will be reasonable. However, for small numbers in particular it can give misleading results.

Bayes' theorem was first used for mark and recapture estimates by Gaskell & George (1972), and allows a rigorous derivation that avoids these shortcomings. In its modern form Bayes'

theorem states that $P(X|Y)P(Y) = P(Y|X)P(X)$ (Sivia 2005), and allows us to write,

$$P(N|N_A, N_B, N_{AB}) = \frac{P(N_A, N_B, N_{AB}|N)P(N)}{P(N_A, N_B, N_{AB})} \tag{5}$$

Repeatedly using $P(X, Y) = P(X|Y)P(Y)$ (Sivia 2005), and conditional independence of N_A ($N_A \leq N$), N_B ($N_B \leq N$), given N, this expands to give,

$$P(N|N_A, N_B, N_{AB}) = \frac{P(N_{AB}|N_A, N_B, N)P(N_A|N)P(N_B|N)P(N)}{P(N_A, N_B, N_{AB})} \tag{6}$$

Equation (6) gives the probability of there being N papers to find, given that author A has found N_A papers, author B has found N_B papers, and N_{AB} of the papers were found by both authors. $P(N)$ is the (prior) probability of there being N papers to be found given no information about the numbers of papers A and B will find, $P(N_A|N)$ is the probability of finding N_A papers given that there are N papers to be found, and equivalently for $P(N_B|N)$. $P(N_{AB}|N_A, N_B, N)$ is the probability of N_{AB} papers being found by both authors, given that there are N papers to find, and that authors A and B each find N_A and N_B papers respectively.

2.1 Searches for every item

Firstly consider $P(N_A|N)$, and assume that all N items are searched for. Given no prior knowledge of how effective author A may be at finding papers, we take $P(N_A|N)$ to be functionally independent of N_A. Correct normalisation requires that $\sum_{N_A=0}^{N} P(N_A|N) = 1$, giving $P(N_A|N) = 1/(N + 1)$, and similarly for $P(N_B|N)$. Equivalently, assume p_A and N are independent, and take $P(N_A|N, p_A)$ as given by (4). Then use marginalisation (Sivia 2005) to write $P(N_A|N) = \int_0^1 P(N_A|N, p_A)P(p_A)dp_A$, assume a uniform prior for $P(p_A)$, and integrate to find the same answer. This latter approach suggests how the method can be generalised if we relax the assumption that all items are equally likely to be found, through modified forms for $P(N_A|N, p_A)$ and $P(p_A)$. $P(N_{AB}|N_A, N_B, N)$ is the probability of there being N_{AB} items found by both A and B, given only the information that A found N_A items, B found N_B items, and that there are N items to find. This can be calculated by using a metaphor of selecting balls from an urn filled with N white balls. The first author picks N_A balls at random, paints them yellow,

and returns them. The second author picks N_B balls, and N_{AB} is the number of yellow balls the second author has picked. This is a well-known problem (*e.g.* Stirzaker (1994, p. 174)), whose solution is the hypergeometric distribution,

$$P(N_{AB}|N_A, N_B, N) = \frac{N_A!N_B!(N-N_A)!(N-N_B)!}{N_{AB}!(N_A-N_{AB})!(N_B-N_{AB})!N!(N-N_f)!} \tag{7}$$

with $N_{AB} \leq N_A \leq N$ and $N_{AB} \leq N_B \leq N$.

Combining the above (6) and (7) with $P(N_A|N) = P(N_B|N) = 1/(N+1)$ we get,

$$P(N|N_A, N_B, N_{AB}) = \frac{(N-N_A)!(N-N_B)!}{N!(N-N_f)!}\frac{P(N)}{(N+1)^2}C \tag{8}$$

where C is functionally dependent on N_A, N_B, and N_{AB}, but not N, and is most easily found by ensuring that $P(N|N_A, N_B, N_{AB})$ is normalised to 1 after summing over N from the total number of different papers found $N_f = N_A + N_B - N_{AB}$, to ∞. This Bayes' theory approach was used by Zucchini & Channing (1986) to derive a similar result, but without the factors of $P(N_A|N)$ and $P(N_B|N)$ that lead to some differences discussed later. Note that because the sum is over N not N_{AB}, the moments are different to those usually associated with the hypergeometric distribution that involve sums over N_{AB}.

2.2 Searching for a predetermined number of items

If authors A and B search for a fixed number of say 10 items each, so that N_A and N_B are now specified in advance, then the previous derivation is modified slightly. As before, $N_{AB} \leq N_A \leq N$ and $N_{AB} \leq N_B \leq N$, but N and N_{AB} can otherwise be assumed independent of N_A and N_B. If I is some prior information, such as the number of items N_A to be searched for by A and the number of items N_B to be searched for by B, then Bayes' theorem gives (Sivia 2005) $P(X|Y,I) = P(Y|X,I)P(X|I)/P(Y|I)$. Substituting N for X, N_{AB} for Y, and N_A, N_B for I, Bayes' theorem gives,

$$P(N|N_A, N_B, N_{AB}) = \frac{P(N_{AB}|N_A, N_B, N)P(N|N_A, N_B)}{P(N_{AB}|N_A, N_B)} \tag{9}$$

If we make the prior assumption that all values of N (greater than or equal to the largest of N_A and N_B), are equally likely, then $P(N|N_A, N_B)$ will not depend on N. This is an "improper",

i.e. un-normalisable, prior. Strictly $P(N|N_A, N_B)$ should be zero for N bigger than the largest conceivable number of items in the set being searched. With this assumption the factor of $P(N|N_A, N_B)$ is replaced with a constant term, leaving,

$$P(N|N_A, N_B, N_{AB}) = \frac{(N - N_A)!(N - N_B)!}{N!(N - N_f)!}K \tag{10}$$

where, as for C in (8), K is functionally dependent on N_A, N_B, N_{AB}, and is most easily found by ensuring that (10) is correctly normalised. This is the equation whose *approximated* moments have been extensively used (Seber 1982, Sutherland 2006, Hook and Regal 1995) and studied (Chapman 1951, Zucchini & Channing 1986, Seber 1970, Wittes 1972, García-Pelayo 2006), and that we will exactly calculate shortly.

3 Results

Given a suitable choice for $P(N)$ or $P(N|N_A, N_B)$ respectively, (8) and (9) provide the full solution to the problem, allowing numerical values for the average and standard deviation to be calculated by summing from $N = N_f$ to $N = \infty$ for different moments of N. The following section takes the prior $P(N|N_A, N_B)$ as being constant, then calculates the moments of (10) exactly. It also gives an (often excellent) approximation for the moments of (8) when the prior $P(N)$ is constant, and suggests a prior for which the calculated moments are exact. Throughout we will use the statistical physics notation of angled brackets, with e.g. $\langle f(N)\rangle$, to denote the expected value of some function $f(N)$, obtained by averaging over the probability density function for N. Firstly we will calculate moments of the extensively studied (10), and compare these exactly calculated moments with existing approximations. Then we will consider the moments of (8), and use these in some applications.

