A gallery of useful discrete probability distributions
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 1st 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 $p$,

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

• For small $p$ the standard deviation $=(1-p)^{0.5}/p \sim = \text{mean}=1/p$

• Very different from Binomial and Poisson, where variance $= \text{mean}$ and standard deviation $= \text{mean}^{1/2}$
Geometric distribution in biology

- 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 that time most mitochondrial genes were transferred into 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) offsprings. 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.
- \( P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N) \exp(-t+1)/N \)
- \( t \) 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:

$$E(T) = \sum_{t=1}^{\infty} t \cdot \left(\frac{1}{N}\right) \exp\left(-\frac{t}{N}\right) = N$$

• Most Recent maternal Common Ancestor (MRCA) of all people living today lived $T_{\text{MRCA}} = 2N$ generations ago
• $T_{\text{MRCA}} = 2 \cdot 3.5\times 10^9$ generations
• If the generation time 20 years it is 140 billion years $> 10$ times the time since the Big Bang.
• Something is wrong here!
Effective size of the ancestral human population

• 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 \( N_e \sim 7500 \) people

• For humans of **European + Asian ancestry** \( N_e \sim 3100 \) people

• **Mito Eve lived** ~

\[ 2 \times \left( \frac{N_e}{2} \right) \times 20 \text{ years} = 7500 \times 20 \text{ years} = 150,000 \text{ years ago} \]

---

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. Visscher 3,5,8

1 Colton 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 3JT, 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 (HCB); and residents from Utah with ancestry from northern and western Europe (CEU). For CEU, JPT, and HCB, 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) and 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
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

• Unlike Mito or Y-chromosome, **nuclear DNA gets mixed with every generation**
  – Each of us gets 50% of nuclear DNA from the father & 50% from the mother
  – Each of us has 2 parents, 4 grandparents, 8 great-grand parents ...

• If one assumes:
  – Well-mixed marriages (not true: mostly local marriages until recently)
  – Constant size population (not true: much smaller in the past)
  – 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 and rare migrations

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.
Archaea were discovered here at UIUC in 1977 by Carl R. Woese (1928-2012) and George E. Fox.
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
  \[
f(x) = C^n_x p^x (1 - p)^{n-x} \quad \text{for } x = 1, 2, \ldots n
  \]

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-Runger.
Negative Binomial Mean & Variance

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

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

(3-12)

• Compare to geometric distribution:

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

(3-10)
Matlab exercise

• Estimate mean, variance, and PMF based on 100,000 random variables drawn from a negative binomial distribution with $p=0.1, r=3$.

• Repeat with negative binomial distribution with $p=0.1, r=100$. 
Matlab: Negative binomomial distribution

- Stats=100000;
- r=3; p=0.1;
- r2=zeros(Stats,1);
- for k=1:Stats
  - n_trials=0;
  - n_successes=0;
  - while n_successes<r
    - if rand<p
      - n_successes=n_successes+1;
    - end;
    - n_trials=n_trials+1;
  - end;
  - r2(k)=n_trials;
- end;
- disp('Observed average value'); disp(sum(r2)./Stats);
- disp('Expected average value'); disp(r./p);
- disp('Observed variance'); disp(sum(r2.^2)./Stats-(sum(r2)./Stats).^2);
- disp('Expected variance'); disp(r.*(1-p)./p^2);
- [a,b]=hist(r2, 1:max(r2));
- p_nb=a./sum(a);
- figure; semilogy(b,p_nb,'ko-');
Cancer is scary!

- Approximately 40% 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>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>

Table from J. Pevsner 3rd edition

  “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.
Cigarette consumptions: U.S. Department of Agriculture, 1900–2007
Probability theory and statistics is a powerful tool to learn new cancer biology.
“Driver genes” theory

- Progression of cancer is caused by accumulation of mutations in a handful of “driver” genes
- Mutations in driver genes boost the growth of a tumor
- Oncogenes: expression needs to be elevated for cancer
- Tumor suppressors (e.g. p53) need to be turned off in cancer

Douglas Hanahan and Robert A. Weinberg
Hallmarks of Cancer: The Next Generation
Cell 144, 2011
Statistics of cancer incidence vs age

Multi-mutation theory of cancer:
Carl O. Nordling (British J. of Cancer, March 1953):

Cancer death rate
\( \sim (\text{patient age})^6 \)

It suggests the existence of
k=7 driver genes

\[
P(T_{\text{cancer}} \leq t) \sim (u_1 t)(u_2 t)\ldots(u_k t) \sim u_1 u_2 \ldots u_k t^k
\]

\[
P(T_{\text{cancer}} = t) \sim \frac{d}{dt} (u_1 t)(u_2 t)\ldots(u_k t) \sim k u_1 u_2 \ldots u_k t^{k-1}
\]

Fig. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.
How many driver gene mutations for different types of cancer?

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

Smokers have 3.23 times more mutations in lungs
Cancer cells carry both “Driver” and “Passengers” mutations

Passenger mutations cause little to no harm (see later for how even little harm matters)

Both are common as cancers elevate mutation rate
Number of passenger+driver mutations follows negative binomial distribution

- 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 ($p = \frac{\text{Genome_target_of_driv}}{\text{Genome_target_of_driv}+\text{Genome_target_of_pass}}$) $(1-p)$ – it is a passenger mutation

$$P(n_p + k \mid p, k) = \binom{n_p + k - 1}{n_p} (1 - p)^{n_p} p^k$$
What if passenger mutations slow down the growth of cancer tumors?

McFarland CD, Mirny L, Korolev KS, PNAS 2014
Can we prove/quantify it using statistics?

Assume: growth rate of cancer \(= \frac{(1+s_d)^{Nd}}{(1+s_p)^{Np}} \)

\( \mu = 10^{-8}, \text{Target}_d = 1,400, \text{Target}_p = 10^7, s_d = 0.05 \) to \( 0.4, \) \( s_p = 0.001 \)

\( s_p/s_d \) for breast: \( 0.0060 \pm 0.0010; \)
melanoma: \( 0.016 \pm 0.003; \) lung: \( 0.0094 \pm 0.0093; \)

Blue - data on breast cancer: incidence; non-synonymous mutations