Discrete Probability Distributions
Geometric and Negative Binomial
illustrated by
Mitochondrial Eve
and
Cancer Driver/Passenger Genes
Binomial Distribution

• Number of successes in $n$ independent Bernoulli trials

• The probability mass function is:

$$P(X = x) = C_x^n p^x (1 - p)^{n-x} \text{ for } x = 0, 1, \ldots, n \quad (3-7)$$
Geometric Distribution

• A series of Bernoulli trials with probability of success = $p$. continued **until the first success**. $X$ is the number of trials.

• Compare to: Binomial distribution has:
  – Fixed number of trials = $n$.
  – Random number of successes = $x$.

• Geometric distribution has reversed roles:
  – Random number of trials, $x$
  – Fixed number of successes, in this case 1.
  – Success always comes in the end: so no combinatorial factor $C^n_x$
  – $P(X=x) = p(1-p)^{x-1}$ where:
    $x-1 = 0, 1, 2, ...$, the number of failures until the $1^{st}$ success.

• **NOTE OF CAUTION:** Matlab, Mathematica, and many other sources use $x$ to denote the number of failures until the first success. We stick with Montgomery-Runger notation.
Geometric Mean & Variance

• If $X$ is a geometric random variable (according to Montgomery-Bulmer) with parameter $p$,

$$
\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \quad \text{(3-10)}
$$

• For small $p$ the standard deviation $\sim =$ mean

• Very different from Poisson, where it is variance = mean and standard deviation = mean$^{1/2}$
Each of our cells has mitochondria with 16.5kb of mtDNA inherited only from our mother.

Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA.

Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC’s Carl R. Woese fame).

Since then most mitochondrial genes were transferred to the nucleus.

Plants also have plastids with genomes related to cyanobacteria.
Time to the last common (maternal) ancestor follows geometric distribution

- **Constant population** of N women
- **Random number** of (female) offspring. Average is 1 (but can be 0 or 2)
- **Randomly pick two women.** Question: how many generations T since their last maternal ancestor?
- T is a random variable. What is its PMF: P(T=t)? Answer: P(T=t) follows a geometric distribution
- Do these two women have the same mother? Yes: “success” in finding their last common ancestor (p=1/N). P(T=1)=1/N.
- No? “failure” (1-p=1-1/N). Go to their mothers and repeat the same question.
- P(T=t)=(1-1/N)^{t-1} (1/N) \approx (1/N) \exp(-T/N)
- T can be inferred from the density of differences on mtDNA =2\mu T

Maddamsetti R, MOBILE GENETIC ELEMENTS, 6, e1137380(2016)
• There are about $N=3.5 \times 10^9$ women living today

• For a random pair of women the average number of generations to the last common maternal ancestor is:

$$E(T) = \sum_{T=1}^{\infty} T \cdot \exp(-T/N) = 1/p = N$$

• Most Recent maternal Common Ancestor (MRCA) of all people living today lived $T_{MRCA} = 2N$ generations ago

• $T_{MRCA} = 2 \cdot 3.5 \times 10^9$ generations

• If the generation time 20 years it is 140 billion years > 10 times the time since the Big Bang.

• Something is wrong here!
• Population is not constant and for a long time was very low

• Change N to “effective” size $N_e$

• Current thinking is that for all of us including people of African ancestry $N_e \approx 7500$ people

• For humans of European + Asian ancestry $N_e \approx 3100$ people

• Mito Eve lived ~

  $2 \times N_e \times 20 \text{ years} = 2 \times 7500 \times 20 \text{ years} = 300,000 \text{ years ago}$

Recent human effective population size estimated from linkage disequilibrium

Albert Tenesa,1,2,3 Pau Navarro,3 Ben J. Hayes,4 David L. Duffy,5 Geraldine M. Clarke,6 Mike E. Goddard,4,7 and Peter M. Visscher3,5,8

1 Colon Cancer Genetics Group, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; 2 MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; 3 Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JL, United Kingdom; 4 Victorian Institute of Animal Science, DPI, Attwood 3049, Australia; 5 Queensland Institute of Medical Research, Royal Brisbane Hospital, Brisbane 4006, Australia; 6 The Wellcome Trust Centre for Human Genetics, The University of Oxford, Oxford OX3 7BN, United Kingdom; 7 Institute of Land and Food Resources, University of Melbourne, Parkville 3010, Australia

Effective population size ($N_e$) determines the amount of genetic variation, genetic drift, and linkage disequilibrium (LD) in populations. Here, we present the first genome-wide estimates of human effective population size from LD data. Chromosome-specific effective population size was estimated for all autosomes and the X chromosome from estimated LD between SNP pairs <100 kb apart. We account for variation in recombination rate by using coalescent-based estimates of fine-scale recombination rate from one sample and correlating these with LD in an independent sample. Phase I of the HapMap project produced between 18 and 22 million SNP pairs in samples from four populations: Yoruba from Ibadan (YRI), Nigeria; Japanese from Tokyo (JPT); Han Chinese from Beijing (CHB); and residents from Utah with ancestry from northern and western Europe (CEU). For CEU, JPT, and CHB, the estimate of effective population size, adjusted for SNP ascertainment bias, was ~3000, whereas the estimate for the YRI was ~7500, consistent with the out-of-Africa theory of ancestral human population expansion and concurrent bottlenecks. We show that the decay in LD over distance between SNPs is consistent with recent population growth. The estimates of $N_e$ are lower than previously published estimates based on heterozygosity, possibly because they represent one or more bottlenecks in human population size that occurred ~10,000 to 200,000 years ago.

