Overdispersed, Beta-binomial, and Markovian Children

Nils Lid Hjort

Department of Mathematics, University of Oslo

Norsk statistisk forening, Stavanger, juni 2019

Take a look around you. To the first order of (good) approximation, we're all the results of independent Bernoulli coin tosses, with $\Pr(\text{girl}) = p$, $\Pr(\text{boy}) = 1 - p$. You need a Big Sample to see that $p \neq 0.50$, and and an Even Bigger Sample to see that the binomial model isn't entirely accurate.

Plan:

A How to spot the difference between 0.500 and 0.485

B Children are (slightly) extrabinomially overdispersed

C The beta-binomial model

D Siblings are (slightly) Markovian

E A few Mysteries of Human Production

A: how to spot the difference between 0.500 and 0.485

You ask n human beings if they are girls or boys, and count z girls. With $z \sim \mathrm{Bin}(n, p)$, you compute

$$\widehat{p} = z/n \quad \text{and} \quad \widehat{p} \pm 1.96 \{\widehat{p}(1-\widehat{p})/n\}^{1/2}.$$

If $\widehat{p} = 0.485$, you need 4265 children for your interval to fall to the left of 0.500.

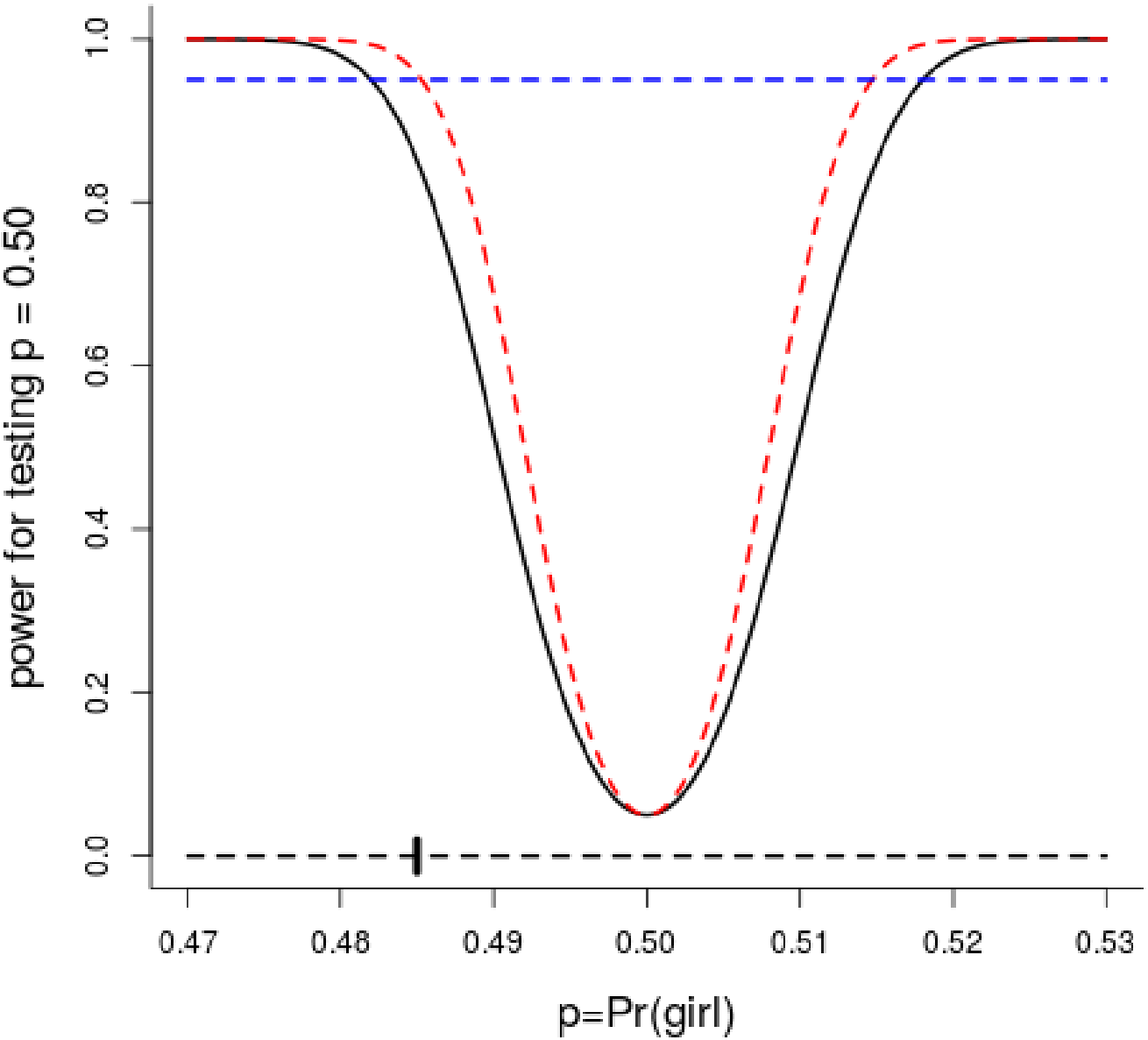
But you need a bigger sample. You're testing $H_0 : p = p_0$ vs. $p \neq p_0$, with $p_0 = 0.50$, and use

