Theoretical Population Biology 92 (2014) 51–54

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

# Within a sample from a population, the distribution of the number of descendants of a subsample's most recent common ancestor

John L. Spouge

*Building 38A, Room 6N603, National Center for Biotechnology Information, Bethesda MD 20894, United States*

ARTICLE INFO

*Article history:*
Received 26 June 2013
Available online 7 December 2013

*Keywords:*
Most recent common ancestor of a subsample
Coalescent theory

ABSTRACT

Sample $n$ individuals uniformly at random from a population, and then sample $m$ individuals uniformly at random from the sample. Consider the most recent common ancestor (MRCA) of the subsample of $m$ individuals. Let the subsample MRCA have $j$ descendants in the sample ($m \leqslant j \leqslant n$). Under a Moran or coalescent model (and therefore under many other models), the probability that $j = n$ is known. In this case, the subsample MRCA is an ancestor of every sampled individual, and the subsample and sample MRCAs are identical. The probability that $j = m$ is also known. In this case, the subsample MRCA is an ancestor of no sampled individual outside the subsample. This article derives the complete distribution of $j$, enabling inferences from the corresponding $p$-value. The text presents hypothetical statistical applications pertinent to taxonomy (the gene flow between Neanderthals and anatomically modern humans) and medicine (the association of genetic markers with disease).



## 1. Introduction

Consider the following hypothetical situation. Within a sample of $n$ individuals, a subsample of $m$ individuals share a morphological character. Upon genetic analysis, the $m$ individuals share some genetic characters with a further $j - m \geq 0$ individuals within the sample. One might desire a $p$-value to test whether $j - m$ is "too small", i.e., to test whether the concentration of the morphological character among individuals with the genetic characters is too excessive to reflect chance alone. This article derives a $p$-value by giving the sampling distribution of $j$. Depending on its context, a small $p$-value might suggest among other possibilities, e.g., that gene flow between the subpopulations represented by the subsample and its complement within the sample is not free (i.e., that the mathematical assumptions underlying the coalescent are violated), or that the genetic characters have a causal influence on the morphological character. The Discussion demonstrates how the $p$-value might be relevant to rejecting the hypothesis of free gene flow between Neanderthals and anatomically modern humans (Krings et al., 1997; Nordborg, 1998; Krings et al., 2000) or to associating a genetic disease or phenotype with a set of DNA markers necessary but not sufficient for it.

To determine the distribution corresponding to the $p$-value, consider Kingman's coalescent (Kingman, 1982a,b), where $n$ individuals are sampled uniformly at random at time $t_0$ from a large population. Kingman examined a haploid population, but coalescent models can also apply to sexual populations (Nordborg, 2004; Pollak, 2004; Wakeley et al., 2012). A pure death process $D_t$ ($t \geq 0$) counts the ancestors of the sample at prior times $t_0 - t$. The process $D_t$ transitions through the states $n \to n-1 \to \cdots \to 2 \to 1$, with the state $D_t = k$ ($k = 2, \ldots, n$) having a sojourn time $\tau_k$ exponentially distributed with parameter $d_k = \frac{1}{2}k(k-1)$, and with the state $D_t = 1$ absorbing.

The sample ancestry can be described using $\mathscr{E}_n$, the set of all equivalence relations on the $n$ individuals. Consider the Markov chain $\mathscr{R}_n \to \mathscr{R}_{n-1} \to \cdots \to \mathscr{R}_2 \to \mathscr{R}_1$, whose state-space is $\mathscr{E}_n$, where $\mathscr{R}_k$ corresponds to having $D_t = k$ ancestors ($k = n, n-1, \ldots, 1$). The variate $\mathscr{R}_k$ partitions the $n$ individuals into $k$ equivalence classes, each equivalence class corresponding to an ancestor and containing the ancestor's descendants at time $t_0$. Define the identity relation $\Delta = \{(i,i) : i = 1, 2, \ldots, n\}$ and the trivial relation $\Theta = \{(i,j) : i, j = 1, 2, \ldots, n\}$. Given $\xi, \eta \in \mathscr{E}_n$, let $\xi \prec \eta$ denote that $\eta$ can be obtained from $\xi$ by combining two equivalence classes in $\xi$, and in fact, $\Delta = \mathscr{R}_n \prec \mathscr{R}_{n-1} \prec \cdots \prec \mathscr{R}_2 \prec \mathscr{R}_1 = \Theta$. The transition probabilities of the Markov chain $\{\mathscr{R}_k\}$ are

$$\mathbb{P}\{\mathscr{R}_{k-1} = \eta | \mathscr{R}_k = \xi\} = \begin{cases} 2/[k(k-1)] & \text{if } \xi \prec \eta \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Kingman shows that if $\xi$ contains $k$ equivalence classes,

$$\mathbb{P}\{\mathscr{R}_k = \xi\} = \frac{(n-k)!k!\,(k-1)!}{n!\,(n-1)!}\lambda_1!\lambda_2!\cdots\lambda_k!, \tag{2}$$

