The Standard Coalescent

Two independent Processes

Continuous: Exponential Waiting Times
Discrete: Choosing Pairs to Coalesce.

\[
\begin{align*}
\text{Waiting} & \quad \text{Coalescing} \\
(1,2) & \rightarrow (3,4,5) \\
(1,2)(3,4,5) & \quad \text{Exp} (\tau) \\
(1,3)(4,5) & \quad \text{Exp} (\tau) \\
(1,2)(3,4,5) & \quad \text{Exp} (\tau) \\
(1,2)(3,4,5) & \quad \text{Exp} (\tau)
\end{align*}
\]

A set of realisations
(from Felsenstein)

The Exponential Distribution

The Exponential Distribution: \( R^+ \sim \text{Exp}(a) \)

Density: \( f(t) = ae^{-at}, \quad P(X > t) = e^{-at} \)

Properties: \( X \sim \text{Exp}(a), \quad Y \sim \text{Exp}(b) \) independent

i. \( P(X > t | X > t_i) = P(X > t - t_i) \quad (t > t_i) \)

ii. \( E(X) = \frac{1}{a}. \)

iii. \( P(X < Y) = a/(a + b). \)

iv. \( \min(X, Y) \sim \text{Exp}(a + b). \)

v. Sums of \( k \) iid \( X \) is \( \Gamma(k, a) \) distributed

\[
\frac{a^k x^{k-1}e^{-ax}}{\Gamma(k)}
\]
Binomial Numbers

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{r!b!}
\]

Binomial Expansion: \((a + b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}\)

Special Cases:

\[
(1+1)^n = \sum_{i=0}^{n} \binom{n}{i} = 2^n
\]

\[
(1-1)^n = \sum_{i=0}^{n} \binom{n}{i}(-1)^i = 0
\]

Counting by Bijection

Tree Counting

Tree: Connected undirected graph without cycles. k nodes (vertices) & k-1 edges. Nodes with one edge are leaves (tips) - the rest are internal.

Labels of internal nodes are permutable without change of biological interpretation. If labels at leaves are ignored we have the shape of a tree.

Ignore root & branch lengths gives unrooted tree topology.

If age ordering of internal nodes are retained this gives the coalescent topology.

Most biological trees are bifurcating. Valency 3 (number of edges touching internal nodes) if made unrooted. Such unrooted trees have n-2 internal nodes & 2n-3 edges.
Trees: Rooted, bifurcating & nodes time-ranked.

Recursion: $T_k = \binom{k}{2} T_{k-1}$
Initialisation: $T_1 = T_2 = 1$

<table>
<thead>
<tr>
<th>k</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\prod_{j=2}^{k} \binom{j}{2}$</td>
<td>3</td>
<td>18</td>
<td>180</td>
<td>2700</td>
<td>5.7 $10^4$</td>
<td>1.5 $10^5$</td>
<td>5.7 $10^5$</td>
<td>2.5 $10^6$</td>
<td>6.9 $10^6$</td>
<td>5.6 $10^9$</td>
</tr>
</tbody>
</table>

Trees: Unrooted & valency 3

Recursion: $T_n = (2n-5) T_{n-1}$
Initialisation: $T_1 = T_2 = T_3 = 1$

Let $l_{1:n}$ be length of $i$'th external branch in an $n$-tree. Obviously $E(\epsilon) = n E(l_{1:n})$

Coalescent versus unrooted tree topologies

4 leaves: 3 unrooted trees & 18 coalescent topologies.
1 unrooted tree topology contains 6 coalescent topologies.

Inner & outer branches

External ($\epsilon$) versus Internal ($i$) Branches.
$E(\epsilon) = 2$ \quad $E(i) = \sum_{i=1}^{n} \binom{n}{j}$ - 2

Red - external. Others internal.
Except for green branch, internal-external corresponds to singlet/non-singlet segregating sites if only one mutation can happen per position.
Probability of hanging Sub-trees.

For a coalescent with \( n \) leaves at time 0, with \( k \) ancestors at time \( t_1 \), let \( \xi \) be the groups of leaves of the \( k \) subtrees hanging from time \( t_1 \). Let \( \lambda_1, \lambda_2 \ldots, \lambda_k \) be the number of leaves of these sub-trees.

\[
P(R_i = \xi) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \lambda_1! \lambda_2! \ldots \lambda_k!
\]

Example: \( n=8, k=3 \). Classes observed : 4, 3, 1

\[
\frac{5!3!2!}{8!7!} = 0.0012
\]

The basal division splits the leaves into \((k,n-k)\) sets with probability: \( 1/(n-1) \).

Nested subsamples

\[ \Pr\{\text{MRCA(sub-sample)} = \text{MRCA(sample)}\} = \frac{(i+1)(j-1)}{(i-1)(j+1)} \]

\[ \Pr\{\text{MRCA(sub-sample)} = \text{MRCA(population)}\} = \frac{(j-1)}{(j+1)} \]

Age of a Mutation

The probability that there are \( k \) differences between two sequences. Going back in time 2 kinds of events can occur (mutations (\( \Theta \)) - or a coalescent (1)). This gives a geometric distribution:

\[
\frac{1}{1+\theta} \frac{\theta}{1+\theta} \\
\]

---
Let $X_0$ be the initial configuration of the initial Urn.

A step: take a random ball the urn and put it back together with an extra of the same colour.

