Definitions:

Sample space: set of all possible outcomes
Event: subset of sample space
Relative frequency interpretation:
long run fraction of success is P(A).
Probability measure: function assigning [0.1] to an event.
Needs to satisfy the following axioms:
1. $P(A) \geq 0$ for all A.
2. $P(S) = 1$
3. If events A, B, C… are disjoint, then P(AUBUC…) = P(A)+P(B)+P(C)…

Fundamental counting principle:
The total number of ways of performing a sequence of k actions is the product of the # of ways to do each action.
Choose = n!/k!/(n-k)!
Inclusion-exclusion proof:
$P(AUB) = P(A,B^C)+P(A^C,B)+P(A,B)$
$P(A,B^C) = P(A) – P(A,B)$ // $P(B,A^C) = P(B) – P(A,B)$
Independence:
P(A,B)=P(A)*P(B) and P(B|A)=P(B)
Conditionality:
P(B|A)=P(A,B)/P(A)
Random Variable:
Numerical feature of a random outcome: function mapping numbers to outcomes.
Distribution/PMF:
$f_X(x) = P\{X=x\}$: List of all possible values and their probabilities
Binomial distribution X~B(n,p):
n independent "trials" of an "experiment"
Each trial is 0 or 1 (success/failure)
P is the probability of success
X is the number of successes among n trials
$P\{X=k\} = (n \text{ choose } k)*p^k(1-p)^{n-k}$ for k in [0,n]
Law of Total Probability:
If S can be partitioned into $A_1, A_2, \dots A_k$

$$P(B) = \sum_{j=1}^{k} P(A_j B) = \sum_{j=1}^{k} P(A_j)P(B \mid A_j)$$

Used to convert T to t in prob calculations:

$$P\{T > U\} = \int P\{T > U | T = t\} f_T(t) dt$$

Bayes' Rule:

$$P(A_i \mid B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)P(B \mid A_j)}$$

Posterior ∝ Prior x Likelihood – P(a|x) ∝ P(a)*P(x|a)
Model: collection of prob measures $\{P_X : x \in X\}$
Parameter: unknown number describing a probability distribution (like theta = proportion of democrats)
Likelihood: Probability of observed data as a function of the unknown parameter, given by the model.
If X ~ B(21,theta)

$$p(X = 18 \mid \theta) = \binom{21}{18} \theta^{18}(1-\theta)^3 \propto \theta^{18}(1-\theta)^3.$$

MLE: highest point, differentiate twice.
Consider ln(L(A)) or other monotonically increasing fxns

Distributions

PDF: nonnegative function.
$P\{a \leq X \leq b\} = \text{integrate}(f_X,a,b)$
CDF: $F(x) = P\{X \leq x\} = \text{integrate}(f_X,-\text{inf},x)$
FTC: $f_X = F'(x)$
Uniform: X ~ U(a,b)
$f_X = 1/(b-a)$ for a<x<b, 0 everywhere else.
$F_X = (x-a)/(b-a)$ for a<x<b, 0 before, 1 after.
Exponential: X ~ Exp(λ)
$f(t) = \lambda e^{-\lambda t}$ for t > 0.
$F(t) = 1 - e^{-\lambda t}$ for t > 0.
Memoryless:

$$P\{T > c + t \mid T > c\} = \frac{P(\{T > c + t\} \cap \{T > c\})}{P\{T > c\}}$$

$$= \frac{P\{T > c + t\}}{P\{T > c\}} = \frac{e^{-\lambda(c+t)}}{e^{-\lambda c}} = e^{-\lambda t} = P\{T > t\}$$

If Y=aX+b, then
$F_Y = P\{Y \leq y\} = P\{aX+b \leq y\}$
$F_Y = P\{X \leq (y-b)/a\} = F_X[y/a-b/a]$
$f_X = F'_X[y/a-b/a] = 1/a * f_X[y/a-b/a]$
$f_Y = 1/|a| * f_X[(y-b)/a]$
Cauchy: $f(x) = 1/(pi(1+x^2))$
Fat tails, no defined mean/variance
Laser Pole:
X = S*tan(A)+B, and A~U(-pi/2,pi/2)
f(x|a)=dcauchy(x-a)

$$F_{X|S,\Theta}(X = x|S = s, \Theta = \theta) = P\{s * \tan(A) + \theta \leq x\}$$

$$F_{X|S,\Theta}(x|s,\theta) = P\left\{A \leq \tan^{-1}\left(\frac{x-\theta}{s}\right)\right\} = F_A\left(\tan^{-1}\left(\frac{x-\theta}{s}\right)\right)$$