*E-mail addresses:* spouge@ncbi.nlm.nih.gov, spouge@nih.gov.

http://dx.doi.org/10.1016/j.tpb.2013.11.004

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the sizes of the equivalence classes of $\xi$. Eq. (1) implies Eq. (2), so Eq. (2) holds for any model imposing Eq. (1) on the ancestry of a sample, in particular the Moran model (Moran, 1962; Kimura and Crow, 1964; Watterson, 1984; Donnelly and Tavare, 1986a,b) (without mutation), or indeed any model of ancestry approximating a coalescent process closely enough.

Now, draw a subsample of $m$ individuals uniformly at random from the sample of $n$ individuals. The subsample has a most recent common ancestor (MRCA). For $1 \le m \le j \le n$, let $p_{n,m;j}$ denote the probability that the subsample MRCA has $j$ descendants within the sample. For $j = n$, e.g., the subsample has the same MRCA as the sample. From Theorem 2 in Saunders et al. (1984) with $l_1 = l_2 = 2$ (or Example 1 in Saunders et al. (1984)),

$$p_{n,m;n} = \frac{m-1}{m+1}\frac{n+1}{n-1}. \tag{3}$$

(See also p. 77 in Hein et al. (2005).)

In a standard notation (Graham et al., 1994), let $n^{\underline{m}} = n(n-1)\cdots(n-m+1)$ denote the falling factorial for $1 \le m \le n$, with $n^{\underline{0}} = 1$. In addition to Eq. (3), we have the trivial boundary cases $p_{n,1;1} = p_{n,n;n} = 1$, so for $m < n$, consider the recursion

$$p_{n,m;m} = \frac{m(m-1)}{n(n-1)}p_{n-1,m-1;m-1} + \frac{(n-m)(n-m-1)}{n(n-1)}p_{n-1,m;m}, \tag{4}$$

which conditions on $\mathscr{R}_2$, the two terms corresponding to coalescences: (1) within the subsample (probability $m(m-1)/[n(n-1)]$); and (2) outside of the subsample (probability $(n-m)(n-m-1)/[n(n-1)]$). Eq. (4) can provide an inductive proof of the formula

$$p_{n,m;m} = \frac{2(m-1)!}{(m+1)(n-1)^{\underline{m-1}}} \tag{5}$$

from (Wiuf and Donnelly, 1999). (See also, e.g., p. 84 in Hein et al. (2005) and Eq. (1) in Rosenberg (2007).)

If $j = m$, Eq. (5) provides a $p$-value $p_{n,m;m}$, to test whether under the assumptions underlying the coalescent, subsample ancestries are likely to coalesce before coalescing with the remainder of the sample (see, e.g., p. 86 in Hein et al. (2005) for examples concerning Neanderthal ancestry (Nordborg, 1998 and Harris and Hey, 1999)). If $j > m$, then the relevant (left-sided) $p$-value becomes a sum $p_{n,m;j,\bullet} = \sum_{i=m}^{j} p_{n,m;i}$. With the motivating applications mentioned in the Introduction, Theorem 1 in the Results section extends the analytic formula for $p_{n,m;j}$ from $j = m$ and $j = n$ to $1 \le m \le j \le n$.

## 2. Theory

**Theorem 1.** *Let $1 \le m \le n$, and consider a sample whose ancestry satisfies Eq. (1). Under the set-up described above, for $m = 1$, definitions show that $p_{n,m;j}$ equals 1 if $j = 1$ and 0 otherwise. For $m > 1$,*

$$p_{n,m;j} = \begin{cases} \dfrac{m-1}{m+1}\dfrac{2(j-2)^{\underline{m-2}}}{(n-1)^{\underline{m-1}}} & \text{for } 2 \le m \le j < n \\ \dfrac{m-1}{m+1}\dfrac{n+1}{n-1} & \text{for } 2 \le m \le j = n, \end{cases} \tag{6}$$

