Population Dynamics and Surname Distribution

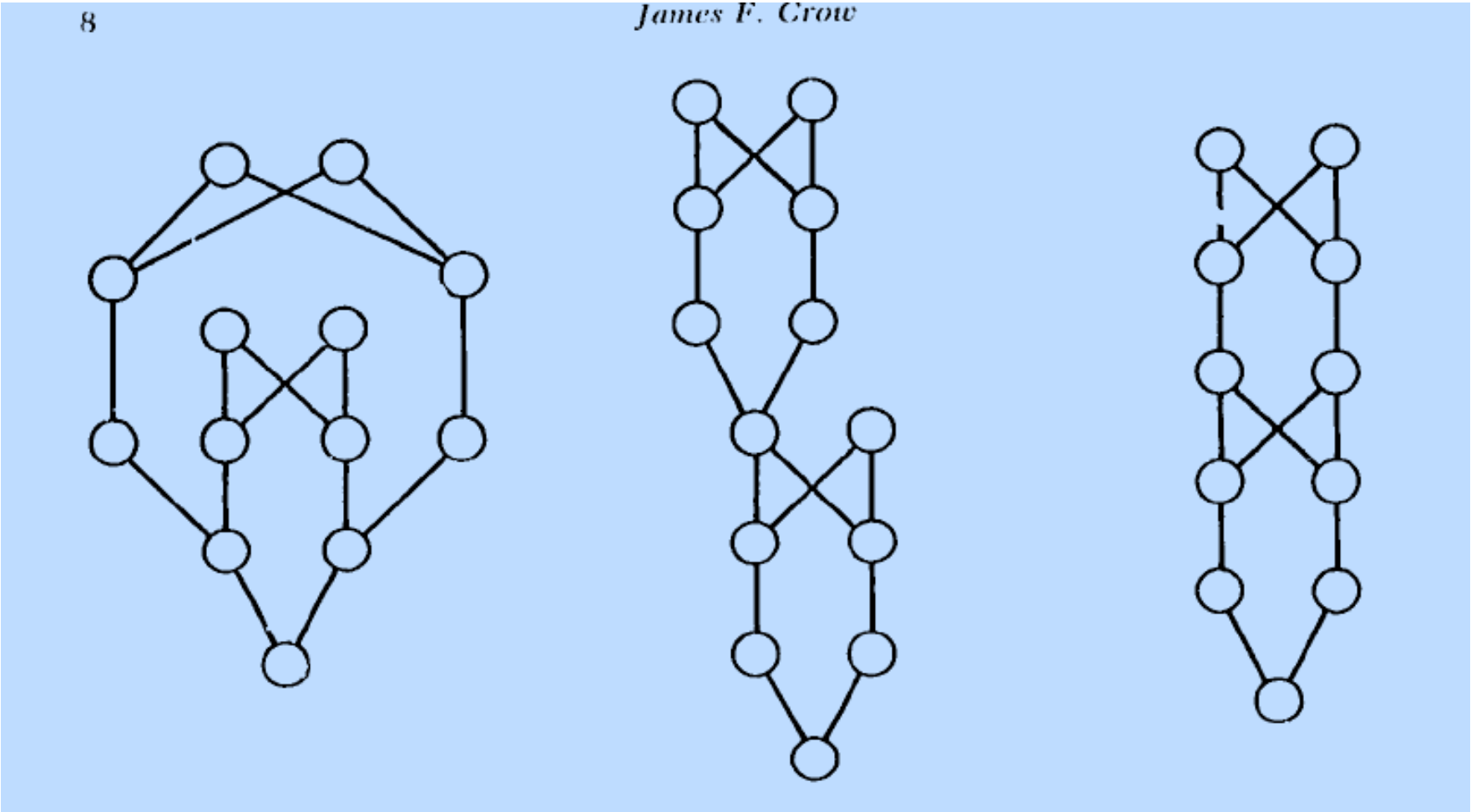
History (1)

- To make a long history short

- Landmarks in the study of isonymy:

- 1875 G. Darwin - Isonymy and consanguinity
- 1965 Crow and Mange - Isonymy and inbreeding F = P/4
- 1977 G.W. Lasker - Relatedness of populations
- 1982 Eugene Symposia - Surnames as markers of inbreeding and migration

