CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 3

Review where we are:

$$P(A|S_B I) = \int_x P(A|S_B x I)\, p(x|I)\, dx$$
$$= \int_x \frac{1}{1+x}\, p(x|I)\, dx$$

We are trying to estimate a parameter

$$x = P(S_B|BC), \quad (0 \le x \le 1)$$

The form of our estimate is a (Bayesian) probability distribution (of the parameter, itself here just happening to be a probability)

This is a sterile exercise if it is just a debate about priors. What we need is data! Data might be a previous history of choices by the jailer in identical circumstances.

BCBCCBCCCBBCBCBCCCCBBCBCCCBCBCBBCCB

$$N = 35, \quad N_B = 15, \quad N_C = 20$$

(What's wrong with: x=15/35=0.43? Hold on…)

We hypothesize (might later try to check) that these are i.i.d. "Bernoulli trials" and therefore informative about *x*

"independent and identically distributed"

As good Bayesians, we now need $P(\text{data}|x)$

$P(\text{data}|x)$ means different things in frequentist vs. Bayesian contexts, so this is a good time to understand the differences (we'll use both ideas as appropriate)

Frequentist considers the universe of what might have been, imagining repeated trials, even if they weren't actually tried, and needs no prior:

since i.i.d. only the $\mathcal{N}$'s can matter (a so-called "sufficient statistic").

$$P(\text{data}|x) = \binom{N}{N_B} \overbrace{x^{N_B}\,(1-x)^{N_C}}^{\text{prob. of exact sequence seen}} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

no. of equivalent arrangements

Bayesian considers only the exact data seen, and has a prior:

$$P(x|\text{data}) \propto x^{N_B}\,(1-x)^{N_C}\; p(x|I)$$

but we might first suppose that the prior it is **uniform**

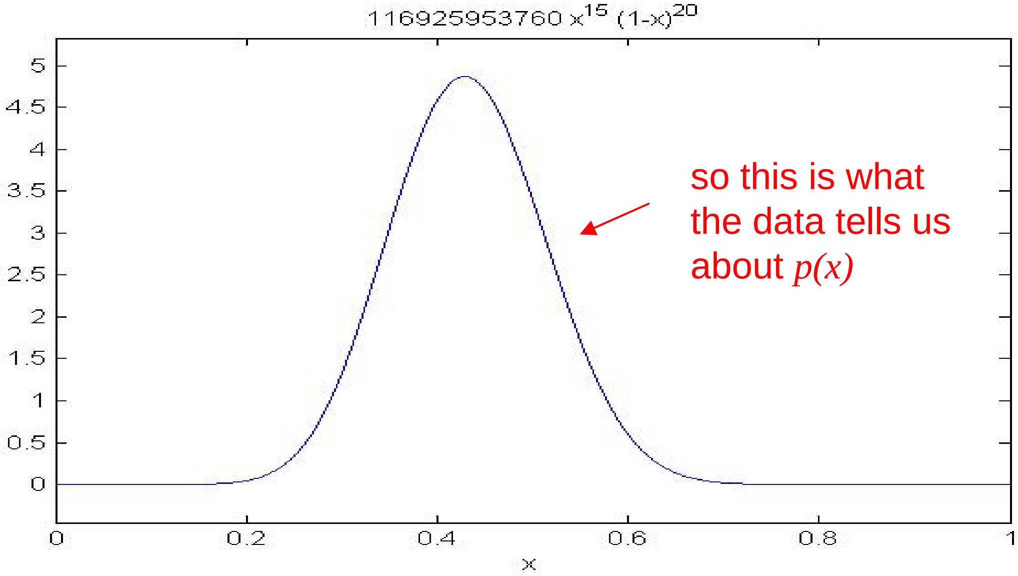
No binomial coefficient, both conceptually and also since independent of x and absorbed in the proportionality. Use only the data you see, not "equivalent arrangements" that you didn't see. This issue is one we'll return to, not always entirely sympathetically to Bayesians (e.g., goodness-of-fit).

Bayes numerator and denominator are:

$$P(x|\text{data}) \propto x^{N_B}(1-x)^{N-N_B} \times 1$$

$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B}(1-x)^{N-N_B}dx = \frac{\Gamma(N_B+1)\Gamma(N-N_B+1)}{\Gamma(N+2)}$$

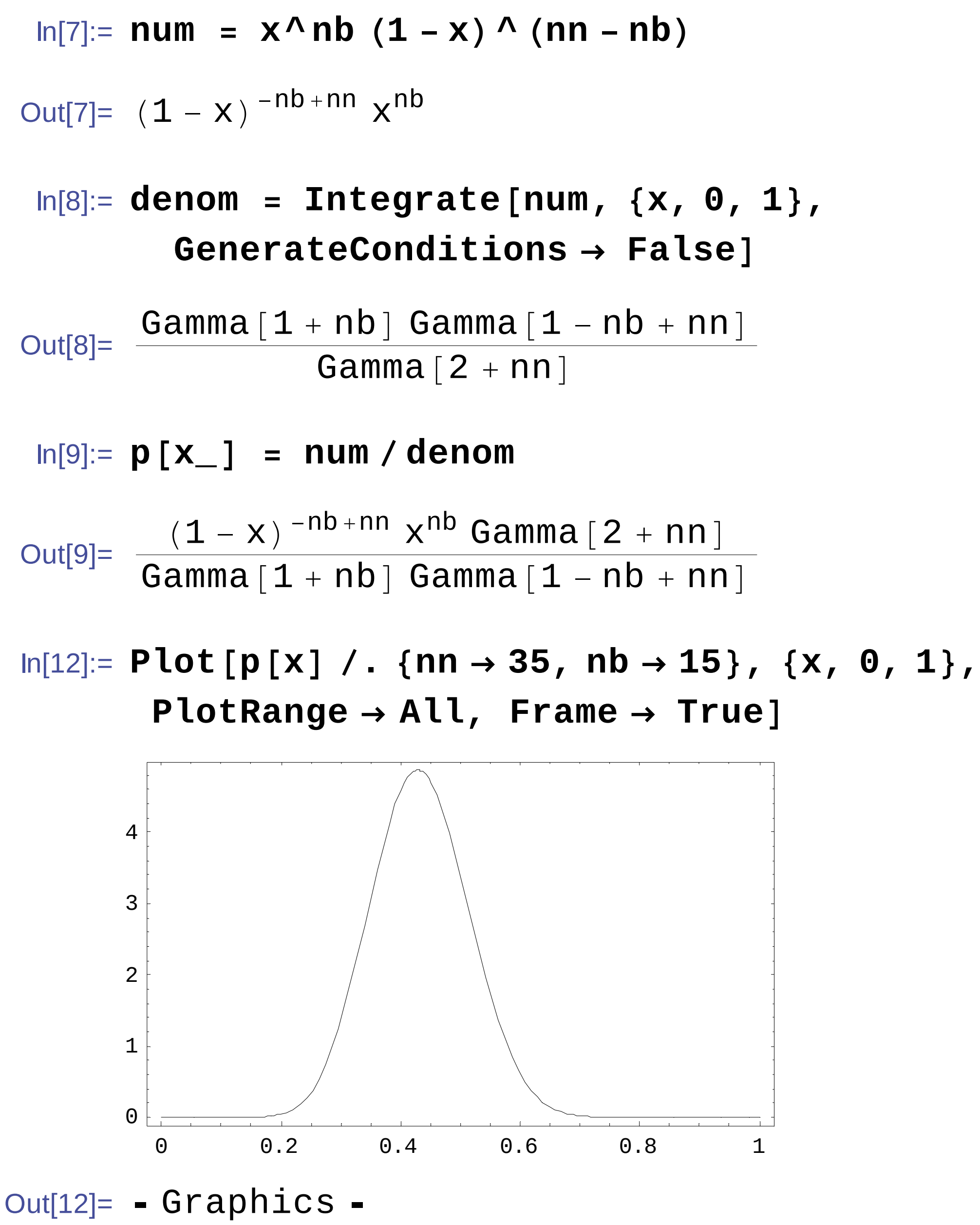
Plot of numerator over denominator for N=35, N_B = 15:



so this is what the data tells us about *p(x)*

You should learn to do calculations like this in MATLAB or Mathematica:

```
syms nn nb x
num = x^nb * (1-x)^(nn-nb)
num =
x^nb*(1-x)^(nn-nb)
denom = int(num, 0, 1)
denom =
gamma(nn-nb+1)*gamma(nb+1)/gamma(nn+2)
p = num / denom
p =
x^nb*(1-x)^(nn-nb)/gamma(nn-
nb+1)/gamma(nb+1)*gamma(nn+2)
ezplot(subs(p,[nn,nb],[35,15]),[0,1])
```



```
In[7]:= num = x^nb (1 - x)^(nn - nb)
```

Out[7]= $(1-x)^{-nb+nn}\,x^{nb}$

```
In[8]:= denom = Integrate[num, {x, 0, 1},
          GenerateConditions → False]
```

Out[8]= $\frac{\text{Gamma}[1+nb]\ \text{Gamma}[1-nb+nn]}{\text{Gamma}[2+nn]}$

```
In[9]:= p[x_] = num / denom
```

Out[9]= $\frac{(1-x)^{-nb+nn}\,x^{nb}\ \text{Gamma}[2+nn]}{\text{Gamma}[1+nb]\ \text{Gamma}[1-nb+nn]}$

```
In[12]:= Plot[p[x] /. {nn → 35, nb → 15}, {x, 0, 1},
          PlotRange → All, Frame → True]
```

```
Out[12]= - Graphics -
```

Find the mean, standard error, and mode of our estimate for x

$$P(x|\text{data}) \propto x^{N_B}(1-x)^{N-N_B}$$

$$\frac{dP(x|\text{data})}{dx} = 0 \;\Rightarrow\; x = \frac{N_B}{N}$$

"maximum likelihood" (ML) answer is to estimate x as exactly the fraction seen

$$\langle x \rangle = \int_0^1 x P(x|\text{data})dx = \frac{N_B+1}{N+2}$$

mean is the 1st moment
notice it's different from ML!

variance involves the 2nd moment,

$$\text{Var}(x) = \langle x^2 \rangle - \langle x \rangle^2 = \int_0^1 x^2 P(x|\text{data})dx - \langle x \rangle^2 = \frac{(N_B+1)(N-N_B+1)}{(N+2)^2(N+3)}$$

This shows how *p(x)* gets narrower as the amount of data increases.

(Let’s leave behind the metaphor of the Jailer and Prisoner A.)

What we are illustrating is called Bernoulli trials:

- two possible outcomes
- i.i.d. events
- a single parameter x (the probability of one outcome)
- a sufficient statistic is the pair of numbers N and N_B

Jacob and Johann Bernoulli

$$P(\text{data}|x) = x^{N_B}(1-x)^{N-N_B}$$ (in the Bayesian sense)

$$P(x|\text{data}) \propto x^{N_B}(1-x)^{N-N_B} \times P(x|I)$$

for uniform prior, the Bayes denominator is, as we’ve seen, easy to calculate:

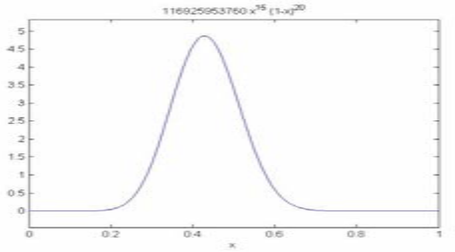
$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B}(1-x)^{N-N_B}dx = \frac{\Gamma(N_B+1)\Gamma(N-N_B+1)}{\Gamma(N+2)}$$

Are there any other mathematical forms for the prior that would still leave the Bayes denominator easy to calculate?

Yes! try

$$P(x|I) \propto x^{\beta}(1-x)^{\alpha}$$

Choose α and β to make any desired center and width.



$$P(x|\text{data}) = x^{N_B}(1-x)^{N-N_B} \times x^{\beta}(1-x)^{\alpha}$$

$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B+\beta}(1-x)^{N-N_B+\alpha}dx$$

$$= \frac{\Gamma(N_B+\beta+1)\Gamma(N-N_B+\alpha+1)}{\Gamma(N+\alpha+\beta+2)}$$

Priors that preserve the analytic form of p(x) are called “conjugate priors”. There is nothing special about them except mathematical convenience.