3.1 The moments of (10)

To calculate the moments we first rewrite (10) in terms of $X = N - N_f$, $X_A = N_A - N_{AB}$, and $X_B = N_B - N_{AB}$, so that $N_f = N_{AB} + X_A + X_B$, and,

$$P(X|X_A, X_B, N_{AB}) = \frac{(X + X_A)!(X + X_B)!}{X!(X + N_f)!} K \tag{11}$$

This gives a probability distribution for the number of papers X that have not been found, with X between 0 and ∞. The moments of (11) are calculated next using a generating function approach. Appendix A contains an alternative (our original) calculation for the moments that is less systematic, but uses simpler mathematical concepts and avoids the use of generating functions. All appendices are available as online supplementary material. The moments of (11) can be written,

$$\langle X^p \rangle = \frac{\left(z\frac{\partial}{\partial z}\right)^p \sum_{X=0}^{\infty} \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!} z^X \Big|_{z=1}}{\sum_{X=0}^{\infty} \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!} z^X \Big|_{z=1}} \tag{12}$$

where the operator $(z\partial/\partial z)^p f(z)|_{z=1}$ represents applying $z \times \partial/\partial z$ to $f(z)$ p times, and then evaluating the result at $z = 1$. The denominator of (12) is simply $1/K$. Equation (12) differs slightly from conventional moment generating functions (Stirzaker 1994), in that the factor of z before $\partial/\partial z$ ensures that repeated application of $(z\partial/\partial z)$ yields the moments, not the "factorial moments" (Stirzaker 1994) that would be obtained by repeatedly applying $(\partial/\partial z)$. The hypergeometric function is defined for $|z| < 1$ by (Arfken 1985),

$${}_2F_1(a+1, b+1, c+1, z) = \frac{c!}{a!b!} \sum_{n=0}^{n=\infty} \frac{(n+a)!(n+b)!}{n!(n+c)!} z^n \tag{13}$$

provided $c \neq 0$, -1, -2, ... It also has an integral representation (Arfken 1985),

$${}_2F_1(a+1, b+1, c+1, z) = \frac{c!}{b!(c-b-1)!} \int_0^1 t^b (1-t)^{c-b-1} (1-tz)^{-a-1} dt \tag{14}$$

that is valid for $|z| < 1$ and $z = 1$ provided $\mathrm{Re}(c+1) > \mathrm{Re}(b+1) > 0$. This standard result (14) is not obviously symmetric with respect to a and b as would be expected from (13), however the expected symmetry is recovered later in (19) and (20) when the calculation is complete. As a consequence of (13), (12) can be written as,

$$\langle X^p \rangle = \frac{\left(z\frac{\partial}{\partial z}\right)^p {}_2F_1(X_A+1, X_B+1, N_f+1, z)\Big|_{z=1}}{{}_2F_1(X_A+1, X_B+1, N_f+1, z)|_{z=1}} \tag{15}$$

with the requirements of $\mathrm{Re}(N_f+1) > \mathrm{Re}(X_B+1) > 0$, clearly satisfied. Equation (15) is easily evaluated. Firstly use (14) to substitute for ${}_2F_1(X_A+1, X_B+1, N_f+1, z)$, then take derivatives, and set $z=1$. The resulting integral can be evaluated using the beta function's identity (Arfken 1985),

$$\int_0^1 t^b(1-t)^{c-b-1}(1-t)^{-a-1}dt = \frac{b!(c-a-b-2)!}{(c-a-1)!} \tag{16}$$

that holds provided $\mathrm{Re}(c+1) > \mathrm{Re}(a+1) + \mathrm{Re}(b+1)$ and $\mathrm{Re}(b+1) > 0$, a requirement that will restrict the values of N_{AB} for which the resulting formulae can be used. This is relatively straightforward because for $t \in (0,1)$ and $|z| \leq 1$, $(1-tz)^{-a-1}$ is continuous with respect to both t and z, and we can bring the derivative with respect to z inside the integral. Then noting that,

$$z\frac{\partial}{\partial z}\left(\frac{1}{(1-tz)}\right)^{a+1} = \frac{a+1}{(1-tz)^{a+2}} - \frac{a+1}{(1-tz)^{a+1}} \tag{17}$$

and applying $z\partial/\partial z$ to (14) p times, we get,

$$\left(z\frac{\partial}{\partial z}\right)^p {}_2F_1(a+1,b+1,c+1,z)\Big|_{z=1} = (a+1)\left(z\frac{\partial}{\partial z}\right)^{p-1}\Big[{}_2F_1(a+2,b+1,c+1,z) - {}_2F_1(a+1,b+1,c+1,z)\Big]\Big|_{z=1} \tag{18}$$

where the use of (17) can be seen by setting $p=1$. Equation (18) can be iterated until the right hand side is a function of ${}_2F_1(a,b,c,1)$, for various a's, b's, and c's, and can be evaluated using (16). For $\langle X \rangle$ this gives the average number of items missed as,

$$\langle X \rangle = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)}{(N_{AB} - 2)} \text{ with } N_{AB} > 2 \tag{19}$$

where X_A, X_B, and N_f have been written in terms of N_A, N_B, and N_{AB}, and $N_{AB} > 2$ arises from the requirement on a, b, and c, that allows (16) to be used. Similarly the standard deviation σ is found from,

$$\sigma^2 = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)(N_A - 1)(N_B - 1)}{(N_{AB} - 2)^2 (N_{AB} - 3)} \text{ with } N_{AB} > 3 \tag{20}$$

Higher moments are also easily calculated and expressions for the skewness and kurtosis are given in the online supplementary material. Equations (19) and (20) are exact under the assumptions for which the prior $P(N|N_A, N_B)$ in (10) does not depend on N. The constraints on the minimum value of N_{AB} for which the expressions hold is a mathematical requirement, and appears to be a requirement for the series to converge. As discussed later, this requirement on N_{AB} can be overcome with a suitably convergent prior distribution $P(N)$. Because both N_A and N_B are greater than or equal to N_{AB}, then $N_{AB} > 2$ will require $N_A > 2$ and $N_B > 2$ also.

3.2 Comparison with Chapman's estimate

Previous approaches have approximated these same average and standard deviation by a combination of conjecture and estimations for the precision and bias (Chapman 1951, Seber 1970, Wittes 1972, Seber 1982). It has been observed (García-Pelayo 2006) that previous (approximate) estimates can be inaccurate for combinations of N_A, N_B, and N_{AB} that cause the hypergeometric distribution to have a 'long tail', for example if $N_A \gg N_B$. These remarks can now be clarified. Chapman's (1951) estimation gives $\langle N \rangle \approx \frac{(N_A+1)(N_B+1)}{(N_{AB}+1)} - 1$, and $\langle X \rangle = \langle N \rangle - N_f$, as,

$$\langle X \rangle \approx \frac{(N_A - N_{AB})(N_B - N_{AB})}{(N_{AB}+1)} \tag{21}$$

Comparing this with (19) (for example by subtracting (21) from (19)), we can see that:

1. it is always less than (19),
2. that this is more pronounced when either or both of $(N_A - N_{AB})$ or $(N_B - N_{AB})$ are large, or when N_{AB} is small, but that conversely,
3. provided neither N_A nor N_B equals N_{AB}, it will give the same (unbiased) estimate if N_{AB} is sufficiently large compared with both $(N_A - N_{AB})$ and $(N_B - N_{AB})$.

