ON THE NUMBER OF SIBLINGS AND p-TH COUSINS IN A LARGE POPULATION SAMPLE

VLADIMIR SHCHUR[1] AND RASMUS NIELSEN[1,2]

1. Abstract

The number of individuals in a random sample with close relatives in the sample is a quantity of interest when designing Genome Wide Association Studies (GWAS) and other cohort based genetic, and non-genetic, studies. In this paper, we develop expressions for the distribution and expectation of the number of p-th cousins in sample from a population of size N under two dioecious generalizations of the Wright-Fisher model. We also develop simple asymptotic expressions for large values of N. For example, the expected proportion of individuals with at least one p-th cousin in a sample of K individuals, for a non-monogamous generalization of the Wright-Fisher model, is approximately $e^{-(2^{2p-1})N/K}$. Our results show that a substantial fraction of individuals in the sample will have at least a second cousin if the sampling fraction (K/N) is on the order of 10^{-2}. This confirms that, for large cohort samples, relatedness among individuals cannot easily be ignored.

2. Introduction

As genomic sequencing and genotyping techniques are becoming cheaper, the data sets analyzed in genomic studies are becoming larger. With an increase in the proportion of individuals in the population sampled, we might also expect an increase in the proportion of related individuals in the sample. For example, Moltke *et al.* (2014) found in a sample of 2,000 Inuit from Greenland that almost half of the sample had one or more close relatives in the sample. The census population size for Greenland Inuit is only about 60,000 individuals and the population size might be substantially lower. Henn *et al.* (2012) found 5000 third-cousin and 30,000 fourth cousin relatives in a sample of 5000 self-reported Europeans, with nearly every individual having a detected cryptic relationship. In Genome Wide Association Mapping Studies (GWAS), related individuals are routinely removed from the sample, but other strategies also exist for using relatedness as a covariate in the statistical analyses (e.g., Visscher *et al.* 2008). These observations raise the following question: given a particular effective population size, how many close relatives would we expect to find in a sample? The answer to this question may help guide study designs and strategies for addressing relatedness in population samples and improve design for GWAS. Of particular interest is the number of individuals in the sample without relatives, i.e. the number of individuals remaining in the sample if individuals with relatives are removed.

Substantial progress has been made on understanding the structure of a pedigree in a population. For example, Chang (1999) showed that the most recent common ancestor of all present-day individuals is expected to have lived $log_2(N)$ generations in the past if N is the population size. A great deal of progress has also been made in understanding the difference between genealogical processes in full diploid pedigree models versus the approximating coalescent process (e.g., Wakeley et al. 2012; Wilton et al. 2016). However, the distribution and expectation of the number of individuals with relatives in a random population sample is still unknown.

In this paper we will address this question by exploring two models that both are diploid and dioecious generalizations of the Wright-Fisher model. We will use these models to derive distributions and expectations of the number of individuals that have, or do not have, siblings, first, second, etc. cousins within a sample.

3. Dioecious Wright-Fisher Model

The Wright-Fisher model (Fisher 1930; Wright 1931) describes the genealogy of a population with constant effective population size N as follows: The first generation $G = \{g_1, g_2, \dots, g_N\}$ contains N individuals. A new generation $H = \{h_1, h_2, \dots, h_N\}$ with N individuals is created. Then for each individual h_i from H a parent g_j is selected randomly and uniformly from G. The process is then applied repeatedly to move forward in time as far as needed.

In our study we will keep track of the two parents of an individual. To model this we work under models similar to the Wright-Fisher model in the sense that generations do not overlap and for each individual we choose parents from the previous generation randomly and uniformly. In these models we assume that the population size in generation G_i is $2N$ and that there are exactly N male and N female individuals. Each individual from generation G_{i-1} (we enumerate generations backward in time starting from 0, i.e. G_0 is the present generation) is assigned to a parent pair (one male and one female) from G_i.

In the first model, which can be considered a monogamous model, we fix the parent pairs, i.e. we assume each male and female is part of exactly one potential parent pair. For each individual in H we choose a parent pair randomly and uniformly from the pool of pairs in G. In the second model, we assume that individuals in H chose male and female parents in G indepdenpendently of each other. We consider this second model to be a model on non-monogamous mating. Both models are similar to each other in that the marginal distribution of the number of offspring of each individual is binomially distributed with mean 2. However, they differ from each other in the correlation structure among parents. In particular, the monogamous model does not allow for half-sibs and full sibs have a very low probability under the non-monogamous model. We note that other generalizaiotns of the Wright-Fisher models could be considered, but most would likely have dynamics that are somewhat intermediate between these two models.

We say that two individuals are siblings if they have at least one common parent. Similarly, we say that two individuals are p-th cousins if they share at least on ancestor in generation G_{p+1}.

Let S be a random sample of individuals from G_0 of size K. In this paper we are interested in the number U_T of individuals in S which do not have $(T-1)$-order cousins ($T=1$ would stand for siblings, $T=2$ for first cousins, etc.) within S and have pedigrees with no cycles (no inbreeding). We will derive the probability distribution of U_1 and expectations of U_T for $T \geq 2$ in terms of Stirling numbers of the second kind.

Notice, that every genealogy has the same probability under the model. Hence our problem is equivalent to counting the number of possible genealogies with certain properties. To enumerate different genealogies, we will use the following approach. Firstly, we divide a sample S into subsets. Then we assume that individuals from the same subset have the same parents, and individuals from different subsets have different parents. This approach is a basis for our analyses and leads us to the proof of formulas for expectations of U_T.

4. Stirling numbers of the second kind and their generalization

As previously mentioned, we provide a formula for expectation of U_T in terms of Stirling numbers of the second kind. In this section we provide definitions and properties of these numbers.

