bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 a sample from a population of size N under two diploid Wright-Fisher models. 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 diploid dioecious Wright-Fisher model, is approximately $1 - e^{-(2^{2p-1})K/N}$. 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 analysed 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 effective population size might be substantially lower. Henn *et al.* (2012) found 5000 pairs of third-cousin and 30,000 pairs of fourth cousin relatives in a sample of 5000 self-reported Europeans, with nearly every individual having a detected cryptic relationship. In Genome Wide Association 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.

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 diploid and dioecious Wright-Fisher models. 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. The model assumes that generations do not overlap. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ and $\hat{\mathcal{G}} = \{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_N\}$ be two successive generations with N individuals in each. Then for each individual $\hat{g}_i$ from $\hat{\mathcal{G}}$ a parent g_j is selected randomly and uniformly from $\mathcal{G}$.

In our study we consider a diploid population where each individual has two parents, one male and one female. Similarly to the original haploid Wright-Fisher model, the dioecious Wright-Fisher model (see e.g. Nagylaki 1997, King *et al.* 2017) assumes that generations do not overlap and, for each individual, the parents are chosen from the previous generation uniformly at random. The difference is that instead of a single parent, in the dioecious case, each individual has two parents, one male and one female, which are drawn independently from the corresponding sets of males and females in the preceding generation. We will refer to this model as the 'non-monogamous Wright-Fisher model' because we will also consider a model in which female and male parents form monogamous pairs. We will refer to the latter model as the 'monogamous Wright-Fisher model'. As we will assume exactly equal proportions of males and females, the monogamous Wright-Fisher model is identical to the bi-parental monoecious model in King *et al.* (2017).

For both the non-monogamous and monogamous models, we assume that there are exactly N male and N female individuals. Each individual from generation $\mathcal{G}_{i-1}$ (we enumerate generations backward in time starting from 0, i.e. $\mathcal{G}_0$ is the present generation and $\mathcal{G}_1$ is the generation of parents of individuals from $\mathcal{G}_0$) is assigned to a parent pair (one male and one female parent) from $\mathcal{G}_i$. As we described above, under the non-monogamous model, male and female parents are chosen independently from each other for every individual. In the monogamous case, the parent pairs are fixed, i.e. we assume each male and female is part of exactly one potential parent pair.

The two diploid 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. The important

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

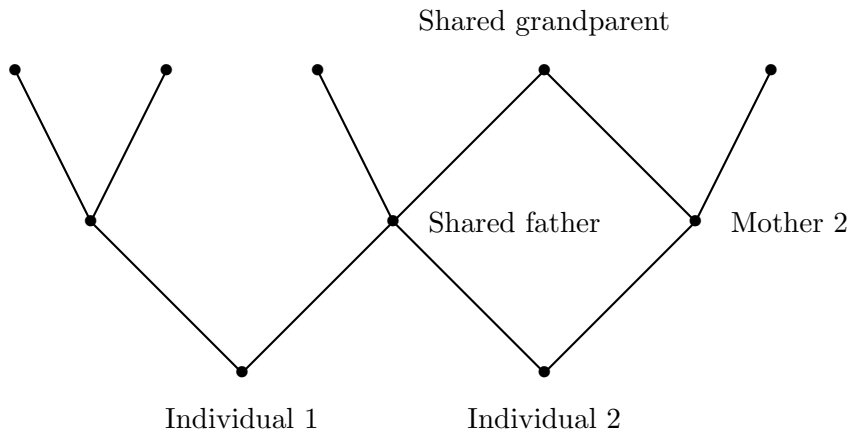


FIGURE 1. The two offspring (Individual 1 and Individual 2) related by this genealogy are half-siblings and first cousins at the same time. Notice that Individual 1 has a tree-like genealogy (no cycles, no inbreeding). The second individual though has inbreeding in its genealogy.

difference between these two models is that the monogamous model does not allow for half-siblings (we say that two individuals are half-siblings if they share only one parent). On the other hand under the non-monogamous model, full siblings (individuals which share both parents) have a very low probability of appearing.

We note that other dioecious versions of the Wright-Fisher models could be considered with varying degree of promiscuity, but most would likely have distributions of relatedness that are somewhat intermediate between these two models, as long as they otherwise maintain Wright-Fisher dynamics. We also note that none of these models probably accurately describe the behaviour of human populations, which likely have a much higher variance in offspring number, variable population sizes, etc.