$$V_n = \frac{z - np_0}{\sqrt{np_0q_0}}, \quad \text{with} \quad q_0 = 1 - p_0.$$

Then you compute the power function

$$\pi_n(p) = \Pr\{|V_n| \geq 1.96 \,|\, p\},$$

and hope $\pi_n(p) \geq 0.95$ for $p = 0.485$. Then you need $n \geq 14433$.



Power function for easy 0.05-level test of $H_0: p = 0.50$, for $n = 10000$ (black curve) and $n = 15000$ (red curve).

B: children are overdispersed

B. Die jedesmal folgenden Geschlechtskombinationen.

Geschlecht.		Zahl der Ehen bez. Mütter.	Summe der Knaben.	Summe der Mädchen	Summe der Kinder.
Knaben.	*Mädchen.*				
6.	7.	8.	9.	10.	11.
		Acht Kinder.			
8	—	342	2736	—	2736
7	1	2092	14644	2092	16736
6	2	6678	40068	13356	53424
5	3	11929	59645	35787	95432
4	4	14959	59836	59836	119672
3	5	10649	31947	53245	85192
2	6	5331	10662	31986	42648
1	7	1485	1485	10395	11880
—	8	215	—	1720	1720
		53680	221023	208417	429440

From the Geißler data, Zeitschrift des königlichen sächsischen statistischen Bureaus, 1889.

I use the Geißler data, from Sachsen 1889. In particular, there's information on the number $N(y)$ of families with y girls and $8-y$ boys, for $n = 38459$ families with at least 8 children.

First, there are $\sum_{y=0}^{8} yN(y) = 149158$ out of $8n = 307928$ children, leading to $\widehat{p}_0 = 0.484$ for $\Pr(\text{girl})$.

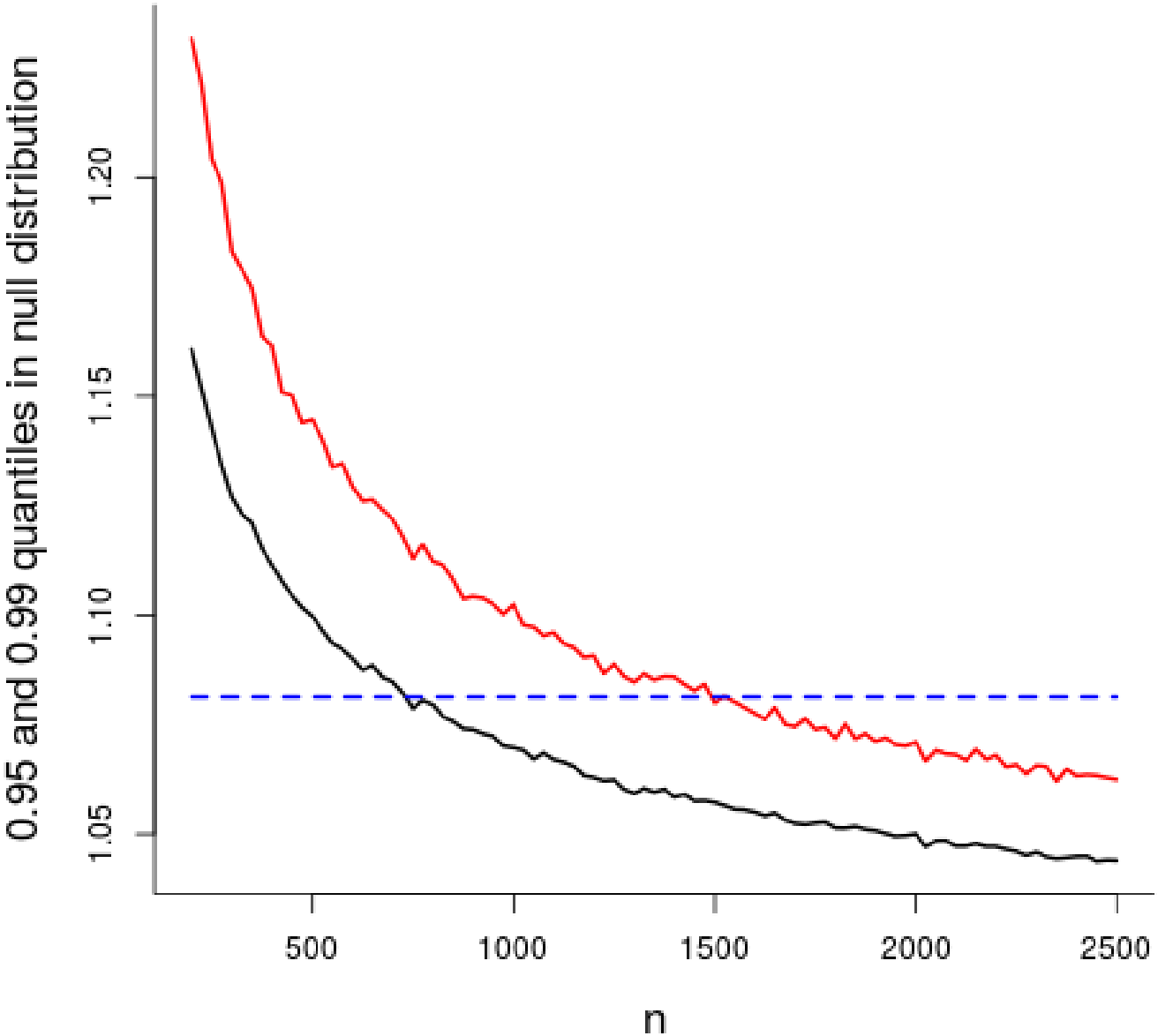
Second, there is overdispersion. I compute