*with $p_{n,m;j} = 0$ unless $2 \le m \le j \le n$.*

**Remark.** Eq. (6) reduces to Eq. (5) in the case $j = m$, as it should.

**Proof.** Note the following identity for $1 \le a \le b$:

$$a\sum_{i=a}^{b}(i-1)^{\underline{a-1}} = \sum_{i=a}^{b}[i-(i-a)](i-1)^{\underline{a-1}} = \sum_{i=a}^{b}\left[i^{\underline{a}} - (i-1)^{\underline{a}}\right] = b^{\underline{a}} \tag{7}$$

where the second equality follows because $i(i-1)^{\underline{a-1}} = i^{\underline{a}}$ and $(i-1)^{\underline{a-1}}(i-a) = (i-1)^{\underline{a}}$. Thus, $\sum_{j=m}^{n} p_{n,m;j} = 1$ for $2 \le m \le n$:

$$\begin{aligned} &\frac{m-1}{m+1}\left[\frac{n+1}{n-1} + \sum_{j=m}^{n-1}\frac{2(j-2)^{\underline{m-2}}}{(n-1)^{\underline{m-1}}}\right] \\ &= \frac{m-1}{m+1}\left[\frac{n+1}{n-1} + \frac{2}{(n-1)^{\underline{m-1}}}\sum_{j=m}^{n-1}(j-2)^{\underline{m-2}}\right] \\ &= \frac{m-1}{m+1}\left[\frac{n+1}{n-1} + \frac{2}{(n-1)^{\underline{m-1}}}\frac{(n-2)^{\underline{m-1}}}{m-1}\right] \\ &= \frac{m-1}{m+1}\left[\frac{n+1}{n-1} + \frac{2}{n-1}\frac{n-m}{m-1}\right] \\ &= 1, \end{aligned} \tag{8}$$

where the second equality follows from Eq. (7).

For $m = 1$, definitions yield $p_{n,1;1} = 1$, and for $m > 1$, $p_{n,m;j} = 0$ unless $m \le j \le n$. To set up an inductive proof of Theorem 1 for the cases in Eq. (6), let $\mathscr{P}_i$ be the proposition that Theorem 1 holds for every $2 \le m \le j \le n \le i$. To start the induction, $\mathscr{P}_2$ is true, because by definition $p_{2,2;2} = 1$, agreeing with Eq. (6) for $2 \le m \le j = n \le 2$ (the other case $2 \le m \le j < n \le 2$ being vacuous).

For the inductive step, assume $\mathscr{P}_{n-1}$ holds for some fixed $n \ge 3$. From Eq. (2) for $k = 2$, the probability that one of the two equivalence classes of $\mathscr{R}_2$ contains all $m$ subsample individuals and has a total of $i$ elements is

$$\left[\frac{(n-2)!2!1!}{n!(n-1)!}i!(n-i)!\right]\frac{(n-m)!}{(n-i)!(i-m)!} = \frac{2}{n-1}\frac{i^{\underline{m}}}{n^{\underline{m}}}, \tag{9}$$

because there are $(n-m)!/[(n-i)!(i-m)!]$ equally probable ways of forming the two equivalence classes of $\mathscr{R}_2$ by placing the $m$ subsample individuals into an equivalence class of $i$ elements.

As usual, let empty sums equal 0. To check that $\mathscr{P}_n$ follows from $\mathscr{P}_{n-1}$, we check first that Eq. (6) holds for $2 \le m \le j < n$, then conclude from $\sum_{j=m}^{n} p_{n,m;j} = 1$ and Eq. (8) that Eq. (6) also holds for $2 \le m \le j = n$. For $2 \le m \le j < n$, then,

$$\begin{aligned} p_{n,m;j} &= \frac{2}{n-1}\sum_{i=j}^{n-1}\frac{i^{\underline{m}}}{n^{\underline{m}}}p_{i,m;j} \\ &= \frac{2}{n-1}\left[\frac{j^{\underline{m}}}{n^{\underline{m}}}\frac{m-1}{m+1}\frac{j+1}{j-1} + \sum_{i=j+1}^{n-1}\frac{i^{\underline{m}}}{n^{\underline{m}}}\frac{m-1}{m+1}\frac{2(j-2)^{\underline{m-2}}}{(i-1)^{\underline{m-1}}}\right] \\ &= \frac{m-1}{m+1}\frac{2}{n-1}\frac{1}{n^{\underline{m}}}\left[j^{\underline{m}}\frac{j+1}{j-1} + 2(j-2)^{\underline{m-2}}\sum_{i=j+1}^{n-1} i\right] \\ &= \frac{m-1}{m+1}\frac{2}{n-1}\frac{1}{n^{\underline{m}}} \\ &\quad\times\left\{j^{\underline{m}}\frac{j+1}{j-1} + 2(j-2)^{\underline{m-2}}\frac{1}{2}\left[n^2-(j+1)^2\right]\right\} \\ &= \frac{m-1}{m+1}\frac{2(j-2)^{\underline{m-2}}}{(n-1)^{\underline{m-1}}}, \end{aligned} \tag{10}$$

where the first equality is justified as follows. One of the two equivalence classes of $\mathscr{R}_2$ must contain all $m$ individuals from