As mentioned above, individuals are siblings if they have the same parents. If individuals share only one parent, we call them half-siblings. We say that two individuals are p-th cousins if there is at least one coalescence between their genealogies in generation $\mathcal{G}_{p+1}$. Of course, the amount of shared genetic material would depend on the number of shared ancestors in a certain generation. For two individuals, the number of shared ancestors is given in the supplementary materials of King *et al.* (2017) (see the discussion below). Notice, that two individuals can have different relations simultaneously. An example of such a situation is given in Figure 1: the individuals related by this genealogy are half-siblings and first-cousins at the same time.

Let $\mathcal{S}$ be a random sample of size K of individuals from the present-day generation $\mathcal{G}_0$ of a population described by either a monogamous or non-monogamous Wright-Fisher models. In this paper we derive the number U_T (notation for monogamous case) or V_T (notation for non-monogamous case) of individuals in $\mathcal{S}$ which do not have $(T-1)$-order cousins ($T = 1$ would stand for (half-)siblings, $T = 2$ for first cousins, etc.) within $\mathcal{S}$ and

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

have genealogy with no cycles. We will derive the probability distribution of U_1 and V_1 and expectations of U_T and V_T for $T \geq 2$ in terms of Stirling numbers of the second kind. Further we present a simple analytical approximation of expectations of U_T and V_T. We derive this approximation as an exponential function of the ratio of the sample size to the effective population size.

The condition that individual's genealogy does not have cycles means that there is no inbreeding in the history of the individual. Indeed, a cycle appears when two mating individuals share an ancestor, hence they are related to each other. On the contrary, if there is no inbreeding within T generations of ancestors of a certain individual, then all the ancestors have different parents, hence in the $\mathcal{G}_k$ ($k \leq T$) there are exactly 2^k ancestors of the individual under consideration.

Notice that the requirement that there is no inbreeding is satisfied as long as 2^T is small compared to the effective population size N. In this paper we are particularly interested in large populations. We will compute the fraction of individuals with siblings ($T = 1$) or p-th cousins ($T = p + 1$) in a sample in the limit of the effective population size N going to infinity. For fixed values of T and the sample size, K , the number of siblings and cousins goes to zero in the limit of large N. However, for a fixed ratio K/N, there is a positive expected number of siblings and offspring, but the expected number of cycles in the genealogy is small compared to K. This observation follows from the fact that the probability that two individuals share a parent is $1/N$, which is a rare event for large N. Hence for large N all the ancestors of an individual are unrelated with high probability. We will, therefore, approximate the number of individuals who have siblings (or p-th cousins) by $K - U_T$ or $K - V_T$ depending on the model. We notice that using this method we cannot characterise, for example, the overlap between the set of individuals who have siblings and the set of individuals who have first-cousins, so we cannot provide an approximation of the number of individuals who have at least some kind of relatives within several generations.

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 $\mathcal{S}$ into subsets of siblings (in case of non-monogamous model, we create two independent partitions of the sample, one of partitions corresponding to shared fathers and the other corresponding to shared mothers). Then we assume that individuals from the same subset have the same parent couple (in the case of the monogamous model) or the same father or mother (in the case of the non-monogamous model), and individuals from different subsets have different parents. This approach is the basis for our analyses and leads us to the proof of formulas for expectations of U_T and V_T.

The combinatorial technique used to obtain exact formulas for expectations of U_T and V_T is very similar to the technique used in King *et al.* (2017) (see supplementary materials S1). In particular, we have to keep track of the number of ancestors at each generation which is the question of interest of the section S1.1 of King *et al.* (2017). Notice, that results in our paper and the result of S1.2 of King *et al.* (2017) complement each other. We find the expected number of individuals in a sample which do not have any relatives with respect to a certain generation, hence we know approximately the number of individuals

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which share at least one ancestor in that generation with at least one more individual from the given sample. However we cannot characterise finer relatedness (e.g. the number of shared ancestors in a given generation,) as more than one coalescence per generation between genealogies of two individuals is possible. The pairwise analysis of individuals can be performed using King *et al.* (2017) results, though it can be computationally challenging. The asymptotic behaviour derivation for $E(U_T)/K$ and $E(V_T)/K$ (for fixed K/N ratio) is a completely new result to the best of our knowledge.