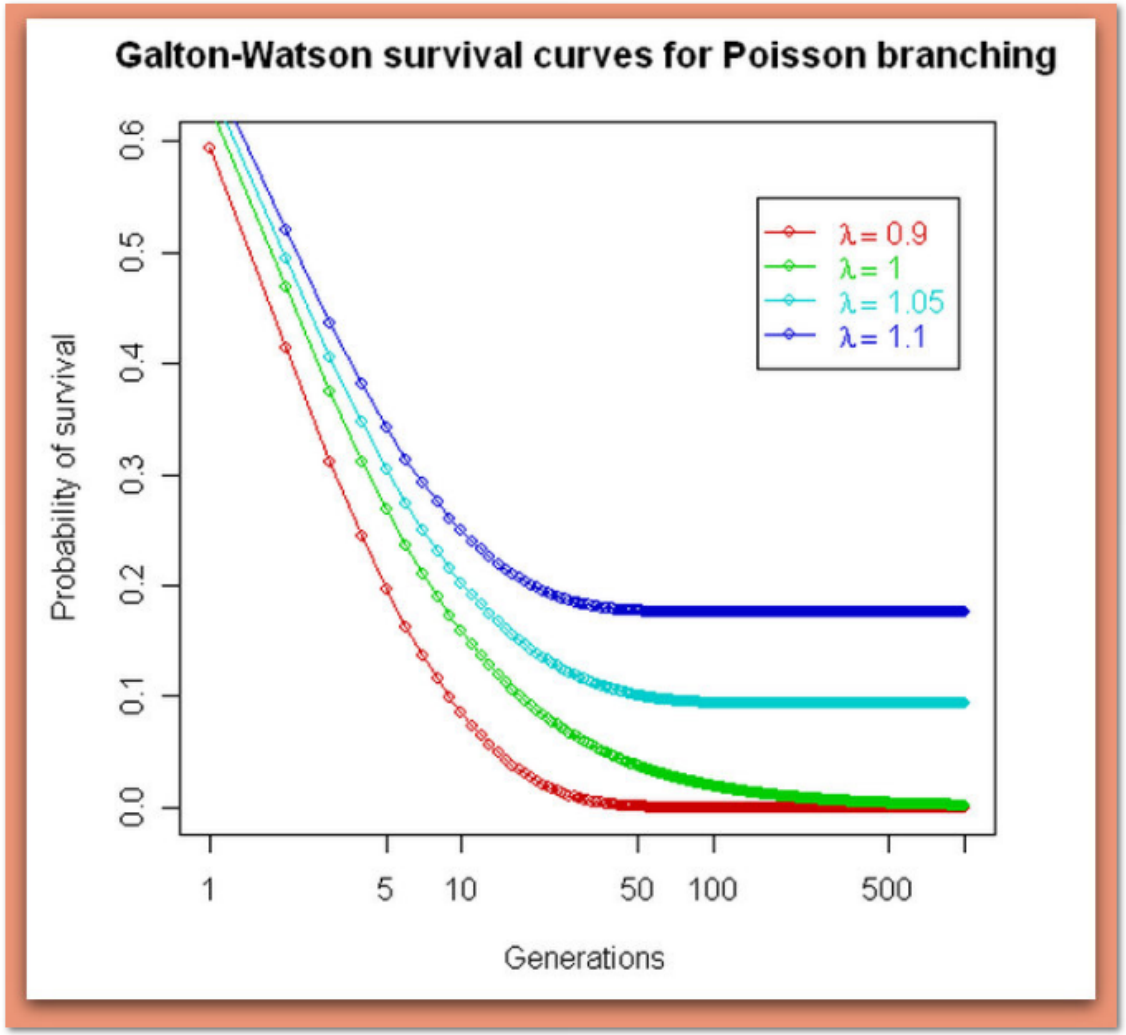
Crow-Mange consanguinity schemes



History (2)

- Landmarks in the study of surname distributions:
- 1874 Galton and Watson - Branching processes: surname extinction
- 1925 G.U. Yule - Mathematical theory of evolution
- 1943 R.A. Fisher - Frequency of individuals according to species
- 1955 H.A. Simon - Skew distribution functions
- 1967 Karlin and McGregor - Neutral mutations
- 1972 W.J. Ewens - Sampling theory of neutral alleles
- 1974 Yasuda and Cavalli-Sforza - Evolution of surnames
- 1983 Fox and Lasker - Frequency distribution of surnames
- 1984 Wijsman et al. - Migration matrices
- 1997 M. Jobling - Y chromosomes and surnames
- 2000 Sykes and Irven - Origin of surmame Sykes
- 2001 S.P. Hubbell - Neutral theory of biodiversity and biogeography

Galton-Watson extinction curves



Fox-Lasker distribution

The distribution of surname frequencies

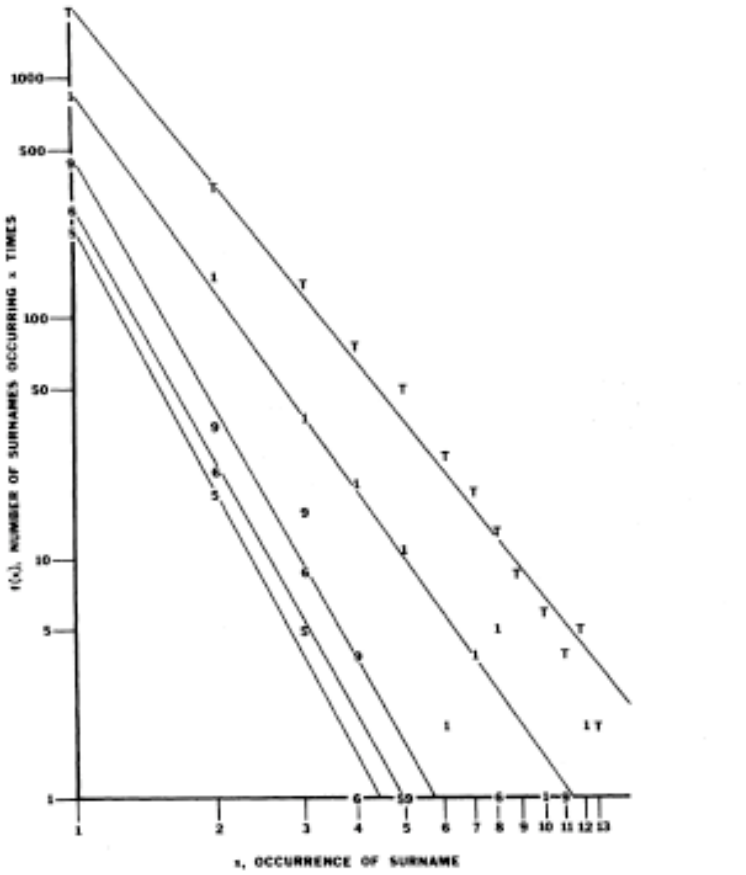


Figure 1. *Number of surnames occurring x times, f(x) plotted against x, on logarithmic scales, with fitted lines of the discrete Pareto distributions (districts, 1, 5, 6, 9 and all districts combined, T).*