Similar remarks apply to the widely used estimate for the variance (Seber 1970), that has

$$\sigma^2 \approx \frac{(N_A+1)(N_B+1)(N_A - N_{AB})(N_B - N_{AB})}{(N_{AB}+1)^2(N_{AB}+2)} \tag{22}$$

and is unbiased for $N_{AB} \gg 1$, but accuracy requires an increasingly large N_{AB} if either $(N_A - N_{AB})$ or $(N_B - N_{AB})$ are small, and in practice it can be inaccurate.

Seber (1970, 1982) has remarked that Chapman's calculations are equivalent to approximating (10) with a Poisson distribution. Appendix B finds this requires both $0 \neq (N_A - N_{AB})/N_{AB} \ll 1$ and $0 \neq (N_B - N_{AB})/N_{AB} \ll 1$, (and implicitly that $N_{AB} \gg 1$). When this is true, the mean of the approximating Poisson distribution coincides with the maximum of (11) with $\langle X \rangle = (N_A - N_{AB})(N_B - N_{AB})/N_f$, and approximates both (19) and (21) (for this limit). Similarly for the variance. In contrast (19) and (20) result from exactly calculating the moments of (11). As noted in Appendix B, this Poisson approximation generalises to the situation studied by (García-Pelayo 2006), in which there are n searches instead of only two.

3.3 The moments of (8)

When all items are searched for by both A and B, the probability distribution for the number of items searched for is given by (8). For the common choice of prior with $P(N)$ constant, Appendix C shows how the moments of (8) can be closely approximated using the moments of (10), and calculates rigorous maximum bounds for the error in the approximation. When $N_f \gg 1$ the error will be small and a good approximation is given by,

$$\langle X \rangle = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)}{N_{AB}} \text{ for } N_{AB} > 0 \tag{23}$$

with an error that is less than $\pm\langle X \rangle/(N_f+1)$. Unfortunately $\sigma^2 = \langle X^2 \rangle - \langle X \rangle^2$ can be arbitrarily small, but the approximation for σ^2 of,

$$\sigma^2 = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)(N_A + 1)(N_B + 1)}{N_{AB}^2\,(N_{AB} - 1)} \text{ with } N_{AB} > 1 \tag{24}$$

has a maximum error that is of order $\langle X^2 \rangle/N_f$. Consequently unless $\langle X^2 \rangle/N_f \ll 1$, (24) is not guaranteed to be a good approximation for σ^2. Often there will be a prior reason to expect that $N \gg 1$. For these cases an alternative approach is to assume the almost constant prior of,

$$P(N) = \kappa \frac{(N+1)}{(N+2)} \tag{25}$$

with κ constant, that may be written as $P(N) = \kappa(1 - 1/(N+2))$, and monotonically increases from $P(0) = \kappa/2$ to $P(\infty) = \kappa$. This prior gives a small bias against low values of N but is approximately constant for larger values of N. For example, $P(N)$ varies by less than ten percent between $N = 8$ and $N = \infty$. For this prior (8) becomes,

$$P(N|N_A, N_B, N_{AB}) = \frac{(N - N_A)!(N - N_B)!}{(N+2)!(N - N_f)!}\kappa \tag{26}$$

Remembering that $N_f = N_A + N_B - N_{AB}$, then rewriting (26) in terms of $(N+2)$, (N_A+2), (N_B+2), and $(N_{AB}+2)$, it will be clear that the change of variables that replaces: $(N+2)$ with N, (N_A+2) with N_A, (N_B+2) with N_B, $(N_{AB}+2)$ with N_{AB}, makes (26) the same form as (10). The condition that $N = N_f$ may be written as $(N+2) = (N_A+2)+(N_B+2)-(N_{AB}+2)$, so after the change of variables the lower limit $N = N_f$ on sums for the moments remains the same. The upper limit of $N = \infty$ is clearly also unchanged. Consequently the exact moments of (26) can be found by replacing N_A with $N_A + 2$, N_B with $N_B + 2$, and N_{AB} with $N_{AB} + 2$, in the exactly calculated moments of (10), with for example (19) and (20) becoming (23) and (24). (An alternative presentation of these remarks can be found in Appendix C.) With the prior (25), (23) and (24) are exact moments of (8), and the error bounds now provide a bound on the maximum possible difference between estimates calculated with this, and with a flat prior. For those cases when it is reasonable to assume this prior, we think it is preferable to explicitly use it along with the exact estimates (23) and (24), in preference to assuming a constant prior and treating (23) and (24) as approximations.

Both (23) and (24) are more similar to the Chapman and Lincoln-Petersen estimates than (19) and (20). This is despite them being approximations to the moments of (8), not (10), that Chapman's calculation is intended to approximate. This might help explain why the discrepancy between Chapman's estimate and (19) is generally overlooked. For many cases of interest the number of items found (N_f) is large, with $N_f \gg 1$, and for these cases (23) provides an accurate estimate for $\langle X \rangle$. Next we consider some examples.

3.4 Examples

When A and B each search for a number of items that is predetermined in advance of their search, then (19) and (20) provide simple estimates for the maximum number of items that could be found by a search for all items, and the precision of the estimate. They are exact moments of (10). When all items are searched for, provided the number of items found (N_f) is much greater than one, then a very good estimate can be made using (23), and if the prior $P(N) = \kappa(N+1)/(N+2)$ is assumed then (23) and (24) are exact moments of (8). Both pairs of estimates can give substantially different estimates to those of Chapman (21) and Lincoln-Petersen (3). For example, Chao et al. (2008) propose a method to combine multiple intersections of lists and the Lincoln-Petersen or Chapman estimator, with the intention of improving the accuracy of epidemiological estimates. The number of items in common between lists is not predetermined, and is anywhere between zero and every item on the shortest list. Their proposed method is illustrated in Section 4 of Chao et al. (2008), and the estimates calculated by the method are given on the top of page 968, where they are calculated from the numbers in their Table 5b using the Chapman and also the Lincoln-Petersen estimate. The results of their calculations are reported in Table 6 on page 968 of their paper, and repeated in part in Table 1. The total number of items (N_f) is much larger than one in all cases, and consequently an accurate estimate is given by (23). An immediate concern is that the Chapman and Lincoln-Petersen estimates are estimators for the moments of (10), that arise from a search procedure for a predetermined number of items, and should not be used. It is a fortunate coincidence that the moments of (8) are closer to the Chapman and Lincoln-Petersen estimates than are the exact moments of (10) that they are intended to approximate. They are also estimates for the most probable population size, and not the expectation of the population size, which can be much larger. For the cases in Table 5b of Chao et al. (2008) where (23) and (24) are defined, we find the revised estimates given in Table 1. Also included are the estimates from Table 6 of Chao et al. (2008), and Seber's estimate for the variance. Our estimates are substantially different, and in some cases N_{AB} is too small to allow them to be used. It is unusual, but not unreasonable, to find distribution functions without a well-defined mean or standard deviation. Without a suitable prior distribution the female list

	N_A	N_B	N_{AB}	$\langle N \rangle$	σ	$\langle N \rangle_C$	$\langle N \rangle_{LP}$	σ_S
Male	323	101	3	11014	7638	8261	10874	3599
Female	21	19	1	438	undefined	219	399	115
Combined	344	120	4	10434	5890	8348	10320	3067

Table 1: Estimates for $\langle N \rangle = N_f + \langle X \rangle$ and σ are calculated using (23), (24), and the numbers in Table 5b of Chao et al. (2008), that are reproduced above as N_A, N_B, and N_{AB}. The estimates from Table 6 of Chao et al. (2008), that use the Chapman ($\langle N \rangle_C$) and Lincoln-Petersen ($\langle N \rangle_{LP}$) estimates for $\langle N \rangle$, and Seber's estimate for the variance (σ_S), are also included. Our estimates, where they are defined, are substantially different to the quoted estimates (Chao et al. 2008) that use the Lincoln-Petersen (3) and Chapman (21) estimates.

for the "shared population" of Chao et al. (2008) will fall into this category. For such cases it is necessary to (explicitly) use a suitable prior if estimates are to be correctly made.

