K. Lachhab

# A stochastic model for the mitochondrial Eve

Bachelorthesis

Supervisor: dr. M.O. Heydenreich

4 August 2014


Mathematical Institute of Leiden University

**Abstract**

The existence of the most recent common maternal ancestor, the mitochondrial Eve, is studied using results from the Galton-Watson process. To accomplish this, ancestral trees are generated according to certain assumptions and progeny distributions.
Results from the Galton-Watson process are first recreated from which the condition for an infinite tree is derived. These results are applied to a model of the mitochondrial Eve. Assuming a number of women contemporary to the mitochondrial Eve, the maximum probability of one lineage surviving is also determined in a specific situation.

# Contents

# 1 Introduction

## 1.1 The mitochondrial Eve

The Out of Africa model [12] proposes the evolution of archaic Homo Sapiens into modern humans as originating in Africa after which earlier human populations were replaced by migration and without genetic mixing. By contrast, the Multiregional hypothesis [11] proposes that all archaic human life forms evolved simultaneously to the modern human, while having enough genetic mixing to ensure the development of a unique human race. Both hypotheses rely in part on fossil evidence but focus also on the evolution of genetic material. One source for such genetic research is mitochondrial DNA (mtDNA), which is not found in cells' nuclei but in organelles called the mitochondria.

Though very short, mtDNA may play a part in attempting to answer these evolutionary questions. As it happens, mtDNA is inherited solely from the mother as opposed to nuclear DNA which is inherited in equal parts from the father and the mother. Due to this peculiar form of inheritance, the mtDNA of an individual would be identical to the mtDNA of every maternal ancestor, provided there are no mutations and natural selection does not play a part.

Cann, Stoneking and Wilson examined mtDNA of 147 living humans drawn from five geographic populations in [3] and stated that the mtDNA of all people could be described as mutations of the mtDNA of a single woman. This most recent common ancestor was named the mitochondrial Eve and would have lived in Africa around 200.000 years ago. More recent research estimates the mitochondrial Eve to have lived 99.000 to 148.000 years ago. [9]

Often viewed as the proof for the validity of the Out of Africa model, research into the mitochondrial Eve does not give conclusive proofs for either the Out of Africa model or the Multiregional hypothesis. While the existence of one common female ancestor is widely accepted, the question remains when and where this ancestor lived and mainly what kind of starter populations are eligible for the existence of a mitochondrial Eve. [4]

The starter population question arises from the fact that the mitochondrial Eve is not equivalent to the biblical Eve. The mitochondrial Eve is only the most recent common ancestor in the inheritance of mtDNA. However, studies into nuclear DNA show that descendants of women contemporary to the mitochondrial Eve are alive today. Since their mtDNA is not distinct, there must have been a time in their ancestry when there were no female descendants thus breaking the mtDNA line.

This bachelor thesis uses a model developed by Neves and Moreiro in [8] to study ancestral trees generated under certain assumptions to find an answer to these questions using results from the Galton-Watson process.

## 1.2 The Galton-Watson process

The mathematical theory utilised in modelling the mitochondrial Eve is that of the Galton-Watson process. The problem of the disappearance of (aristocratic) family names was first posed by F. Galton in 1873 and subsequently solved by H.W. Watson in 1874. However, Watson had concluded the inevitable extinction of all surnames, a fault which was not corrected until 1930 when J.F. Steffensen gave a detailed proof. [6]

In the Galton-Watson process the inheritance of surnames is described from father to son. Assuming that a man has $r$ sons with probability $q_r$, $r \in \mathbb{N}_0$, the existence of a surname through the generations can be modelled as a random tree generated by these probabilities. Such a tree starts with one particle (the father) in generation 0, branching a number of particles in the next generation (his sons) according to a given probability distribution. The question of the survival of a surname then becomes the probability of such a tree being of infinite length.

The application of the Galton-Watson process to the problem of the mitochondrial Eve can thus be explained as replacing the father to son inheritance to that of a mother to daughter inheritance. When assuming each woman has probability $q_r$ of having $r$ daughters, a similar ancestral tree can be generated. To be able to make any statement about the existence of the mitochondrial Eve, the

probability of an infinite tree is of course of interest. Due to the fact that the mitochondrial Eve was not the only woman living at that time, also of interest are the conditions and the probability for the existance of only one such infinite tree.

# 2 The Galton-Watson model

The Galton-Watson process revolves around generating random trees according to certain probability distributions. To translate the problem of surname extinction to a mathematical model, certain definitions and assumptions are in order.

### Number of generations

The model used is discrete in time and focusses on the number of particles in each generation. Let $Z_n$ be the number of particles in the $n$-th generation of a tree, $n \in \mathbb{N}_0$. Each particle in generation $n$ will die out in generation $n+1$, so generations are non-overlapping. Three assumptions must be made regarding the sequence $(Z_n)_{n\in\mathbb{N}_0}$:

1. The number of particles in a generation is only dependent on the number of particles in the previous generation. The random variables $Z_0, Z_1, Z_2, \ldots$ thus form a Markov chain meaning

   $$P(Z_{n+1} = z | Z_1 = z_1, Z_2 = z_2, \ldots, Z_n = z_n) = P(Z_{n+1} = z | Z_n = z_n)$$

   for any $z, z_1, \ldots, z_n \in \mathbb{N}_0$.