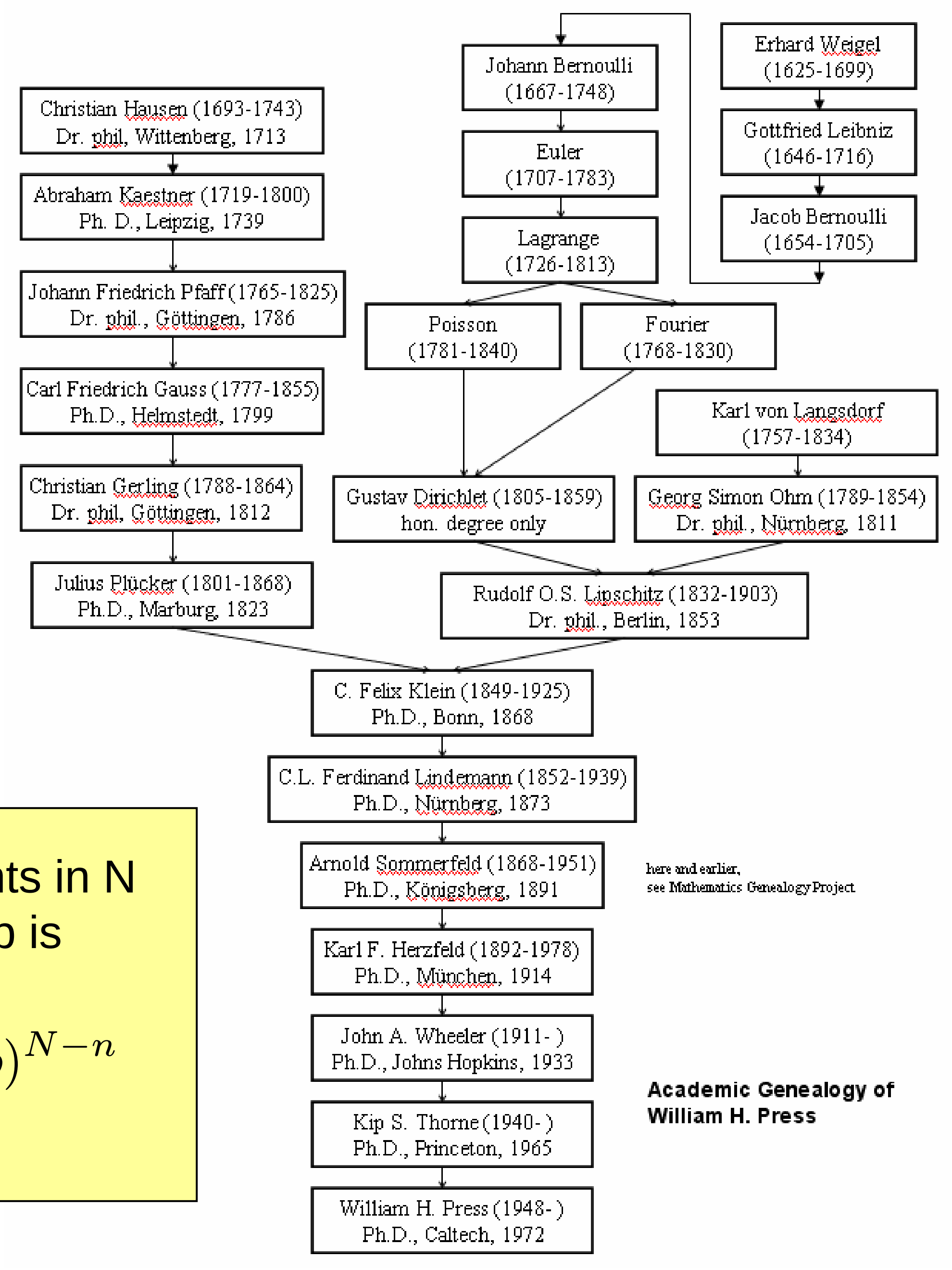
If you start with a conjugate prior, you’ll also be able to assimilate new data trivially, just by changing the parameters of your estimate. This is because every posterior is in the right analytic form to be the new prior!

By the way, if I show a special love of Bernoulli trials, it might be because I am an academic descendent of the Bernoulli brothers!

Actually, this is not a very exclusive club: Gauss and the Bernoullis each have ~50,000 recorded descendents in the Mathematics Genealogy database, and probably many times more unrecorded.

The probability of getting n events in N tries, each with i.i.d. probability p is

$$\mathrm{bin}(n, N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$



Johann Bernoulli (1667-1748)

Euler (1707-1783)

Lagrange (1726-1813)

Erhard Weigel (1625-1699)

Gottfried Leibniz (1646-1716)

Jacob Bernoulli (1654-1705)

Poisson (1781-1840)

Fourier (1768-1830)

Christian Hausen (1693-1743)
Dr. phil, Wittenberg, 1713

Abraham Kaestner (1719-1800)
Ph. D., Leipzig, 1739

Johann Friedrich Pfaff (1765-1825)
Dr. phil., Göttingen, 1786

Carl Friedrich Gauss (1777-1855)
Ph.D., Helmstedt, 1799

Christian Gerling (1788-1864)
Dr. phil, Göttingen, 1812

Julius Plücker (1801-1868)
Ph.D., Marburg, 1823

Karl von Langsdorf (1757-1834)

Gustav Dirichlet (1805-1859)
hon. degree only

Georg Simon Ohm (1789-1854)
Dr. phil., Nürnberg, 1811

Rudolf O.S. Lipschitz (1832-1903)
Dr. phil., Berlin, 1853

C. Felix Klein (1849-1925)
Ph.D., Bonn, 1868

C.L. Ferdinand Lindemann (1852-1939)
Ph.D., Nürnberg, 1873

Arnold Sommerfeld (1868-1951)
Ph.D., Königsberg, 1891

Karl F. Herzfeld (1892-1978)
Ph.D., München, 1914

John A. Wheeler (1911-)
Ph.D., Johns Hopkins, 1933

Kip S. Thorne (1940-)
Ph.D., Princeton, 1965

William H. Press (1948-)
Ph.D., Caltech, 1972

here and earlier,
see Mathematics Genealogy Project

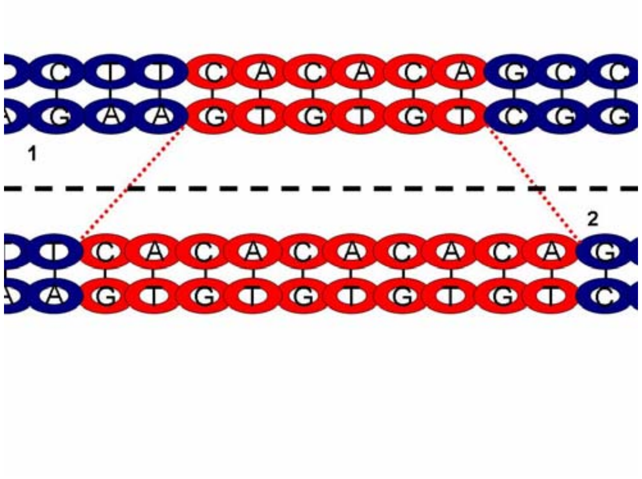
Academic Genealogy of William H. Press

Next example (with some biology):

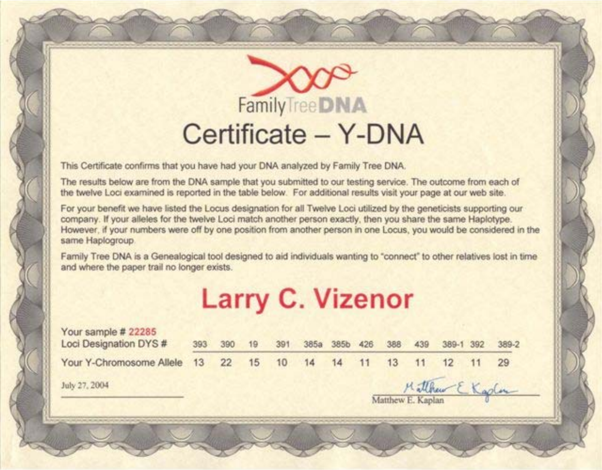
Individual identity, or ancestry, can be determined by “variable length short tandem repeats” (STRs) in the genome.