Surnames in Europe

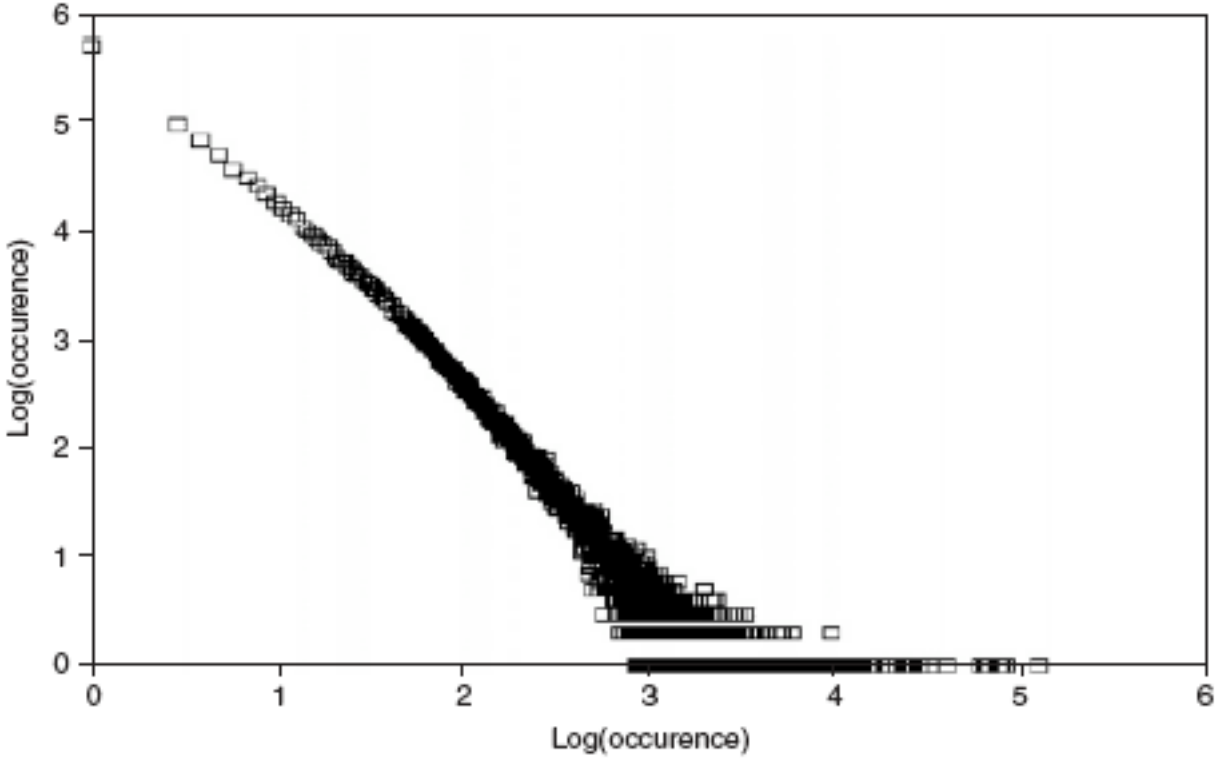


Fig. 2. The log-log distribution of the frequency of occurrence of surnames in Western Europe.

Yule processes

- N(k;s) = number of families with k individuals at time step s

- $N(k; s + 1) = N(k; s) + b(s)\,[(k-1)\,N(k-1; s) - k\,N(k; s)]$;
- $N(1; s + 1) = N(1; s) + a - b(s)\,N(1; s)$;
- $N(s) = N(0) + s$; $b(s) = (1-a)/N(s)$.

- Approximation: $N(k; s + 1) - N(k; s) = N(k; s)/N(s) = P(k)$
- Reduced equation $P(k) + (1-a)\,k\,P(k) = (1-a)(k-1)\,P(k-1)$

- Solution: $P(k) = c\,(k-1)!\,\Gamma(c+1)/\Gamma(k+c+1)$
- Zero-truncated Yule distribution (Beta function) with $c = 1/(1-a)$

- Asymptotic behavior : power law with exponent $-(c+1)$

Branching processes

- Panaretos: equivalence with Yule process
- Consul: Geeta distribution from a branching process
- Reed and Hughes (2002): a Galton-Watson branching process with mutation and/or immigration predicts an exponent -(2+b/d)

  where b is the probability of mutation

  d is the growth ratio of the population

Master equation

- Baek et al (2007):
- P(j;s;k;t) = probability for a family to have k members at time t if it had j members at time s

- $dP(j;s;k;t)/dt = L(k-1;t)P(j;s;k-1;t)+[M(k+1;t)+B(k+1;t)]P(j;s;k+1;t) - [L(k;t)+M(k;t)+B(k;t)]P(j;s;k;t)$

  L(k;t) = birth rate, M(k;t) = death rate, B(k;t) surname creation rate

  In absence of mutations the exponent is -1 (China, Korea)

  In presence of mutations the exponent is –(2+ b/d)

Renormalization Group approach

- Fock space formalism for classical objects was introduced by Doi
- Even in absence of self-organized criticality, RG naturally leads to scale invariance and scaling behavior
- Galton-Watson branching processes may be represented in a properly defined Hilbert space.
- Reproduction governed by chance is seen as a decay process described by a non-Hermitian Hamiltonian
- All the predictions of the Master Equation approach may be recovered and confirmed

A disturbing aspect

- In all approaches when mutations are taken into account the (opposite of the) exponent is larger than 2
- Experimental evidence concerning all countries favors exponents that are definitely smaller than 2

- Bartley et al. Considered a model with birth, death and creation of surnames and approximated it with a continuum equation of the Fokker-Planck type for the distribution of surname frequencies.