The Stirling number of a second kind $S(n,k)$ is the number of ways to partition a set of size n into k non-empty disjoint subsets. These numbers can be computed using the recursion (Abramowitz and Stegun 1972)

$$S(n,k) = kS(n-1,k) + S(n-1,k-1),$$

with $S(0,0) = S(n,0) = S(0,n) = 0$ for $n > 0$. Notice that $S(n,n) = 1$.

An $r-$associated Stirling number of the second kind, $S_r(n,k)$ (Comtet 1974), is the number of partitions of a set of size n into k non-empty subsets of size at least r. These numbers obey a recursion formula (Comtet 1974) similar to that for Stirling numbers of second kind

$$S_r(n+1,k) = kS_r(n,k) + \binom{n}{r-1} S_r(n-1,k-1)$$

with $S_r(n,0) = S_r(1,1) = 0$. In particular, for $r=2$

$$S_2(n+1,k) = kS_2(n,k) + nS_2(n-1,k-1).$$

4.1. Uniformly valid approximation for $S_2(n,k)$.

The following useful approximation of Stirling numbers of the second kind is established by Temme (1993)

$$S_2(n,k) = \left(1 + O\left(\frac{1}{n}\right)\right) \sqrt{\frac{t_0}{(1+t_0)(x_0 - t_0)}} e^A k^{n-k} \binom{n}{k}, \tag{1}$$

where $t_0 = n/k - 1$, $x_0 \neq 0$ is the non-zero root of the equation

$$\frac{k}{n} x = 1 - e^{-x}, \tag{2}$$

and

$$A = -n \ln x_0 + k \ln(e^{x_0} - 1) - kt_0 + (n-k) \ln t_0.$$

The following form of this approximation is known

$$S_2(n,k) = \left(1+O\left(\frac{1}{n}\right)\right)\sqrt{\frac{n-k}{n(1-G)}}\frac{1}{G^k\ (\frac{n}{k}-G)^{n-k}}\left(\frac{n-k}{e}\right)^{n-k}\binom{n}{k}, \tag{3}$$

with $G = -W_0(-n/ke^{-n/k})$, where W_0 is the main branch of Lambert W-function (Olver et al. 2010).

We did not find a reference for this formula in the literature, so we provide briefly the proof. Notice that $-1/e < -n/ke^{-n/k} < 0$, hence $G \in (0,1)$. Let us show that $x_0 = n/k - G$ is the non-zero root of equation 2

$$1-e^{-x_0} = 1-e^{-\frac{n}{k}}e^{-W_0(-n/ke^{-n/k})} = 1-e^{-\frac{n}{k}}\frac{W_0(-n/ke^{-n/k})}{-n/ke^{-n/k}} = \frac{k}{n}\left(\frac{n}{k}+W_0(-n/ke^{-n/k})\right) = \frac{k}{n}x_0,$$

where the second equality is due to the Lambert function property $e^{-W(x)} = W(x)/x$. Substituting t_0 and x_0 in approximation 1 by their values and simplifying the formula, one gets the needed result. Obviously,

$$\sqrt{\frac{t_0}{(1+t_0)(x_0-t_0)}} = \sqrt{\frac{n-k}{n(1-G)}}.$$

Now consider $e^A k^{n-k}$

$$\begin{aligned}
e^A k^{n-k} &= (n/k-G)^{-n}(e^{n/k-G}-1)^k e^{-k(n/k-1)}(n/k-1)^{n-k}k^{n-k} = \\
&= (n/k-G)^{-n}\left(e^{\frac{n}{k}}\frac{-n/ke^{-n/k}}{W_0(-n/ke^{-n/k})}-1\right)^k\left(\frac{n-k}{e}\right)^{n-k} = \\
&= (n/k-G)^{-n}\left(\frac{n/k}{G}-1\right)^k\left(\frac{n-k}{e}\right)^{n-k} = \\
&= (n/k-G)^{-n+k}G^{-k}\left(\frac{n-k}{e}\right)^{n-k},
\end{aligned}$$

which finished the proof of equivalence of approximations (1) and (3).

5. Probability distribution U_1

We say that two individuals are siblings if they have a common parent. In this section we study the number of individuals U_1 without siblings within a sample of a population. We derive both the probability distribution and expectation of U_1.

Theorem 1. *Let U_1 be a random variable representing the number of individuals in a sample S of size K without siblings in S under monogamous dioecious Wright-Fisher model. Then*

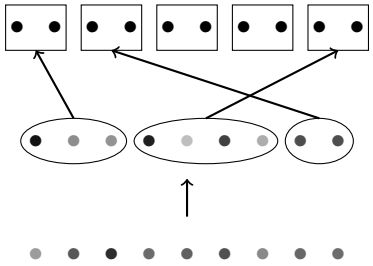


FIGURE 1. Illustration to the proof of Theorem 1: the process of partitioning a sample into subsets and then assigning them to different couples of parents.

- *the probability distribution of* U_1 *is*
$$\mathbb{P}(U_1 = u) = \frac{\binom{K}{u} \sum_{t=1}^{\lfloor \frac{K-u}{2} \rfloor} S_2(K-u,t)\binom{N}{u+t}(u+t)!}{\sum_{t=1}^{m} S(K,t)\binom{N}{t}t!};$$
- *the expectation of* U_1 *is*
$$\mathbb{E}(U_1) = K(1-1/N)^{K-1};$$
- *if* $N = \alpha K$
$$\lim_{K\to\infty} \frac{\mathbb{E}(U_1)}{K} = e^{-1/\alpha}.$$

Proof. We begin the proof by computing the number of possible partirtions of S into u subsets of size 1 and t subsets of size ≥ 2. Each such subset corresponds to the descendants in S of the same couple from G_1. There are $\binom{K}{u}S_2(K-u,t)$ such partitions (see figure 1). Here the first multiplier corresponds to the number of choices of the first u individuals and the second multiplier corresponds to the number of partitions of the remaining $K-u$ individuals into t subsets.