Substitute into the CDF: $F_A(\alpha) = \frac{\alpha}{\pi} + \frac{1}{2}$

$$F_{X|S,\Theta}(x|s,\theta) = \frac{1}{\pi}\left(\tan^{-1}\left(\frac{x-\theta}{s}\right)\right) + \frac{1}{2}$$

Joint distribution: pmf/pdf of a random vector, 2D surface or higher dimension.
$f(x,y) \geq 0$ for all (x,y)

$$f_{X,Y}(x,y) = P\{(X,Y) = (x,y)\} = P\{X = x, Y = y\}$$

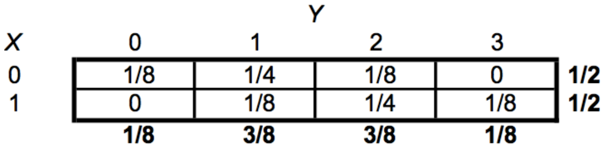
$$P\{(X,Y) \in A\} = \iint_{(x,y) \in A} f(x,y)\, dx\, dy$$

$$P\{(X \leq a, b \leq Y \leq c\} = \int_{y=b}^{c} \int_{x=-\infty}^{a} f(x,y)\, dx\, dy$$

Marginal distribution:
Distribution of a single random variable from a joint distribution.

$$f_X(x) = P\{X = x\} = \sum_y P\{X = x,\, Y = y\} = \sum_y f_{X,Y}(x,y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \qquad f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$



	Y				
X	0	1	2	3	
0	1/8	1/4	1/8	0	**1/2**
1	0	1/8	1/4	1/8	**1/2**
	1/8	**3/8**	**3/8**	**1/8**	

Expectation:

LLN: mean/E[X] is limiting long-run average of iid r.v.'s

$$E(X) = \sum_x x f_X(x) \qquad E(X) = \int x f_X(x)\,dx$$

LOTUS: If Y = g(X),

$$E(Y) = E\left(g(X)\right) = \int g(x) f_X(x)\;dx$$

$$E\left(g(X)\right) = \sum_x g(x) P\{X = x\}$$

Example:

Define $Y = g(U) = U^2$ where $U \sim U(0,1)$

$$F_Y(y) = P\left\{U^2 \le y\right\} = P\left\{U \le y^{1/2}\right\} = y^{1/2} \text{ for } 0 < y < 1$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{2}y^{-1/2}$$

$$E(Y) = \int_0^1 y f_Y(y)\;dy = \int_0^1 y\frac{1}{2}y^{-1/2}\;dy$$

$$E(Y) = \int u^2 f(u)\;du = \int_0^1 u^2\;du = \frac{1}{3}$$

Linearity of expectation (proof):

Let $Z = g(X,Y) = X + Y$

$$\begin{aligned} E(Z) &= \sum_{x,y}(x+y)P\{X = x, Y = y\} \\ &= \sum_x x \sum_y P\{X = x, Y = y\} + \sum_y y \sum_x P\{X = x, Y = y\} \\ &= \sum_x x P\{X = x\} + \sum_y y P\{Y = y\} = E(X) + E(Y) \end{aligned}$$

Expectations of products and independence:

$$\begin{aligned} E(Z) &= \sum_{x,y}(xy)\;\underbrace{P\{X = x, Y = y\}}_{P\{X=x\}P\{Y=y\} \text{ if } X,Y \text{ indep}} \\ &= \sum_x x P\{X = x\} \sum_y y P\{Y = y\} \\ &= E(X)E(Y) \end{aligned}$$

Indicator variables:

I(A) = 1 when A occurs and 0 otherwise.