“Mitochondrial Eve” lived in Africa between 100,000 and 150,000 years ago (or 240,000?)

“Adam” and “Eve” are both out of Africa

- "Mitochondrial Eve" lived in Africa between 100,000 and 150,000 years ago (or 240,000?)
- "Y-chromosome Adam" also lived in Africa between 120,000 and 160,000 years ago
Mitochondrial Eve (maternally transmitted ancestry) Y-chromosome Adam (paternally transmitted ancestry) lived ~200,000 years ago.

When lived the latest common ancestor shared by all of us based on nuclear DNA?

A. 1 million years ago
B. 200,000 years ago
C. 3400 years ago
D. 660 years ago
E. Yesterday, I really have no clue

Get your i-clickers
Last common ancestor in nuclear (non Y-chr) DNA is another matter

• Nuclear DNA gets mixed with every generation
  – Each of us gets 50% of nuclear DNA from father & 50% from mother
  – Each has 2 parents, 4 grandparents, 8 great-grand parents ....

• If one assumes:
  – Well-mixed marriages (not true: mostly local until recently)
  – Constant size population (not true: much smaller)
  – In 33 generations the number of ancestors:
    \[2^{33} = 8 \text{ billion} > 7 \text{ billion people living today}\]

• Every pair of us living today should have at least one shared ancestor who lived
  – 33 generations * 20 years/generation=660 years ago ~1300 AD
Corrected for mostly local marriages

Modeling the recent common ancestry of all living humans

Douglas L. T. Rohde¹, Steve Olson² & Joseph T. Chang³

¹Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
²7609 Seaborg Road, Bethesda, Maryland 20817, USA
³Department of Statistics, Yale University, New Haven, Connecticut 06520, USA

With 5% of individuals migrating out of their home town, 0.05% migrating out of their home country, and 95% of port users born in the country from which the port emanates, the simulations produce a mean MRCA date of 1,415 BC and a mean IA date of 5,353 BC.

Figure 2: Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.
Mitochondrial Eve (maternally transmitted ancestry) lived ~200,000 years ago.

Y-chromosome Adam (paternally transmitted ancestry) lived ~200,000 years ago.

When lived the latest common ancestor shared by all of us based on nuclear DNA?

A. 1 million years ago
B. 200,000 years ago
C. 3400 years ago
D. 660 years ago
E. Yesterday, I really have no clue

Get your i-clickers
Last Universal Common Ancestor (LUCA)

Archaea were discovered here at UIUC in 1977 by Carl R. Woese (1928-2012) and George E. Fox

LUCA: 3.5-3.8 Billion years old
Earth is 4.5 Byr old
Negative Binomial Definition

• In a series of independent trials with constant probability of success, $p$, let the random variable $X$ denote the number of trials until $r$ successes occur. Then $X$ is a negative binomial random variable with parameters: $0 < p < 1$ and $r = 1, 2, 3, \ldots$

• The probability mass function is:

$$f(x) = C_{r-1}^{x-1} p^r (1 - p)^{x-r} \quad \text{for } x = r, r + 1, r + 2 \ldots \quad (3-11)$$

• Compare it to binomial

$$f(x) = C_x^n p^x (1 - p)^{n-x} \quad \text{for } x = 1, 2, \ldots n$$

NOTE OF CAUTION: Matlab, Mathematica, and many other sources use $x$ to denote the number of failures until one gets $r$ successes. We stick with Montgomery-Rung.
Negative Binomial Mean & Variance

• If $X$ is a negative binomial random variable with parameters $p$ and $r$,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2} \quad (3-12)$$

• Compare to geometric distribution:

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \quad (3-10)$$
Cancer is scary!