~0.5% mutation prob per STR per generation (though highly variable)

if use Y chromosome only, get paternal ancestry



There are companies that sell “certificates” with your genotype. A bit opportunistic, since in a few years your whole genome will be sequenced by your health plan.



Family Tree DNA

Certificate – Y-DNA

This Certificate confirms that you have had your DNA analyzed by Family Tree DNA.

The results below are from the DNA sample that you submitted to our testing service. The outcome from each of the twelve Loci examined is reported in the table below. For additional results visit your page at our web site.

For your benefit we have listed the Locus designation for all Twelve Loci utilized by the geneticists supporting our company. If your alleles for the twelve Loci match another person exactly, then you share the same Haplotype. However, if your numbers were off by one position from another person in one Locus, you would be considered in the same Haplogroup.

Family Tree DNA is a Genealogical tool designed to aid individuals wanting to “connect” to other relatives lost in time and where the paper trail no longer exists.

Larry C. Vizenor

Your sample # 22285

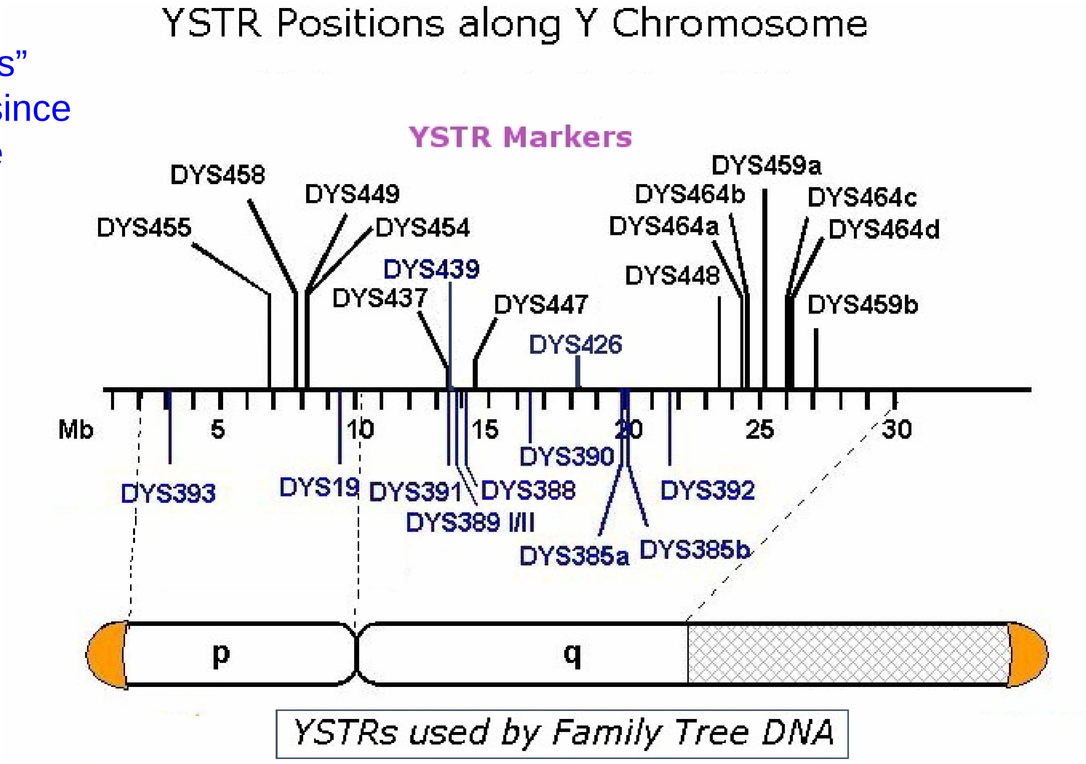
Loci Designation DYS #	393	390	19	391	385a	385b	426	388	439	389-1	392	389-2
Your Y-Chromosome Allele	13	22	15	10	14	14	11	13	11	12	11	29