We remind the reader that the Stirling number of the second kind $S(n,k)$ is the number of ways to partition a set of size n into k non-empty disjoint subsets. A generalisation of this is the $r-$associated Stirling number of the second kind, $S_r(n,k)$ (Comtet 1974), which is the number of partitions of a set of size n into k non-empty subsets of size at least r. We provide more detailed information on the Stirling numbers of the second kind in the Appendix.

4. Probability distribution U_1

We say that two individuals are siblings if they have the same parents. 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 $\mathcal{S}$ of size K without siblings in $\mathcal{S}$ under monogamous dioecious Wright-Fisher model. Then*

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

Proof. We begin the proof by computing the number of possible partitions of $\mathcal{S}$ into u subsets of size 1 and t subsets of size greater than or equal to 2. Each subset of such a partition corresponds to the descendants in $\mathcal{S}$ of the same couple of parents from $\mathcal{G}_1$. There are $\binom{K}{u}S_2(K-u,t)$ such partitions (see figure 2). 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 disjoint subsets.

Now we need to assign $u+t$ subsets to different couples of parents from $\mathcal{G}_1$. There are $\binom{N}{u+t}$ possibilities for choosing couples that have descendants in $\mathcal{S}$ and $(u+t)!$ permutations which assign these particular couples to different subsets of the given partitions of $\mathcal{S}$.

Finally, summing over all possible values of t we get

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

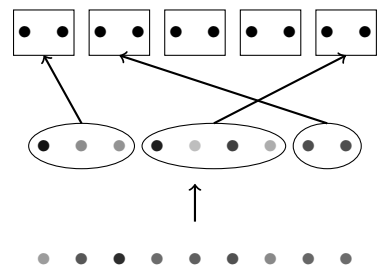


FIGURE 2. Illustration to the proof of Theorem 1. Each dot correspond to an individual. The bottom set of points corresponds to the individuals in the sample $\mathcal{S}$. This sample is divided in disjoint subsets (the set of points in the middle): this partition corresponds to sets of siblings, or in other words individuals from each subset will be assigned to the same couple of parents. The top row corresponds to the set of couples in the parent generation. Subsets of siblings (from the middle row) are assigned to different couples of parents (from the top row).

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

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 $\mathcal{I}$ does not have any siblings in $\mathcal{S}$ is $\pi_1 = (1 - 1/N)^{K-1}$, because all other individuals from $\mathcal{S} \setminus \{\mathcal{I}\}$ can be assigned to any couple of parents except for the parents of the individual $\mathcal{I}$. 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 of the theorem it is enough to rewrite

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

because $K/N = \alpha$ by definition. Now notice that

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

Hence the last statement of the theorem is proved

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

□

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. 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_2 be a random variable representing the number of individuals in a sample $\mathcal{S}$ of size K without first cousins in $\mathcal{S}$ under a monogamous dioecious Wright-Fisher model. Then the expectation of U_2 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 $\mathcal{S}$. Then the expectation $\mathbb{E}(U_2) = K\pi_2$. Denote individuals from $\mathcal{G}_T$ which have descendants in $\mathcal{S}$ by $\mathcal{S}^T$.

Choose an individual $s_0 \in \mathcal{S}$, 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 $\mathcal{G}_2$ and those couples do not have other descendants in $\mathcal{S}^1$.

Similarly to derivation of distribution of U_1, we first partition $\mathcal{S}$ into m disjoint subsets. We choose m couples from $\mathcal{G}_1$ 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 $\mathcal{S}^1$. 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 $K/N = \alpha$ for some constant $0 \le \alpha \le 1$ and we consider the limit of $\mathbb{E}(U_2)/K$ for $K \to \infty$.

Theorem 3. *Let $0 \le \alpha \le 1$ and set $K = \alpha N$. Then*

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

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 $K = \alpha N$ for some $0 \le \alpha \le 1$ and set $\beta = (2\ln 2)^{-1}$. Then*

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

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proof. Denote

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

First, notice that

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

We will show that for $\beta = (2\ln 2)^{-1} < 1/2$

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

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 following ratio for $m \le \lfloor \beta K\rfloor$

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

by applying approximation (11). Here $G_1 = G(K,m)$ and $G_2 = G(K,m+1)$.