   The transition probabilities are time- and population size-independent.

2. Each particle in generation $Z_n$ generates a number of other particles in the next generation, independently and identically distributed.

3. A tree always starts with one particle, i.e. $Z_0 = 1$.

The first assumption can prove to be a restriction if a man with few brothers would be more likely to have fewer sons than a man with many brothers. In the second assumption, the case in which different particles interact with one another is excluded. The third assumption can be easily adjusted if $Z_0 \neq 1$ since families of initial particles develop independently, so any other starter population can be considered by taking a tree with $Z_0 = 1$ and starting from a different generation.

### Progeny distribution and generating function

Let numbers $q_0, q_1, \ldots$ be given such that $q_r \geq 0$ for $r \in \mathbb{N}_0$ and $\sum_{r=0}^{\infty} q_r = 1$. Let $P(Z_1 = k) = q_k$ and let all following $Z_i$ be distributed as the sum of $Z_{i-1}$ independent random variables identically distributed as $Z_1$, $i \geq 2$. Furthermore, when a tree has gone extinct in generation $n$ it will remain extinct for any following generation so $P(Z_{n+1} = 0 | Z_n = 0) = 1$ for all $n \in \mathbb{N}_0$.

Let $S(x)$ be the probability generating function of $q_r$:

$$S(x) = \sum_{r=0}^{\infty} q_r x^r, \ x \in \mathbb{R}$$

with iterations

$$S_0(x) = x, \ S_1(x) = S(x) \text{ and } S_{n+1}(x) = S(S_n(x)), \ n \in \mathbb{N}_0.$$

Since the generating function is a power series with a radius of convergence greater than or equal to 1 and due to the normilisation of $q_r$, $S(x)$ is continuous on the unit interval. $S(x)$ is also differentiable on the unit interval if $S'(1) < \infty$.

To avoid triviality and to ensure the existence of a solution to the problem, the following assumptions should be made regarding the progeny distribution:

1. The values of $p_r$ are time- and population size-independent.

2. To avoid the trivial case of a tree existing solely of each particle always producing $i$ particles in the next generation: $p_i \neq 1$ for any $i \in \mathbb{N}_0$. Such a tree will be finite if $i = 0$ or infinite if $i \geq 1$.

3. To ensure the strict convexity of the generating function $S$ on the unit interval: $p_0 + p_1 < 1$.

4. The expected value $E(Z_1)$ is finite, from this it follows that $S'(1) < \infty$.

The process described can be thought of as a genealogical tree of one father, the particle in $Z_0$, who has a number of sons generated by the probabilities $q_r$. In succession, his sons have again, all independently, a number of sons generated by the same distribution thus allowing the surname to be passed on through the generations. This process continues infinitely or until there are no more sons born and the surname becomes extinct.

## 2.1 Extinction probability

Galton posed a question about the disappearance of family names. Of interest is then of course the probability that such a rooted tree is finite or the event that the extinction probability is positive. Let therefore $E_n$ be the set of all rooted trees that end in no more than $n$ generations, so $E_n = \{T \text{ rooted tree} : Z_n = 0\}$. Using this, the extinction probability $\bar{\theta}$ can be defined.

**Definition 2.1.** Let $\bar{\theta}_n = P(E_n)$ be the probability of extinction in at most $n$ generations, with $\bar{\theta}_0 = q_0$ and $\theta_n = 1 - \bar{\theta}_n$. Let $\bar{\theta} = \lim_{n\to\infty} \bar{\theta}_n$.

The last element needed is a well known given of continuous and differentiable functions, the Mean Value Theorem, a proof of which can be found in [1], section 2.8, theorem 11. This theorem will be used in this section as well as in section 2.2.

**Theorem 2.2** (Mean Value Theorem)**.** *Let $a < b$ and $f\colon [a,b] \to \mathbb{R}$ continuous on $[a,b]$ and differentiable on $(a,b)$. Then there exists some $c \in (a,b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

### Main theorem

It is now possible to formulate and prove the main theorem regarding the Galton-Watson process which gives the condition for the existence of an infinite tree.

**Theorem 2.3.** *Let $m = \sum_{r=1}^{\infty} r q_r$ be the mean number of sons, then:*

*(i) $\bar{\theta}$ is the smallest solution of $S(x) = x$ for $x \in [0,1]$;*

*(ii) $\bar{\theta} = 1$ if $m \leq 1$ and $\bar{\theta} < 1$ if $m > 1$.*

*Proof.* Any tree in $E_n$ is extinct in at most $n$ generations. If such a tree has $r$ particles in the first generation, then there has to be a tree in $E_{n-1}$ attached to each of these $r$ particles. It follows that

$$\bar{\theta}_n = P(E_n) = \sum_{r=0}^{\infty} q_r P(E_{n-1})^r = S(\bar{\theta}_{n-1}). \tag{2.1}$$

Taking the limit on both sides and using the fact that $S$ is a continuous function on the unit interval and $\bar{\theta} \in [0,1]$ gives

$$\bar{\theta} = S(\bar{\theta}). \tag{2.2}$$

So the extinction probability $\bar{\theta}$ must be a fixed point of $S$ in $[0,1]$. Due to the normilisation of the probabilities $q_r$, it is clear that 1 is a fixed point of $S$.[1]

Assume there exists $a \in [0,1)$ with $a = S(a)$. The generating function $S$ is continuous and differentiable on the unit disk, so $S$ is also continuous on $[a,1]$ and differentiable on $(a,1)$. With theorem 2.2 it follows that there must exist $b \in (a,1)$ such that

$$S'(b) = \frac{S(1) - S(a)}{1 - a} = 1$$

[1] Watson believed this to be the only solution of the equation and thus concluded the inevitable disappearance of all surnames. However, he overlooked the possibility of there being a second solution less than 1.

using that $a = S(a)$ and $S(1) = 1$.