July 27, 2004

Matthew E. Kaplan

Matthew E. Kaplan

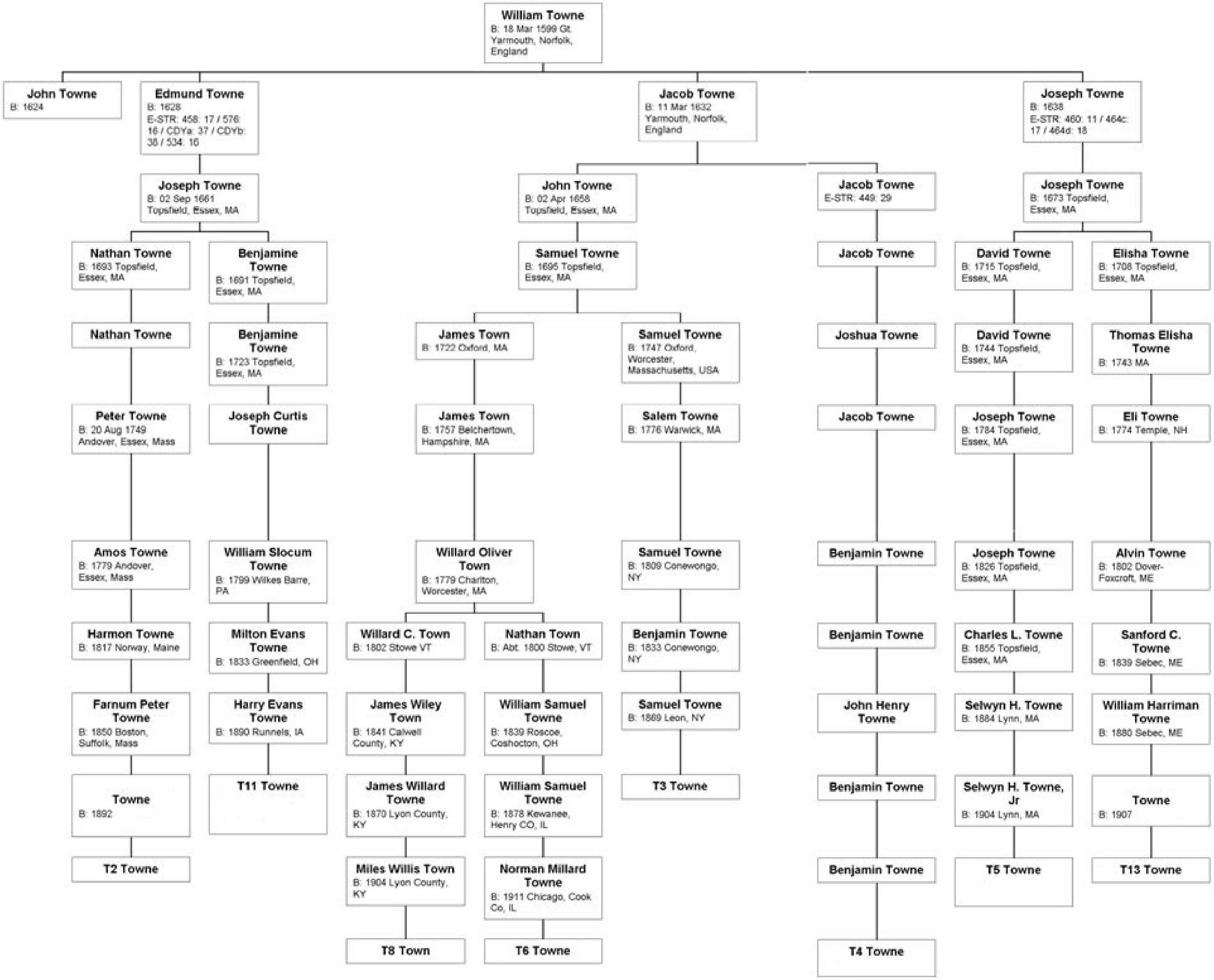
YSTR Positions along Y Chromosome



YSTRs used by Family Tree DNA

Margaret, my ex-wife, is really into the Towne family.
(And, she’s neither a biologist nor a Towne.)

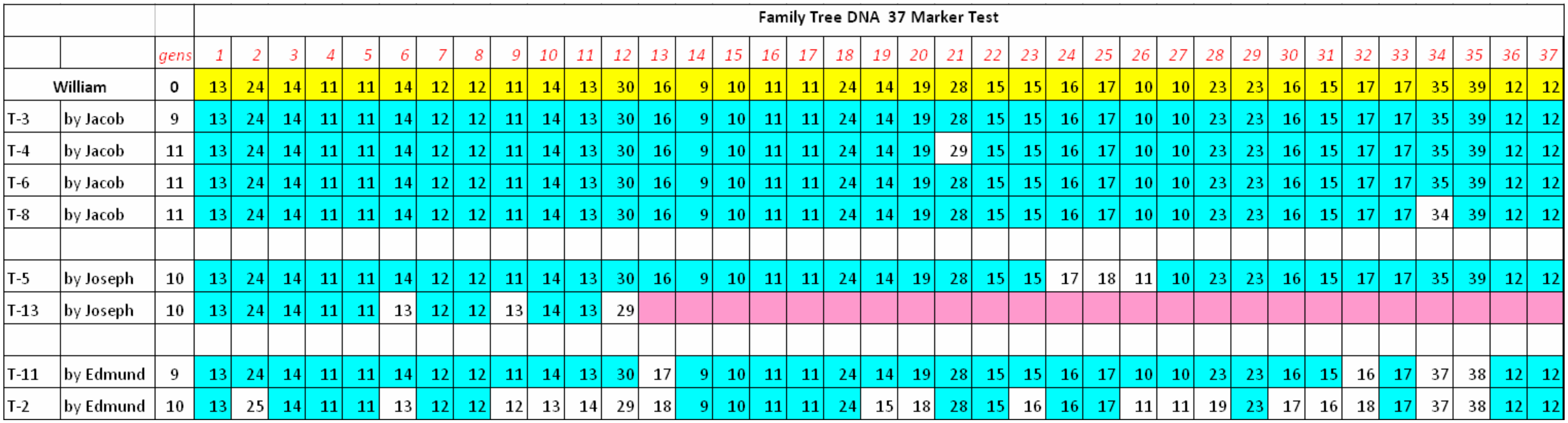
Descendant Chart for
William Towne



- William Towne B: 18 Mar 1599 Gt. Yarmouth, Norfolk, England
  - John Towne B: 1624
  - Edmund Towne B: 1628 E-STR: 458: 17 / 576: 16 / CDYa: 37 / CDYb: 38 / 534: 16
    - Joseph Towne B: 02 Sep 1661 Topsfield, Essex, MA
      - Nathan Towne B: 1693 Topsfield, Essex, MA
        - Nathan Towne
          - Peter Towne B: 20 Aug 1749 Andover, Essex, Mass
            - Amos Towne B: 1779 Andover, Essex, Mass
              - Harmon Towne B: 1817 Norway, Maine
                - Farnum Peter Towne B: 1850 Boston, Suffolk, Mass
                  - Towne B: 1892
                    - T2 Towne
      - Benjamine Towne B: 1691 Topsfield, Essex, MA
        - Benjamine Towne B: 1723 Topsfield, Essex, MA
          - Joseph Curtis Towne
            - William Slocum Towne B: 1799 Wilkes Barre, PA
              - Milton Evans Towne B: 1833 Greenfield, OH
                - Harry Evans Towne B: 1890 Runnels, IA
                  - T11 Towne
  - Jacob Towne B: 11 Mar 1632 Yarmouth, Norfolk, England
    - John Towne B: 02 Apr 1658 Topsfield, Essex, MA
      - Samuel Towne B: 1695 Topsfield, Essex, MA
        - James Town B: 1722 Oxford, MA
          - James Town B: 1757 Belchertown, Hampshire, MA
            - Willard Oliver Town B: 1779 Charlton, Worcester, MA
              - Willard C. Town B: 1802 Stowe VT
                - James Wiley Town B: 1841 Calwell County, KY
                  - James Willard Towne B: 1870 Lyon County, KY
                    - Miles Willis Town B: 1904 Lyon County, KY
                      - T8 Town
              - Nathan Town B: Abt. 1800 Stowe, VT
                - William Samuel Towne B: 1839 Roscoe, Coshocton, OH
                  - William Samuel Towne B: 1878 Kewanee, Henry CO, IL
                    - Norman Millard Towne B: 1911 Chicago, Cook Co, IL
                      - T6 Towne
        - Samuel Towne B: 1747 Oxford, Worcester, Massachusetts, USA
          - Salem Towne B: 1776 Warwick, MA
            - Samuel Towne B: 1809 Conewongo, NY
              - Benjamin Towne B: 1833 Conewongo, NY
                - Samuel Towne B: 1869 Leon, NY
                  - T3 Towne
    - Jacob Towne E-STR: 449: 29
      - Jacob Towne
        - Joshua Towne
          - Jacob Towne
            - Benjamin Towne
              - Benjamin Towne
                - John Henry Towne
                  - Benjamin Towne
                    - Benjamin Towne
                      - T4 Towne
  - Joseph Towne B: 1638 E-STR: 460: 11 / 464c: 17 / 464d: 18
    - Joseph Towne B: 1673 Topsfield, Essex, MA
      - David Towne B: 1715 Topsfield, Essex, MA
        - David Towne B: 1744 Topsfield, Essex, MA
          - Joseph Towne B: 1784 Topsfield, Essex, MA
            - Joseph Towne B: 1826 Topsfield, Essex, MA
              - Charles L. Towne B: 1855 Topsfield, Essex, MA
                - Selwyn H. Towne B: 1884 Lynn, MA
                  - Selwyn H. Towne, Jr B: 1904 Lynn, MA
                    - T5 Towne
      - Elisha Towne B: 1708 Topsfield, Essex, MA
        - Thomas Elisha Towne B: 1743 MA
          - Eli Towne B: 1774 Temple, NH
            - Alvin Towne B: 1802 Dover-Foxcroft, ME
              - Sanford C. Towne B: 1839 Sebec, ME
                - William Harriman Towne B: 1880 Sebec, ME
                  - Towne B: 1907
                    - T13 Towne