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 (2). 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 the ratio (2) has the following approximation

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

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

$$H(x) = \frac{G(x)^{\frac{1}{x}}(x-G(x))^{\frac{x-1}{x}}\left(\ln(x-G(x)) - \ln G(x)\right)}{x^2}. \tag{3}$$

$H(x)$ has one real root $x = 2\ln 2$ if $x \geq 1$.The derivative $H(x)$ is positive for $x > 2\ln 2$, 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 sufficiently large K we obtain the following upper bound for (2)

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

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

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

Now we use the obtained inequality to prove limit (1)

$$\lim_{N\to\infty} \frac{\sum_{m=1}^{\lfloor \beta K \rfloor - 1} T_{K,N}(m)}{T_{K,N}(\lfloor \beta K \rfloor)} \leq \lim_{N\to\infty} \sum_{m=1}^{\lfloor \beta K \rfloor - 1} \left(\frac{A}{N^2}\right)^{\lfloor \beta K \rfloor - m} = \lim_{N\to\infty} \frac{A}{N^2}\frac{(1-A/N^2)^{\lfloor \beta K \rfloor - 1}}{1-A/N^2} = 0, \tag{4}$$

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

Lemma 2. *Let $K = \alpha N$ for some $0 \leq \alpha \leq 1$, set $\beta = (2\ln 2)^{-1}$. Then for any m such that $\lfloor \beta K \rfloor \leq m < K$*

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

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

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

Notice that $xe^x = -1/e + O((x-1)^2)$ 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 (Olver *et al.* (2010)) 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.

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

(5)

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

where $H(x)$ is given by expression (3). 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 (5) 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

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

with some constant C_0, which depend on α. □

Now we are ready to prove the theorem.

Proof. Firstly, notice that

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the right part trivially converges to $e^{-4\alpha}$ with $N \to \infty$ (we remind that $K = \alpha N$ for some constant $0 \le \alpha \le 1$).

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

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

because from Lemma 2 it follows

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

□

6. General case: expectation of U_p for $p \ge 2$

Similarly to the expectation of U_2, we can find the probability of the expected numbers U_p ($p \ge 2$) of individuals which do not have $(p-1)$-cousins and with pedigrees without cycles.

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

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

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

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

Remark 1. For $k = 1$, 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.

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Theorem 4. • *For any natural $p \geq 1$ the expectation of U_p is*

$$\mathbb{E}(U_p) = K \frac{\sum_{m_1=1}^{K} R_1 \sum_{m_2=2}^{2m_1} R_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} R_{p-1} N^{2m_{p-1}} W(p)}{\sum_{m_1=1}^{K} R'_1 \sum_{m_2=2}^{2m_1} R'_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} R'_{p-1} N^{2m_{p-1}}}, \tag{6}$$

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

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

$$R'_j = S(2m_{j-1}, m_j) \binom{N}{m_j} m_j!$$

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

$$\lim_{N \to \infty} \frac{\mathbb{E}(U_p)}{K} = \lim_{N \to \infty} \left(1 - \frac{2^{p-1}\alpha}{K}\right)^{2^{p-1}K} = e^{-(2^{2p-2})\alpha}. \tag{7}$$

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 (7) 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. □

7. 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 $\mathcal{S}$ 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 $\mathcal{S}$ 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 fixed ratio between K and the population sizes N.

Theorem 5. • *For any natural $p \geq 1$, the expectation of V_p is*

$$\mathbb{E}(V_p) = K \frac{\sum_{m_1=1}^{K} P_1 \sum_{m_2=2}^{2m_1} P_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P_{p-1} N^{2m_{p-1}} W^2(p)}{\sum_{m_1=1}^{K} P'_1 \sum_{m_2=2}^{2m_1} P'_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P'_{p-1} N^{2m_{p-1}}}, \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where we assume $m_0 = K$ *and*

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

$$P'_j := \sum_{n=2^{j-1}}^{m_j-2^{j-1}} S(m_{j-1},n)S(m_{j-1},m_j-n)\binom{N}{n}\binom{N}{m_j-n}n!(m_j-n)!$$

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* $K = \alpha N$*, then*

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

The proof of the theorem is similar to the case of the monogamous model. The function P_j counts the number of possibilities to have exactly m_j parents (male plus female)