Assumption 3 on the progeny distribution ensures the strict convexity of $S$ on $[0, 1]$, so the first derivative $S'(x)$ is a strictly increasing function on $[0, 1]$. If there exists $b \in [0, 1)$ with $S'(b) = 1$ it must then follow that $S'(1) > 1$. Since $S'(1) = \sum_{r=1}^{\infty} rq_r = m$ this ensures the existence of an solution in $[0, 1)$ only when $m > 1$.

Assume there are two solutions in $[0, 1)$, $\theta_1$ and $\theta_2$ with $0 \leq \theta_1 < \theta_2 < 1$. $S(x)$ is continuous on $[\theta_1, \theta_2]$ and differentiable on $(\theta_1, \theta_2)$, using theorem 2.2 there exists $\epsilon_1 \in (\theta_1, \theta_2)$ such that

$$S'(\epsilon_1) = \frac{S(\theta_2) - S(\theta_1)}{\theta_2 - \theta_1} = 1.$$

$S(x)$ is also continuous on $[\theta_2, 1]$ and differentiable on $(\theta_2, 1)$ so there exists $\epsilon_2 \in (\theta_2, 1)$ such that

$$S'(\epsilon_2) = \frac{S(1) - S(\theta_2)}{1 - \theta_1} = 1.$$

Since $S'(x)$ is a strictly increasing function on $[0, 1]$, this contradicts the existence of two points on $[0, 1]$ with the same derivative. So any solution of $\bar{\theta} = S(\bar{\theta})$ on the unit interval will be unique.

This leads to two distinct situations:

1. If $m \leq 1$, the only fixed point of $S$ in [0,1] is 1. This means the extinction probability $\bar{\theta}$ is equal to 1; all trees will become finite;
2. If $m > 1$ there exists another unique fixed point less than 1, so the extinction probability $\bar{\theta}$ is less than 1; there is a positive probability for the existence of an infinite tree;

and thus concludes the proof. □

In the *subcritical* case $m < 1$ all trees will become finite at one point. The same goes for $m = 1$ but since this value is the turning point for the extinction probability $\bar{\theta}$ to go from 1 to strictly less than 1, this case is called *critical*. When $m > 1$ there exists a positive probability of any tree becoming infinite: this is known as the *supercritical* case.

### 2.1.1 Approximation of $\theta$

In the supercritical case, the extinction probability $\bar{\theta}$ is strictly less than 1 so the survival probability $\theta$ is strictly positive. The value of $\theta$ can be exactly determined solving (2.2). However, this can prove to be very complicated depending on the distribution of $q_r$.

It is however possible to approximate the value of $\theta$ using only the mean and variance of the progeny distribution. This approximation $\theta_a$ of $\theta$ is known as the *small survival probability approximation*.

**Definition 2.4.** Let $v = \sum_{r=0}^{\infty}(r - m)^2 q_r$ be the variance of the number of sons. Assume that $v \in \mathbb{R}$ and $v < \infty$.

**Lemma 2.5.** *For $m > 1$, $S'''(1) < \infty$ and certain $\zeta \in [\bar{\theta}, 1]$*

$$\theta = \theta_a + \tilde{R}(\theta)$$

*with*

$$\theta_a = \frac{2(m-1)}{v + m(m-1)} \text{ and } \tilde{R}(\theta) = \frac{2\theta^2}{3!} \frac{S'''(\zeta)}{v + m(m-1)}.$$

*Proof.* Determine the second order Taylor polynomial of the generating function $S(x)$ around $x = 1$

with error term $R(x)$:

$$\begin{aligned} S(x) &= S(1) + (x-1)S'(1) + \frac{1}{2}(x-1)^2 S''(1) + R(x) \\ &= \sum_{r=0}^{\infty} q_r + (x-1)\sum_{r=1}^{\infty} r q_r + \frac{1}{2}(x-1)^2 \sum_{r=2}^{\infty} r(r-1)q_r + R(x) \\ &= 1 + (x-1)m + \frac{1}{2}(x-1)^2 \sum_{r=2}^{\infty} r(r-1)q_r + R(x) \end{aligned}$$

with $R(x) = \frac{S'''(\zeta)}{3!}(x-1)^3$ for certain $\zeta$ between $x$ and 1.
Since $\bar{\theta} = S(\bar{\theta})$ it follows that

$$\begin{aligned} \bar{\theta} &= 1 + (\bar{\theta}-1)m + \frac{1}{2}(\bar{\theta}-1)^2 \sum_{r=2}^{\infty} r(r-1)q_r + R(\bar{\theta}) \\ 1-\theta &= 1 + (-\theta)m + \frac{1}{2}(-\theta)^2 \sum_{r=2}^{\infty} r(r-1)q_r + R(1-\theta) \\ \theta &= m\theta - \frac{1}{2}\theta^2 \sum_{r=2}^{\infty} r(r-1)q_r - R(1-\theta) \\ \theta &= \frac{2(m-1)}{\sum_{r=2}^{\infty} r(r-1)q_r} - \frac{2R(1-\theta)}{\theta \sum_{r=2}^{\infty} r(r-1)q_r}. \end{aligned} \tag{2.3}$$