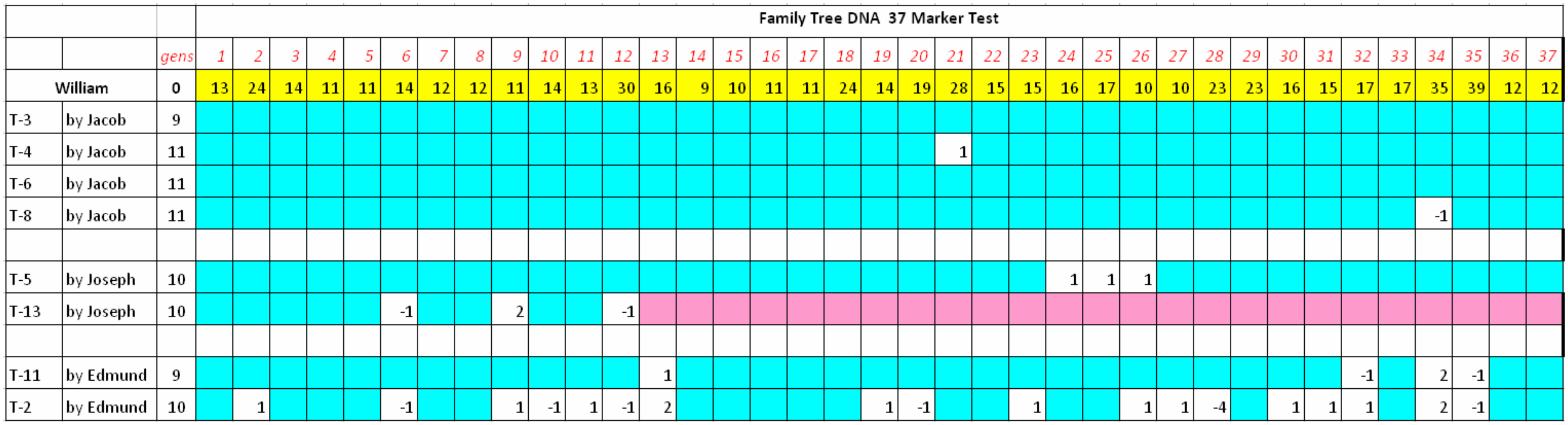
as of March 6, 2008

Here's data from Margaret on 8 recent Townes (identified only by T code).
(We'll use this data several times in the next few of lectures.)



			Family Tree DNA 37 Marker Test																																				
		gens	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
William		0	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-3	by Jacob	9	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-4	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	29	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-6	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-8	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	34	39	12	12
T-5	by Joseph	10	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	17	18	11	10	23	23	16	15	17	17	35	39	12	12
T-13	by Joseph	10	13	24	14	11	11	13	12	12	13	14	13	29																									
T-11	by Edmund	9	13	24	14	11	11	14	12	12	11	14	13	30	17	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	16	17	37	38	12	12
T-2	by Edmund	10	13	25	14	11	11	13	12	12	12	13	14	29	18	9	10	11	11	24	15	18	28	15	16	16	17	11	11	19	23	17	16	18	17	37	38	12	12

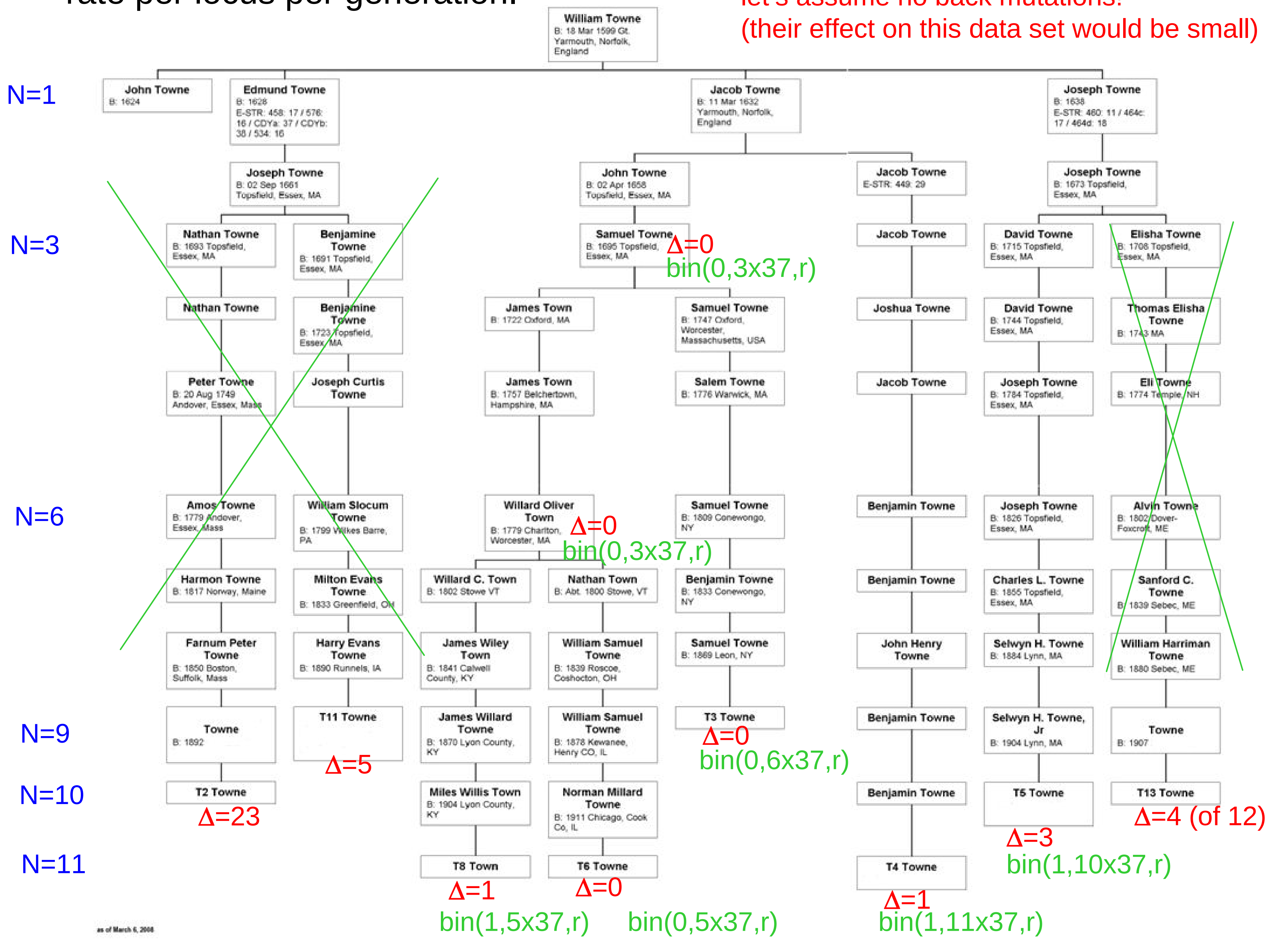
or, just showing the changes:



			Family Tree DNA 37 Marker Test																																				
		gens	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
William		0	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-3	by Jacob	9																																					
T-4	by Jacob	11																					1																
T-6	by Jacob	11																																					
T-8	by Jacob	11																																		-1			
T-5	by Joseph	10																								1	1	1											
T-13	by Joseph	10						-1			2			-1																									
T-11	by Edmund	9													1																			-1		2	-1		
T-2	by Edmund	10		1				-1			1	-1	1	-1	2						1	-1			1			1	1	-4		1	1	1		2	-1		

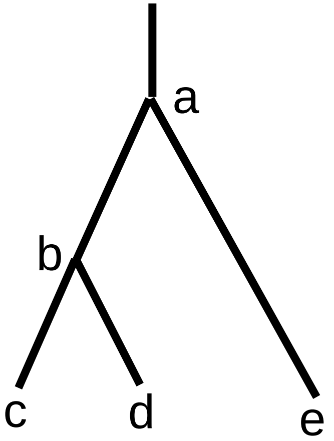
Let’s do a Bayesian estimation of the parameter r, the mutation rate per locus per generation.

let’s assume no back mutations!
(their effect on this data set would be small)



Unraveling dependencies

We used “ancestry” as an intuitive example of how dependencies work, because its causal mechanism is well known! Later, we’ll do more general “Bayesian networks” to capture models with more complicated causal dependencies.



$$\begin{aligned} P(abcde) &= P(e|a\cancel{bcd})P(abcd) \\ &= P(e|a)P(c|\cancel{a}b\cancel{d})P(abd) \\ &= P(e|a)P(c|b)P(d|\cancel{a}b)P(ab) \\ &= P(e|a)P(c|b)P(d|b)P(b|a)P(a) \end{aligned}$$

Another important idea is “conditional independence”

Example: b and e are “conditionally independent given a”

$$\begin{aligned} P(be|a) &= P(b|\cancel{e}a)P(e|a) \\ &= P(b|a)P(e|a) \end{aligned}$$

while b and d are not conditionally independent given a:

$$P(bd|a) = P(b|da)P(d|a)$$

?

So we have a statistical model for the data,
that is, a way to compute $P(\text{data}|\text{parameters})$

It is not "exact", but statistical models rarely (never?) are.

neglects backmutations
assumes single probability for all loci
etc.

The model is:

$$P(\text{data}|r) = \text{bin}(0, 3\times 37, r)\,\text{bin}(0, 3\times 37, r)\,\text{bin}(1, 5\times 37, r)\,\text{bin}(0, 5\times 37, r)$$
$$\times\,\text{bin}(0, 6\times 37, r)\,\text{bin}(1, 11\times 37, r)\,\text{bin}(3, 10\times 37, r)$$

Bayes estimation of the parameter:

$$P(r|\text{data}) \propto P(\text{data}|r) \times P(r) \propto P(\text{data}|r) \times \frac{1}{r}$$

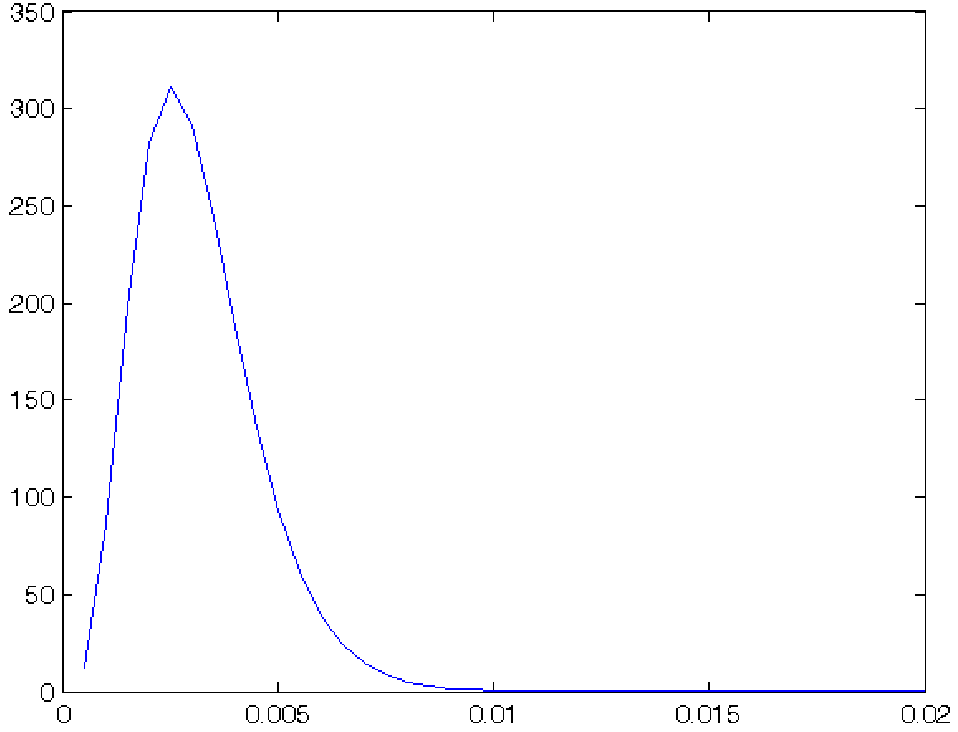
What kind of prior is this???
It is called "log-uniform"