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*

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

8. Numerical results

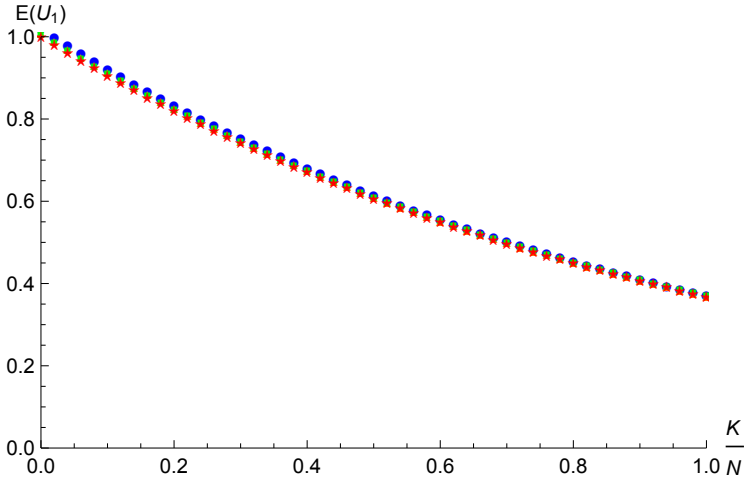
In this section we present numerical results for expectations of U_p and V_p, $p = 1, 2, 3$. Every plot of figures 3 and 4 represents the behaviour of $\mathbb{E}(U_p)/K$ or $\mathbb{E}(V_p)/K$ for a particular $p = 1, 2, 3$. Those values are computed by formulas (6) or (8) for different values of N ($N = 20, 100, 200$) 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.

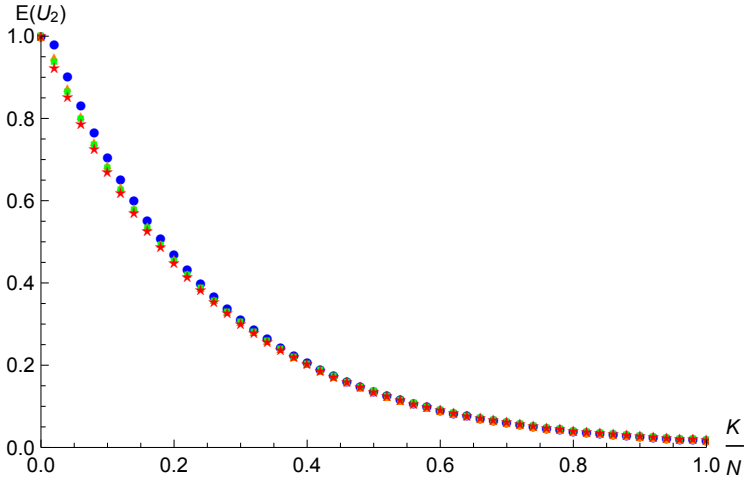
9. Discussion

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 used two extensions of Wright-Fisher model which keeps track of the two parents of an individual.

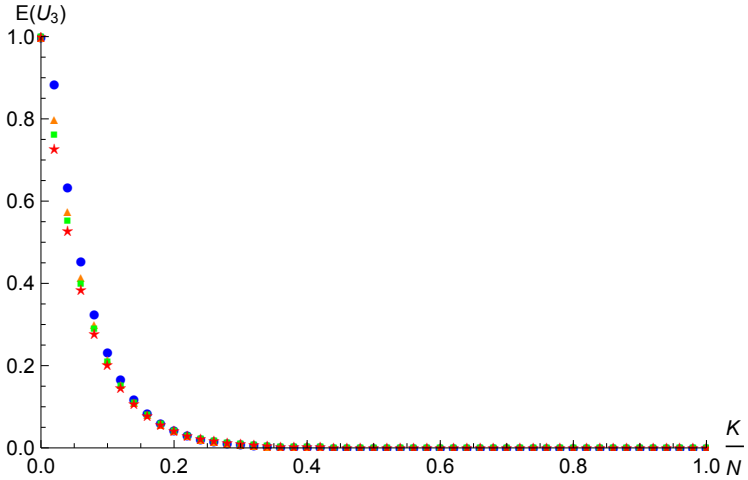
bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(A) p=1, individuals without siblings.