Smaller deviations from the usual Lincoln-Petersen and Chapman estimates are expected when N_{AB} is sufficiently large compared to N_A and N_B. For example, in a recent review by May et al. (2011), there were 177 relevant papers found by author A, 265 papers found by author B, and 171 of these papers found by both authors (K.E. May, private communication). Using (23) and (24), we find $\langle X \rangle \simeq 3.9$ and $\sigma = 2.5$. Therefore whereas 271 papers were found, our estimate gives between 1 and 6 missed papers. Putting it another way, the estimate is that between 97.6% and 99.5% of the papers searched for from within the total sample of just over 8 thousand papers were found. The standard estimates (Chapman 1951, Seber 1970) give $\langle X \rangle = 3.3$ and $\sigma = 2.3$, and are somewhat smaller despite the reasonably large value of $N_{AB} = 171$. Another literature search example (Spoor et al. 1996) found $N_A = 150$, $N_B = 123$, and $N_{AB} = 115$, for which (23) and (24) give $\langle X \rangle = 2.8$ and $\sigma = 2.0$. These compare with the standard estimates (Chapman 1951, Seber 1970) that give, $\langle X \rangle = 2.4$ and $\sigma = 1.8$.

3.5 Limitations of the model

Underlying the calculation is the assumption that all items are equally likely to be found. Clearly there will be cases where some items are more difficult to find. However even in those cases, some (lower bound) estimate for the number of items missed is better than no estimate at all. The method will fail most dramatically if there is a sub-population that is much more difficult to find; it is possible that both searchers could miss all or most of that sub-population, and will overestimate the accuracy of their search. These limitations should be considered before applying these estimates, and when reporting them. If there is a (prior) reason to think the assumptions are inappropriate, one way that modified assumptions can be included is through different priors for $P(N_A|N)$ and $P(N_B|N)$ as was discussed in Section 2.1. In general this will give distribution functions that are most easily calculated numerically.

4 Bayesian corrections and other search procedures

An advantage of the Bayesian approach is that the assumptions are explicit at the outset and the resulting answers are exact, with no additional free parameters. Before concluding we consider two easily evaluated examples that illustrate how different prior assumptions and different search procedures affect the estimates.

4.1 One partial and one comprehensive search

Firstly imagine a situation where one author (*e.g.* A) searches for a fixed number of papers so that $P(N_A|N)$ no longer appears in (8), but the other author (B) searches for as many papers as possible with $P(N_B|N) = 1/(N+1)$, with no prior knowledge of the number of papers searched for other than it being finite ($P(N)$ constant). For this case (10) is modified by the factor $1/N!$ becoming $1/(N+1)!$. In Section 3.3 it was explained how a suitable change of variables could transform (26) into the same form as (10), allowing the moments of (26) to be calculated from those of (10) by a simple change of variables. The same is true here, the change of variables that replaces: $(N+1)$ with N, (N_A+1) with N_A, (N_B+1) with N_B, $(N_{AB}+1)$ with N_{AB}, leads

to the same form of $P(N|N_A, N_B, N_{AB})$ as (10). Similarly to Section 3.3, because the equation $N = N_f = N_A + N_B - N_{AB}$ may be written as $(N+1) = (N_A+1) + (N_B+1) - (N_{AB}+1)$, the lower limit on the range of summation for the moments remains unchanged by the change of variables, as does the $N = \infty$ upper limit. Consequently the exact moments can be found by replacing N_A by $N_A + 1$, N_B by $N_B + 1$, N_{AB} by $N_{AB} + 1$, in (19) and (20), giving,

$$\langle X \rangle = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)}{(N_{AB} - 1)} \text{ with } N_{AB} > 1 \tag{27}$$

and,

$$\sigma^2 = \frac{(N_A - N_{AB} + 1)(N_B - N_{AB} + 1)(N_A)(N_B)}{(N_{AB} - 1)^2 (N_{AB} - 2)} \text{ with } N_{AB} > 2 \tag{28}$$

Interestingly, for this search procedure the standard capture-recapture estimate conjectured by Chapman of $\langle N \rangle \approx \frac{(N_A+1)(N_B+1)}{(N_{AB}+1)} - 1$, approximates the "most probable" value of N, where $P(N|N_A, N_B, N_{AB})$ is a maximum. The maximum can be approximated by setting $P(N|N_A, N_B, N_{AB}) = P(N-1|N_A, N_B, N_{AB})$ and solving for N (Chapman 1951, García-Pelayo 2006). For the stated prior assumptions this gives,

$$\frac{(N - N_A)!(N - N_B)!}{(N+1)!(N - N_A - N_B + N_{AB})!} = \frac{(N - N_A - 1)!(N - N_B - 1)!}{N!(N - N_A - N_B + N_{AB} - 1)!} \tag{29}$$

whose solution for N is exactly Chapman's conjectured estimate. (Strictly this estimate is only an approximation to the most probable value of N: a more precise value can be found using Stirling's approximation for the factorials and differentiating with respect to N to find the maximum of $P(N|N_A, N_B, N_{AB})$.)

4.2 The influence of a proper prior

To illustrate the effect of $P(N)$, consider the normalisable prior $P(N) = \kappa(N+1)/(N+2)(N+3)(N+4) \sim \kappa/N^2$, with κ constant, and let both A and B search for as many items as possible with $P(N_A|N) = P(N_B|N) = 1/(N+1)$. For this example (10) is modified by $1/N!$ becoming $1/(N+4)!$. Following a similar change of variables as discussed above and in Section 3.3, but now with: $(N+4)$ replaced by N, (N_A+4) with N_A, (N_B+4) with N_B, $(N_{AB}+4)$ with N_{AB}, then $P(N|N_A, N_B, N_{AB})$ becomes the same form as in (10). Consequently modified estimates

can be found by substituting N_A with N_A+4, N_B with N_B+4, and N_{AB} with $N_{AB}+4$, in (19) and (20), leading to a reduced estimate for $\langle X \rangle$.