The log-uniform prior has equal probability in each order of magnitude.

$$\int_r^{10r} P(r)dr = \int_r^{10r} \frac{1}{r}dr = \log 10$$

It is often taken as the non-informative prior when you don't even know the order of magnitude of the (positive) quantity.
It is an "improper prior" since its integral is infinite.
This is almost always ok, but it is possible to construct paradoxes with improper priors (e.g., the "marginalization paradox")

Here is the plot of the (normalized) $P(r|\text{data})$

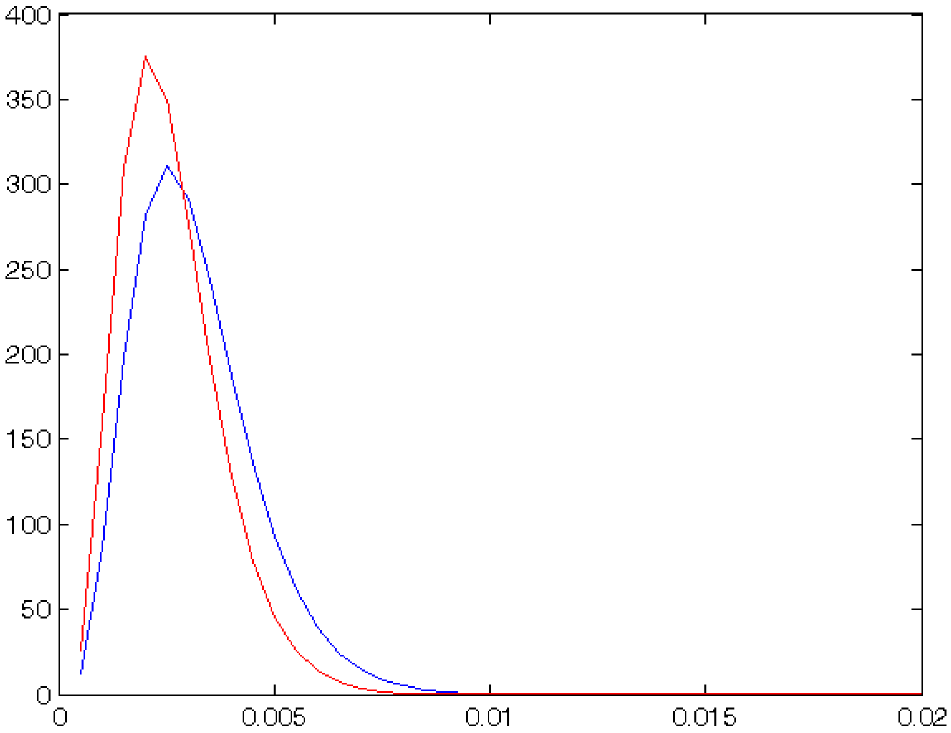


This is (almost) real biology. We've measured the mutation probability, per locus per generation of Y chromosome STRs. This tells us something about the actual DNA replication machinery!

It really did matter (a bit) that we sorted out the conditional dependencies correctly.
Here's a comparison to doing it wrong by assuming all data independent:

The true dependencies allow somewhat larger values of r, because we don't wrongly count the Δ=0 branches multiple times

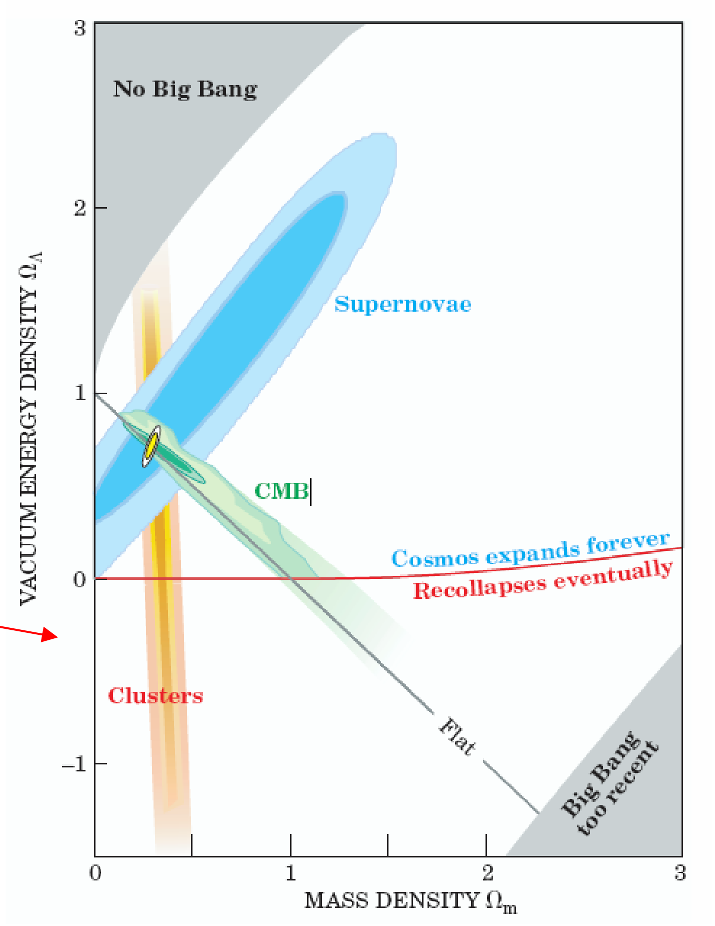
We'll come back to the Towne family for some fancier stuff later!



Ignoring conditional dependencies and just multiplying the probabilities of the data as if they were independent is called naïve Bayes. People often do this. It is mathematically incorrect, but sometimes it is all you can do!

The basic paradigm of Bayesian parameter estimation :

- Construct a statistical model for the probability of the observed data as a function of all parameters
  - treat dependency in the data correctly
- Assign prior distributions to the parameters
  - jointly or independently as appropriate
  - use the results of previous data if available
- Use Bayes law to get the (multivariate) posterior distribution of the parameters
- Marginalize as desired to get the distributions of single (or a manageable few multivariate) parameters



Cosmological models are typically fit to many parameters. Marginalization yields the distribution of parameters of interest, here two, shown as contours.