- They showed that the asymptotic regime is the standard one, but for smaller values of the family size the distribution may be described by an approximate power law with exponent less than 2.

Effects of sampling (1)

- Finite size effects and sampling may very well alter the observed pattern even for models predicting scaling in the N-> infinity limit.

- The (expected) frequency distribution in a sample is in general different from the frequency distribution of the full system
- For sufficiently large samples of a system with frequency distribution N(k) the expected values are <n(l)> = Sum N(k) P(kl)

  where P(k,l) is the binomial distribution

  We define special generating functions

  G(z) = Sum N(k) (1-z/N)^k, g(z) = Sum <n(l)> (1-z/n)^l

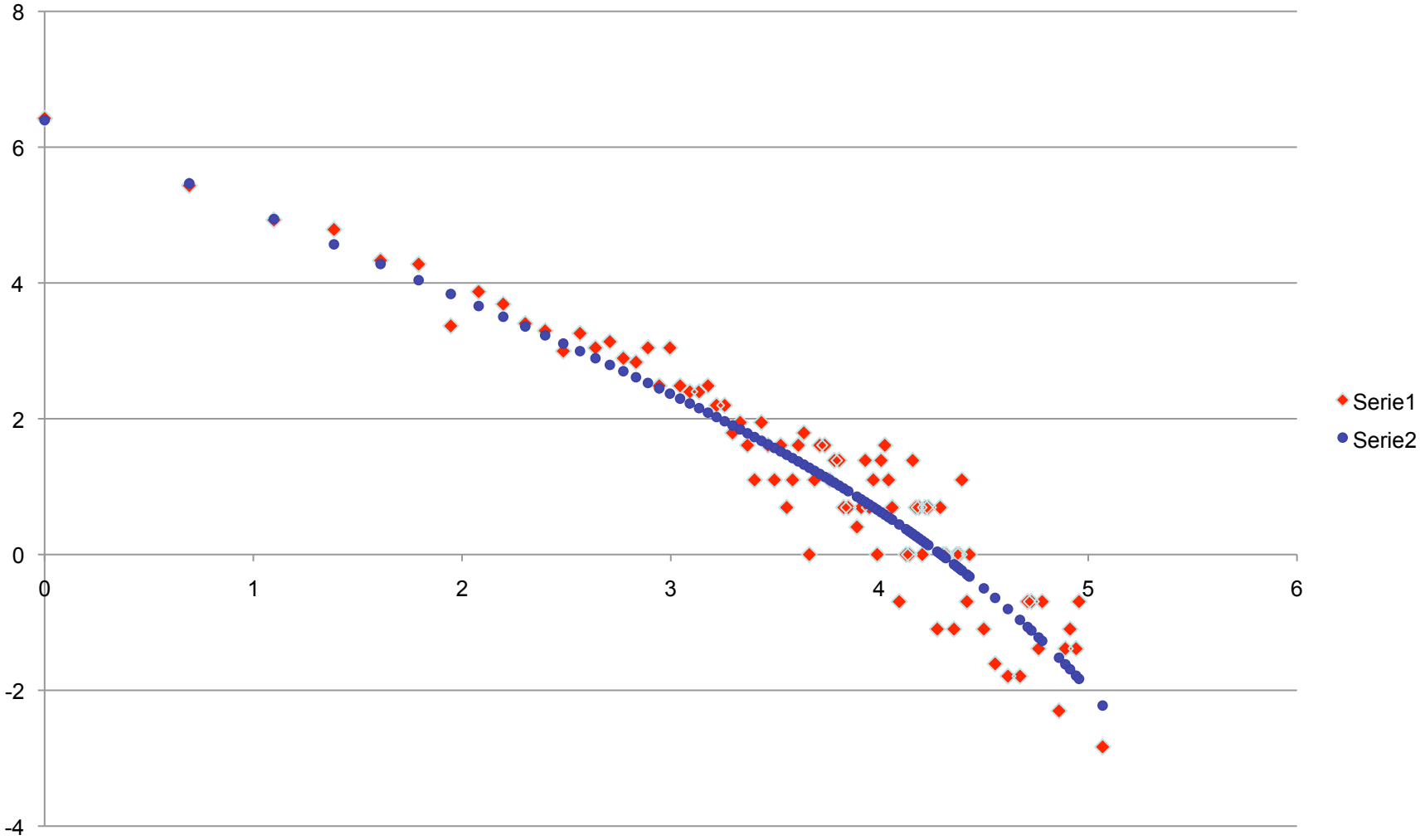
  and we can prove that g(z) = G(z) for all samples of n elements out of a system containing N elements

Effects of sampling (2)

- As a consequence it is possible to define a wide set of expectation values that are independent of the size of the sample
- The simplest example is
- $M2 = \Sigma\, k(k-1)\, N(k)/N^2 = \Sigma\, l(l-1)\langle n(l)\rangle /n^2 = 1/\alpha$ (isonymy)