$$W_n = \frac{S_n^2}{8\widehat{p}_0(1-\widehat{p}_0)},$$

i.e. the data variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{y=0}^{8} N(y)(y-\bar{y})^2$$

divided by the binomial variance $8\widehat{p}_0(1-\widehat{p}_0)$. I find $W_n = 1.081$, which looks modest, but which is extremely significant, for our high n.



Is $W_n = 1.081$ much bigger than 1? Answer: it depends on n.
Compute the p-value, $p_n = \Pr\{W_n \geq 1.081 \,|\, H_0\}$.

So there's clear overdisperson (visible if we have tons of data). This is also seen by comparing the observed $N(y)$ to the binomially expected

$$E(y) = nf(y, \widehat{p}_0) = n\binom{8}{y}\widehat{p}_0^y(1-\widehat{p}_0)^{8-y} \quad \text{for } y = 0, 1, \ldots, 8.$$

The Pearson residuals

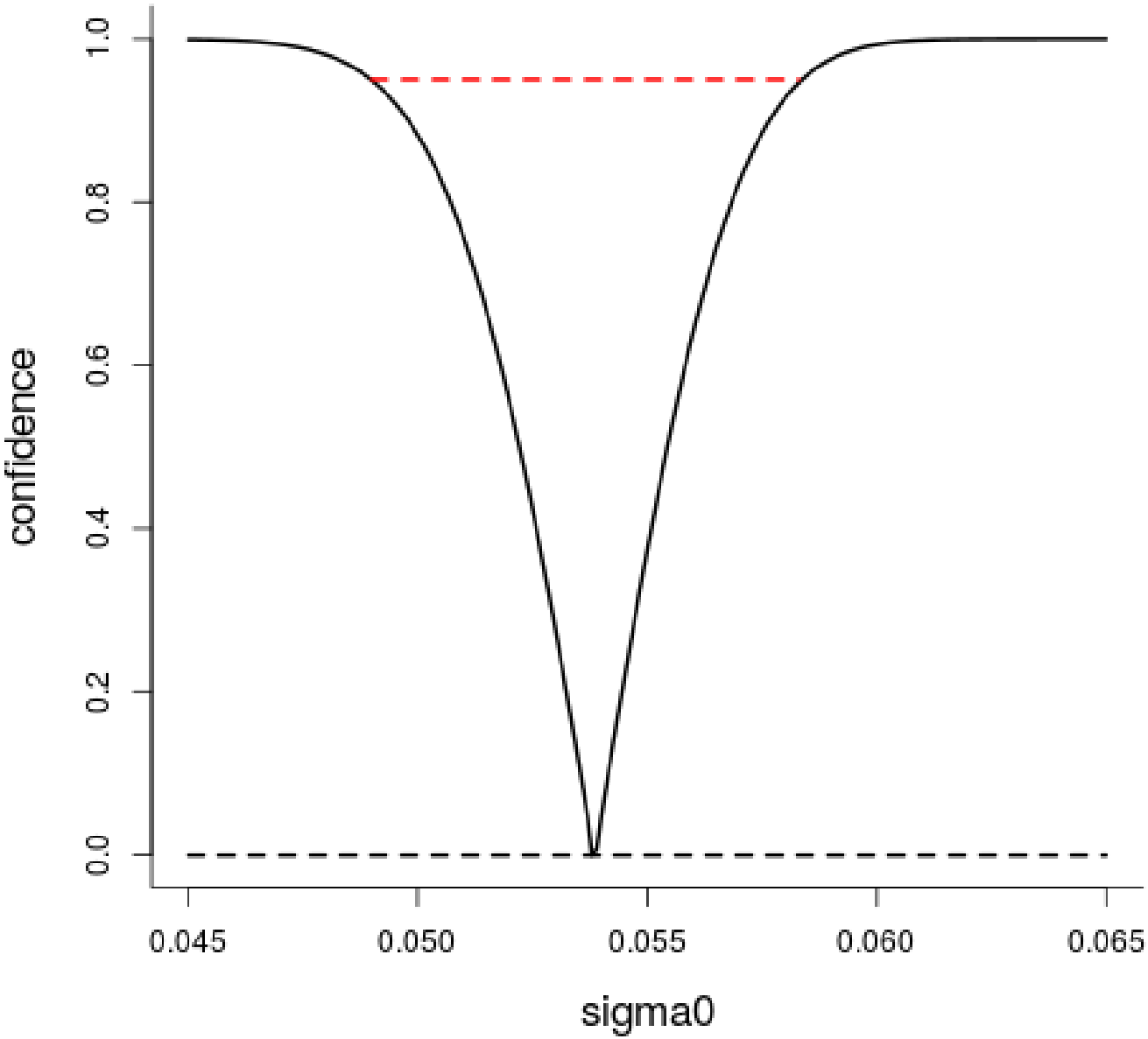
$$r(y) = \{N(y) - E(y)\}/E(y)^{1/2}$$

are not all inside $(-3, 3)$, and the chi-squared is 159.41, far too high.

Suppose now that p varies in the population, with standard deviation σ_0, and with $Y \mid p \sim \mathrm{Bin}(m, p)$. Then

$$\mathrm{E}\, Y = mp_0 \quad \text{and} \quad \mathrm{Var}\, Y = mp_0(1-p_0) + m(m-1)\sigma_0^2.$$

This is only visible if $m \geq 2$. I find $\widehat{\sigma}_0 = 0.054$, along with a full confidence curve $\mathrm{cc}(\sigma_0)$. 95% of all couples have their p inside $[0.38, 0.58]$.



Confidence curve $\mathrm{cc}(\sigma_0)$ for the overdispersion.

C: a beta-binomial model