This approximation for $\theta$ can be exclusively expressed in terms of the mean $m$ and the variance $v$. It holds that

$$\begin{aligned} \sum_{r=2}^{\infty} r(r-1)q_r &= \sum_{r=1}^{\infty} r(r-1)q_r \\ &= \sum_{r=1}^{\infty} r^2 q_r - \sum_{r=1}^{\infty} r q_r \\ &= \sum_{r=1}^{\infty} r^2 q_r - m + m^2 - m^2 \\ &= \sum_{r=1}^{\infty} r^2 q_r - 2m^2 + m^2 + m(m-1) \\ &= \sum_{r=1}^{\infty} r^2 q_r - 2m\sum_{r=1}^{\infty} r q_r + m^2 \sum_{r=1}^{\infty} q_r + m(m-1) \\ &= \sum_{r=1}^{\infty} (r-m)^2 q_r + m(m-1) \\ &= v + m(m-1). \end{aligned} \tag{2.4}$$

Substituting (2.4) in (2.3) gives

$$\theta = \frac{2(m-1)}{v + m(m-1)} - \frac{2R(1-\theta)}{\theta(v + m(m-1))} = \theta_a + \tilde{R}(\theta).$$

So the small survival probability approximation $\theta_a$ is given by

$$\theta_a = \frac{2(m-1)}{v + m(m-1)}$$

and the error is given by

$$\begin{aligned}\tilde{R}(\theta) &= -\frac{2R(1-\theta)}{\theta(v+m(m-1))} \\ &= -\frac{2}{3!\theta}\frac{S'''(\zeta)}{v+m(m-1)}(1-\theta-1)^3 \\ &= \frac{2\theta^2}{3}\frac{S'''(\zeta)}{v+m(m-1)}\end{aligned}$$

for certain $\zeta \in [1-\theta, 1]$. This means dat $\zeta \in [0,1]$ and $S'''(\zeta)$ reaches its maximum on the unit interval for $\zeta = 1$. Since $S'''(1) < \infty$ it follows that the error in the approximation is bound. □

The error term will always be bound but when imposing a maximum on $r$, it will also be small. Imposing a maximum on the value of $r$ will not be too great a restriction on the model in a biological sense, since this is the amount of sons a man would have. It will however determine how small the error term $\tilde{R}(\theta)$ is.

## 2.2 Convergence rates

### 2.2.1 Convergence of $\theta$

The survival probability of a tree can be viewed when looking at a finite number of generations, or $\theta_n$. More interesting however is the survival probability of a tree when viewing an infinite amount of possible generations with offspring, or $\theta$. The question remains in how many generations the sequence $(\theta_n)_{n\in\mathbb{N}_0}$ approaches its limit $\theta$. This convergence rate is determined in two seperate cases, namely $m \neq 1$ (super- and subcritical) and $m = 1$ (critical). In determining this convergence rate the following definition is used.

**Definition 2.6.** Let $(a_n)_{n\in\mathbb{N}_0}$ and $(b_n)_{n\in\mathbb{N}_0}$ be two sequences. Then

$$a_n \sim b_n \text{ if } \lim_{n\to\infty}\frac{a_n}{b_n} = 1.$$

**Super- and subcritical**

**Lemma 2.7.** *For $n \in \mathbb{N}_0$ and $S'(\bar{\theta}) \neq 1$*

$$|\theta_n - \theta| \leq c \cdot e^{-n/\xi}$$

*for certain constant $c \in \mathbb{R}$ and*

$$\xi = \frac{-1}{\ln[S'(1-\theta)]} \tag{2.5}$$

*the correlation time.*

*Proof.* Let $n_0 \in \mathbb{N}_0$. $S(x)$ is continuous and differentiable on the unit interval so with theorem 2.2 there exist $\alpha \in (\bar{\theta}_{n_0-1}, \bar{\theta})$ such that

$$S'(\alpha) = \frac{S(\bar{\theta}) - S(\bar{\theta}_{n_0-1})}{\bar{\theta} - \bar{\theta}_{n_0-1}} = \frac{\bar{\theta} - \bar{\theta}_{n_0}}{\bar{\theta} - \bar{\theta}_{n_0-1}} \text{ so } \bar{\theta} - \bar{\theta}_{n_0} = S'(\alpha)(\bar{\theta} - \bar{\theta}_{n_0-1})$$

using (2.1) and (2.2). In the same way there exists $\alpha' \in (\bar{\theta}_{n_0-2}, \bar{\theta})$ such that

$$\bar{\theta} - \bar{\theta}_{n_0-1} = S'(\alpha')(\bar{\theta} - \bar{\theta}_{n_0-2})$$

so

$$\bar{\theta} - \bar{\theta}_{n_0} = S'(\alpha)S'(\alpha')(\bar{\theta} - \bar{\theta}_{n_0-2}).$$

