A gallery of useful
discrete probability distributions
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:

\[ 0 < p < 1 \text{ and } r = 1, 2, 3, \ldots \]

• The probability mass function is:

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

• Compare it to binomial

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

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 $r$,

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

- 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>1</td>
<td>All causes</td>
<td>2,468,435</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>Diseases of heart</td>
<td>597,689</td>
<td>24.2</td>
</tr>
<tr>
<td>3</td>
<td>Malignant neoplasms</td>
<td>574,743</td>
<td>23.3</td>
</tr>
<tr>
<td>4</td>
<td>Chronic lower respiratory diseases</td>
<td>138,080</td>
<td>5.6</td>
</tr>
<tr>
<td>5</td>
<td>Cerebrovascular diseases</td>
<td>129,476</td>
<td>5.2</td>
</tr>
<tr>
<td>6</td>
<td>Accidents (unintentional injuries)</td>
<td>120,859</td>
<td>4.9</td>
</tr>
<tr>
<td>7</td>
<td>Alzheimer’s disease</td>
<td>83,494</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>Diabetes mellitus</td>
<td>69,071</td>
<td>2.8</td>
</tr>
<tr>
<td>9</td>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
<td>50,476</td>
<td>2.0</td>
</tr>
<tr>
<td>10</td>
<td>Influenza and pneumonia</td>
<td>50,097</td>
<td>2.0</td>
</tr>
<tr>
<td></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
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 $\sim (\text{patient age})^6$
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{i}, Martin A. Nowak\textsuperscript{d}, Giovanni Parmigiani\textsuperscript{n}, and Bert Vogelstein\textsuperscript{j,g,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{i}Ludwig Center for Cancer Genetics and Therapeutics and \textsuperscript{j}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)

\[ P(T_{\text{cancer}} \leq t) \sim (u_1 t)(u_2 t) \ldots (u_n t) = u_1 u_2 \ldots u_n t^n \]

\[ P(T_{\text{cancer}} = t) \sim (u_1 t)(u_2 t) \ldots (u_n t) = u_1 u_2 \ldots u_n t^{n-1} \]

Smokers have 3.23 times more mutations
- "Drivers" carry "Passengers"
- "Passenger" mutations cause little to no harm
- "Passengers" are common as cancers elevate mutation rate

**FIGURE 21.10** Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein et al. (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizontal bars indicate the 25% and 75% quartiles. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC: esophageal adenocarcinomas.

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*a,b,c,1, and Kirill S. Korolev*b,d,1

*Graduate Program in Biophysics, Harvard University, Boston, MA 02115; †Department of Physics and Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and ‡Department of Physics and Program in Bioinformatics, Boston University, Boston, MA 02215

Edited* by Herbert Levine, Rice University, Houston, TX, and approved August 7, 2014 (received for review March 7, 2014)

• 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

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

• Find mean, variance, and PMF based on 100,000 geometrically-distributed numbers with $p=0.1$
• Repeat with negative binomial distribution with $p=0.1$, $r=3$
• Hint: Use help page for `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-');