Now we need to assign $u+t$ subsets to different couples of parents from G_1. There are $\binom{N}{u+t}$ possibilities for choosing couples that have descendants in S and $(u+t)!$ permutations which assign these particular couples to different subsets of the given partitions of S.

Finally, summing over all possible values of t we get

$$\mathbb{P}(U_1 = u) = \frac{\binom{K}{u} \sum_{t=1}^{\lfloor \frac{K-u}{2} \rfloor} S_2(K-u,t)\binom{N}{u+t}(u+t)!}{\sum_{t=1}^{m} S(K,t)\binom{N}{t}t!},$$

where $\lfloor\cdot\rfloor$ stands for the floor integer part.

The expression for expectation of U_1 is much simpler. The probability π_1 that an individual does not have any siblings in S is $\pi_1 = (1-1/N)^{K-1}$, as all other individuals can be assigned to any couple except for the parents of the given individual. By linearity, the expectation of U_1 is

$$\mathbb{E}(U_1) = K\pi_1 = K(1-1/N)^{K-1}.$$

To prove the last statement it is enough to rewrite

$$\frac{\mathbb{E}(U_1)}{K} = (1-1/N)^{-1}\left((1-1/N)^N\right)^{K/N} = (1-1/N)^{-1}\left((1-1/N)^N\right)^{1/\alpha},$$

because $N = \alpha K$. Now notice that

$$\lim_{N\to\infty}(1-1/N)^N = e^{-1}.$$

Hence the last statement of the theorem is proved

$$\lim_{K\to\infty}\frac{\mathbb{E}(U_1)}{K} = e^{-1/\alpha}.$$

□

6. Expectation of U_2

In this section we will provide an expression for expectation of the number U_2 of individuals in a sample which do not have first cousins in this sample. We will also establish a limit for $\mathbb{E}(U_2)/K$ in the case of a fixed ratio between K and N.

Theorem 2. *Let U_1 be a random variable representing the number of individuals in a sample S of size K without siblings in S under a monogamous dioecious Wright-Fisher model. Then the expectation of U_1 is*

$$\mathbb{E}(U_2) = K\frac{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N(N-1)(N-2)^{2m-2}}{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N^{2m}}.$$

Proof. Similarly to the case of $\mathbb{E}(U_1)$, we need to find the probability π_2 for a single individual not to have first cousins within S. Then the expectation $\mathbb{E}(U_2) = K\pi_2$. Denote individuals from G_T which have descendants in S by S^T.

Choose an individual $s_0 \in S$, let p_1^0 and p_2^0 be parents of s_0. If s_0 does not have first cousins, then p_1^0 and p_2^0 are assigned to different couples from G_2 and those couples do not have other descendants in S^1.

Similarly to derivation of distribution of U_1, we first partition S into m subsets. We choose m couples from G_1 and establish a one-to-one correspondence between the subsets and the couples. There are N possibilities to choose a couple of parents for p_1^0, $N-1$ choices for p_2^0 and $(N-2)$ choices for all other $2m-2$ individuals from S^1. Summing over m we get

$$\mathbb{E}(U_2) = K\frac{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N(N-1)(N-2)^{2m-2}}{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N^{2m}}.$$

□

Our next goal is to find the limit of $\mathbb{E}(U_2)/K$ for a fixed ratio of sample size to the population size. We assume that $N = \alpha K$ for some constant $\alpha \geq 1$ and we consider the limit of $\mathbb{E}(U_2)/K$ for $K \to \infty$.

Theorem 3. *Let $\alpha \geq 1$ and set $N = \alpha K$. Then*

$$\lim_{K\to\infty} \mathbb{E}(U_2) = e^{-\frac{4}{\alpha}},$$

The following lemma states that the sum of the first βK terms of the series in the formula for $\mathbb{E}(U_2)$ is small for large values of K. This makes it possible to make further approximations under the hypothesis that $m = O(K)$.

Lemma 1. *Let $N = \alpha K$ for some $\alpha > 1$ and set $\beta = (2\ln 2)^{-1}$. Then*

$$\lim_{K\to K} \frac{\sum_{m=1}^{\lfloor \beta K \rfloor - 1} S(K,m)\binom{N}{m} m! N^{2m} \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2}}{\sum_{m=1}^{K} S(K,m)\binom{N}{m} m! N^{2m}} = 0.$$

For simplicity of notations we henceforth drop the integer brackets $\lfloor \cdot \rfloor$.

Proof. Denote

$$T_{K,N}(m) = S(K,m)\binom{N}{m} m! N^{2m}.$$

First, notice that

$$0 \leq T_{K,N}(m)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2} \leq T_{K,N}(m)$$

We will show that for $\beta = (2\ln 2)^{-1} < 1$ there exists a constant $0 < \beta < 1/2$ such that

$$\lim_{K\to\infty} \frac{\sum_{m=1}^{\beta K - 1} T_{K,N}(m)}{T_{K,N}(\beta K)} = 0, \tag{4}$$

which will immediately prove the statement of the Lemma.

Our goal is to prove that

$$T_{K,N}(m) \gtrsim c_1 e^{c_2 m} N^{2m}$$

for some constants c_1, c_2 and K large enough.