So let's do $y \mid p \sim \mathrm{Bin}(m, p)$ with $p \sim \mathrm{Beta}(a, b)$:

$$f_2(y, a, b) = \int_0^1 f(y, p) g(p, a, b) \, \mathrm{d}p$$
$$= \binom{m}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+y)\Gamma(b+m-y)}{\Gamma(a+b+m)},$$

for $y = 0, 1, \ldots, m$. Can do ML, or minimum chi-squared, minimising

$$Q_n(a, b) = \sum_{y=0}^{m} r_2(y)^2 = \sum_{y=0}^{m} \frac{\{N(y) - E_2(y, a, b)\}^2}{E_2(y, a, b)},$$

with $E_2(y, a, b) = n f_2(y, a, b)$ the expected number of families with y girls. Answer: a Beta with mean 0.484 and standard deviation $\sigma_0 = 0.0538$. The fit is very good, with chi-squared lowered from 159.41 to 13.55; much better fit for all-girls and all-boys families.

		binomial		beta-binomial		Markov	
y	$N(y)$	$E_1(y)$	pears_1	$E_2(y)$	pears_2	$E_3(y)$	pears_3
0	264	192.32	5.17	255.54	0.53	255.96	0.50
1	1655	1445.38	5.51	1656.79	−0.04	1654.75	0.01
2	4948	4752.36	2.84	4909.50	0.55	4906.78	0.59
3	8498	8928.90	−4.56	8683.46	−1.99	8685.61	−2.01
4	10263	10484.95	−2.17	10025.35	2.37	10034.41	2.28
5	7603	7879.79	−3.12	7736.16	−1.51	7734.77	−1.50
6	3951	3701.20	4.11	3896.34	0.88	3893.65	0.92
7	1152	993.42	5.03	1171.05	−0.56	1168.12	−0.47
8	161	116.65	4.11	160.81	0.01	160.98	−0.00
	38495	38495	159.41	38495	13.55	38495	13.17

Pearson residuals are $r(y) = \{N(y) - E(y)\}/E(y)^{1/2}$, with goodness-of-fit statistic $K = \sum_{y=0}^{8} r(y)^2$.

