application: error correcting codes
Codes are all around us
Goal: send a 4-bit message over a noisy communication channel.  
Say, 1 bit in 10 is flipped in transit, independently.  
What is the probability that the message arrives correctly?  

Let \( X = \) # of errors; \( X \sim \text{Bin}(4, 0.1) \)  
\[
P(\text{correct message received}) = P(X=0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561
\]

Can we do better? Yes: error correction via redundancy.  
E.g., send every bit in triplicate; use majority vote.  

Let \( Y = \) # of errors in one trio; \( Y \sim \text{Bin}(3, 0.1) \); \( P(\text{a trio is OK}) = \)  
\[
P(Y \leq 1) = \binom{3}{0}(0.1)^0(0.9)^3 + \binom{3}{1}(0.1)^1(0.9)^2 = 0.972
\]

If \( X' = \) # errors in triplicate msg, \( X' \sim \text{Bin}(4, 0.028) \), and  
\[
P(X' = 0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168
\]
The Hamming(7,4) code:
Have a 4-bit string to send over the network (or to disk)
Add 3 “parity” bits, and send 7 bits total
If bits are $b_1b_2b_3b_4$ then the three parity bits are
\[ \text{parity}(b_1b_2b_3), \text{parity}(b_1b_3b_4), \text{parity}(b_2b_3b_4) \]
Each bit is independently corrupted (flipped) in transit with probability 0.1
\[ Z = \text{number of bits corrupted} \sim \text{Bin}(7, 0.1) \]
The Hamming code allow us to correct all 1 bit errors.
(E.g., if $b_1$ flipped, 1st 2 parity bits, but not 3rd, will look wrong; the only single bit error causing this symptom is $b_1$. Similarly for any other single bit being flipped. Some, but not all, multi-bit errors can be detected, but not corrected.)
\[ P(\text{correctable message received}) = P(Z \leq 1) \]
Using Hamming error-correcting codes: $Z \sim \text{Bin}(7, 0.1)$

$$P(Z \leq 1) = \binom{7}{0}(0.1)^0(0.9)^7 + \binom{7}{1}(0.1)^1(0.9)^6 \approx 0.8503$$

Recall, uncorrected success rate is

$$P(X = 0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561$$

And triplicate code error rate is:

$$P(X' = 0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168$$

Hamming code is nearly as reliable as the triplicate code, with $5/12 \approx 42\%$ fewer bits. (& better with longer codes.)
Sending a bit string over the network
\[ n = 4 \text{ bits sent, each corrupted with probability } 0.1 \]
\[ X = \# \text{ of corrupted bits, } X \sim \text{Bin}(4, 0.1) \]
In real networks, large bit strings (length \( n \approx 10^4 \))
Corruption probability is very small: \( p \approx 10^{-6} \)

Extreme \( n \) and \( p \) values arise in many cases
- \# bit errors in file written to disk
- \# of typos in a book
- \# of elements in particular bucket of large hash table
- \# of server crashes per day in giant data center
- \# facebook login requests sent to a particular server
Poisson random variables

Suppose “events” happen, independently, at an average rate of \( \lambda \) per unit time. Let \( X \) be the actual number of events happening in a given time unit. Then \( X \) is a Poisson r.v. with parameter \( \lambda \) (denoted \( X \sim \text{Poi}(\lambda) \)) and has distribution (PMF):

\[
P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}
\]

Examples:
- # of alpha particles emitted by a lump of radium in 1 sec.
- # of traffic accidents in Seattle in one year
- # of babies born in a day at UW Med center
- # of visitors to my web page today

See B&T Section 6.2 for more on theoretical basis for Poisson.
X is a Poisson r.v. with parameter $\lambda$ if it has PMF:

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$$

Is it a valid distribution? Recall Taylor series:

$$e^\lambda = \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \cdots = \sum_{0 \leq i} \frac{\lambda^i}{i!}$$

So

$$\sum_{0 \leq i} P(X = i) = \sum_{0 \leq i} e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{0 \leq i} \frac{\lambda^i}{i!} = e^{-\lambda} e^\lambda = 1$$
expected value of Poisson r.v.s

\[ E[X] = \sum_{0 \leq i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \]

\[ = \sum_{1 \leq i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \]

\[ = \lambda e^{-\lambda} \sum_{1 \leq i} \frac{\lambda^{i-1}}{(i - 1)!} \]

\[ = \lambda e^{-\lambda} \sum_{0 \leq j} \frac{\lambda^j}{j!} \]

\[ = \lambda e^{-\lambda} e^\lambda \]

\[ = \lambda \]