We begin by approximating the ratio for $m \leq \beta K$

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} = \frac{\left(1+O\left(\frac{1}{K}\right)\right)}{\left(1+O\left(\frac{1}{K}\right)\right)} \sqrt{\frac{K-m}{K(1-G_1)}\frac{K(1-G_2)}{K-m-1}} \frac{G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{K-m-1}}{G_1^m\left(\frac{K}{m}-G_1\right)^{K-m}} \left(\frac{K-m}{e}\right)^{K-m} \left(\frac{e}{K-m-1}\right)^{K-m-1} \frac{\binom{K}{m}}{\binom{K}{m+1}} \frac{1}{N-m}\frac{1}{N^2}, \tag{5}$$

by applying approximation 3.

Notice that $0 < G_1 < G_2 < -W_0(-2e^{-2}) < 1/2$. The following term is bounded by a constant (we remind the reader that $0 < m \le \beta K < K/2$)

$$\sqrt{\frac{1-G_2}{1-G_1}} \le \frac{1}{\sqrt{2(1+W_0(-1/2e^{-1/2}))}}.$$

After simplification, all the factorials in the formula are of the form $(constK)!$, hence they can be approximated uniformly in K by Stirling's approximation

$$n! = \left(1+O\left(\frac{1}{n}\right)\right)\sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

For simplicity of notations we drop all terms $1+O(1/K)$ in 5. We also notice that

$$\left(\frac{K-m}{K-m-1}\right)^{K-m-1} = \left(1+\frac{1}{K-m-1}\right)^{K-m-1} = e + O(1/K).$$

So for K large enough (5) has following approximation

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} \approx \sqrt{\frac{K-m}{K-m-1}\frac{1-G_2}{1-G_1}}\frac{G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{K-m-1}}{G_1^m\left(\frac{K}{m}-G_1\right)^{K-m}}\frac{m+1}{N-m}\frac{1}{N^2}.$$

The derivative of $G(x)^{1/x}(x-G(x))^{1-1/x}$ $(x \ge 1)$ with respect to x is

$$H(x) = \frac{G(x)^{\frac{1}{x}}(x-G(x))^{\frac{x-1}{x}}\left(\ln(x-G(x)) - \ln G(x)\right)}{x^2}. \tag{6}$$

$H(x)$ has one real root $x = 2\ln 2$ if $x \ge 1$. For $x > 2\ln 2$, $H(x) > 0$, so $G(x)^{1/x}(x-G(x))^{1-1/x}$ is an increasing function of x for $x > 2\ln 2$. Hence as soon as $K/m > 2\ln 2$, or $m < K/(2\ln 2)$, the following inequality holds

$$\frac{G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{K-m-1}}{G_1^m\left(\frac{K}{m}-G_1\right)^{K-m}} < 1.$$

Consequently, for sufficently large K we obtain the following upper bound for (5)

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} \le \frac{1}{\sqrt{2(1+W_0(-1/2e^{-1/2}))}}\frac{\beta}{\alpha-\beta}\frac{1}{N^2} =: \frac{A}{N^2}$$

Hence, by recursion for $m < \beta K$

$$T_{K,N}(m) \le \left(\frac{A}{N^2}\right)^{\beta K - m} T_{K,N}(\beta N).$$

Now we use the obtained inequality to prove limit (4)

$$\lim_{K\to\infty}\frac{\sum_{m=1}^{\beta K-1} T_{K,N}(m)}{T_{K,N}(\beta K)} \le \lim_{K\to\infty}\sum_{m=1}^{\beta K-1}\left(\frac{A}{N^2}\right)^{\beta K-m} = \lim_{K\to\infty}\frac{A}{N^2}\frac{(1-A/N^2)^{\beta K-1}}{1-A/N^2} = 0, \tag{7}$$

where the second equality holds by summing over the geometric progression. □

Lemma 2. *Let $N = \alpha K$ for some $\alpha > 1$, set $\beta = (2\ln 2)^{-1}$. Then for any m such that $\beta K \le m < K$*

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} \le O\left(\frac{1}{K}\right).$$

Proof. From the proof of Lemma 1, for K large enough and for $\beta \le m/K \le 1$

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} \le C\sqrt{\frac{1-G_2}{1-G_1}}\frac{G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{K-m-1}}{G_1^m\left(\frac{K}{m}-G_1\right)^{K-m}}\frac{1}{N^2}.$$

Notice that $xe^x = -1 + O((x-1)^2)$ near $x = -1$. Hence $1 - G(x) = O(|x-1|)$ and $x - G(x) = O(|x-1|)$ for $x \to 1$. By definition, the Lambert W-function is the inverse function of xe^x. If $x_1 > -1$ and $x_2 < -1$ are two points in the neighbourhood of -1 such that $x_1e^{x_1} = x_2e^{x_2}$, then $|x_1 - x_2| = O(|x_1 - 1|) = O(|x_2 - 1|)$. For $x > 1$, $-xe^{-x} \in [-1/e; 0]$. The value of the main branch, $W_0(xe^x)$, is in the interval $[-1, 0]$. So $-x$ and $W_0(-xe^{-x})$ correspond to x_1 and x_2.