Notice that for this latter example the requirement that $N_{AB} > 3$ in (20) becomes (with N_{AB} replaced by $N_{AB}+4$), $N_{AB} > -1$, and the estimates hold for all N_A, N_B, and N_{AB}. The conclusion is that whereas (19) and (20) can only be used when N_{AB}, N_A, and N_B are sufficiently large (> 3), when all items are searched for (resulting in the extra factor of $1/(N+1)^2$ in $P(N|N_A, N_B, N_{AB})$), the equations apply for a greater range of values. In fact unless N_{AB} is sufficiently large, then estimates can *only* be calculated with a sufficiently convergent (*i.e.* realistic) prior for a given search strategy (such as searching for a fixed number of items, or for all the items). In summary, it is important to ensure that the assumptions upon which any given estimate depends are consistent with the problem being studied.

5 Conclusions

The original purpose of this calculation was to consider two authors A and B searching a finite set of papers for those to include in a literature survey, and to use the number of papers found by authors A (N_A) and B (N_B), along with the number found by *both* authors (N_{AB}), to estimate how accurate the search was. Bayes' theorem is used to rigorously formulate this "mark-recapture" problem for two different search procedures. The first procedure corresponds to A and B searching for all of the items, the second corresponds to A and B each searching for a predetermined number of items, before comparing their results to allow an estimate for N. For the latter case, exact calculations lead to simple formulae for the average number of items missed from the search (19), and the standard deviation (20). The skewness and kurtosis of the probability distribution are given within the appendices in the online supplementary information, and higher moments may be calculated in a similar way.

Equations (19) and (20) are exact moments of the widely-studied probability distribution (10) from Chapman's 1951 paper, which is shown here to result from a procedure in which A and B each search for a *predetermined* number of items. Previous estimates using this distribution have been derived using a combination of conjecture and approximations. Chapman's conjectured

estimate is found (under suitable assumptions) to be an approximation to the most probable value of N. This provides a good approximation to (19) if N is large and both searchers individually find the majority of the items searched for, but is increasingly bad if either searcher finds substantially more (or fewer) items than their partner, which can often be the case.

For many cases such as the literature search application, all items are searched for by both A and B, which leads to a modified probability distribution (8). If a constant prior is assumed then the moments of (8) can be closely approximated provided the number of items found (N_f) is much greater than one, which will very often be the case. When this is the case, an excellent approximation for the number of items missed is given by (23). Alternately if there is a prior reason to think $N \gg 1$, then it is reasonable to use the almost constant prior $P(N) = \kappa(N+1)/(N+2)$, and the calculation for the estimates of (23) and (24) becomes exact. For estimates arising from this search procedure, there is a smaller difference between them and Chapman's estimate (which we have shown here does not apply, and in principle should not be used), but it can still be substantial. We recommend using the improved estimates given by (19), (20), (23), and (24), as is appropriate to the search procedure.

The formulae apply to an enormously wide variety of problems with two independent searches in which the number of items found by searcher A (N_A), searcher B (N_B), and the number found by both (N_{AB}), can be determined. By "independent", we mean that A finding an item does not affect the probability of B finding it (*e.g.* for mark-and-recapture, animals do not become "shy" or "tame" after handling). Finally we caution against an assumption used in the calculation – that all objects searched for are equally likely to be found. This will fail if there is a sub-population that is much more difficult to find, for which case both searchers will appear to have found the majority of items and will over-estimate the accuracy of their search. These issues are beyond the intended scope of this paper. Nonetheless even when the assumption is only approximately true (often the assumption will be good), these improved estimates (19), (20), (23), and (24) will hopefully provide a valuable standard tool for literature searches and more generally.

References

Arfken, G. (1985), *Mathematical Methods for Physicists*, Academic Press Inc., San Diego, CA.

Bennett, D. A., Latham, N. K., Stretton, C. & Anderson, C. S. (2004), Capture-recapture is a potentially useful method for assessing publication bias, *Journal of Clinical Epidemiology* **57**, 349–357. (doi:10.1016/j.jclinepi.2003.09.015)

Booth, A. (2010), How much searching is enough? comprehensive versus optimal retrival for technology assessments, *International Journal of Technology Assessment in Health Care* **26**, 431–435. (doi:10.1017/S0266462310000966)

Chao, A., Pan H.-Y. & Chiang, S.-C. (2008), The Petersen-Lincoln estimator and its extension to estimate the size of a shared population, *Biometrical Journal* **50**, 957-970. (doi:10.1002/bimj.200810482)

Chapman, D. G. (1951), Some properties of the hypergeometric distribution with applications to zoological census, *University of California Public. Stat.* **1**, 131–160.

Edwards, P., Clarke, M., DiGuiseppi, C., Pratap, S., Roberts, I. & Wertz, R. (2002), Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records, *Statist. Med.* **21**, 1635–1640. (doi:10.1002/sim.1190)

García-Pelayo, R. (2006), A Bayesian, combinatorial approach to capture-recapture, *Theoretical Population Biology* **70**, 336–351. (doi:10.1016/j.tpb.2006.06.008)

Gaskell, T. J. & George, B. J. (1972), A Bayesian modification of the Lincoln index, *J. Appl. Ecol.* **9**, 377–384.

Higgins, J. P. T. & Green, S., eds (2011), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*, The Cochrane Collaboration, www.cochrane-handbook.org.

Hook, E. B. & Regal, R. R. (1995), Capture-recapture methods in epidemiology: methods and limitations, *Epidemiologic Reviews* **17**, 243–264.

Kastner, M., Straus, S. E., McKibbon, K. A. & Goldsmith, C. H. (2009), The capture mark-recapture technique can be used as a stopping rule when searching in systematic reviews, *Journal of Clinical Epidemiology* **62**, 149–157. (doi:10.1016/j.jclinepi.2008.06.001)

Lax, E. (2004), *The mould in Dr Florey's coat*, Little, Brown Book Group, London.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J. & Moher, D. (2009), The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration, *British Medical Journal* **339**, b2700. (doi:10.1136/bmj.b2700)

Lincoln, F. C. (1930), Calculating waterfowl abundance on the basis of banding returns, *U.S. Dept. Agric. Circ.* **118**, 1–4.

Masters, D. (1946), *Miracle drug: the inner history of penicillin*, Erye and Spottiswoode, London.

May, K. E., Villar, J., Kirtley, S., Kennedy, S. H. & Becker, C. M. (2011), Endometrial alterations in endometriosis: a systematic review of putative biomarkers, *Human Reproduction Update* **17**(5), 637–653. (doi:10.1093/humupd/dmr013)

Petersen, C. G. J. (1896), The yearly immigration of young plaice into the Limfjord from the German Sea, *Rep. Danish Biol. Sta.* **6**, 1–48.

Poorolajal, J., Haghdoost, A. A., Mahmoodi, M., Majdzadeh, R., Nasseri-Moghaddam, S. & Fotouhi, A. (2010), Capture-recapture method for assessing publication bias, *Journal of Research in Medical Sciences* **15**, 107–115.

Sackett, D. L., Rosenburg, W. M. C., Gray, J. A. M., Haynes, R. B. & Richardson, W. S. (1996), Evidence based medicine: what it is and what it isn't, *British Medical Journal* **312**, 71–72. (doi:10.1136/bmj.312.7023.71)

Seber, G. A. F. (1970), The effects of trap response on tag recapture estimates, *Biometrics* **26**, 13–22.

Seber, G. A. F. (1982), *Estimates of animal abundance*, 2nd edn, Charles Griffin & Company Ltd., London & High Wycombe.