As expected, given definition in terms of “average rate \( \lambda \)”

\( \text{(Var}[X] = \lambda, \text{too; proof similar, see B&T example 6.20)} \)
binomial random variable is Poisson in the limit

Poisson approximates binomial when $n$ is large, $p$ is small, and $\lambda = np$ is “moderate”

Formally, Binomial is Poisson in the limit as $n \to \infty$ (equivalently, $p \to 0$) while holding $np = \lambda$
\[ X \sim \text{Binomial}(n, p) \]

\[ P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i} \]

\[ = \frac{n!}{i!(n-i)!} \left( \frac{\lambda}{n} \right)^i \left( 1 - \frac{\lambda}{n} \right)^{n-i} \]

\[ = \frac{n(n-1) \cdots (n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1 - \lambda/n)^n}{(1 - \lambda/n)^i} \]

\[ \approx \frac{\lambda^i}{i!} \cdot e^{-\lambda} \]

I.e., Binomial \( \approx \) Poisson for large \( n \), small \( p \), moderate \( i, \lambda \).
sending data on a network, again

Recall example of sending bit string over a network
Send bit string of length $n = 10^4$
Probability of (independent) bit corruption is $p = 10^{-6}$
$X \sim \text{Poi}(\lambda = 10^4 \cdot 10^{-6} = 0.01)$
What is probability that message arrives uncorrupted?

$$P(X = 0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-0.01} \frac{0.01^0}{0!} \approx 0.990049834$$

Using $Y \sim \text{Bin}(10^4, 10^{-6})$:

$$P(Y=0) \approx 0.990049829$$
binomial vs Poisson

- Binomial(10, 0.3)
- Binomial(100, 0.03)
- Poisson(3)
Recall: if $Y \sim \text{Bin}(n,p)$, then:

$$E[Y] = pn$$

$$\text{Var}[Y] = np(1-p)$$

And if $X \sim \text{Poi}(\lambda)$ where $\lambda = np$ ($n \to \infty$, $p \to 0$) then

$$E[X] = \lambda = np = E[Y]$$

$$\text{Var}[X] = \lambda \approx \lambda(1-\lambda/n) = np(1-p) = \text{Var}[Y]$$

Expectation and variance of Poisson are the same ($\lambda$)

Expectation is the same as corresponding binomial

Variance almost the same as corresponding binomial

Note: when two different distributions share the same mean & variance, it suggests (but doesn’t prove) that one may be a good approximation for the other.
In a series $X_1, X_2, \ldots$ of Bernoulli trials with success probability $p$, let $Y$ be the index of the first success, i.e.,

$$X_1 = X_2 = \ldots = X_{Y-1} = 0 \& X_Y = 1$$

Then $Y$ is a geometric random variable with parameter $p$.

Examples:

- Number of coin flips until first head
- Number of blind guesses on LSAT until I get one right
- Number of darts thrown until you hit a bullseye
- Number of random probes into hash table until empty slot
- Number of wild guesses at a password until you hit it

$$P(Y=k) = (1-p)^{k-1}p; \text{ Mean } 1/p; \text{ Variance } (1-p)/p^2$$
Draw $d$ balls (without replacement) from an urn containing $N$, of which $w$ are white, the rest black.

Let $X =$ number of white balls drawn

$$P(X = i) = \binom{w}{i} \binom{N-w}{d-i} \binom{N}{d}, \quad i = 0, 1, \ldots, d$$

(note: $n \choose k = 0$ if $k < 0$ or $k > n$)

$E[X] = dp$, where $p = w/N$ (the fraction of white balls)

proof: Let $X_j$ be 0/1 indicator for j-th ball is white, $X = \sum X_j$

The $X_j$ are dependent, but $E[X] = E[\sum X_j] = \sum E[X_j] = dp$

$Var[X] = dp(1-p)(1-(d-1)/(N-1))$
$N \approx 22500$ human genes, many of unknown function
Suppose in some experiment, $d = 1588$ of them were observed (say, they were all switched on in response to some drug)

A big question: What are they doing?

One idea: The Gene Ontology Consortium (www.geneontology.org) has grouped genes with known functions into categories such as “muscle development” or “immune system.” Suppose 26 of your $d$ genes fall in the “muscle development” category.

Just chance?

Or call Coach & see if he wants to dope some athletes?