- Properties of the frequency distributions of samples suggest the use of a parametrization based on the negative binomial distribution.
- Main features:
- - the sampled distribution has the same structure for all sample sizes
- - the distribution depends on two parameters, whose one is just the exponent c of the asymptotic distribution and the other contains the dependence on the sample size
- - the invariant moments are easily computed, depend only on alpha and c and have simple scaling properties

Surnames in Pisa



Statistics of genealogical trees

- The neutral theory of evolution suggested the creation and study of stochastic models of reproduction and evolution
- Derrida et al. (1999) studied the statistical properties of ancestors' tables and found a RG equation for the generating function g (G;z) of the moments of the distribution of ancestors' repetitions in the G-th generation
- $$g(G+1;z) = \exp [m\, g(G; z/m) - m]$$

  where m is the average number of descendants of a couple

  For a fixed size of the population the fraction of individuals having asymptotically no descendants is about 20%

Ancestors' repetitions

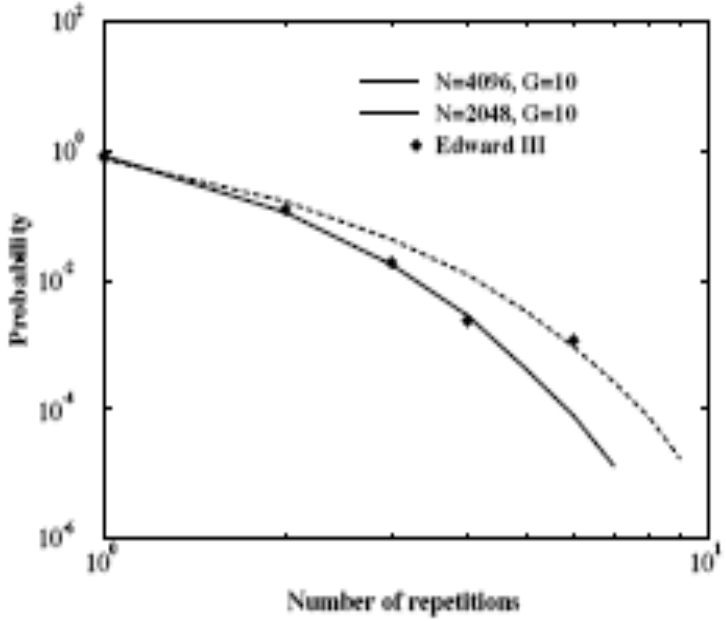


FIG. 1. Probability of ancestor repetitions in the genealogical tree of the king Edward III [5]. The continuous and dashed lines represent the results of simulations of $F(r)$ in a closed population with 2^{11} and 2^{12} individuals for our model. Averages have been performed over the ten first generations of 10^3 independent trees.

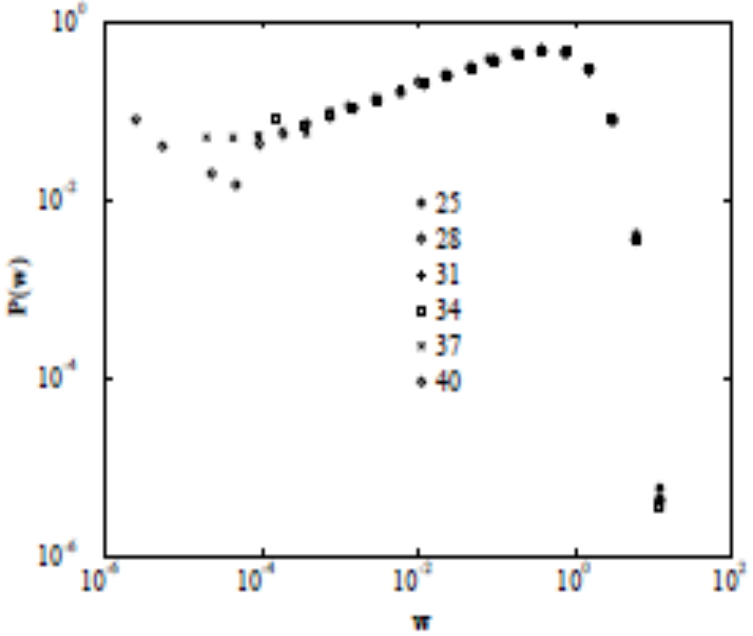


FIG. 3. Data collapse for the rescaled distribution of repetitions $P(w)$ after the transient period. Averages have been performed over 10^3 independent trees for a population size $N = 2^{20}$.