I(A) has distribution {1 w.p. P(A), 0 w.p. 1-P(A)}

E[I(A)] = p, Var(I(A)) = p(1 – p)

$I(A^C) = 1 – I(A)$

I(A,B) = I(A)*I(B)

Indicator trick- write X as a sum of indicators

$X = I_1 + I_2 + I_3 \ldots I_n$ by $I_k = I\{k^{th}$ trial is a success}

E[X] = P(1)+P(2)+P(3)…P(n)

Variance:

$Var(X) = E[X^2] – E[X]^2$

$Var(X) = E[(x – E[X])^2]$

$Var(cX) = c^2\ Var(X)$

Var(X + Y) = Var(X) + Var(Y) + 2cov(X,Y)

Integration by Parts:

$$\int t^2 e^{-3t}\;dt \quad \Rightarrow$$

+	t^2	e^{-3t}
−	$2t$	$(-1/3)e^{-3t}$
+	2	$(1/9)e^{-3t}$
−	0	$(-1/27)e^{-3t}$

$$\int t^2 e^{-3t}\,dt = +t^2(-1/3)e^{-3t} + (-2t)(1/9)e^{-3t} + 2(-1/27)e^{-3t}.$$

Integration Tricks:

$k! = \int_0^\infty x^k e^{-x} dx$

Variance of Uniform:

"$U(a,b) \sim a + U(0,b-a)$" ← explain

$$\text{var}\left(U(a,b)\right) = \text{var}\left(U(0,b-a)\right)$$

$$U(0,b-a) \sim (b-a)U(0,1)$$

$$\text{var}\left(U(0,b-a)\right) = (b-a)^2\,\text{var}\left(U(0,1)\right)$$

$$E(U^2) = \int_0^1 u^2 du = \frac{1}{3}. \quad E(U) = \frac{1}{2}.$$

$$\text{var}(U) = E\left(U^2\right) - (EU)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

$Var(U(a,b))=(b – a)^2/12$

Variance of Binomial:

$$\text{var}(X) = \underbrace{\text{var}(I_1)}_{p(1-p)} + \cdots + \underbrace{\text{var}(I_n)}_{p(1-p)} = np(1-p)$$

Counting Example:

The number of ways event B can occur (only one match) can be counted as follows:

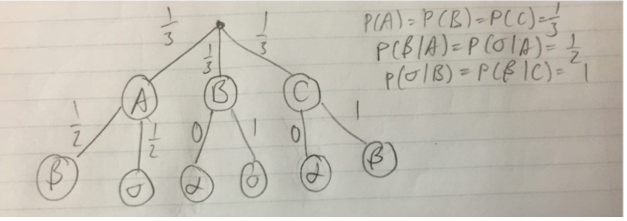
a) Choose the 2 people who will have the same birthday (k choose 2)
b) Choose k-1 birthdays: (365 choose [k-1])
c) Count the number of ways to distribute the k-1 birthdays chosen (k-1)!
d) Divide by the total number of possible birthdays: 365^k

Choosing k white balls and a not-red ball:

1. Choose the k winning balls: 5 choose k
2. Choose the 5-k losing balls: 54 choose 5-k

3a. Choose a losing red ball: 34 choose 1

Monty Hall



Distribution of a max:

$F(v) = P\{V \le v\} = P\{\max\{U_1,U_2,U_3 \le v\} = P\{U_1, v \le U_2, v \le U_3 \le v\}$

Assuming U_1,U_2, and U_3 are all independent, $F(v) = P\{U_1 \le v\}*P\{U_2 \le v\}*P\{U_3 \le v\}$

Since U_i are all drawn from identical uniform distributions, $F(v) = P\{U \le v\}^3$

Note that the cdf of the uniform is $F(u) = P\{U \le u\}$. Thus, $F(v) = F(u)^3$

Distribution of a median:

$$\begin{aligned} F(w) &= P\{W \le w\} = P\{\text{median}\{U_1,U_2,U_3\} \le w\} = \\ &\quad P\{U_1 \le w, U_2 \le w, U_3 \le w\} \\ &+ P\{U_1 > w, U_2 \le w, U_3 \le w\} \\ &+ P\{U_1 \le w, U_2 > w, U_3 \le w\} \\ &+ P\{U_1 \le w, U_2 \le w, U_3 > w\} \end{aligned}$$

Assuming U_1,U_2, and U_3 are independent and interchangeable

$$\begin{aligned} F(w) &= P\{U_1 \le w\}*P\{U_2 \le w\}*P\{U_3 \le w\} \\ &\quad + 3*P\{U_1 > w\}*P\{U_2 \le w\}*P\{U_3 \le w\} \end{aligned}$$

$$\begin{aligned} F(w) &= P\{U \le w\}^3 + 3*P\{U > w\}*P\{U \le w\}^2 \\ &= P\{U \le w\}^3 + 3*(1 - P\{U \le w\})*P\{U \le w\}^2 \end{aligned}$$