D: Markovian siblings

My mother first had four boys – was her $\Pr(\text{boy})$ affected by this, for her fifth child? With 0 for boy and 1 for girl, consider a Markov chain of children, with transition matrix

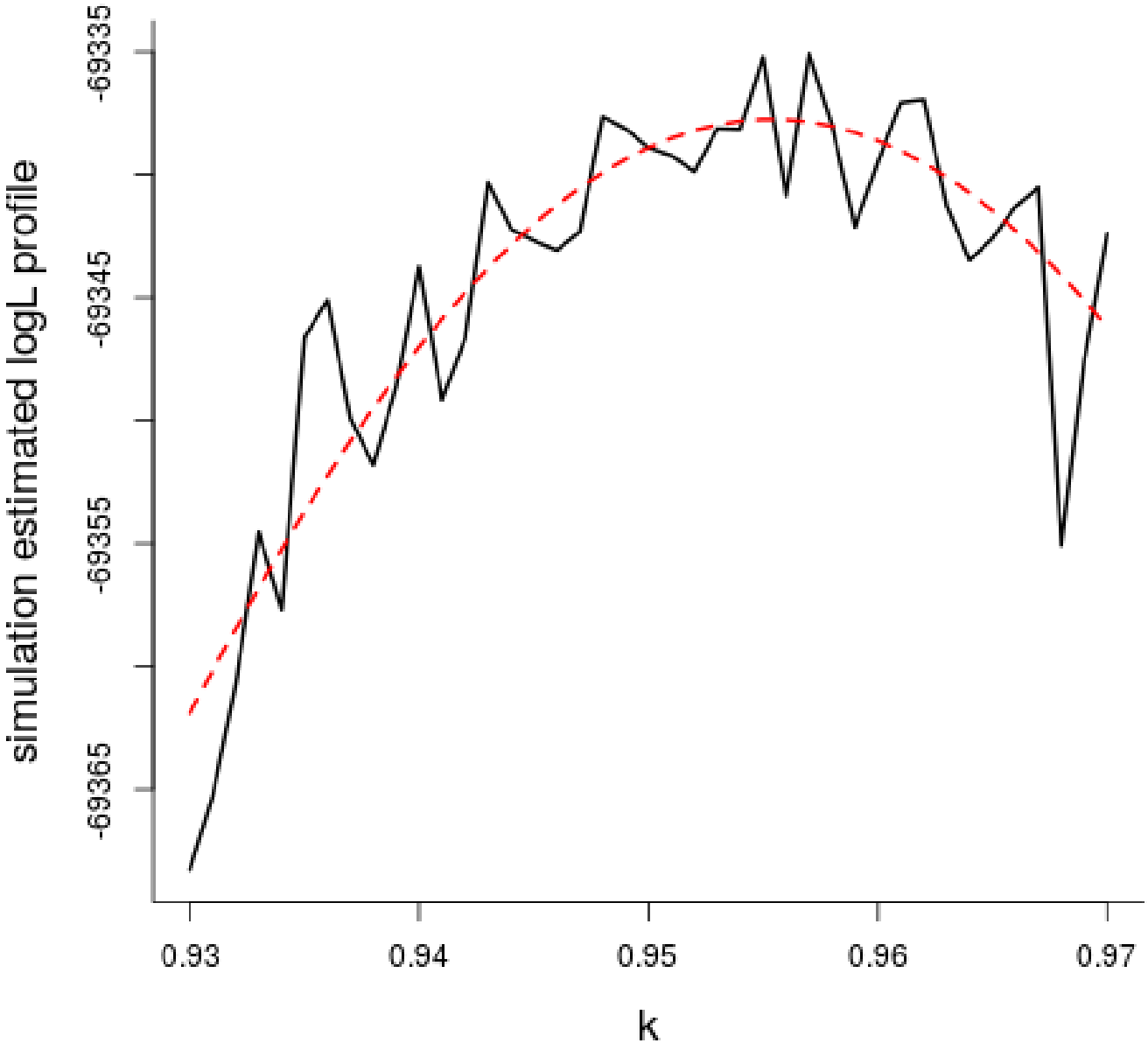
$$P = \begin{pmatrix} 1 - kp_0 & kp_0 \\ kq_0 & 1 - kq_0 \end{pmatrix}.$$

The chain has $(q_0, p_0) = (1 - p_0, p_0)$ as equlibrium. The correlation between (boy, boy), and also between (girl, girl), is $1 - k$. If $k = 1$ we're back to ordinary independence.