Hypergeometric: GO has 116 genes in the muscle development category. If those are the white balls among 22500 in an urn, what is the probability that you would see 26 of them in 1588 draws?
A differentially bound peak was associated to the closest gene (unique Entrez ID) measured by distance to TSS within CTCF flanking domains. OR: ratio of predicted to observed number of genes within a given GO category. Count: number of genes with differentially bound peaks. Size: total number of genes for a given functional group. Ont: the Geneontology. BP = biological process, MF = molecular function, CC = cellular component.

<table>
<thead>
<tr>
<th>GOID</th>
<th>Term</th>
<th>P Value</th>
<th>OR</th>
<th>Count</th>
<th>Size</th>
<th>Ont</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005856</td>
<td>cytoskeleton</td>
<td>2.05E-11</td>
<td>2.40</td>
<td>94</td>
<td>490</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0043292</td>
<td>contractile fiber</td>
<td>6.98E-09</td>
<td>5.85</td>
<td>22</td>
<td>58</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0030016</td>
<td>myofibril</td>
<td>1.96E-08</td>
<td>5.74</td>
<td>21</td>
<td>56</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0044449</td>
<td>contractile fiber part</td>
<td>2.58E-08</td>
<td>5.97</td>
<td>20</td>
<td>52</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0030017</td>
<td>sarcomere</td>
<td>4.95E-08</td>
<td>6.04</td>
<td>19</td>
<td>49</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0008092</td>
<td>skeletal muscle development</td>
<td>2.50E-08</td>
<td>4.13</td>
<td>21</td>
<td>65</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0015629</td>
<td>actin cytoskeleton</td>
<td>4.73E-06</td>
<td>3.08</td>
<td>27</td>
<td>111</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0003779</td>
<td>actin binding</td>
<td>7.99E-07</td>
<td>3.58</td>
<td>159</td>
<td></td>
<td>MF</td>
</tr>
<tr>
<td>GO:0006936</td>
<td>collagen fibril</td>
<td>9.14E-06</td>
<td>1.84</td>
<td>29</td>
<td>52</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0044430</td>
<td>cytoskeleton protein</td>
<td>5.07E-05</td>
<td>2.69</td>
<td>26</td>
<td>116</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0031674</td>
<td>I band</td>
<td>2.27E-05</td>
<td>5.67</td>
<td>12</td>
<td>32</td>
<td>CC</td>
</tr>
<tr>
<td>GO:003012</td>
<td>muscle system process</td>
<td>2.54E-05</td>
<td>4.11</td>
<td>16</td>
<td>52</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0030029</td>
<td>actin filament-based process</td>
<td>2.89E-05</td>
<td>2.73</td>
<td>27</td>
<td>119</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0007517</td>
<td>muscle development</td>
<td>5.06E-05</td>
<td>2.69</td>
<td>26</td>
<td>116</td>
<td>BP</td>
</tr>
</tbody>
</table>

The probability of seeing this many genes from a set of this size by chance according to the hypergeometric distribution is approximately \(2.05 \times 10^{-11}\).
Supreme Court case: Berghuis v. Smith

If a group is underrepresented in a jury pool, how do you tell?

Justice Breyer [Stanford Alum] opened the questioning by invoking the binomial theorem. He hypothesized a scenario involving “an urn with a thousand balls, and sixty are red, and nine hundred forty are black, and then you select them at random... twelve at a time.” According to Justice Breyer and the binomial theorem, if the red balls were black jurors then “you would expect... something like a third to a half of juries would have at least one black person” on them.

- Justice Scalia’s rejoinder: “We don’t have any urns here.”
• Should model this combinatorially
  ▪ Ball draws not independent trials (balls not replaced)
• Exact solution:
  \[ P(\text{draw 12 black balls}) = \frac{\binom{940}{12}}{\binom{1000}{12}} \approx 0.4739 \]
  \[ P(\text{draw } \geq 1 \text{ red ball}) = 1 - P(\text{draw 12 black balls}) \approx 0.5261 \]
• Approximation using Binomial distribution
  ▪ Assume \( P(\text{red ball}) \) constant for every draw = 60/1000
  ▪ \( X = \# \text{ red balls drawn} \). \( X \sim \text{Bin}(12, 60/1000 = 0.06) \)
  ▪ \( P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0.4759 = 0.5240 \)

*In Breyer’s description, should actually expect just over half of juries to have at least one black person on them*
Often care about 2 (or more) random variables simultaneously measured $X = \text{height}$ and $Y = \text{weight}$
$X = \text{cholesterol}$ and $Y = \text{blood pressure}$
$X_1, X_2, X_3 = \text{work loads on servers A, B, C}$