By the logic in part (a), $P\{w \le U\} = w$

$F(w) = w^3 + 3*(1 – w)*w^2 = w^3 + 3w^2 – 3w^3$

Likelihood Calculations:

$$L(\lambda) = \prod_{i=1}^{3} f(T_i|\lambda) = f(2.3|\lambda) * f(1.7|\lambda) * \int_{-\infty}^{5.0} f(x|\lambda)\,dx$$

PDF to PDF:

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \\ 0 \end{cases}$$

Covariance:
Cov(X,Y) = E[(X-μ_x)(Y-μ_y)] = E[XY] – E[X] E[Y]
Uncorrelated: Cov(X,Y) = 0. Indep implies uncorrelated
Cov(aX+c, bY+d) = ab*Cov(X,Y) = ab*Cov(Y,X)
FOIL: Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)

Sum of iids:
SD(Xbar) = $\sigma/\sqrt{n}$

Geometric distribution T ~ Geom(p):
T is total trials, p is P(stop condition)
P{T=k} = $p(1-p)^{k-1}$ for k=1,2,3…
E[T] = 1/p

Markov Inequality:
P{Z ≥ c} ≤ E[Z]/c -if Z is a r.v. > 0 and c > 0

$$c\,I\{Z \geq c\} = \begin{cases} c & \text{if } Z \geq c \\ 0 & \text{otherwise} \end{cases}, \text{ so } c\,I\{Z \geq c\} \leq Z$$

Take expected values: cP{Z ≥ c} ≤ E[Z].

Chebychev Inequality:
P{|Y – μ_y| ≥ cσ_y} ≤ 1/c^2 -if c > 0
Prob that Y differs from its mean by > c SDs is 1/c^2.

$$P\left\{\left|Y-\mu_y\right| \geq c\,\sigma_y\right\} = P\left\{\frac{\left(Y-\mu_y\right)^2}{\sigma_y^2} \geq c^2\right\} = P\left\{Z \geq c^2\right\}$$

By Markov's inequality: $P\left\{Z \geq c^2\right\} \leq \frac{E(Z)}{c^2} = \frac{1}{c^2}$

Weak Law of Large Numbers

THEOREM: Let $X_1, X_2, \ldots$ be *iid* with mean μ and variance σ^2, and define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
$\bar{X}_n \to \mu$ "in probability" as $n \to \infty$. That is, for each $\varepsilon > 0$, $P\{|\bar{X}_n - \mu| < \varepsilon\} \to 1$ as $n \to \infty$

Cauchy violation because it has no mean. The average of n Cauchy rv's is distributed like a single observation.
Proof: apply Chebychev's inequality to Y = Xbar_n to get:

$$P\left\{\left|\bar{X}_n - \mu\right| \geq c\frac{\sigma}{\sqrt{n}}\right\} \leq \frac{1}{c^2}, \quad c = \frac{\varepsilon\sqrt{n}}{\sigma}, \quad P\left\{\left|\bar{X}_n - \mu\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{\varepsilon^2 n} \to 0$$

Take complements to give the theorem.

Normal Distribution:

$$N(\mu,\sigma^2) \text{ density: } f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

68, 95, 99.7 rule:
x% of the distribution lies within 1, 2, 3 SDs of the mean

Statistics Terms
Samples: iid rv's drawn from a population distribution
Statistic: computable function of data, e.g. Xbar, s^2

Convolution (Z = X+Y, X ind. Y)

$$f_Z(z) = \sum_x f_X(x)\,f_Y(z-x) \qquad f_Z(z) = \int f_X(x)\,f_Y(z-x)\,dx$$

Sum of independent normally distrib rv's is also normal.

Central Limit Theorem
The sum of many iid rv's is distributed nearly normally.