Estimation is non-trivial. I do ML, with

$$\ell_n(k, p_0) = \sum_{i=1}^{n} \log f_3(y_i, k, p_0) = \sum_{y=0}^{m} N(y) \log f_3(y, k, p_0),$$

but with no formula for $f_3(y, k, p_0) = \Pr(Y = y \mid k, p_0)$, where $Y = \sum_{i=1}^{m} X_i$ from the Markov chain. I manage via simulated likelihood (and heavy computations).



Simulated log-likelihood profile $\ell_{n,\mathrm{prof}}(k)$, and $\widehat{k} \doteq 0.95$, gender correlation 0.05. (So yes, my parents' 5th child was also a boy.)

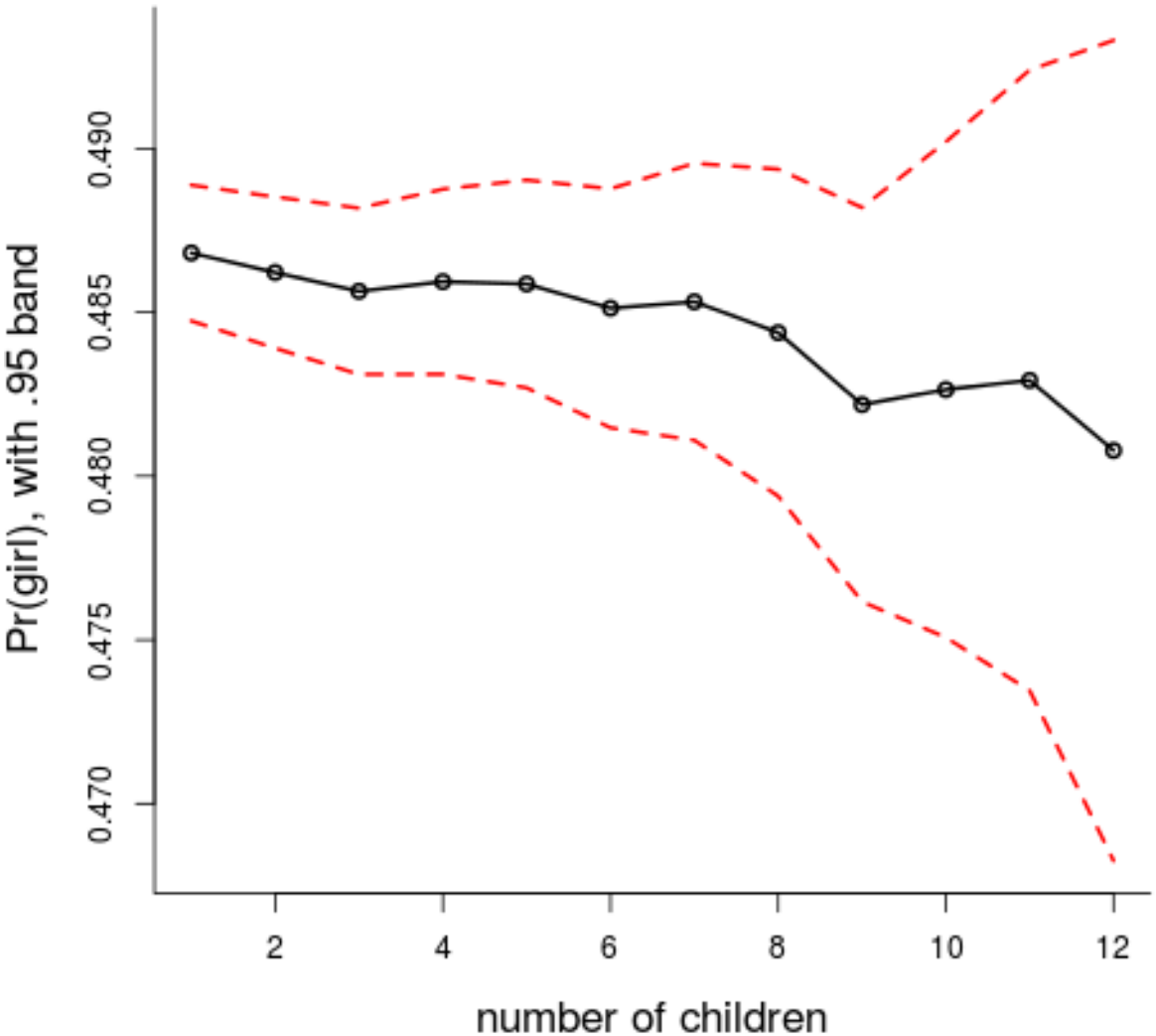
E: mysteries of human reproduction

Erna: "Jeg tror ikke jeg trenger å forklare hvordan dette gjøres. Jeg skal heller ikke komme med noen pålegg." With 1.62 kids per woman, Norway will be go from 5.2 millions to just half a million in 300 years (and just a few thousand in 2999 a.C.).