Hence

$$\sqrt{\frac{1-G_2}{1-G_1}} = \frac{1-K/(m+1)}{1-K/m} = O(1).$$

Now we use mean value theorem to approximate

(8)
$$\left|G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{1-\frac{m+1}{K}} - G_1^m\left(\frac{K}{m}-G_1\right)^{1-\frac{m}{K}}\right| \le \left|\frac{K}{m}-\frac{K}{m+1}\right| \max_{[K/(m+1),K/m]}|H(x)|,$$

where $H(x)$ is given by expression (6). Denote $\Delta x = |x-1|$, and notice that

$$\hat{H}(x) := \frac{H(x)}{\ln(x-G(x)) - \ln G(x)} = \frac{G(x)^{\frac{1}{x}}(x-G(x))^{\frac{x-1}{x}}}{x^2}$$

and $\ln G(x)$ are continuous near $x = 1$ and $\hat{H}(1) = 1$, $\ln G(1) = 0$. So for small Δx

$$H(1+\Delta x) = O(\ln \Delta x),$$

and hence

$$\max_{[K/(m+1),K/m]}|H(x)| = |H(K/(m+1))| = O(\ln K)$$

which leads to the approximation of (8) with $m = O(K)$

$$\left|G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{1-\frac{m+1}{K}} - G_1^m\left(\frac{K}{m}-G_1\right)^{1-\frac{m}{K}}\right| \le \frac{K}{m(m+1)}|H(K/(m+1))| \lesssim \frac{\ln K}{K}$$

We use this estimate and the Taylor expansion of logarithm to get

$$\frac{G_2^{m+1}\left(\frac{K}{m+1}-G_2\right)^{K-m-1}}{G_1^m\left(\frac{K}{m}-G_1\right)^{K-m}} \lesssim \left(1+\frac{\ln K}{K}\right)^K \approx K.$$

Finally, we estimate the ratio $T_{K,N}(m)/T_{K,N}(m+1)$ for K large enough

$$\frac{T_{K,N}(m)}{T_{K,N}(m+1)} \leq \frac{C_0}{K}$$

with some constant C_0, which depend on α. □

Now we are ready to prove the theorem.

Proof. Firstly, notice that

$$1 \geq \left(1-\frac{2}{\alpha K}\right)^{2m} \geq \left(1-\frac{2}{\alpha K}\right)^{2K} \geq \left(1-\frac{2}{\alpha K}\right)^{\frac{4}{\alpha}\frac{\alpha K}{2}} \geq e^{-\frac{4}{\alpha}},$$

and $(1-1/N)(1-2/N)^2 \to 1$ as $N \to \infty$. Hence, the lower bound is valid for any α and K

$$\frac{E_2(\alpha,K)}{K} = \frac{\sum_{m=1}^{K} T_{K,N}(m)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2}}{\sum_{m=1}^{K} T_{K,N}(m)} \geq \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{-2} e^{-4/\alpha},$$

where the right part trivially converges to $e^{-4/\alpha}$ with $K \to \infty$ (we remind that $N = \alpha K$ for some constant α).

Now we prove that this bound is sharp by applying subsequently Lemmas 1 and 2

$$\begin{aligned}
\lim_{K\to\infty} \frac{E_2(\alpha,K)}{K} &= \lim_{K\to\infty} \frac{\sum_{m=1}^{K} T_{K,N}(m)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2}}{\sum_{m=1}^{K} T_{K,N}(m)} \\
&= \lim_{K\to\infty} \frac{\sum_{m=\beta K}^{K} T_{K,N}(m)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2}}{\sum_{m=\beta K}^{K} T_{K,N}(m)} \\
&\leq \lim_{K\to\infty} \frac{\sum_{m=\beta K}^{K-1} T_{K,N}(m) + T_{K,N}(K)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)^{2m-2}}{\sum_{m=\beta K}^{K-1} T_{K,N}(m) + T_{K,N}(K)} = e^{-4/\alpha},
\end{aligned}$$

because from Lemma 2 it follows

$$0 \leq \lim_{K\to\infty} \frac{\sum_{m=\beta K}^{K-1} T_{K,N}(m)}{T_{K,N}(K)} \leq \lim_{K\to\infty} \frac{\sum_{m=\beta K}^{K-1} \left(\frac{C_0}{K}\right)^{K-m} T_{K,N}(K)}{T_{K,N}(K)} = \lim_{K\to\infty} \frac{C_0}{K} \frac{1-\left(\frac{C_0}{K}\right)^{K}}{1-\frac{C_0}{K}} = 0.$$

□

7. General case: expectation of U_p for $p \geq 1$

Similarly to the expectation of U_2, we can find the probability the expected numbers U_3 and U_4 of individuals which do not have second and third cousins and with pedigrees without cycles.

Lemma 3. *Let S be a set and $S' \subset S$ be a subset of size $|S'| = p$. The number of partitions of a set S of size N into M disjoint subsets such that all elements of S' are in different subsets is*

$$Q_p(N, M) = \sum_{t=0}^{p} \binom{p}{t} S(N-p, M-t) \binom{M-t}{p-t}.$$

Proof. Let $S'' \subset S'$, $S'' = \{e_1, e_2, \ldots, e_t\}$, such that each element, $e_i \in S''$, makes its own subset $P_i = \{e_i\}$ in the partition of S. If $t = |S''|$ there are $\binom{p}{t}$ ways to choose such a subset. Then, $S \setminus S'$ should be split into $M - t$ non-empty subsets, $P_{t+1}, P_{t+2}, \ldots, P_M$, to obtain a partition of S into exactly M subsets. There are $S(N-p, M-t)$ possible ways of doing that. Each of the $p - t$ elements of $S' \setminus S''$ are then added to distinct subsets among the remaining $M - t$ subsets, $P_i, i > t$, which can be done in $\binom{M-t}{p-t}$ ways.

Summing over all possible values of t we prove the statement. □

Remark 1. For $p = 1$, Lemma 3 turns into the well-known recursive formula for Stirling numbers of the second kind.

The next theorem establishes the expression for the expectation of U_p and its limit for fixed K to N ratio in the general case. Due to the size of the formula we had to introduce additional notations for readability.

Theorem 4.