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



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

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

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The first extension corresponds to a monogamous population and the second to a non-monogamous population. The two models represent two extremes in terms of degree of promiscuity, and we might expect that in most other dioecious versions of the Wright-Fisher model, with intermediate degrees of promiscuity, the number of individuals without siblings or p-th cousins is somewhere in between those two regimes - as long as the models otherwise maintain Wright-Fisher dynamics.

Under both models we derived expressions for these expectations under the hypothesis that the pedigrees have no cycles (except for the one appearing 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 point-wise 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 cousin 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 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.

In the study of Danish population structure, Athanasiadis *et al.* (2016) discovered 3 pairs of first cousins and one pair of second cousins in a sample of just 406 individuals. Based on their estimate of an effective population size of $500,000$, we would expect to find 1.32 individuals with first cousins under the monogamous model and 2.63 individuals with first cousins under the non-monogamous model. The empirical number of 3 first-cousins in the sample is therefore not significantly different of the expected number of 1.32 under the monogamous model assumption. It is also not statistically significantly different from the expected number of 2.63 under the non-monogamous model. The expected number of second cousins in the sample is 5.24 and 10.41 under the monogamous and non-monogamous models, respectively. The inferred number of 1 is much smaller than this, likely because it is difficult to infer second cousins empirically. We would in general expect that the true number of second cousins is larger than the true number of first cousins.

Notice, that the probability for two individuals to be p-th cousins is approximately $\hat{\pi}_p(N) = c_m(2^{p+1}-1)2^p/N$, where c_m is 1 for monogamous model and 2 for non-monogamous model. Hence, the expected number of pairs of p-th cousins in a sample of size K is approximately $\hat{\pi}_p(N)K^2/2$. Henn *et al.* (2012) found approximately 5000 pairs of third cousins and 30000 pairs of fourth cousins in a sample of only 5000 individuals with European ancestry, which would be expected for effective population sizes of $2 \cdot 10^5 - 3 \cdot 10^5$ under the

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

monogamous model and twice that $(4 \cdot 10^5 - 6 \cdot 10^5)$ under the non-monogamous model. These numbers are roughly compatible with estimates of effective population sizes obtained for modern European populations (e.g., Athanasiadis *et al.* (2016)). We note that effective population size is a tricky concept for a spatially distributed population such as European humans, but the breeding structure observed in these samples suggest that the degree of relatedness in the sample is compatible with population sizes on the order of $10^5 - 10^6$.

10. Appendix: Stirling numbers of the second kind and their generalisation

In this section we provide definitions and properties of Stirling numbers of the second kind.

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 $r = 2$

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

10.1. Uniformly valid approximation for $S(n,k)$.

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

$$S(n,k) = \left(1 + O\left(\frac{1}{n}\right)\right) \sqrt{\frac{t_0}{(1+t_0)(x_0 - t_0)}} e^A k^{n-k} \binom{n}{k}, \tag{9}$$

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

$$\frac{k}{n} x = 1 - e^{-x}, \tag{10}$$

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

$$S(n,k) = \left(1 + O\left(\frac{1}{n}\right)\right) \sqrt{\frac{n-k}{n(1-G)}} \frac{1}{G^k \, (\frac{n}{k} - G)^{n-k}} \left(\frac{n-k}{e}\right)^{n-k} \binom{n}{k}, \tag{11}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We did not find a reference for the formula (11) 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 (10)

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

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 (9) 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 (9) and (11).

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] Athanasiadis G, Cheng JY, Vilhjlmsson BJ, et al (2016) Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity. Genetics 204(2):711-722. doi:10.1534/genetics.116.189241
[3] Chang JT (1999) Adv Appl Prob 31:1002-1026
[4] Comtet L (1974) Advanced Combinatorics. Reidel, Dordrecht, Holland
[5] Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
[6] 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
[7] King L, Wakeley J, Carmi S (2017) A non-zero variance of Tajima?s estimator for two sequences even for infinitely many unlinked loci. Theor Popul Biol. doi:10.1016/j.tpb.2017.03.002
[8] 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
[9] Nagylaki T (1997) Multinomial-Sampling Models for Random Genetic Drift. Genetics 145:485-491

bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] Olver FWJ et al (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge, p 131