• It hit my family twice last year
• Approximately 39.6 percent of men and women will be diagnosed with cancer at some point during their lifetimes (source: NCI website)


<table>
<thead>
<tr>
<th>Rank</th>
<th>Cause of death</th>
<th>Number</th>
<th>Percent of all deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>–</td>
<td>All causes</td>
<td>2,468,435</td>
<td>100.0</td>
</tr>
<tr>
<td>1</td>
<td>Diseases of heart</td>
<td>597,689</td>
<td>24.2</td>
</tr>
<tr>
<td>2</td>
<td>Malignant neoplasms</td>
<td>574,743</td>
<td>23.3</td>
</tr>
<tr>
<td>3</td>
<td>Chronic lower respiratory diseases</td>
<td>138,080</td>
<td>5.6</td>
</tr>
<tr>
<td>4</td>
<td>Cerebrovascular diseases</td>
<td>129,476</td>
<td>5.2</td>
</tr>
<tr>
<td>5</td>
<td>Accidents (unintentional injuries)</td>
<td>120,859</td>
<td>4.9</td>
</tr>
<tr>
<td>6</td>
<td>Alzheimer’s disease</td>
<td>83,494</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes mellitus</td>
<td>69,071</td>
<td>2.8</td>
</tr>
<tr>
<td>8</td>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
<td>50,476</td>
<td>2.0</td>
</tr>
<tr>
<td>9</td>
<td>Influenza and pneumonia</td>
<td>50,097</td>
<td>2.0</td>
</tr>
<tr>
<td>10</td>
<td>Intentional self-harm (suicide)</td>
<td>38,364</td>
<td>1.6</td>
</tr>
</tbody>
</table>


• “War on Cancer” – president Nixon 1971.
  “Moonshot to Cure Cancer” – vice-president Joe Biden 2016
“War on Cancer” progress report

**Figure 2**

*Cancer Death Rates* by Sex, U.S., 1975–2005

**Figure 3**

*Tobacco Use in the U.S., 1900–2005*

*Age-adjusted to the U.S. 2000 standard population.


Statistics of Cancer

- Bert Vogelstein et al: Cancer is caused by accumulation of “driver” gene mutations
  - Oncogenes: ↑
  - Tumor suppressors: ↓ (may need 2 mutations)
  - 7 strikes and you are out

Cancer death rate ~ (patient age)^6
Ongoing discussion: how many strikes?

Only three driver gene mutations are required for the development of lung and colorectal cancers

Cristian Tomasetti\textsuperscript{a,b,1}, Luigi Marchionni\textsuperscript{a}, Martin A. Nowak\textsuperscript{d}, Giovanni Parmigiani\textsuperscript{a}, and Bert Vogelstein\textsuperscript{a,9,1}

\textsuperscript{a}Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and \textsuperscript{b}Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; \textsuperscript{c}Cancer Biology Program, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; \textsuperscript{d}Program for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; \textsuperscript{e}Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215; and \textsuperscript{f}Ludwig Center for Cancer Genetics and Therapeutics and \textsuperscript{g}Howard Hughes Medical Institute, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Contributed by Bert Vogelstein, November 21, 2014 (sent for review July 31, 2014; reviewed by Zvia Agur)

\[ P(T_{\text{cancer}} \leq t) \sim (u_1t)(u_2t)\ldots(u_nt) = u_1u_2\ldots u_nt^n \]

\[ P(T_{\text{cancer}} = t) \sim (u_1t)(u_2t)\ldots(u_nt) = u_1u_2\ldots u_{n-1}t^{n-1} \]

Smokers have 3.23 times more mutations

![Graph showing comparison of LUAD incidence fold increase in smokers and never-smokers](image)
“Drivers” carry “Passengers”

“Passenger” mutations cause little to no harm

“Passengers” are common as cancers elevate mutation rate
Passenger mutations: negative binomial distribution

Tug-of-war between driver and passenger mutations in cancer and other adaptive processes

Christopher D. McFarland, Leonid A. Mirny, and Kirill S. Korolev

What is the probability to have \( n_p \) passenger mutations or \((n_p+k)\) total mutations by the time you are diagnosed with cancer requiring \( k \) driver mutations?

Let \( p \) is the probability that a mutation is a driver \((1-p)\) – it is a passenger

\[
P(n_p + k \mid p, k) = \binom{n_p + k - 1}{n_p} (1 - p)^{n_p} p^k
\]
McFarland CD, Mirny L, Korolev KS, PNAS 2014
Matlab exercise

• Find mean, variance, and histogram of 100,000 geometrically-distributed numbers with p=0.1

• Hint: Use help page for `random` command on how to generate geometrically-distributed random numbers
Matlab: Geometric distributions

• Stats=100000;
• p=0.1;
• r2=random('Geometric',p,Stats,1);
• r2=r2+1;
• disp(mean(r2));
• disp(var(r2));
• disp(std(r2));
• [a,b]=hist(r2, 1:max(r2));
• p_g=a./sum(a);
• figure; semilogy(b,p_g,'ko-');
Matlab: Negative binomial distributions

- Stats=100000;
- r=3; p=0.1;
- r2=random('Negative Binomial',r,p,Stats,1);
- r2=r2+r;
- disp(mean(r2));
- disp(var(r2));
- disp(std(r2));
- [a,b]=hist(r2, 1:max(r2));
- p_nb=a./sum(a);
- figure; semilogy(b,p_nb,'ko-');