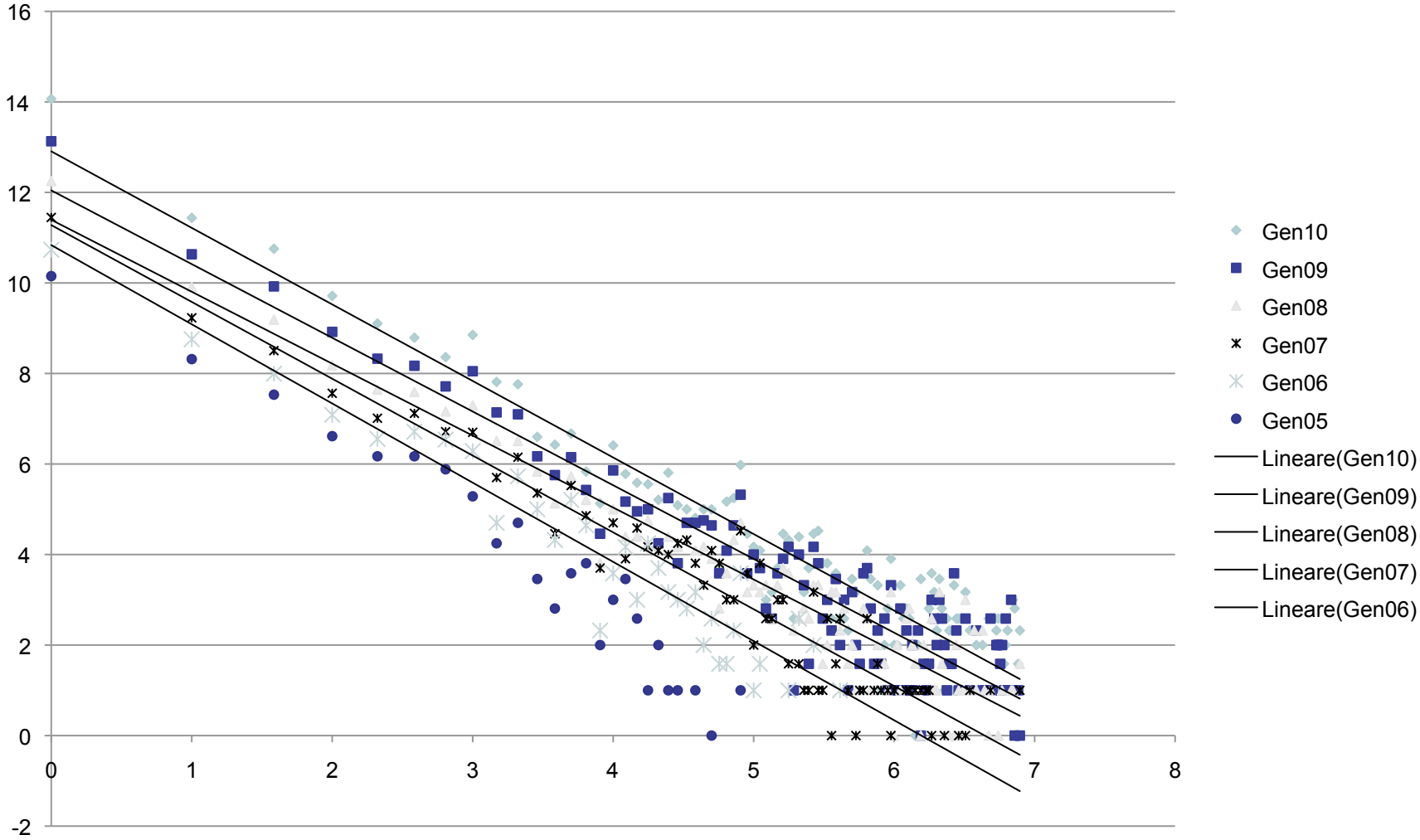
MRCA, IAP, and all that

- A strictly related issue is the estimate of the Most Common Recent Ancestor (MRCA) of a given group (or of all humanity)
- Computer simulations indicate a rather small distance from present times, order of Log(N) generations for N individuals
- One may also define the Identical Ancestors Point (IAP), a time characterized by a set of individuals that are ancestors either of everybody living now or of nobody still living.
- Computer simulations indicate an IAP at about 1.77 Log(N) generations back in time
- No genetic relevance of these concepts beacause of gene dilution in bisexual reproduction.

Empirical study of Ancestors'tables (1)

- Ancestors' tables for about 100 (noble) individuals living in the year 1800, reconstructed up to the 10-th generation with limited number of missing entries.
- In principle about 200,000 individuals, in practice less than 27,000 because of repetitions (only 11,000 fully identified)
- Results:
- - Evidence of universality, but no onset of Derrida scaling
- - MRCA around year 1550 (Wilhelm I Graf von Nassau-Dillingen)
- - Family correlations, and possible taxonomy of noble families