Iterating this argument, it follows that for all $n \in \mathbb{N}_0$ there are $\alpha_j \in (\bar{\theta}_{n_0+j-1}, \bar{\theta})$ with $j \in \{1, \ldots, n\}$ such that

$$\bar{\theta} - \bar{\theta}_{n_0+n} = \prod_{j=1}^{n} S'(\alpha_j)(\bar{\theta} - \bar{\theta}_{n_0}).$$

This statement is true for all $n_0 \in \mathbb{N}_0$ so take $n_0 = 0$. It then follows that

$$\bar{\theta} - \bar{\theta}_n = \prod_{j=1}^{n} S'(\alpha_j)(\bar{\theta} - \bar{\theta}_0) \text{ and}$$

$$|\bar{\theta} - \bar{\theta}_n| = (\bar{\theta} - \bar{\theta}_0) \prod_{j=1}^{n} S'(\alpha_j)$$

using the fact that $\bar{\theta} \geq \bar{\theta}_n$ for all $n \in \mathbb{N}_0$.

$S'(x)$ is a strictly increasing function on [0,1]. Since $\alpha_j \in (\bar{\theta}_{j-1}, \bar{\theta})$ and $\lim_{n_0 \to \infty} \alpha_j = \bar{\theta}$ it follows that $\alpha_j \leq \bar{\theta}$ for all $j \in \{1, \ldots, n\}$. Due to the strict increase, this means that $S'(\alpha_j) \leq S'(\bar{\theta})$ for all $j$. Because $\alpha_j, \bar{\theta} \geq 0$ and $S'(x)$ is a power series with non-negative coefficients, it also follows that $S'(\alpha_j), S(\bar{\theta}) \geq 0$.

Applying this in the previous equation gives

$$\begin{aligned} |\bar{\theta} - \bar{\theta}_n| &= (\bar{\theta} - \bar{\theta}_0) \prod_{j=1}^{n} S'(\alpha_j) \\ &\leq (\bar{\theta} - \bar{\theta}_0) \prod_{j=1}^{n} S'(\bar{\theta}) \\ &= (\bar{\theta} - \bar{\theta}_0)[S'(\bar{\theta})]^n. \end{aligned}$$

Substituting $\bar{\theta} = 1 - \theta$ finally concludes

$$\begin{aligned} |\theta_n - \theta| &\leq (\theta_0 - \theta)[S'(1-\theta)]^n \\ &= c \cdot \mathrm{e}^{n \ln[S'(1-\theta)]} \\ &= c \cdot \mathrm{e}^{-n/(-1/\ln[S'(1-\theta)])} \\ &= c \cdot \mathrm{e}^{-n/\xi} \end{aligned}$$

with

$$c = (\theta_0 - \theta) = (q_0 - \theta) \text{ and } \xi = \frac{-1}{\ln[S'(1-\theta)]}.$$

□

**Critical**

To determine the convergence rate of $\theta_n$ to 0 when $m = 1$, a slightly different but equivalent definition of the survival probability $\theta$ is in order.

**Definition 2.8.** The probability that a tree doesn't end in $n$ generations is given by $\theta_n = P(Z_n > 0)$.

This definition yields the following result

$$\theta_n = P(Z_n > 0) = 1 - P(Z_n \leq 0) = 1 - P(Z_n = 0) = 1 - S_n(0). \tag{2.6}$$

This can be used together with lemma 2.9, the full phrasing and proof of which can be found in [5], chapter 1, section 10.2, lemma 10.1.

**Lemma 2.9.** *For $m = 1$, $S'''(1) < \infty$ and $X$ all the points $x$ that either*

*(i) are interior to the unit circle or*

*(ii) lie on the segment of the unit circle with $-\theta_0 \leq arg\ x \leq \theta_0$ excluding $x = 1$:*

$$\frac{1}{1 - S_n(x)} = \frac{1}{1-x} + \frac{nS''(1)}{2} + O(\log n), \quad x \in X, n \to \infty.$$

Lemma 2.9. can be immediately used to determine the convergence rate of $\theta$ in the critical case.

**Corollary 2.10.** *For $S'(\bar{\theta}) = 1$ and $S'''(1) < \infty$*

$$\theta_n \sim \frac{2}{nS''(1)}.$$

*Proof.* Substitute $x = 0$ in lemma 2.9 then

$$\begin{aligned} \frac{1}{1 - S_n(0)} &= 1 + \frac{nS''(1)}{2} + O(\log n) \\ \frac{1}{\theta_n} &= 1 + \frac{nS''(1)}{2} + O(\log n). \end{aligned}$$

By using (2.6) it then follows that

$$\begin{aligned} \lim_{n\to\infty} \frac{\theta_n}{\frac{2}{nS''(1)}} &= \lim_{n\to\infty} \theta_n \frac{nS''(1)}{2} \\ &= \lim_{n\to\infty} \theta_n (\frac{1}{\theta_n} - 1 - O(\log n)) \\ &= \lim_{n\to\infty} 1 - \theta_n - \theta_n O(\log n) \\ &= \lim_{n\to\infty} 1 - \theta_n - \lim_{n\to\infty} \theta_n O(\log n) \\ &= \lim_{n\to\infty} \bar{\theta}_n - \lim_{n\to\infty} \theta_n O(\log n) \\ &= 1 - 0 = 1. \end{aligned}$$