Sivia, D. S. (2005), *Data Analysis: A Bayesian Tutorial*, Oxford University Press, Oxford.

Spoor, P. A., Airey, M., Bennett, C., Greensill, J. & Williams, R. (1996), Use of the capture-recapture technique to evaluate the completeness of systematic literature searches, *British Medical Journal* **313**, 342–343. (doi:10.1136/bmj.313.7053.342)

Stirzaker, D. (1994), *Elementary Probability*, Cambridge University Press, Cambridge.

Sutherland, W. J. (2006), *Ecological census techniques*, 2nd edn, Cambridge University Press, Cambridge, UK.

Wittes, J. T. (1972), On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate, *Biometrics* **28**, 592–597.

Zucchini, W. & Channing, A. (1986), Bayesian estimation of animal abundance in small populations using capture-recapture information, *South African Journal of Science* **82**, 137–140.

Acknowledgements

Thanks to Dr Katie Webster (previously Dr Katie May) for recording and supplying the numbers from the recent literature search described in May et al. (2011), for emphasising the potential use of this technique in literature searches, and for numerous helpful discussions. Thanks to Professor Walter Zucchini for supplying a copy of Zucchini & Channing (1986), and Martin O'Brien for helpful discussions and comments. Thanks also to the editor of The American Statistician, and the associate editor in particular, for numerous helpful comments and suggestions.

A The moments

Here we briefly present our original derivation of the moments of (10), that uses simpler mathematical concepts, but is less conventional and systematic than the generating function approach presented in the main text. Repeating (10) here for convenience, with,

$$P(X|X_A, X_B, N_{AB}) = \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!}K \tag{30}$$

and X between 0 and ∞. Next define,

$$S(X_A, X_B, N_f) = \sum_{X=0}^{\infty} \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!} \tag{31}$$

where we note that $N_f = N_{AB} + X_A + X_B$, and also that $K = 1/S(X_A, X_B, N_f)$. The aim is to express the moments in terms of the function $S(X_A, X_B, N_f)$, evaluate $S(X_A, X_B, N_f)$ using an identity due to Gauss, then combine the results to obtain explicit expressions for the moments in terms of X_A, X_B, and N_f.

Starting with $\langle X \rangle$, notice that,

$$\begin{aligned}
\sum_{X=0}^{\infty} X \frac{(X+X_A)!(X+X_B)!}{X!\big(X+N_f\big)!} &= \sum_{X=1}^{\infty} \frac{X}{X!} \frac{(X-1+X_A+1)!(X-1+X_B+1)!}{\big(X-1+N_f+1\big)!} \\
&= \sum_{X=1}^{\infty} \frac{1}{(X-1)!} \frac{((X-1)+(X_A+1))!((X-1)+(X_B+1))!}{\big((X-1)+(N_f+1)\big)!} \\
&= \sum_{X=0}^{\infty} \frac{(X+X_A+1)!(X+X_B+1)!}{X!\big(X+N_f+1\big)!} \\
&= S(X_A+1, X_B+1, N_f+1)
\end{aligned} \tag{32}$$

Hence,

$$\langle X \rangle = \frac{S(X_A+1, X_B+1, N_f+1)}{S(X_A, X_B, N_f)} \tag{33}$$

Similarly for $\langle X^2 \rangle$,

$$\begin{aligned}
\sum_{X=0}^{\infty} X^2 \frac{(X+X_A)!(X+X_B)!}{X!\big(X+N_f\big)!} &= \sum_{X=1}^{\infty} \frac{X}{X!} (X-1+1) \frac{(X-1+X_A+1)!(X-1+X_B+1)!}{\big(X-1+N_f+1\big)!} \\
&= \sum_{X=1}^{\infty} \frac{((X-1)+1)}{(X-1)!} \frac{((X-1)+(X_A+1))!((X-1)+(X_B+1))!}{\big((X-1)+(N_f+1)\big)!} \\
&= \sum_{X=0}^{\infty} (X+1) \frac{(X+X_A+1)!(X+X_B+1)!}{X!\big(X+N_f+1\big)!}
\end{aligned} \tag{34}$$

Repeating the same trick to remove the factor of X then gives,

$$\langle X^2 \rangle = \frac{S(X_A+2, X_B+2, N_f+2)}{S(X_A, X_B, N_f)} + \frac{S(X_A+1, X_B+1, N_f+1)}{S(X_A, X_B, N_f)} \tag{35}$$

Similarly but with more algebra for the higher order moments, e.g.

$$\langle X^3\rangle = \frac{S(X_A+3,X_B+3,N_f+3)}{S(X_A,X_B,N_f)} + 3\frac{S(X_A+2,X_B+2,N_f+2)}{S(X_A,X_B,N_f)} + \frac{S(X_A+1,X_B+1,N_f+1)}{S(X_A,X_B,N_f)} \tag{36}$$

and,

$$\begin{aligned}\langle X^4\rangle &= \tfrac{S(X_A+4,X_B+4,N_f+4)}{S(X_A,X_B,N_f)} + 6\tfrac{S(X_A+3,X_B+3,N_f+3)}{S(X_A,X_B,N_f)} \\ &\quad +7\tfrac{S(X_A+2,X_B+2,N_f+2)}{S(X_A,X_B,N_f)} + \tfrac{S(X_A+1,X_B+1,N_f+1)}{S(X_A,X_B,N_f)}\end{aligned} \tag{37}$$

To evaluate $S(X_A,X_B,N_f)$, we firstly note that the hypergeometric function has for $|z|<1$ and $c\neq 0,\ -1,\ -2,\ ...$ (Arfken 1985),

$${}_2F_1\left(a+1,b+1,c+1,z\right) = \frac{c!}{a!b!}\sum_{n=0}^{\infty}\frac{(a+n)!(b+n)!}{(c+n)!}\frac{z^n}{n!} \tag{38}$$

For $z=1$ an identity due to Gauss gives (Arfken 1985),

$${}_2F_1\left(a+1,b+1,c+1,1\right) = \frac{\Gamma(c+1)\Gamma(c-a-b-1)}{\Gamma(c-a)\Gamma(c-b)}\ ,\ \mathrm{Re}(c) > \ \mathrm{Re}(a+b)+1 \tag{39}$$

with $c\neq 0,\ -1,\ -2,\ ...$, as above. Equations (38) and (39) may be combined to give (for $z=1$),

$$\sum_{n=0}^{\infty}\frac{(a+n)!(b+n)!}{(c+n)!}\frac{1}{n!} = \frac{a!b!}{c!}\frac{\Gamma(c+1)\Gamma(c-a-b-1)}{\Gamma(c-a)\Gamma(c-b)}\ ,\ \mathrm{Re}(c) > \ \mathrm{Re}(a+b)+1 \tag{40}$$

Therefore with the replacements of $n=X$, $c=N_f$, $a=X_A$, and $b=X_B$ (so that $c=a+b+N_{AB}>(a+b)+1$ for $N_{AB}>1$), we get,

$$\begin{aligned}S(X_A,X_B,N_f) &= \textstyle\sum_{X=0}^{\infty}\frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!} \\ &= X_A!X_B!\tfrac{(N_f-X_A-X_B-2)!}{(N_f-X_A-1)!(N_f-X_B-1)!}\ ,\ N_{AB}>1\end{aligned} \tag{41}$$

