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

John L. Spouge

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 \( n \rightarrow n - 1 \rightarrow \cdots \rightarrow 2 \rightarrow 1 \), with the state \( D_t = k \) \((k = 2, \ldots, n)\) having a sojourn time \( t_k \) 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 \( \phi_n \), the set of all equivalence relations on the \( n \) individuals. Consider the Markov chain \( \mathcal{A}_n \rightarrow \mathcal{A}_{n-1} \rightarrow \cdots \rightarrow \mathcal{A}_2 \rightarrow \mathcal{A}_1 \), whose state-space is \( \phi_n \), where \( \mathcal{A}_k \) corresponds to having \( D_t = k \) ancestors \((k = n, n - 1, \ldots , 1)\). The variate \( \mathcal{A}_k \) 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 \( \Delta = \{(i, i) \mid i = 1, 2, \ldots, n\} \) and the trivial relation \( \Theta = \{(i, j) \mid i, j = 1, 2, \ldots, n\} \). Given \( \xi, \eta \in \phi_n \), let \( \xi \prec \eta \) denote that \( \eta \) can be obtained from \( \xi \) by combining two equivalence classes in \( \xi \), and in fact, \( \Delta = \mathcal{A}_n \prec \mathcal{A}_{n-1} \prec \cdots \prec \mathcal{A}_2 \prec \mathcal{A}_1 = \Theta \). The transition probabilities of the Markov chain \( \{\mathcal{A}_k\} \) are

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

Kingman shows that if \( \xi \) contains \( k \) equivalence classes,

\[
\mathbb{P}\{\mathcal{A}_k = \xi\} = \frac{(n-k)!k!}{n!(n-1)!} \lambda_1 \lambda_2 \cdots \lambda_k !.
\]
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 $1 \leq m \leq j \leq n$, let $p_{n,m,j}$ denote the probability that the subsample MRCA has $j$ descendents 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)),

$$ P_{n,m,n} = \frac{m - 1 + 1}{m + 1} n - 1. $$

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

In a standard notation (Graham et al., 1994), let $n^m = n(n - 1) \cdots (n - m + 1)$ denote the falling factorial for $1 \leq m \leq n$, with $n^0 = 1$. 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

$$ p_{n,m,m} = \frac{m(m - 1)}{n(n - 1)} p_{n-1,m-1,m-1} + \frac{(n - m)(n - m - 1)}{n(n - 1)} p_{n-1,m,m}. $$

(4)

which conditions on $\mathcal{P}_2$, 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

$$ p_{n,m,m} = \frac{2(m - 1)!}{(m - 1)!(n - 1)!} $$

(5)

from (Wiu 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 and Harris and Hey, 1999)). If $j > m$, then the relevant (left-sided) $p$-value becomes a sum $p_{n,m,j} = \sum_{i=m}^{n} p_{n,m,i}$. 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 $1 \leq m \leq j \leq n$.

2. Theory

**Theorem 1.** Let $1 \leq m \leq n$, 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 $m > 1$,

$$ p_{n,m,j} = \begin{cases} \frac{m - 1}{m + 1} - \frac{(j - 2)m - 1}{(n - 1)m - 1} & \text{for } 2 \leq m \leq j < n \\ \frac{m - 1}{m + 1} & \text{for } 2 \leq m \leq j = n, \end{cases} $$

(6)

with $p_{n,m,j} = 0$ unless $2 \leq m \leq j \leq n$.

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

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

$$ a \sum_{i=a}^{b} (i - 1)^{2m-1} = b \sum_{i=a}^{b} (i - a) (i - 1)^{2m-1} $$

$$ = \sum_{i=a}^{b} (i^2 - (i - 1)^2) = b^2 $$

(7)

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

$$ m - 1 \sum_{j=m}^{n} \left[ \frac{n + 1 - i - m}{n - 1 + j - m} \right] $$

$$ = m - 1 \left[ \frac{n + 1}{n - 1} + \frac{2}{(n - 1)m - 1} \sum_{j=m}^{n} (j - 2)^{m-2} \right] $$

$$ = m - 1 \left[ \frac{n + 1}{n - 1} + \frac{2}{(n - 1)m - 1} \right] $$

$$ = m - 1 \left[ \frac{n}{n - 1} + \frac{2(n - m)}{n - 1} \right] $$

$$ = 1, $$

(8)

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

For $m = 1$, let $P_{n,1,1}$ be the case, and for $m > 1$, $p_{n,m,j} = 0$ unless $m \leq j \leq n$. To set up an inductive proof of Theorem 1 for the cases in Eq. (6), let $\mathcal{P}_n$ be the proposition that $\mathcal{P}_1$ holds for every $2 \leq m \leq j \leq n \leq i$. To start the induction, $\mathcal{P}_2$ is true, because by definition $p_{2,2,2} = 1$, agreeing with Eq. (6) for $2 \leq m \leq j = n \leq 2$ (the other case $2 \leq m \leq j < n \leq 2$ being vacuous).