The last equality follows from the fact that in the critical case, all trees will eventually die out so $\bar{\theta}_n$ converges to 1. In the second term, $O(\log n)$ converges to $\infty$ in logarithmic time while $\theta_n$ converges to 0. Using (2.6) once again it follows that $\theta_n = 1 - S_n(0)$. Since $S_n(0)$ is a power series, $\theta_n$ converges faster to 0 than in logarithmic time. It then follows that the product of $\theta_n$ and $O(\log n)$ will converge to 0 for $n \to \infty$ which proves the collorary. □

### 2.2.2 Number of generations

Another area of interest is the number of generations in a finite tree, in the critical and subcritical case. It is known that this sequence will converge to 0, but at what rate? To determine this, define

$$p_n = \begin{cases} q_0, & n = 0 \\ P(E_n \backslash E_{n-1}), & n \geq 1 \end{cases}$$

as the probability that a tree terminates in exactly $n$ generations.

**Lemma 2.11.** *For $n \in \mathbb{N}_0$ and $S'(\bar{\theta}) \neq 1$*

$$p_n \leq \tilde{c} \cdot e^{-n/\xi}$$

*for certain constant $\tilde{c} \in \mathbb{R}$ and $\xi$ the correlation time as in (2.5).*

*Proof.* A tree with $r$ particles in the first generation can only terminate in exactly one generation when all these particles produce zero progeny in the second generation. So

$$\begin{aligned} p_1 &= \sum_{i=1}^{\infty} q_i q_0^i \\ &= S(q_0) - q_0 \\ &= S(p_0) - p_0. \end{aligned}$$

Subsequently, a tree that ends in exactly $n$ generations can be thought of as a tree consisting of a first generation with a subtree connected to each particle in this first generation. This substree is at most $n-1$ generations long, otherwise it wouldn't terminate in $n$ generations. If there are $i$ such particles, then at least one of these $i$ subtrees must not be in $E_{n-2}$ otherwise the whole tree would terminate in less than $n$ generations. This translates in counting the subtrees in $E_{n-1}$ and then substracting the subtrees in $E_{n-2}$:

$$\begin{aligned} p_n &= \sum_{i=1}^{\infty} q_i (P(E_{n-1})^i - P(E_{n-2})^i) \\ &= S(P(E_{n-1})) - S(P(E_{n-2})) \\ &= S(\sum_{r=0}^{n-1} p_r) - S(\sum_{r=0}^{n-2} p_r). \end{aligned}$$

Applying the Mean Value Theorem as in lemma 2.7 once again yields that for all $n_0, n \in \mathbb{N}_0$ there exist $\beta_j \in (\sum_{r=0}^{n_0+j} p_r, \sum_{r=0}^{n_0+n-1} p_r)$, $j \in \{1, \ldots, n\}$ such that

$$p_{n_0+n} = \prod_{j=1}^{n} S'(\beta_j) p_{n_0}$$

and for $n_0 = 0$

$$p_n = p_0 \prod_{j=1}^{n} S'(\beta_j).$$

Since $\bar{\theta} = \sum_{r=0}^{\infty} p_r = \lim_{n\to\infty} \sum_{r=0}^{n-1} p_r$ and $\beta_j \in (\sum_{r=0}^{j} p_r, \sum_{r=0}^{n-1} p_r)$ it follows that $\beta_j \leq \bar{\theta}$ for all $j \in \{1, \ldots, n\}$.
Since $S'(x)$ is a stricly increasing function on [0,1] this means that $S'(\beta_j) \leq S'(\bar{\theta})$ for all $j$. $S'(x)$ is also a power series with non-negative coefficients, combining this with $\beta_j, \bar{\theta} \geq 0$ it follows that $S'(\beta_j), S'(\bar{\theta}) \geq 0$ for all $j \in \{1, \ldots, n\}$.

Applying these inequalities to the previous equation leads to

$$\begin{aligned} p_n &= p_0 \prod_{j=1}^{n} S'(\beta_j) \\ &\leq p_0 \prod_{j=1}^{n} S'(\bar{\theta}) \\ &= p_0 \cdot [S'(1-\theta)]^n \\ &= \tilde{c} \cdot \mathrm{e}^{-n/\xi} \end{aligned}$$

with $\tilde{c} = p_0 = q_0$ and the correlation time $\xi$ as in (2.5). □

# 3 The mitochondrial Eve model

The results gained from the Galton-Watson model can now be used for the mitochondrial Eve model. Theorem 2.3 states a condition for the existence of an infinite tree depending on the mean number of sons. In the model for the mitochondrial Eve however results are gained in terms of the distribution of a woman's children, regardless of gender. This model depends on four assumptions:

1. There is no overlap between generations. (Meaning that a particle in generation $n$ dies in generation $n+1$.)
2. Each women has $r$ children with probability $Q_r$, $r \in \mathbb{N}_0$ independent and identically distributed. The values for $Q_r$ are time-and population size-independent.
3. A child is female with probability $p$, and male with probability $1-p$. The value for $p$ is time- and population size-independent.
4. There always exist enough males to procreate.

According to these assumptions, a genealogical tree can be constructed for each woman, being the root of the tree, and her descendants. Define an edge to be open if it connects a mother and her daughter and all other edges to be closed. The question if there exists an infinite lineage of mtDNA corresponds with the existence of an infinite path of open edges starting at the root.