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \textit{ converges in distribution to } N(0,1) \text{ as } n \to \infty$$

$$\lim_{n\to\infty} P\left\{\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right\} = P\{N(0,1) \leq z\}$$

Normal approximation to binomial:
X ~ Bin(n,p). X is the sum of n (many) iid rv's.
If p is X/n, then p ~ N(p, pq/n)
For conservative estimates, p=0.5

Continuity correction:
P(k or less) = N(μ, σ) ≤ k
P(less than k+1) = N(μ, σ) < k+1.
Actual answer closest to N(μ, σ) ≤ k + 0.5

Markov Chain Definitions:

Markov Chain: sequence of rv's
Sample space: set of all {possible states}
Accessible: j is accessible from i if $P^t(i,j) > 0$ for any t.
i and j **communicate** if they are mutually accessible.
Irreducible: all states communicate.
Distribution: $\pi_{t+1}(j) = P(X_{t+1} = j) = \sum_i \pi_t(i)P(i,j)$
$\pi_{t+1} = \pi_t * P = \pi_0 * P^{t+1}$

Probability transition matrix (P):
$P(i,j) = P\{X_{t+1} = j \mid X_t = i\}$
$P^t(i,j) = P\{X_t = j \mid X_0 = i\}$
Time homogeneity: transition probabilities ind of t.
Markov Property: X_{t+1} depends only on current X_t.
$P\{X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0\} = P\{X_{t+1} = x_{t+1} \mid X_t = x_t\}$

Limiting/Steady-state probabilities:
At large t, all rows are the same, each giving the limiting probability at each state. The stationary distribution π must satisfy $\boldsymbol{\pi} = \boldsymbol{\pi P}$ and sum(π) = 1.
A finite MC has at least 1 stationary distribution.
An irreducible MC has at most 1 stationary distribution.
Every distrib is stationary on P = identity.

Time reversibility (implies stationary):
If distrib μ satisfies $\mu(i)P(i,j) = \mu(j)P(j,i)$ for all i & j, then μ is a stationary distrib for a time-reversible MC. The long-run transitions from i to j are equal to the long-run transitions from j to i.

Ergodic Theorem:
If P is irreducible with a stationary distribution π,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{X_t = j\} = \pi(j) \text{ for all states } j$$

As t increases, a time average converges to an expectation
A sample path is a sample from π

Metropolis Method
Use MC to sample from a distribution without the normalizing constant. Sampling from the posterior allows you to express probabilities as fractions.

Metropolis-Hastings:
Start at state x and density f as the stationary distribution.
Step 1: choose a candidate state y according to any transition matrix Q with $P(propose\ y) = Q(x,y)$
Step 2: $P(accept\ y) = \min\left\{1, \frac{f(y)Q(y,x)}{f(x)Q(x,y)}\right\}$
All in all, $P(x,y) = P(propose\ y) * P(accept\ y)$
- $f(x)P(x,y) = f(y)P(y,x)$ bc its expansion is x,y-symmetric, so f is time-reversible and stationary for P.
-If Q is symmetric (Q(x,y) = Q(y,x), then they cancel).

Conditional expectation:
E[g(Y) | x] = $\sum_y g(y) * P\{y|x\}$
E[Y | X] = $\int y * f_Y(y|x)\, dy = \int y * f_{X,Y}(x,y)/f_X(x)\, dy$
Law of Total Expectation:
E[Y] = $\sum_y E[Y|X=x] * P\{X=x\}$
E[Y] = $\int E[Y \mid X=x] * f_X(x)dx$

Conjugate Priors
Definition: family of distrib. If the prior is a member, the posterior is also a member with different parameters. Simple for paper-pencil calculations.
Beta
$Beta(\alpha,\beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$
With Bern(θ), post is: $Beta(\alpha+S, \beta+n-S)$
with S successes out of n trials:

$$\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]\left[\theta^{S}(1-\theta)^{n-S}\right] = \theta^{\alpha+S-1}(1-\theta)^{\beta+n-S-1}$$

mean $\dfrac{\alpha}{\alpha+\beta}$, and variance $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Weighted average of prior mean and data mean:
With k successes out of n trials:

$$\frac{\alpha+k}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{\text{prior mean}} + \frac{n}{\alpha+\beta+n}\underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}}$$

Normal
Posterior mean

$$\frac{\text{prec}_{\text{prior}}}{\text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}}\mu_0 + \frac{\text{prec}_{\bar{X}_n}}{\text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}}\bar{X}_n$$

Posterior precision
$(1/\sigma_0^2) + (n/\sigma^2) = \text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}$
Draw out the distributions!