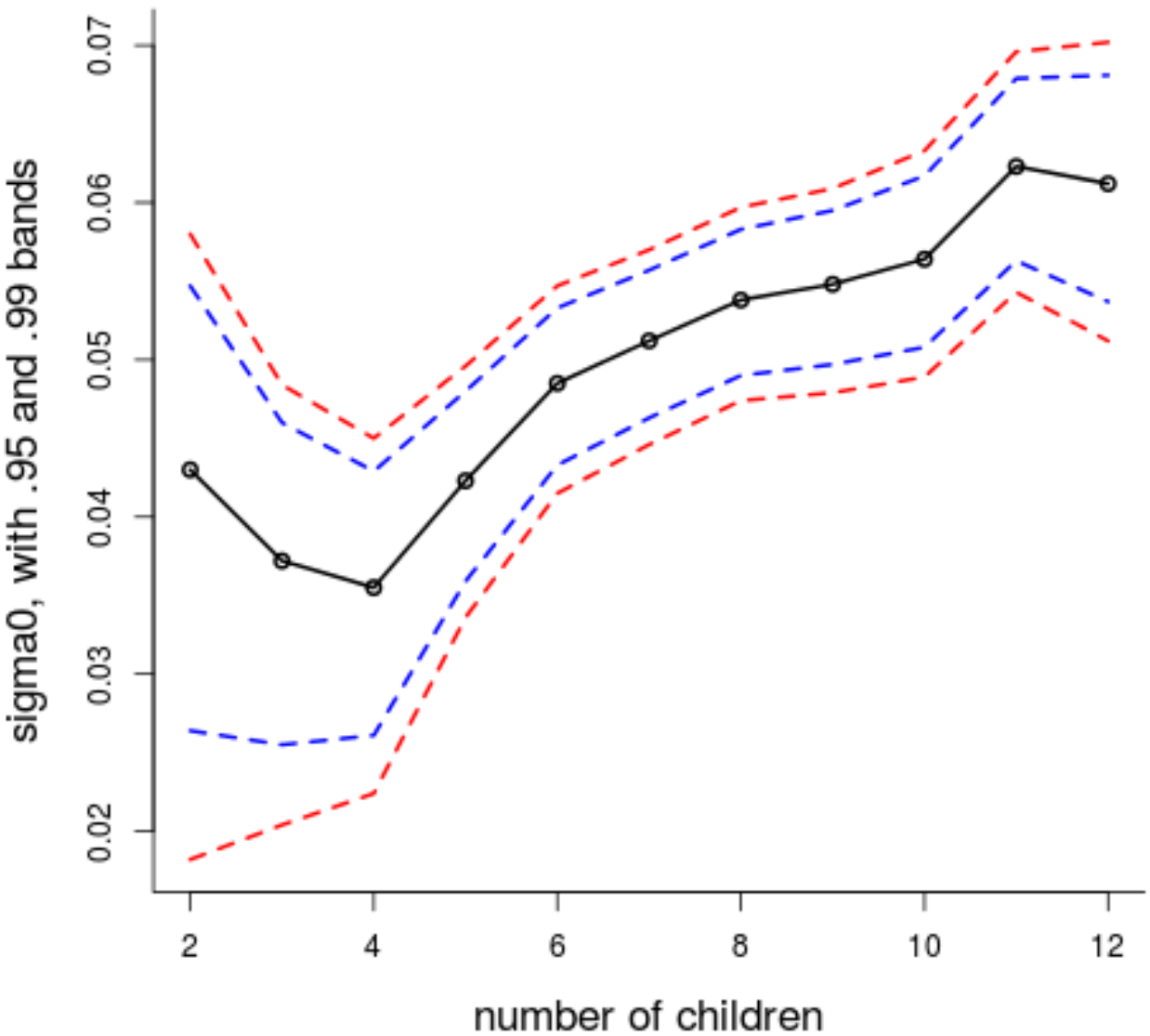
Why (close to) 0.50-0.50? R.A. Fisher (1930) argues via parental expenditure and evolutionary biology that the the sex ratio 'should' be 1:1. At least we're close. Glorious Science may change this.