However, in this genealogical tree not all edges may be statistically independent, which is a condition for the Galton-Watson process. This problem can be solved by defining the female genealogical tree (FGT) as the original tree after removing all male descendants. The question now becomes the probability of an infinite FGT, in which the number of daugthers $q_r$ is distributed by

$$q_r = \sum_{k=r}^{\infty} Q_k \binom{k}{r} p^r (1-p)^{k-r}. \tag{3.1}$$

To determine with which probability the mitochondrial Eve exists and under which conditions she exists, two questions are of importance namely:

1. What is the probability $\theta$ of an infinite FGT?
2. What is the probability of there existing only one infinite FGT?

The theory developed by Galton and Watson as shown in the previous sections can be used to answer the first question, whereas the second question requires some extra work.

In this section the same symbols will be used in the same way as in the Galton-Watson model. So $\theta$ and $\bar{\theta}$ will be used for the survival and extinction probability, respectively, of an FGT. The letters $m$ and $v$ will be used for the mean and variance of the number of daugthers. Lastly, the generation function $S(x)$ will also be used as the generating function of the distribution of $q_r$ for $r$ daugthers as opposed to $r$ sons.

## 3.1 Extinction probability

From (3.1) the mean and variance of the number of daugthers can be determined. Let $\bar{N} = \sum_{r=1}^{\infty} rQ_r$ the mean number of children of either gender. The number of daugthers is then binomially distributed with parameters $\bar{N}$ and $p$. The mean and variance are thus given by

$$m = \bar{N}p \text{ and } v = \bar{N}p(1-p). \tag{3.2}$$

Using theorem 2.3 and (3.2) the condition for the existence of an infinite FGT follows immediately.

**Theorem 3.1.** *Let* $p_c = \frac{1}{\bar{N}}$. *Then*

$$\theta > 0 \text{ if } p > p_c \ .$$

*Proof.* From theorem 2.3 it follows that the survival probability $\theta$ is positive only when the mean number of daugthers is larger than 1. Using (3.2) it shows that

$$\bar{N}p > 1 \text{ so } p > \frac{1}{\bar{N}} = p_c.$$

This concludes the proof and gives the necessary condition for the existence of an infinite FGT. □

### 3.1.1 Approximation of $\theta$

In the mitochondrial Eve model it can also prove difficult to determine the exact value of $\theta$. Since this model is based on the distribution of all children, regardless of gender, this could be even more difficult than in the Galton-Watson model. However, it is again possible to estimate this value using only the mean and variance of the number of daugthers, in the small survival probability approximation.

**Lemma 3.2.** *For $m > 1$, $S'''(1) < \infty$ and certain $\zeta \in [\bar{\theta}, 1]$*

$$\theta = \theta_a + \tilde{R}(\theta)$$

*with*

$$\theta_a = \frac{2p_c(p-1)}{p^2(1-p_c)} \text{ and } \tilde{R}(\theta) = \frac{2p_c^2\theta^2}{3!}\frac{S'''(\zeta)}{p^2(1-p_c)}.$$

*Proof.* Using lemma 2.5 and (3.2) it follows that

$$\begin{aligned}
\theta_a &= \frac{2(\bar{N}-1)}{\bar{N}p(1-p) + \bar{N}p(\bar{N}p-1)} \\
&= \frac{2(\bar{N}p-1)}{-\bar{N}p^2 + \bar{N}^2p^2} \\
&= \frac{2(\frac{p}{p_c}-1)}{p^2(\frac{1}{p_c^2}-\frac{1}{p_c})} \\
&= \frac{2(\frac{p}{p_c}-1)}{p^2(\frac{1-p_c}{p_c^2})} \\
&= \frac{2p_c(p-1)}{p^2(1-p_c)}.
\end{aligned}$$

The error term can also be determined immediately using lemma 2.5. It holds that

$$\begin{aligned}
\tilde{R}(\theta) &= \frac{2\theta^2}{3!}\frac{S'''(\zeta)}{\bar{N}p(1-p) + \bar{N}p(\bar{N}p-1)} \\
&= \frac{2\theta^2}{3!}\frac{S'''(\zeta)}{p^2(\frac{1-p_c}{p_c^2})} \\
&= \frac{2p_c^2\theta^2}{3!}\frac{S'''(\zeta)}{p^2(1-p_c)}
\end{aligned}$$

for certain $\zeta \in [\theta, 1]$. As in lemma 2.5 it follows that $S'''(\zeta)$ reaches its maximum on the interval for $\zeta = 1$ and since $S'''(1) < \infty$ the error term is bound. □

As seen in the Galton-Watson model, to ensure that the error term is small a maximum can be imposed on $r$. Likewise, this will not be too great a restriction on the mitochondrial Eve model.