[11] Temme NM (1993) Asymptotic Estimates of Stirling Numbers. Stud Appl Math 89:233-243. doi: 10.1002/sapm1993893233

[12] 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

[13] 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

[14] 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. Theoretical Population Biology 115:1-12 doi:10.1101/054957

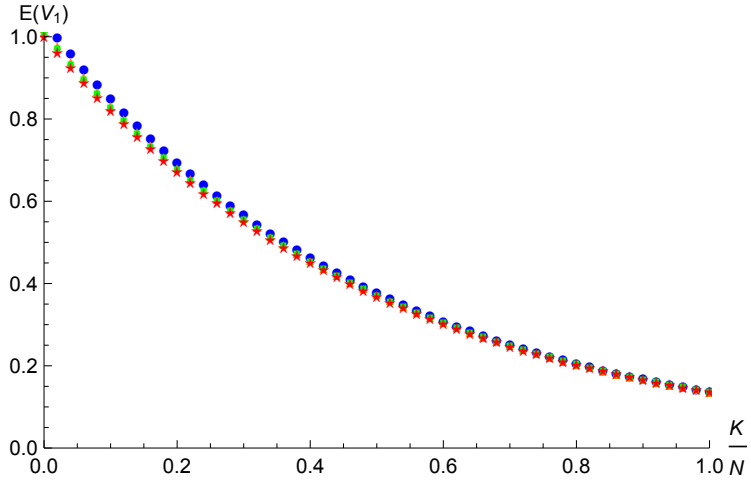
[15] Wright S (1931) Evolution in Mendelian populations. Genetics. 16:97-159

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

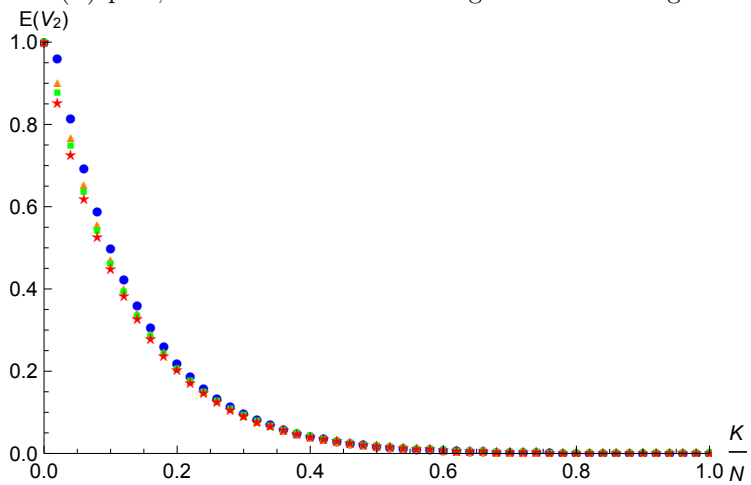
[2]Museum of Natural History, University of Copenhagen, Øster Voldgade 5-7, 1350 København K, Denmark

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

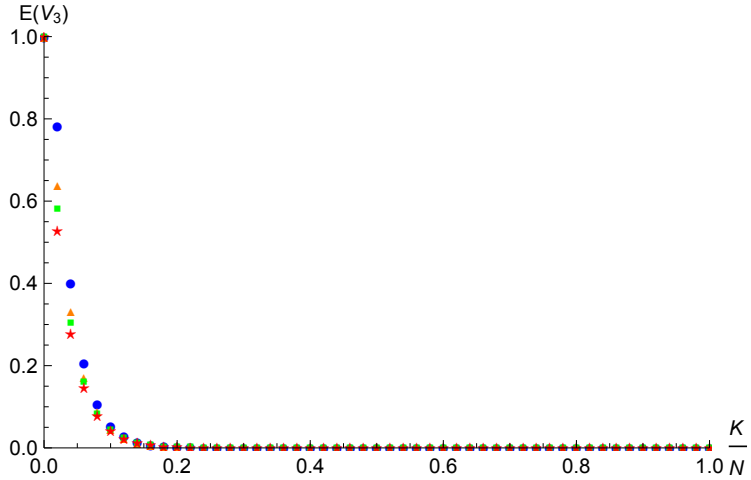
bioRxiv preprint doi: https://doi.org/10.1101/145599; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



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