Hence substituting into (33) gives,

$$\langle X\rangle = \frac{(X_A+1)(X_B+1)}{(N_f-X_A-X_B-2)} = \frac{(X_A+1)(X_B+1)}{(N_{AB}-2)}\ ,\ N_{AB}>2 \tag{42}$$

where the inequality follows from the requirement that $N_f+1>(X_A+1)+(X_B+1)+1$ with $N_f=N_{AB}+X_A+X_B$. Similarly,

$$\begin{aligned}\langle X^2\rangle &= \frac{(X_A+1)(X_A+2)(X_B+1)(X_B+2)}{(N_{AB}-2)(N_{AB}-3)} \\ &\quad + \frac{(X_A+1)(X_B+1)}{(N_{AB}-2)}\ ,\ \text{with } N_{AB}>3\end{aligned} \tag{43}$$

$$\begin{aligned}\langle X^3\rangle &= \frac{(X_A+1)(X_A+2)(X_A+3)(X_B+1)(X_B+2)(X_B+3)}{(N_{AB}-2)(N_{AB}-3)(N_{AB}-4)} \\ &+ 3\frac{(X_A+1)(X_A+2)(X_B+1)(X_B+2)}{(N_{AB}-2)(N_{AB}-3)} \\ &+ \frac{(X_A+1)(X_B+1)}{(N_{AB}-2)} \text{ , with } N_{AB}>4 \end{aligned} \tag{44}$$

$$\begin{aligned}\langle X^4\rangle &= \frac{(X_A+1)(X_A+2)(X_A+3)(X_A+4)(X_B+1)(X_B+2)(X_B+3)(X_B+4)}{(N_{AB}-2)(N_{AB}-3)(N_{AB}-4)(N_{AB}-5)} \\ &+ 6\frac{(X_A+1)(X_A+2)(X_A+3)(X_B+1)(X_B+2)(X_B+3)}{(N_{AB}-2)(N_{AB}-3)(N_{AB}-4)} \\ &+ 7\frac{(X_A+1)(X_A+2)(X_B+1)(X_B+2)}{(N_{AB}-2)(N_{AB}-3)} \\ &+ \frac{(X_A+1)(X_B+1)}{(N_{AB}-2)} \text{ , with } N_{AB}>5 \end{aligned} \tag{45}$$

These may be used to calculate various statistical quantities. The standard deviation $\sigma = \sqrt{\langle X^2\rangle - \langle X\rangle^2}$, which using (42) and (43), simplifies to give,

$$\sigma = \sqrt{\frac{(X_A+1)(X_B+1)(N_{AB}+X_A-1)(N_{AB}+X_B-1)}{(N_{AB}-2)^2(N_{AB}-3)}} \text{ , } N_{AB}>3 \tag{46}$$

The skewness $\gamma = \langle (X-\langle X\rangle)^3\rangle / \langle X^2\rangle^{3/2}$, which expands to give,

$$\gamma = \frac{\langle X^3\rangle - 3\langle X^2\rangle\langle X\rangle + 2\langle X\rangle^3}{\langle X^2\rangle^{3/2}} \tag{47}$$

and may be evaluated using (42) to (44). The kurtosis is given by $\kappa = \langle (X-\langle X\rangle)^4\rangle / \langle X^2\rangle^2$, which expands to give,

$$\kappa = \frac{\langle X^4\rangle - 4\langle X^3\rangle\langle X\rangle + 6\langle X^2\rangle\langle X\rangle^2 - 3\langle X\rangle^4}{\langle X^2\rangle^2} \tag{48}$$

and may be evaluated using (42) to (45). Replacing $X = N - N_f$, $X_A = N_A - N_{AB}$, and $X_B = N_B - N_{AB}$ in (42) and (46), gives (19) and (20) of the main text.

B Poisson approximation

Starting from (11) in the main text, use the approach of Chapman (1951) and García-Pelayo (2006) to find X for which $P(X|X_A, X_B, N_{AB})$ is maximum, from $P(X^*|X_A, X_B, N_{AB}) =$

$P(X^* - 1|X_A, X_B, N_{AB})$. This gives $X^* = X_A X_B/N_f = (N_A - N_{AB})(N_B - N_{AB})/N_f$. When both $X_A/N_{AB} \ll 1$ and $X_B/N_{AB} \ll 1$, then both $X^* \ll X_A$ and $X^* \ll X_B$, and because $N_f = N_{AB} + X_A + X_B$ is larger than either X_A or X_B then $X^* \ll N_f$ also.

Next note that $(X + X_A)! \equiv X_A! X_A^X \exp\{\sum_{y=1}^{X} \log(1 + y/X_A)\}$, as may be seen from expanding $(X + X_A)!$,

$$\begin{aligned}(X + X_A)! &= (X_A + X)(X_A + X - 1)...(X_A + 1)X_A! \\ &= X_A! \exp\left\{\sum_{y=1}^{X} \log(y + X_A)\right\}\end{aligned} \tag{49}$$

where the last line repeatedly used $AB = \exp(\log(AB)) = \exp(\log(A) + \log(B))$. Then write,

$$\begin{aligned}X_A! \exp\left\{\sum_{y=1}^{X} \log(y + X_A)\right\} &= X_A! \exp\left\{\sum_{y=1}^{X} \log(X_A(1 + y/X_A))\right\} \\ &= X_A! \exp\left\{X \log(X_A) + \sum_{y=1}^{X} \log(1 + y/X_A)\right\} \\ &= X_A! \exp\left\{\log(X_A^X)\right\} \exp\left\{\sum_{y=1}^{X} \log(1 + y/X_A)\right\} \\ &= X_A! {X_A}^X \exp\left\{\sum_{y=1}^{X} \log(1 + y/X_A)\right\}\end{aligned} \tag{50}$$

as originally stated. Similarly expanding $(X + X_B)!$ and $(X + N_f)!$, gives,

$$\begin{aligned}P(X|X_A, X_B, N_f) &= K \frac{X_A! X_B!}{N_f!} \frac{1}{X!} \left(\frac{X_A X_B}{N_f}\right)^X \times \\ &\exp\left\{\sum_{i=1}^{X} \log(1 + i/X_A) + \sum_{j=1}^{X} \log(1 + j/X_B) - \sum_{k=1}^{X} \log(1 + k/N_f)\right\}\end{aligned} \tag{51}$$

The above expression is exact, and can be used as the starting point for a variety of approximations. It is composed of the product of a Poisson distribution ${X^*}^X/X!$ with $X^* = X_A X_B/N_f$, a constant term that ensures (51) is correctly normalised, and an exponential term whose exponent is a function of X. As X becomes small relative to X_A, X_B, and N_f, the exponential's exponent tends to zero, and (51) asymptotes to a Poisson distribution. However, because $X_A < N_f$ and $X_B < N_f$, the exponential term's exponent is a strictly increasing function of X. Consequently a good approximation to (51) by a Poisson distribution is only ever possible over a limited range of X. An approximation with a Poisson distribution to (51) can be found by approximating the exponential term in (51) near $X = X^*$. The rate of change of the exponential's exponent near $X = X^*$ can be estimated by considering the difference in its value between X^* and $(X^* - 1)$, which is simply $\log[(1+X^*/X_A)(1+X^*/X_B)/(1+X^*/N_f)]$. Provided this rate of change is small, then a