Universality of surname distribution

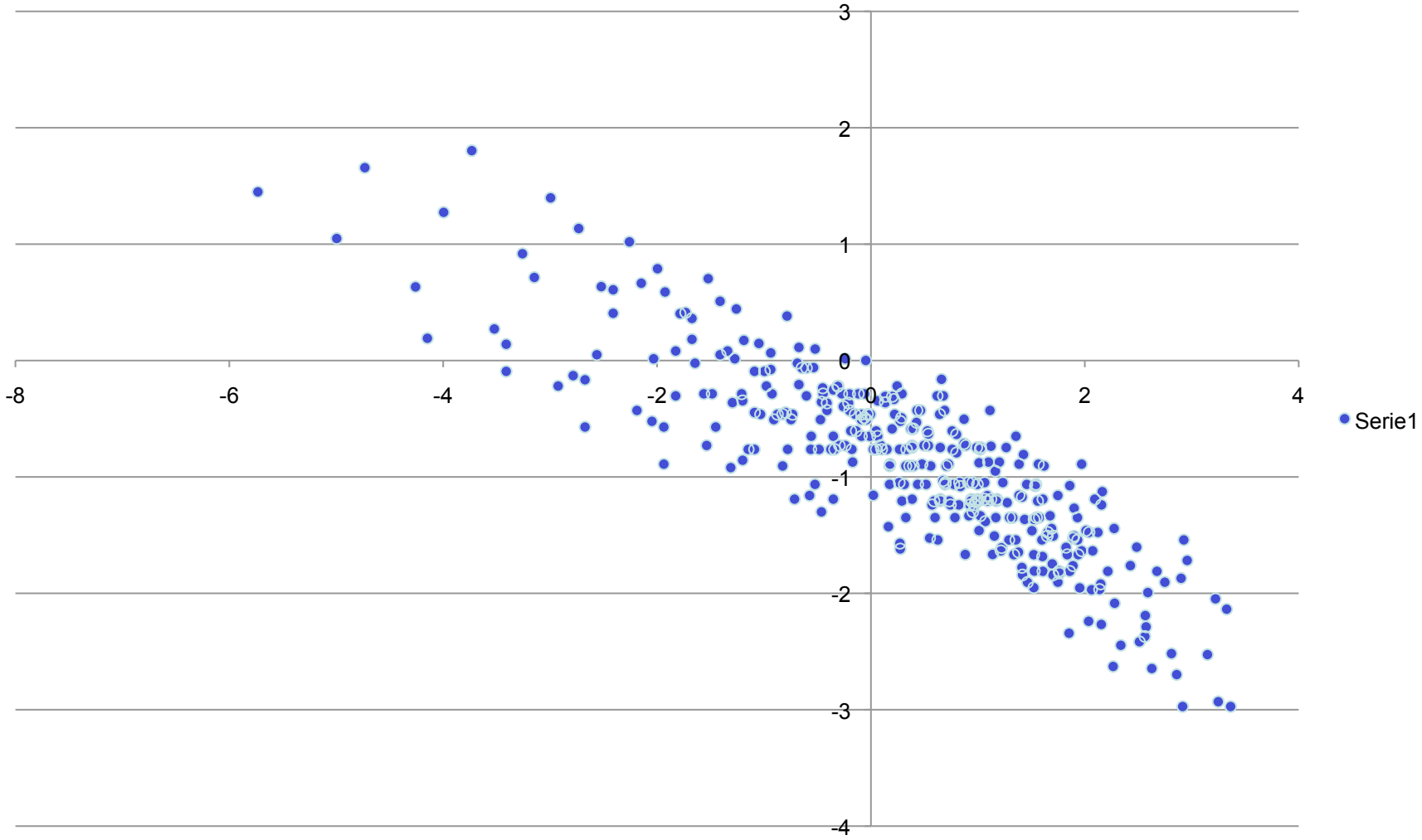


The MRCA of German High nobility

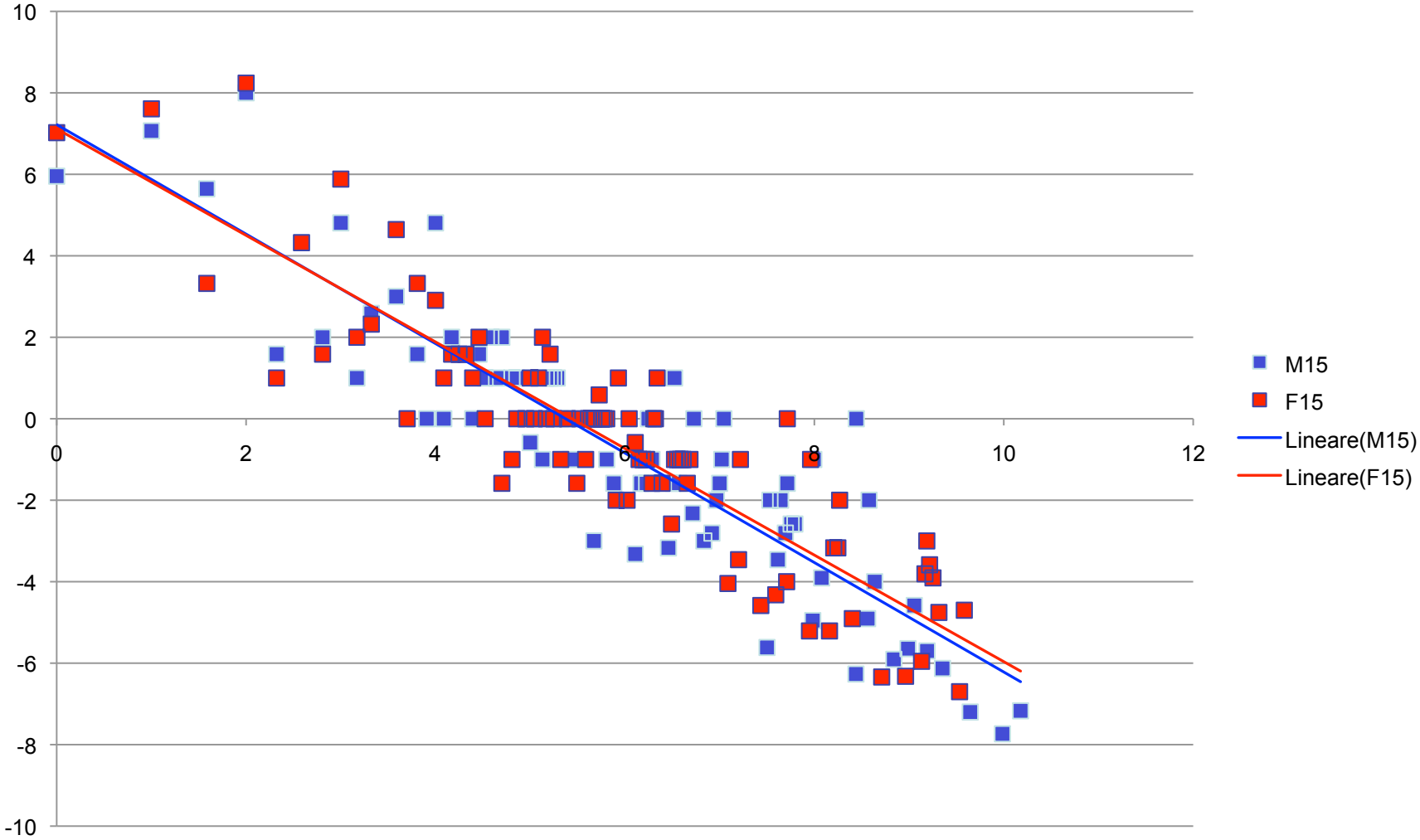
Empirical study of ancestors' tables (2)