$X_k$ be the content after the $k$'th step. Let $Y_k$ be the colour of the $k$'th picked ball.

i. $P(Y_k = j) = P(Y_1 = j)$.

ii. Sequences $Y_1, \ldots, Y_k$ resulting in the same $X_k$ - has the same probability.

Labelling, Polya Urns & Age of Alleles

An Urn: A ball is picked proportionally to its weight. Ordinary balls have weight 1.

- If the initial $\Theta$-size ball is picked, it is replaced together with a completely new type.
- If an ordinary ball is picked, it is replaced together with a copy of itself.

There is a simple relationship between the distribution of "the alleles labeled with age ranking" is the same as "the alleles labeled with size ranking"
Stirling Numbers

n-1 items - k classes:  \( (n-1,k-1) : \{..\},\{..\},..\{..\} \)

{} \{..\},\{..\},..\{..\} \( n \) \( k \)

Basic Recursion: \( S_{n,k} = kS_{n-1,k} + S_{n-1,k-1} \)

Initialisation: \( S_{n,1} = S_{n,n} = 1 \).

Ewens' formula

\( P_n(a_1,a_2,..,a_n) = \frac{n!}{\theta^{(\theta+1)} \prod_j (\theta + j - 1)^{a_j}} \prod_j \theta^{a_j} \)

\( E_s(k \text{ types}) \)

\( P_n(a_1,a_2,..,a_n,k) = \frac{n!}{S_{n,1}^k \prod j a_j!} \)

k is a minimal sufficient statistic for \( \theta \), the probability of the data conditioned on k is \( \theta \)-less and there is no simpler such statistic.

Ancestry to Ancestry

\( h_{i,j} = \text{probability that } i \text{ individuals has } j \text{ ancestors after time } t. \)

\[ h_{i,j} = \sum_{k=j}^{n} e^{-\frac{1}{2} \left( \frac{1}{2} \right)} \frac{(2k-1)(-1)^{k-j} f_{k-j}^j}{j!(k-j)!} i_{[k]} = i(i-1)..(i-k+1) \quad i_{(k)} = i(i+1)..(i+k-1) \]

Example: Disappearance of 7 lineages.
**3 methods of solution:**

i. Sum of different independent exponential distributions:
\[
\{ Y = j \} = \{ \exp \{ \frac{j}{2} \} + \exp \{ \frac{j+1}{2} \} + \ldots + \exp \{ \frac{j+k}{2} \} > t \}
\]

ii. Distribution in markov chain:

iii. Combination of known probabilities:
   a. Probability that i alleles has i/less ancestors.
   b. This probability is the same for all i-sets
   c. No coalescence within a set, implies no coalescence within all subsets.

**The exclusion-inclusion principle.**

**Venn Diagrams:**

\[
\{1 + II\} - \{I\} + \{II\} + \{I\&II\} = 0
\]

\[
\{I + II + III\} = \{I\} + \{II\} + \{III\} - ((I,II) + (I,III) + (II,III)) + \{I,II,III\}
\]

**3 Ancestors to 2 Ancestors:**

\[
(3/2)(e^t - e^{3t})
\]

Exactly one coalescence:

\[
3(e^t - (e^t - e^{3t})/2) - e^{3t}
\]

Jordan’s Sieve:

\[
A_1 : 3e^t
\]

\[-2A_2 : 2 \quad ((e^t + e^{3t})/2)
\]

\[+3A_3 : 3 \quad e^{3t}\]

**Exclusion-inclusion & Jordan’s Sieve**

\[S_j j=1,..,r \text{ the given sets, } A_k \text{ - sum of intersection of k sets}\]

Total number:

\[
\sum_{k=1}^{r} (-1)^{k+1} A_k
\]

In exactly m sets:

\[
\sum_{k=m}^{r} (-1)^{k-m} \binom{k}{m} A_k
\]

Example: the elements above:

in 1 sets \[A_1 - 2A_2 + 3A_3 - 4A_4\]

in 2 sets \[A_2 - 3A_3 + 6A_4\]

in 3 sets \[A_3 - 4A_4\]

in 4 sets \[A_4\]

in some set \[A_1 - A_2 + A_3 - A_4\]
Surviving Lineages

Which probability statements can be made? Let $\mathcal{S}$ be subset of $\{1,2,..,i\}$ and $S(\mathcal{S})$ be the event that no coalescence has happened to $\mathcal{S}$. Additionally, if $\mathcal{S}'$ is a subset of $\mathcal{S}$, then $S(\mathcal{S})$ implies $S(\mathcal{S}')$.

There are $r=\binom{i}{j}$ sets. We want events member of only one of them.

$$
\sum_{k=1}^{r} (-1)^{k-j} \binom{k}{j} A_k \quad \text{where} \quad A_k = \sum_{j} |S_j|
$$

Summation is over all k-subsets of $\{1,..,r\}$ and intersection is between the k sets chosen.

Summary

Tree Counting & Tree Properties.

Basic Combinatorics.

Allele distribution.

Polya Urns + Stirling Numbers.

Number of ancestral lineages after time $t$.

Inclusion-Exclusion Principle.
Recommended Literature

Ewens (1989)
Feller (1968+71) Probability Theory and its Applications I+II. Wiley
Griffiths (1980)
Griffiths & Tavaré (1998) "The Age of a mutation on a general coalescent tree"
Griffiths & Tavaré (1999) "The ages of mutations in gene trees"
Griffiths & Tavaré (2001) "The genealogy of a neutral mutation"
Kingman (1982)
Mörtler
Möhle
Schweinsberg
Simonsen & Churchill (1997)