- *For any natural $p \geq 1$ the expectation of U_p is*

$$\mathbb{E}(U_p) = K \frac{\underbrace{\sum_{m_1=1}^{K} R_1 \sum_{m_2=2}^{2m_1} R_2 \ldots \sum_{m_{p-1}=4}^{2m_{p-2}} R_{p-1}}_{(p-1)\ \textit{nested summations}} N^{2m_{p-1}} W(p)}{\underbrace{\sum_{m_1=1}^{K} R_1 \sum_{m_2=2}^{2m_1} R_2 \ldots \sum_{m_{p-1}=4}^{2m_{p-2}} R_{p-1}}_{(p-1)\ \textit{nested summations}} N^{2m_{p-1}}}, \tag{9}$$

  where by convention we assume $2m_0 := K$,

$$R(j) = Q_{2^{j-1}}(2m_{j-1}, m_j) \binom{N}{m_j} m_j!,$$

  and

$$W(p) = \left(1 - \frac{2^{p-1}}{N}\right)^{2m_{p-1} - 2^{p-1}} \prod_{s=1}^{2^{p-1}} \left(1 - \frac{s}{N}\right)$$

- *If $N = \alpha K$ $(i = 1, 2, \ldots, p)$, then*

$$\lim_{K \to \infty} \frac{\mathbb{E}(U_p)}{K} = \lim_{K \to \infty} \left(1 - \frac{4}{\alpha K}\right)^{4K} = e^{-(2^{2p-2})/\alpha}. \tag{10}$$

Proof. To prove the first statement, we apply repeatedly the same arguments as used for Theorem 2: for each generation, we split the ancestors of the sample into subsets of siblings while controlling that ancestors of the given individual are not in the same subsets.

The proof of (10) is similar to the proof of Theorem 3. First we can show that we can substitute summations over $m_i > \beta K$ for some constant β (see Lemma 1). Then we use estimations for Q_i that are similar to those obtained in Lemma 2. □

8. Non-monogamous Wright-Fisher model

Similar results to those obtained for the monogamous case also hold for the non-monogamous dioecious Wright-Fisher model. However, in contrast to the monogamous case, the probability that two individuals are full siblings or full p-th cousins (i.e. sharing two ancestors) is rather small. Most familial relationships would involve sharing only one common ancestor at a given generation, i.e. related individuals would typically be half siblings or half p-th cousins.

Let V_p be a random variable representing the number of individuals in a sample S of size K without half siblings or full siblings ($p = 1$) or half p-th cousins or full p-th cousins ($p \geq 2$) in S under the non-monogamous Wright-Fisher model. The next theorem established the expression for the expectation of V_p and its limit for $K \to \infty$ in the case of fixe ratio between K and the population sizes N.

Theorem 5.

- *For any natural $p \geq 1$, the expectation of V_p is*

$$\mathbb{E}(V_p) = K \frac{\underbrace{\sum_{m_1=1}^{K} P_1 \sum_{m_2=2}^{2m_1} P_2 \dots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P_{p-1}\, N^{2m_{p-1}} W^2(p)}_{(p-1)\ \textit{nested summations}}}{\underbrace{\sum_{m_1=1}^{K} P_1 \sum_{m_2=2}^{2m_1} P_2 \dots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P_{p-1}\, N^{2m_{p-1}}}_{(p-1)\ \textit{nested summations}}}, \tag{11}$$

where we assume $m_0 = K$ and

$$P_j := \sum_{n=2^{j-1}}^{m_j - 2^{j-1}} Q_{2^{j-1}}(m_{j-1}, n) Q_{2^{j-1}}(m_{j-1}, m_j - n) \binom{N}{n} \binom{N}{m_j - n} n!(m_j - n)!$$

and

$$W(p) = \left(1 - \frac{2^{p-1}}{N}\right)^{m_{p-1} - 2^{p-1}} \prod_{s=1}^{2^{p-1}-1} \left(1 - \frac{s}{N}\right).$$

- *If population sizes $N = \alpha K$, then*

$$\lim_{K \to \infty} \frac{\mathbb{E}(V_p)}{K} = e^{-(2^{2p-1})/\alpha}.$$

The proof of the theorem is similar to the case of monogamous model.

In particular,

$$\mathbb{E}(V_1) = K(1 - 1/N)^{2(K-1)}.$$

Corollary 1. *The qualitative behaviour of U_i and V_i is the same, more precisely*

$$\lim_{K\to\infty} \frac{\mathbb{E}(V_i)}{K} = \left(\lim_{K\to\infty} \frac{\mathbb{E}(U_i)}{K}\right)^2.$$

9. Numerical results

In this section we present numerical results for expectations of U_p and V_p, $p = 1, 2, 3$. Every plot of figures 2 and 3 represents the behaviour of $\mathbb{E}(U_p)/K$ or $\mathbb{E}(V_p)/K$ for a particular $p = 1, 2, 3$. Those values are computed by formulas (9) or (11) for different values of N ($N = 20, 100, 200$) as a function of the ratio K/N. We also add corresponding limiting distribution to every plot to illustrate the convergence.

Because the effective population sizes are typically rather large (at least thousands of individuals) we might expect a satisfactory approximation of $E(U_p)$ and $E(V_p)$ by its limiting distribution even for relatively small K/N ratios. One can also check that in our proofs the errors in the estimates are of the order of $1/N$, hence for the desired ratio we can estimate the absolute error for smaller values of K, N numerically and then increase N to get the desired precision.