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

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

(9)

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

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

$$ p_{n,m,j} = \frac{2}{n - 1} \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} \left( \frac{m - 1}{m + 1} \right)^{j-1} $$

$$ = \frac{2}{n - 1} \left[ \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} + \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{i-1} \right] $$

$$ = m - 1 \left[ \frac{2}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} + \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{i-1} \right] $$

$$ = \frac{m - 1}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} + \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{i-1} \right] $$

$$ = \frac{m - 1}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} + \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{i-1} \right] $$

$$ = \frac{m - 1}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{j-1} + \sum_{i=m}^{n} \frac{m!}{m + 1} \frac{m!}{m + 1} \left( \frac{m - 1}{m + 1} \right)^{i-1} \right] $$

(10)

where the first equality is justified as follows. One of the two equivalence classes of $\mathcal{P}_2$ must contain all $m$ individuals from
the subsample. Let the equivalence class of \( \mathcal{A}_j \) containing the subsample have size \( i \), where \( i \in \{ j, j + 1, \ldots, n - 1 \} \). In each case, Eq. (9) gives the probability of the size \( i \), the weight for \( p_{n,m} \) in the right side of the first equality in Eq. (10).

Now, for \( m \geq 2 \), every \( j \) with \( p_{n,m} > 0 \) satisfies \( 2 \leq m \leq j \leq n \). Because Eq. (10) applies to the cases \( 2 \leq m \leq j < n \), Eq. (8) shows that Eq. (3) for \( p_{n,m,n} \) holds. Thus, Eq. (6) also holds for \( 2 \leq m \leq j = n \), and \( \mathcal{A}_n \) follows from \( \mathcal{A}_{n-1} \), completing the induction and the proof.

**Corollary 1.** In the set-up of Theorem 1, the left-sided \( p \)-value

\[
p_{n,m,j} = \sum_{i=m}^j p_{n,m,i} \]

is equal to 1 for \( n \geq 1 \) and

\[
p_{n,m,j} = \begin{cases} 
\frac{2}{m + 1} \frac{(j - 1)^{m-1}}{(n - 1)^{m-1}} & \text{for } 2 \leq m \leq j < n \\
1 & \text{for } 2 \leq m \leq j = n.
\end{cases}
\]

**Proof.** For \( 2 \leq m \leq j < n \),

\[
p_{n,m,j} = \sum_{i=m}^j p_{n,m,i} = \frac{2}{m + 1} \frac{(j - 1)^{m-1}}{(n - 1)^{m-1}} \sum_{i=m}^j (i - 2)^{m-2} = \frac{2}{m + 1} \frac{(j - 1)^{m-1}}{(n - 1)^{m-1}} \sum_{i=m}^{j-1} (i' - 1)^{m-2} = \frac{2}{m + 1} \frac{(j - 1)^{m-1}}{(n - 1)^{m-1}} \frac{j - m + 1}{j},
\]

where the third equality changes the index of summation from \( i \) to \( i' = i - 1 \), and the final equality follows from the identity in Eq. (7) with \( a = m - 1 \) and \( b = j \). Theorem 1 completes the proof, because for \( 2 \leq m \leq j = n \), \( p_{n,m,n} = \sum_{i=m}^n p_{n,m,i} = 1 \).

**Corollary 2.** In the set-up of Theorem 1, and in the limit \( n \to \infty \) and \( jn^{-1} \to f \) with \( m \) fixed, the left-sided \( p \)-value

\[
p_{n,m,j} = \sum_{i=m}^j p_{n,m,i}
\]

satisfies

\[
\lim_{n \to \infty} p_{n,m,j} = \begin{cases} 
\frac{2}{m + 1} f^{m-1} & \text{for } 2 \leq m \leq j < n \\
1 & \text{for } 2 \leq m \leq j = n.
\end{cases}
\]

3. Numerical results

In Fig. 1, the solid circles and lines correspond to \( n = 8 \); the open circles/triangles and dotted lines correspond to \( n = 16 \). Points on a curve share a common value of \( m \), given in the appropriate color on the left of the curve. Each point on the curve corresponds to \( (j, p_{n,m,j}) \), plotted for the values of \( n \) and \( m \) common to the points on the curve.