- 15 (almost) complete generations of ancestors for Henri, comte de Paris (1908-1999): 65535 individuals in principle, 4257 in practice: essentially all European nobility back to year 1400.
- Results:
- - First hints of Derrida scaling
- - Evidence for "decreasing" population starting from about 1,000 individuals living in the year 1400 (restriction to higher nobility)
- - Surname distribution of ancestors, with some evidence of a Fox-Lasker distribution

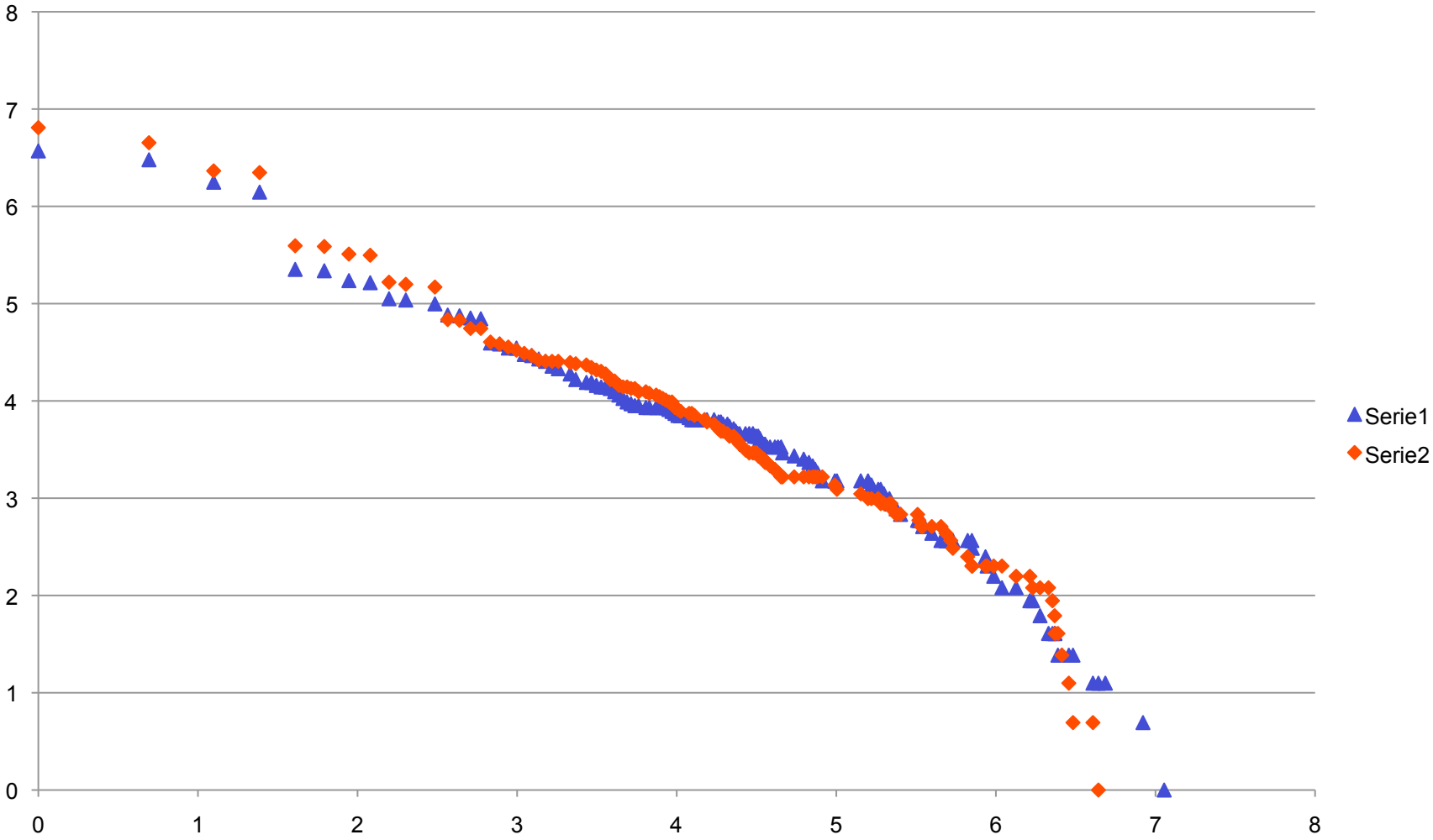
Distribution of repetitions for Henri's Ancestors



Male and Female surname frequencies (1)



M&F surname frequencies (2)



A theoretical result for ancestors' surnames

- m (G;k) = distribution of repetitions of individuals in the G-th generation of ancestors
- M(G;k) = surname distribution of ancestors in the G-th generation
- D(k) distribution of surnames in the full population

- The relationship between the generating functions of the above distributions takes the form
- $M(G+1;z) = D(1 - p + p\, m(G;z))\, M(G;z)$

  where p is the ratio between the number of different ancestors and the size of the population.

  Important corollary is the relationshio between the numbers of surnames in each generation

  $$C(G+1) = 2\, C(G)\, C^*/(C(G)+C^*)$$

  C^* is the total number of surnames in the population (fixed point)

rossi@df.unipi.it