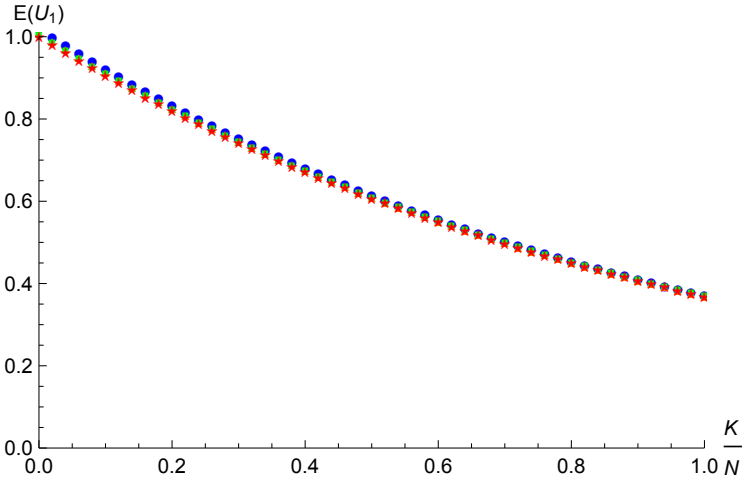
10. Discussion

In this paper we analysed the expected values of the number of individuals without siblings and p-th cousins in a large sample of a population. To do that we suggested two extensions of Wright-Fisher model which keeps track of two parents of an individual. The first extension corresponds to a monogamous population and the second to a non-monogamous population. The two models represent two extremes, and we might expect that in most other dioecious generalizations of the Wright-Fisher model, the number of individuals without siblings or p-the cousins is somewhere in between those two regimes.

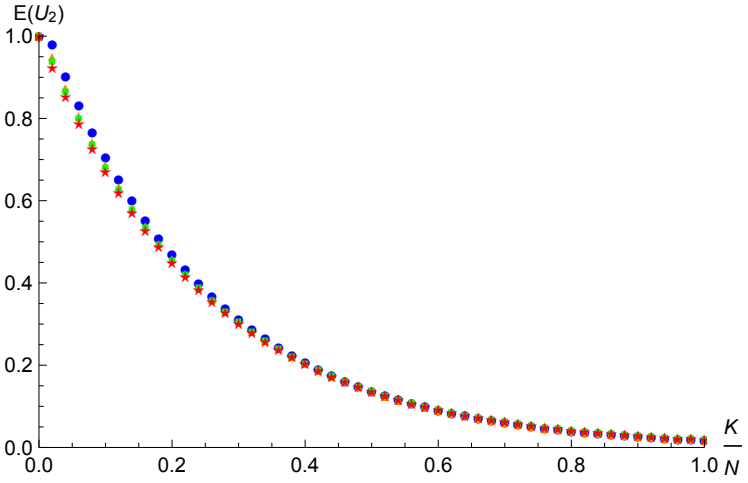
Under both models we derived expressions for these expectations under the hypothesis that the pedigrees have no cycles (except for the one appearign in full sibs). Notice that this restriction is not too strong, because one can easily show that the chance that an individual has a pedigree with a cycle is a second-order effect as soon as the number of ancestors ($\leq 2^p$) in a generation is much smaller than the effective population size N.

The important result of the paper is the limiting distributions for $\mathbb{E}(U_p)/K$ and $\mathbb{E}(V_p)/K$. It turns out that $\mathbb{E}(U_p)/K$ and $\mathbb{E}(V_p)/K$ converge pointwise to $e^{-cK/N}$ where the constant c is 2^{2p-2} for U_p and 2^{2p-1} for V_p.

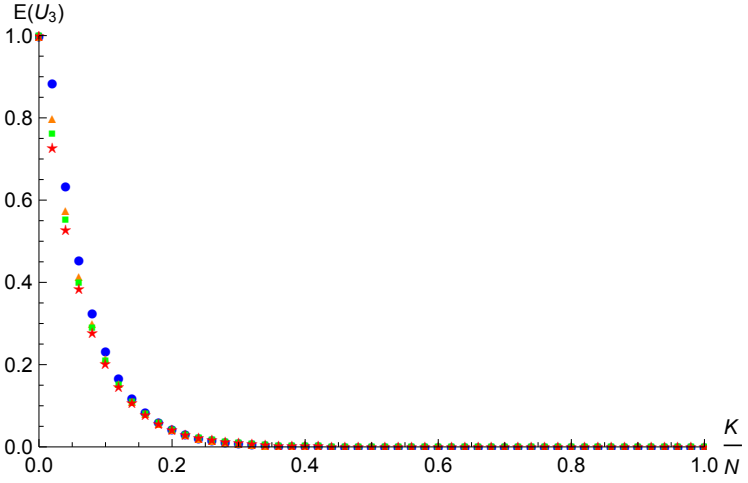
We notice that even when the sampling fraction is relative low, the proportion of individuals in the sample with no close relatives can be small. For example, for the non-monogamous model and a sampling faction of 5%, the proportion of individuals with at least a second cousing is approx. 70% if the population size is at least $N = 200$. For a sampling fraction of 2% the proportion in individuals with at least a second cousin is



(A) p=1, individuals without siblings.



(B) p=2, individuals without first cousins.



(C) p=3, individuals without second cousins.

FIGURE 2. $\mathbb{E}(U_p)/K$ as a function of the K/N ratio for $N = 50$ (●), 100 (▲), 200 (■) and the corresponding limiting distribution (⋆).

close to 50% for reasonably large population sizes in case of random mating population or almost 30% in case of monogamous population. For sampling fractions on the order of 0.01 or larger, we expect a large proportion of individuals to have at least one other individual in the sample to which they are closely related. This fact should be taken into account in all genetic, and non-genetic, epidemiological studies working on large cohorts.

11. Acknowledgement

The work was supported by the UCOP Catalyst Award CA-16-376437.