Typically, as Eq. (3) suggests, much of the probability mass of \( \{p_{n,m,j} : j = m, m + 1, \ldots, n\} \) occurs at \( p_{n,m,n} \). To explore the corresponding left-sided \( p \)-values \( p_{n,m,j} = \sum_{i=m}^j p_{n,m,i} \) numerically, \( p_{n,m,n} = 1 \), and

\[
p_{n,m,j} = \frac{j - m + 1}{j} \]

suggesting that typical left-sided \( p \)-values \( p_{n,m,j} \) \( (m \leq j \leq n) \) decrease rapidly as \( j \) decreases from \( j = n \) to \( j = m \).

In Fig. 1, e.g., \( p_{16,8,12} \approx 0.01 \), suggesting that inference on samples as small as \( n = 16 \) can be surprisingly strong, even for \( j > m \). For large samples \( (n \to \infty) \), Corollary 2 confirms that small subsamples of \( m \) individuals can produce strong inferences, even if \( j \) is much larger than \( m \).

4. Discussion

Theorem 1 is related to combinatorial results on unique event polymorphisms and the sub-trees corresponding to a mutation (Wiuf and Donnelly, 1999). More specifically, a special case of Corollary 1 (Eq. (5)) has been presented as a \( p \)-value for inferring monophony (e.g., Eq. (1) in Rosenberg (2007)), and as a \( p \)-value for inferring non-random mating (e.g., in the informal discussion of the phylogenetic relationship of Neanderthals and modern humans on p. 84 of Hein et al., 2005). Because Corollary 1 generalizes Eq. (5), it has similar applications in taxonomy.

As a hypothetical example, consider the phylogenetic tree presented in Krings et al. (2000) (the following description of the tree suffices for present purposes). The tree was consistent with reciprocal monophony of Neanderthals and modern humans, but contained too few Neanderthals to conclude reciprocal monophony at \( p \leq 0.05 \) from tree topology alone (e.g., Rosenberg, 2007). An alternative statistic, estimated times to most recent common ancestor (TMRC) (Nordborg, 1998), effectively excluded random mating, but not the possibility of some gene flow. The statistical conclusions based on TMRCs, however, require assumptions about the entire history of human population sizes (e.g.,
Tang et al., 2002), whereas conclusions based on tree topology alone require less restrictive assumptions, ones about the size of the human population co-existing with Neanderthals. In general, less restrictive assumptions yield more robust statistical tests, so an inference based on tree topology alone is more robust than an inference based on estimated TMRCAs.

As Neanderthal sequences accumulate, the present state of knowledge does not exclude inferences on a future genetic sample of Neanderthals and modern humans that generates a hypothetical tree with \( n - m \) human ancestors co-existing with \( m \) Neanderthals. The hypothetical tree might contain \( j - m > 0 \) human ancestors sharing a most recent common ancestor with the Neanderthals, the remaining human ancestors lying on a second lineage. With the use of the \( p \)-value in Corollary 1, the tree topology on its own (despite displaying gene flow) could suffice to reject (and reject more strongly than the present data permit) random mating between Neanderthals and human ancestors.

As another hypothetical example of Corollary 1, consider a study of a human genetic disease sampling \( n \) individuals under restrictions justifying a statistical analysis with coalescent theory. Let \( m \) individuals within the sample display the disease, along with a genetic marker suspected as necessary for the disease. Because of incomplete genetic penetrance or the absence of additional but unknown genetic factors required for disease, a further \( j - m > 0 \) individuals might display the relevant marker without displaying the disease. As in the coalescent theory of unique event polymorphisms (Griffiths and Tavare, 1998; Wiuf and Donnelly, 1999; Griffiths and Tavare, 2003; Tavare, 2004), assume that the mutation generating the marker occurred only once. Corollary 1 could test if the association of the marker with the disease reaches statistical significance.

Any statistical test using Eq. (6) or Corollary 1 is based solely on Eq. (1), an assumption common to most coalescent models. Accordingly, such a test is more robust than tests based on more specific coalescent models. Although the test loses power through its sparse assumptions, the Numerical Results section suggests that nonetheless, tests based on Corollary 1 can be surprisingly powerful.

Acknowledgments

It is my pleasure to acknowledge helpful conversations with Drs László Székely, Eva Czabarka, Susanta Tewari, Dave Erickson, and Peter Rogan. I also thank the Editor and two anonymous reviewers for their helpful suggestions, which greatly improved the article. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References