Bivariate Normal

$$f(x,y) \propto \exp\left(\frac{-1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right)$$

5 parameters: x replaced by $(x-\mu_x)/\sigma_x$, same with y
X~N( ) is correlated to Y~N( ) by r.
Step 1: standardize X to get the z-score. $\dfrac{\hat{y}-\bar{y}}{s_y} = r\left(\dfrac{x-\bar{x}}{s_x}\right)$
Step 2: r * z-score of X → z-score of Y
Regression fallacy: failure to see that regression toward the mean comes from natural statistical variability.
Conditional distribution of Y given X=x is Normal with mean from the regression line and std dev from p or r.

$$E(Y \mid X=x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

SD given by $\sqrt{1-\rho^2}\,\sigma_y$

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\text{SD}(X)\,\text{SD}(Y)}$$

$$r(X,Y) = \frac{\widehat{\text{cov}}(X,Y)}{\widehat{\text{SD}}(X)\widehat{\text{SD}}(Y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

Model Selection
Deviance Information Criterion (DIC): penalizes complexity. Defines deviance: -2 x log(likelihood). Bigger deviance = worse fit. Prefer smaller DIC.
(min deviance) + 2 × [(average deviance over MCMC run) − (min deviance)]
Walks on Graphs
Degree - d(i) = number of edges touching node i.
P(i,j) = 1/d(i) for each neighboring pair of i and j.
More on Markov Chains
Example: Ehrenfest chain (diffusion, or dogs and fleas)

This is a model that arose in physics as a simple conceptual model of "mixing", for example, two volumes of gas connected by a small hole. One can use it to think about questions like, "Why is it that we don't hear more often about people dying because all the molecules of air in the room happened to go over into one corner, or maybe up near the ceiling?"

n balls in two urns; *x* balls in urn 1, $n-x$ balls in urn 2.

At each time, choose one of the *n* balls at random (each ball equally likely), and move that ball to the other urn.

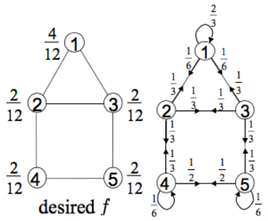
The state X_t is the number of balls in urn 1 at time *t*.

$$\begin{pmatrix}\pi_{t+1}(1) & \pi_{t+1}(2) & \pi_{t+1}(3)\end{pmatrix} = \begin{pmatrix}\pi_t(1) & \pi_t(2) & \pi_t(3)\end{pmatrix}\begin{pmatrix}P(1,1) & P(1,2) & P(1,3)\\ P(2,1) & P(2,2) & P(2,3)\\ P(3,1) & P(3,2) & P(3,3)\end{pmatrix}$$

Why Metropolis-Hastings Works:

$$f(x)P(x,y) = f(x)Q(x,y)\min\left\{1, \frac{f(y)Q(y,x)}{f(x)Q(x,y)}\right\} = \min\left\{f(x)Q(x,y),\ f(y)Q(y,x)\right\}$$

Note this is symmetric in x and y. So $f(x)P(x,y) = f(y)P(y,x)$.
We know this is a sufficient condition for f to be stationary for the matrix *P*.
Why it's useful: We don't need to calculate individual values of f(x), which requires normalization. We just need ratios of the form f(y)/f(x), which is easier.



E.g. the $P(1,3) = 1/6$ comes from the calculation

$$P(1,3) = Q(1,3)\min\left\{1, \frac{f(3)Q(3,1)}{f(1)Q(1,3)}\right\} = \frac{1}{2}\min\left\{1, \frac{(2/12)(1/3)}{(4/12)(1/2)}\right\} = \frac{1}{2}\left\{\frac{1}{3}\right\} = \frac{1}{6}$$

Gibbs Sampler: iteratively perform the following: Leaving all variables but 1 fixed, sample from the conditional distribution of that variable, given the fixed values for the rest.
Special case of MH with "smart" proposals. This explores the space more effectively and produces a good sample from the desired distribution in fewer iterations.
MH acceptance probability is 1 because:

$$\min\left\{1, \frac{f(x,y')\,Q\big((x,y'),(x,y)\big)}{f(x,y)\,Q\big((x,y),(x,y')\big)}\right\} = \min\left\{1, \frac{f(x,y')\,f(y \mid x)}{f(x,y)\,f(y' \mid x)}\right\} = 1$$

The right side simplifies to 1:
f(y|x)/f(x,y) = 1/f(x), f(x,y')/f(y'|x) = f(x).