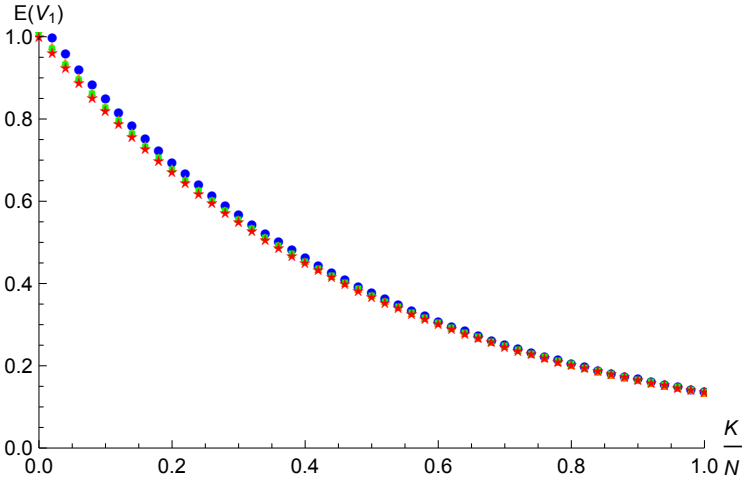
References

[1] Abramowitz M, and Stegun IA, (1972) Handbook of mathematical functions with formulas, graphs and mathematical tables. Dover, New York, p 825
[2] Comtet L (1974) Advanced Combinatorics. Reidel, Dordrecht, Holland
[3] Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
[4] Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, et al. (2012) Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLoS ONE 7(4):e34267. doi:10.1371/journal.pone.0034267
[5] Moltke I, Fumagalli M, Korneliussen TS, Crawford JE, et al (2015) Uncovering the genetic history of the present-day Greenlandic population. Am J Hum Genet 96:54-69. doi:10.1016/j.ajhg.2014.11.012
[6] Nagylaki T (1997) Multinomial-Sampling Models for Random Genetic Drift. Genetics 145:485-491
[7] Olver FWJ et al (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge, p 131
[8] Temme NM (1993) Asymptotic Estimates of Stirling Numbers. Stud Appl Math 89:233-243. doi: 10.1002/sapm1993893233
[9] Visscher PM, Andrew R, Nyholt DR (2008) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. Eur J Hum Genet 16:387-390. doi:10.1038/sj.ejhg.5201990
[10] Wakeley J, King L, Low BS, Ramachandran S (2012) Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingmans Coalescent. Genetics 190: 1433-1445. doi: 10.1534/genetics.111.135574
[11] Wilton PR, Baduel P, Landon MM, and Wakeley J (2016) Population structure and coalescence in pedigrees: comparisons to the structured coalescent and a framework for inference. doi:10.1101/054957
[12] Wright S (1931) Evolution in Mendelian populations. Genetics. 16:97-159

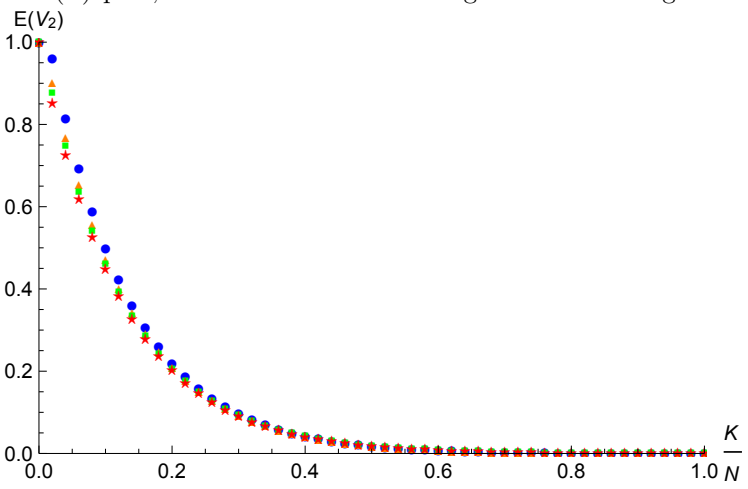
[1]Department of Integrative Biology, University of California Berkeley 4098 Valley Life Sciences Building (VLSB) Berkeley, CA 94720-3140

[2]Museum of Natural History, University of Copenhagen, Øster Voldgade 5-7, 1350 København K, Denmark

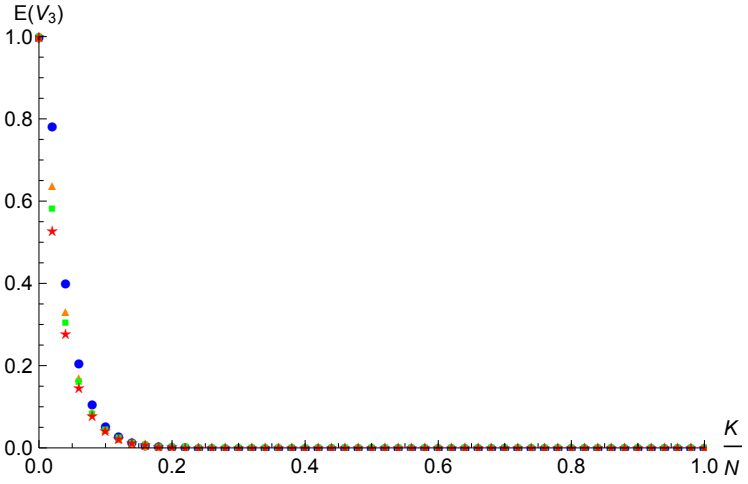
Corresponding author is V.Shchur, vlshchur@gmail.com



(A) p=1, individuals without siblings and half siblings.



(B) p=2, individuals without first-cousins and half first cousins.



(C) p=3, individuals without second cousins and half second cousins.

FIGURE 3. $\mathbb{E}(V_p)/K$ as a function of the K/N ratio for $N = 50$ (●), 100 (▲), 200 (■) and the corresponding limiting distribution (⋆).