**Joint** probability mass function:
$f_{XY}(x, y) = P(X = x \& Y = y)$

**Joint** cumulative distribution function:
$F_{XY}(x, y) = P(X \leq x \& Y \leq y)$
Two joint PMFs

\[
P(W = Z) = 3 \times \frac{2}{24} = \frac{6}{24}
\]

\[
P(X = Y) = \frac{(4 + 3 + 2)}{24} = \frac{9}{24}
\]

Can look at arbitrary relationships between variables this way
Two joint PMFs

**Marginal distribution of one r.v.:** sum over the other:

\[ f_Y(y) = \sum_x f_{XY}(x,y) \]
\[ f_X(x) = \sum_y f_{XY}(x,y) \]

**Question:** Are \( W \) & \( Z \) independent? Are \( X \) & \( Y \) independent?

<table>
<thead>
<tr>
<th>( W )</th>
<th>( Z )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>( f_W(w) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2/24</td>
<td>2/24</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2/24</td>
<td>2/24</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>2/24</td>
<td>2/24</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>2/24</td>
<td>2/24</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>( f_Z(z) )</td>
<td>8/24</td>
<td>8/24</td>
<td>8/24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>( f_X(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>4/24</td>
<td>1/24</td>
<td>1/24</td>
<td>6/24</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>3/24</td>
<td>3/24</td>
<td>6/24</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td>4/24</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4/24</td>
<td>0</td>
<td>2/24</td>
<td>6/24</td>
</tr>
<tr>
<td>( f_Y(y) )</td>
<td>8/24</td>
<td>8/24</td>
<td>8/24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
sampling from a (continuous) joint distribution

- Top row: independent variables
- Bottom row: dependent variables

\[ \text{var}(x) = 1, \text{var}(y) = 1, \text{cov} = 0, n=1000 \]

\[ \text{var}(x) = 1, \text{var}(y) = 3, \text{cov} = 0, n=1000 \]

\[ \text{var}(x) = 1, \text{var}(y) = 3, \text{cov} = 0, n=100 \]

\[ \text{var}(x) = 1, \text{var}(y) = 3, \text{cov} = 0.8, n=1000 \]

\[ \text{var}(x) = 1, \text{var}(y) = 3, \text{cov} = 1.5, n=1000 \]

\[ \text{var}(x) = 1, \text{var}(y) = 3, \text{cov} = 1.7, n=1000 \]
A function \( g(X,Y) \) defines a new random variable.

Its expectation is:

\[
E[g(X, Y)] = \sum_x \sum_y g(x, y) f_{XY}(x,y)
\]

Expectation is linear. I.e., if \( g \) is linear:

\[
E[g(X, Y)] = E[a X + b Y + c] = a E[X] + b E[Y] + c
\]

Example:

\( g(X,Y) = 2X-Y \)

\[
E[g(X,Y)] = 72/24 = 3
\]

\( E[g(X,Y)] = 2 \cdot 2.5 - 2 = 3 \)
**random variables – summary**

**RV**: a numeric function of the outcome of an experiment

*Probability Mass Function* \( p(x) \): prob that RV = x; \( \sum p(x) = 1 \)

*Cumulative Distribution Function* \( F(x) \): probability that RV \( \leq x \)

Concepts generalize to *joint* distributions

**Expectation**:

- of a random variable: \( E[X] = \sum x \cdot p(x) \)
- of a function: if \( Y = g(X) \), then \( E[Y] = \sum g(x) \cdot p(x) \)

**Linearity**:

\[
E[aX + b] = aE[X] + b
\]

\[
E[X+Y] = E[X] + E[Y]; \text{ even if dependent}
\]

*This interchange of “order of operations” is quite special to linear combinations. E.g. } E[XY] \neq E[X] \cdot E[Y], \text{ in general (but see below)}*
**random variables – summary**

**Variance:**
\[ \text{Var}[X] = E[(X-E[X])^2] = E[X^2] - (E[X])^2 \]

**Standard deviation:** \[ \sigma = \sqrt{\text{Var}[X]} \]

\[ \text{Var}[aX+b] = a^2 \text{Var}[X] \]

If \( X \) & \( Y \) are **independent**, then
\[ E[X \cdot Y] = E[X] \cdot E[Y]; \]
\[ \text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] \]

(These two equalities hold for **indp** rv’s; but not in general.)