Are there really more boys born in times of war? Earlier: bah! But Sir David Spiegelhalter says yes, complete with an argument. (If you're old enough, read his Sex by Numbers.)

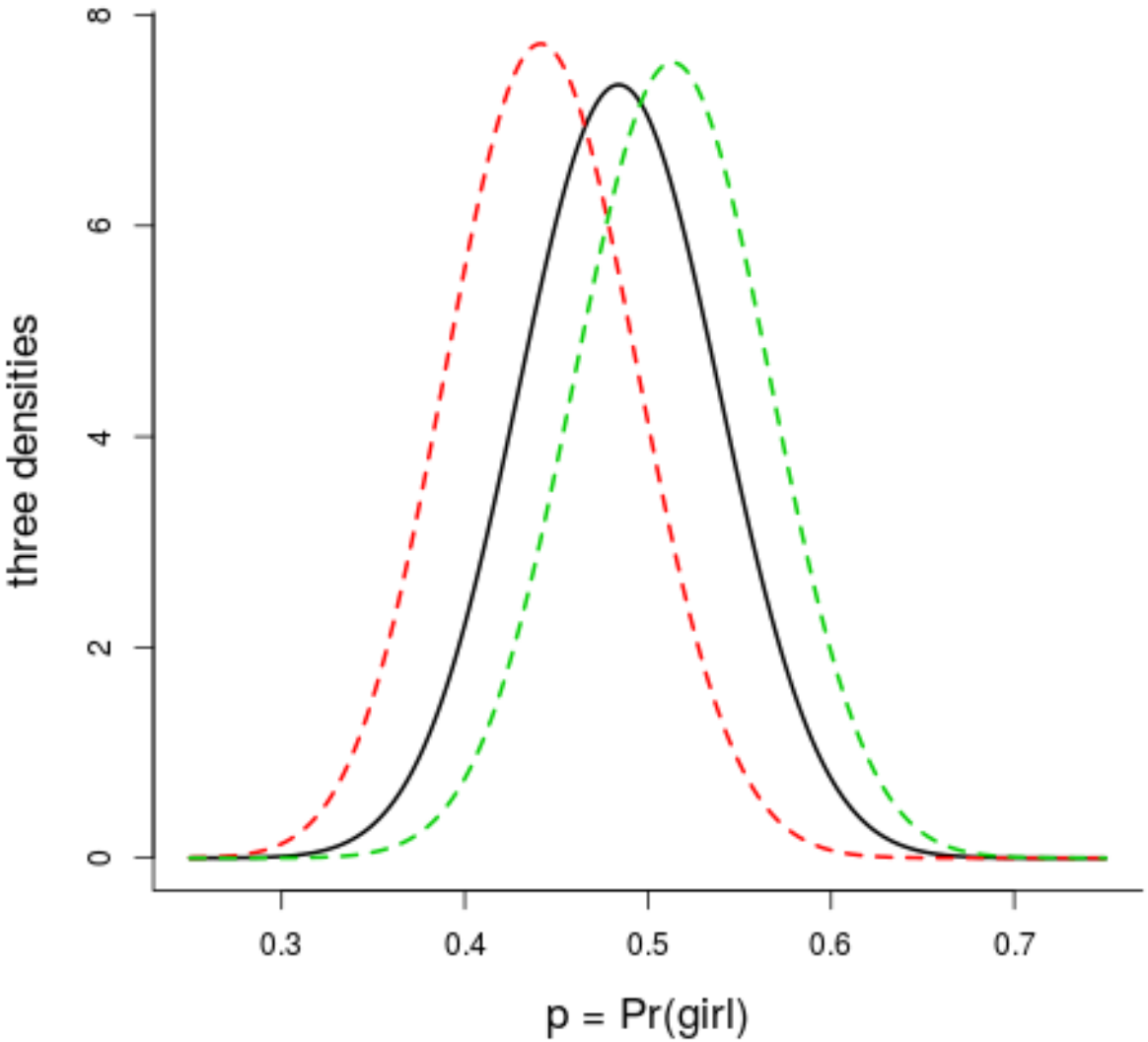
Other models? Feel free, combining Markov with beta-binomial, etc., using Geissler and e.g. Amish and Mennonites data. But for Homo Sapiens this is now less interesting than it has been over the past six thousand year.



Are there more and more boys, in bigger and bigger families?



Is there more and more p-variability, in bigger and bigger families?



Bayes' theorem for given families: red left for Kristin Lavrandsdatter; green right for my mother's parents; black middle for the overall population.