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:

\[
x-1 = 0, 1, 2, \ldots, \text{the number of failures until the 1}\text{st success}.
\]

• 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.
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:

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

• Compare it to binomial

\[ f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x = 1, 2, \ldots n \]

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 \( r \),

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

• Compare to geometric distribution:

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


  “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

- Men
- Both Sexes
- Women

Per capita cigarette consumption

Figure 3

Tobacco Use in the U.S., 1900–2005

- Male lung cancer death rate
- Female lung cancer death rate
- Age-Adjusted Lung Cancer Death Rates*

Sources:
- Cigarette consumptions: U.S. Department of Agriculture, 1900–2007

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

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

It suggests the existence of $k=7$ driver genes

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

\[
P(T_{\text{cancer}} = t) \sim \frac{d}{dt} (u_1 t)(u_2 t) \cdots (u_k t) \sim k u_1 u_2 \cdots u_k t^{k-1}
\]
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 ($p = \frac{\text{Genome}\_\text{target}\_\text{of}\_\text{driv}}{\text{Genome}\_\text{target}\_\text{of}\_\text{driv}+\text{Genome}\_\text{target}\_\text{of}\_\text{pass}}$) $(1-p)$ – 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?
Can we prove/quantify it using statistics?

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

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

$s_p/s_d$ for breast: $0.0060\pm0.0010$;
melanoma: $0.016\pm0.003$; lung: $0.0094\pm0.0093$;
Blue - data on breast cancer: incidence; non-synonymous mutations
Matlab exercise

• Find mean, variance, and PMF based on 100,000 geometrically-distributed numbers with $p=0.1$

• Repeat with negative binomial distribution with $p=0.1$, $r=3$

• Repeat with negative binomial distribution with $p=0.1$, $r=100$

• Hint: Use help page for \texttt{random} Matlab command on how to generate random numbers with different PMFs
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-');
Matlab: Negative binomial distributions

- Stats=100000;
- r=100; 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-');