## 3.2 Number of lineages

To answer the second question of when there exists only one infinite FGT, let $W$ be the number of women contemporary to the mitochondrial Eve and $Y_n$ the number of lineages remaining after $n$ generations. The distribution of $Y_n$ is then given by

$$P(Y_n = l) = \binom{W}{l} \theta_n^l (1 - \theta_n)^{W-l}. \tag{3.3}$$

The existence of the mitochondrial Eve is then compatible with the survival of only one lineage. By looking at the survival of two or more lineages it was shown in [2] that the existence of the mitochondrial Eve would only be possible in a stable-sized population while in [8] it was concluded that this would also be possible in a exponentially growing population. Both articles however rely on assumptions of values for $p$ and $W$, and assumptions on the progeny distribution. To give an exact probability for the existence of only one lineage thus proves to be quite difficult. Of course it depends on the survival probability over $n$ generations, as seen in (3.3). It is mentioned earlier that this value depends solely on the progeny distribution and is not exactly known in non-trivial cases.

### Poisson approximation

To be able to say something about the existence of only one lineage, it is possible to determine the probability in one particular situation. This can be done using the fact that a binomial distribution with parameters $n$ and $p$ can be approximated by a Poisson distribution with parameter $np$ if $n$ is large enough and $p$ is small enough.

The mean number of lineages is given by $\mathbb{E}(Y_n) = W\theta_n$. For the existence of the mitochondrial Eve an infinite amount of generations is needed, so in this case $\lim_{n\to\infty} \mathbb{E}(Y_n) = W\theta$.
Assume that this value is equal to 1, so $W\theta = 1$. The value of $W$ can be estimated from variability in nuclear DNA of modern humans as in [10]. While an exact value is not known, it is known that $W \gg 1$. It then follows that $\theta \ll 1$, so the conditions for the Poisson approximation are in place. For $\mathbb{E}(Y_n) = 1$ it holds that

$$\lim_{n\to\infty} P(Y_n = l) = \frac{(W\theta)^l \mathrm{e}^{-W\theta}}{l!} = \frac{\mathrm{e}^{-1}}{l!}.$$

This gives a probability of $\mathrm{e}^{-1}$ for the existence of only one lineage.

# 4 Conclusion

In this thesis the probability of the existence of the mitochondrial Eve is studied using results from the Galton-Watson process. Given the many comparisons between these subjects, the condition for a positive survival probability in the mitochondrial Eve model can be immediately determined. The exact value of $\theta$ can be difficult to determine, but it is possible to approximate this value in the small survival probability approximation.

The existence of the mitochondrial Eve however not only translates to the existence of an infinite tree, but to the existence of only one such infinite tree. Given the number of women contemporary to the mitochondrial Eve, the question is whether or not only one lineage will survive after an infinite number of generations. The maximum probability of one lineage surviving is determined in one specific case by using a Poisson approximation. Further studies into the subject could attempt to generalise and determine the probability of one lineage surviving with different progeny distributions.

The mitochondrial Eve could help answer questions about the early origins of the human species. Another entity which could play a role is the Y-chromosomal Adam, which would be a most recent paternal common ancestor. The dating of this ancestor is much less clear compared to the mitochondrial Eve, varying from 120.000 to 388.000 years ago. [7],[9]

It would seem that the results developed for the mitochondrial Eve model could be used for a Y-chromosomal Adam model. One objection however is raised in the fourth assumption in the mitochondrial Eve model, which is the existence of enough males to procreate. The equivalence of this statement would prove to be an objection due to the birth rate of males being slightly higher than that of females. When viewing a large number of generations this will definitely make a difference. Another continuation could then be the study of the Y-chromosal Adam using an adapted model, or a biparental model in which lineages of men are modelled alongside those of women.

It is clear that the theory first finalised by Galton and Watson in 1874 is still applicable to a number of situations. Using and adapting this theory could provide answers on ancestry and evolution.

# References


[1] Adams, R.A., Essex, C., *Calculus: A Complete Course*. (7th ed.) Pearson Education Canada, 2009

[2] Avise, J.C., Neigel, J.E., Arnold, J., *Demographic influences on mitochondrial DNA lineage survivorship in animal populations*. Journal of Molecular Evolution 20.2 (1984): p. 99-105.

[3] Cann, R. L., Stoneking, M., Wilson, A.C., *Mitochondrial DNA and Human Evolution*. Nature, vol. 325 (1987): p. 31-36

[4] Gibbons, A., *Mitochondrial Eve: wounded, but not dead yet*. Science, vol. 257 (1992): p. 873-875

[5] Harris, T.E., *The theory of branching processes*. Courier Dover Publications, 2002.

[6] Kendall, D.G., *Branching processes since 1873*. Journal of the London Mathematical Society 1.1 (1966): p. 385-406.

[7] Mendez, F.L., et al. *An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree*. The American Journal of Human Genetics 92.3 (2013): p. 454-459.

[8] Neves, A.G.M., Moreira, C.H.C., *Applications of the Galton-Watson process to human DNA evolution and demography*. Physica A: Statistical Mechanics and its Applications, Volume 368, Issue 1 (2006): p. 132-146

[9] Poznik, G.D., Henn, B.M., Yee, M., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A., Bustamante, C.D., *Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females*. Science, vol. 341, (2013): p. 562-565

[10] Takahata N., *Allelic genealogy and human evolution*. Molecular Biology and Evolution 10.1 (1993): p. 2-22.

[11] Thorne, A.G., Wolpoff, M.H., *The multiregional evolution of humans*. Scientific American 266,4 (1992) p. 76-83

[12] Wilson, A.C., Cann, R.L., *The recent african genesis of humans*. Scientific American 266.4 (1992): p. 68-73