Poisson distribution will provide a good approximation near the maximum of (51). If $X^*/X_A \ll 1$ and $X^*/X_B \ll 1$ (implying $X^*/N_f \ll 1$), then $\log[(1+X^*/X_A)(1+X^*/X_B)/(1+X^*/N_f)]$ will be small, and the exponent will be approximately constant near X^*. Therefore if $X^*/X_A \ll 1$ and $X^*/X_B \ll 1$, the Poisson distribution provides a good approximation near the maximum of (51). If a precise and accurate approximation for the moments of (51) only requires a sufficiently precise approximation to (51) near $X = X^*$ (we do not claim to show this here), then the Poisson distribution will provide a good approximation for the moments of (51). These remarks are consistent with the observations in the main text that: (19) is always greater than (21), but provided that $X^*/X_A \ll 1$ and $X^*/X_B \ll 1$ (implying $X^*/N_f \ll 1$), the exact (19) and approximated moments (21), are approximately the same (for a Poisson distribution, $\langle X \rangle = X^*$ and $\sigma^2 = X^*$, *e.g.* see Stirzaker (1994)). The above calculation easily generalises to the case studied by García-Pelayo (2006) with n-persons searching, consequently similar remarks apply to that problem also.

C Relation between (8) and (10)

Here the relationship between (8) and (10) is discussed. Firstly write (8) in terms of $X = N - N_f$, $X_A = N_A - N_{AB}$, $X_B = N_B - N_{AB}$, and $N_f = N_{AB} + X_A + X_B$, to give,

$$P(X|X_A, X_B, N_f) = \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f)!} \frac{P(X+N_f)}{(X+N_f+1)^2} C \tag{52}$$

Throughout this section we will only consider the case where $P(X+N_f) = P(N)$ is constant. The moments of (52) are then,

$$\langle X^p \rangle = \frac{\sum_{X=0}^{\infty} \frac{X^p}{(X+N_f+1)} \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+1)!}}{\sum_{X=0}^{\infty} \frac{1}{(X+N_f+1)} \frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+1)!}} \tag{53}$$

where one of the factors of $1/(X+N_f+1)$ has been incorporated into $1/(X+N_f+1)!$. Note that,

$$\frac{1}{X+N_f+1} > \frac{1}{X+N_f+2} \tag{54}$$

and that,

$$\begin{aligned}\frac{1}{X+N_f+1} &= \frac{1}{X+N_f+2}\left(\frac{1}{1-\frac{1}{X+N_f+2}}\right)\\ &= \frac{1}{X+N_f+2}\sum_{k=0}^{\infty}\left(\frac{1}{X+N_f+2}\right)^k\\ &< \frac{1}{X+N_f+2}\sum_{k=0}^{\infty}\left(\frac{1}{N_f+2}\right)^k = \frac{1}{X+N_f+2}\left(\frac{N_f+2}{N_f+1}\right)\end{aligned} \tag{55}$$

Using these bounds (54) and (55) in the numerators and denominators of (53) as appropriate (with (55) used for the sum in the numerator and (54) for the sum in the denominator to give the upper bound, and vice versa for the lower bound), we find,

$$\begin{aligned}&\frac{1}{\left(\frac{N_f+2}{N_f+1}\right)}\frac{\sum_{X=0}^{\infty}X^p\frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+2)!}}{\sum_{X=0}^{\infty}\frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+2)!}} < \langle X^p\rangle\\ &< \left(\frac{N_f+2}{N_f+1}\right)\frac{\sum_{X=0}^{\infty}X^p\frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+2)!}}{\sum_{X=0}^{\infty}\frac{(X+X_A)!(X+X_B)!}{X!(X+N_f+2)!}}\end{aligned} \tag{56}$$

where the factors of $1/(X+N_f+2)$ have been incorporated into the factors of $1/(X+N_f+2)!$. Using $\langle X^p\rangle_0[N_f+2]$ to refer to moments of (30), but with N_f replaced by N_f+2, or equivalently noting that $N_f = X_A + X_B + N_{AB}$, by replacing N_{AB} by $N_{AB}+2$, keeping X_A and X_B fixed everywhere else. With this notation in (56), and using $-1/(N_f+1) < -1/(N_f+2)$ to make the left hand side of the inequality symmetric with the right, we find,

$$\left(1-\frac{1}{N_f+1}\right)\langle X^p\rangle_0[N_f+2] < \langle X^p\rangle < \left(1+\frac{1}{N_f+1}\right)\langle X^p\rangle_0[N_f+2] \tag{57}$$

Or equivalently,

$$\langle X^p\rangle = \langle X^p\rangle_0[N_f+2]\left(1\pm\frac{1}{N_f+1}\right) \tag{58}$$

where the factor of $\pm 1/(N_f+1)$ gives a maximum error bound. Improved bounds can be found on a case by case basis, by considering $\langle X^p\rangle - \langle X^p\rangle_0[N_f+2]$, simplifying as far as possible, then using (54) and (55) to express the sums in a form that can be evaluated using (41). Returning to (58), if $N_f \gg 1$ then an excellent approximation to $\langle X^p\rangle$ that is correct to within $\pm 100/(N_f+1)$ percent, is given by $\langle X^p\rangle_0[N_f+2]$. This approximation for the moments of (52) is equal to the exact moments of (30) with N_{AB} replaced by $N_{AB}+2$, keeping X_A and X_B fixed. Consequently using (42) we have,

$$\langle X\rangle = \frac{(X_A+1)(X_B+1)}{N_{AB}} \text{ with } N_{AB} > 0 \tag{59}$$

with a maximum error of $\pm\langle X\rangle/(N_f+1)$, which with the substitutions $X_A = N_A - N_{AB}$ and $X_B = N_B - N_{AB}$, is (23) of the main text. Similarly using (42) and (46) an approximation for σ^2 is,

$$\sigma^2 = \frac{(X_A+1)(X_B+1)(N_{AB}+X_A+1)(N_{AB}+X_B+1)}{N_{AB}^2\,(N_{AB}-1)} \text{ with } N_{AB} > 1 \qquad (60)$$

which with the substitutions $X_A = N_A - N_{AB}$ and $X_B = N_B - N_{AB}$, is (24) of the main text. Unfortunately whereas (59) has a maximum error of order $\langle X\rangle/N_f$, which is much less than $\langle X\rangle$ if $N_f \gg 1$, $\sigma^2 = \langle X^2\rangle - \langle X\rangle^2$ can be arbitrarily small, but the maximum possible error remains of order $\langle X^2\rangle/N_f$. Therefore unless $\langle X^2\rangle/N_f \ll 1$, (60) will not be guaranteed to give a good approximation for σ^2. As is noted in the main text, an alternative approach is to use the prior $P(N) = \kappa(N+1)/(N+2)$, for which (59) and (60) are the exactly calculated moments. For that case this calculation gives the maximum difference between the moments with this, and with a prior that is independent of N.