Probability & Statistics

> Author: Jose 胡冠洲 @ ShanghaiTech & Staff of Harvard STAT-110

Full-ver. Cheatsheet

See below (page 2-3).

Harvard Stat-110 Cheatsheet

See below (page 4-13).

Harvard Stat-110 Review Notes

See below (page 14-122).

Dis Dist.	PMF	E	Var	PGF	MGF
Bern(p)	$P(X=1)=p, P(X=0)=q$	p	pq	$pt+q$	pe^t+q
Bin(n,p)	$\binom{n}{x}p^xq^{n-x}$, $x\in[0,n]$	np	npq	$(pt+q)^n$	$(pe^t+q)^n$
HGeom(w,b,n)	$\frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}}$, $0\leq k\leq w, 0\leq n-k\leq b$	$\frac{nw}{w+b}$	/	/	/
Geom(p)	q^kp, $x\geq 0$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qt}$	$\frac{p}{1-qe^t}$
FS(p)	$q^{k-1}p$, $x\geq 1$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pt}{1-qt}$	$\frac{pe^t}{1-qe^t}$
NBin(r,p)	$\binom{n+r-1}{r-1}p^rq^n$, $n\geq 0$	$r\frac{q}{p}$	$r\frac{q}{p^2}$	$(\frac{p}{1-qt})^r$	$(\frac{p}{1-qe^t})^r$
Pois(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$, $k\geq 0$	λ	λ	$e^{\lambda(t-1)}$	$e^{\lambda(e^t-1)}$

Cont Dist.	PDF	CDF	E	Var	MGF
Unif(a,b)	$\frac{1}{b-a}$, $a<x<b$	$\cdots\frac{x-a}{b-a}$, $a<x<b\cdots$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
$\mathcal{N}(0,1)$	$\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$	$\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$	0	1	$e^{\frac{t^2}{2}}$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(z-\mu)^2/(2\sigma^2)}$	$\int_{-\infty}^{z}\frac{1}{\sigma\sqrt{2\pi}}e^{-(t-\mu)^2/(2\sigma^2)}dt$	μ	σ^2	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$
Expo(λ)	$\lambda e^{-\lambda x}$, $x>0$	$1-e^{-\lambda x}$, $x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Beta(a,b)	$\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}$	/	$\frac{a}{a+b}$	/	/
Gamma(a,λ)	$\frac{1}{\Gamma(a)}(\lambda y)^ae^{-\lambda y}\frac{1}{y}$	/	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	/

Bayes' Rule	Y dis	Y cont
X dis	$\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X cont	$\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

LOTP	Y dis	Y cont
X dis	$\sum_y P(X=x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty}P(X=x\|Y=y)f_Y(y)dy$
X cont	$\sum_y f_X(x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty}f_{X\|Y}(x\|y)f_Y(y)dy$

Counting	Order	No Order
Replace	n^k	$\binom{n+k-1}{n-1}$
No Replace	$n\cdots(n-k+1)$	$\binom{n}{k}$

Bose-Eistein: $x_1+\cdots+x_r=n, x_i>0$, $\binom{n-1}{r-1}$ kinds.

Identities: $n\binom{n-1}{k-1}=k\binom{n}{k}$, $\binom{m+n}{k}=\sum_{j=0}^{k}\binom{m}{j}\binom{n}{k-j}$

Definitions

- PMF: $P(X=x)$ - Nonnegative; Sums to 1
- joint PMF: $P(X=x,Y=y)=P(X=x|Y=y)P(Y=y)$
- marginal PMF: $P(X=x)=\sum_y P(X=x,Y=y)$
- PDF: $f_X(x)=F'_X(x)$ - Nonnegative; integrates to 1
- joint PDF: $f_{X,Y}(x,y)=\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)$
- marginal PDF: $f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy$
- CDF: $F_X(x)=P(X\leq x)$ - Nondecreasing; Right-continuous; $\lim_{x\to-\infty}F(x)=0$, $\lim_{x\to\infty}F(x)=1$
- joint CDF: $F_{X,Y}(x,y)=P(X\leq x,Y\leq y)$
- c.e.: $E(Y|A)=\sum yP(Y=y|A)$ / $\int_{-\infty}^{\infty}yf(y|A)dy$
- c.e.: $E(Y|X)$ with $E(Y|X=x)=g(x)$
- c.v.: $Var(Y|X)=E((Y-E(Y|X))^2|X)=E(Y^2|X)-(E(Y|X))^2$

Expectation / Mean

- 1. $E(X)=\sum_x xP(X=x)$
  2. survival: $E(X)=\int_{-\infty}^{0}(G(x)-1)dx+\int_0^{\infty}G(x)dx$
  3. LOTUS: $E(g(X))=\sum_x g(x)P(X=x)$
  4. indicators: $X=I_1+I_2+\cdots+I_n$ (or other partitions), then $E(X)=p_1+p_2+\cdots+p_n$
  5. by PGF / MGF: $E(X)=g'_X(1)$ or $E(X)=M'_X(0)$
  6. 2D-L: $E(g(X,Y))=\sum_x\sum_y g(x,y)P(X=x,Y=y)$
  7. LOTE: $E(X)=\sum_{i=1}^{n}E(Y|A_i)P(A_i)$
  8. Adam's Law
- Properties:
  1. Linearity
  2. Monotonicity: $X\geq Y$ w.p. $1\Rightarrow E(X)\geq E(Y)$

Variance

- 1. $Var(X)=E((X-E(X))^2)=E(X^2)-(E(X))^2\geq 0$
  2. LOTUS: $E(X^2)=\sum_x x^2P(X=x)$
  3. indis: $X=I_1+I_2+\cdots+I_n$, then $E(\binom{X}{2})=\sum_{i<j}I_iI_j$
  4. PGF / MGF: $E(X^2-X)=g''_X(1)$ or $E(X^2)=M''_X(0)$
  5. Eve's Law
- Properties:
  1. $Var(X+c)=Var(X)$, $Var(cX)=c^2Var(X)$
  2. $Var(X+Y)=Var(X)+Var(Y)$ *iff independent!*

Proposition: $f_{X,Y}(x,y)=g(x)h(y)\Rightarrow$ independent and if g is valid PDF, both are valid marginal PDFs of X and Y

Independence

1. $F_{X,Y}(x,y)=F_X(x)F_Y(y)$
2. $f_{X,Y}(x,y)=f_X(x)f_Y(y)$
3. $f_{X|Y}(x|y)=f_X(x)$

Symmetry

- $X\sim\text{Bin}(n,p)\Rightarrow n-X\sim\text{Bin}(n,q)$
- $X\sim\text{HGeom}(w,b,n)\Leftrightarrow\text{HGeom}(n,w+b-n,w)$
- $Z\sim\mathcal{N}(0,1)\Rightarrow\phi(z)=\phi(-z)$
- $Z\sim\mathcal{N}(0,1)\Rightarrow\Phi(z)=1-\Phi(-z)$
- $Z\sim\mathcal{N}(0,1)\Rightarrow -Z\sim\mathcal{N}(0,1)$

Memoryless Property: Geom/Expo $\Leftrightarrow P(X \geq n+k | X \geq k) = P(X \geq n)$; $G(s+t) = G(s)G(t)$

Dist. Transform and Connections *independence!!*

- sum of Bin: $X \sim \text{Bin}(n,p)$, $Y \sim \text{Bin}(m,p) \Rightarrow X+Y \sim \text{Bin}(n+m,p)$
- sum of Pois: $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \Rightarrow X+Y \sim \text{Pois}(\lambda_1+\lambda_2)$
- sum of Geom: $X_j \sim \text{Geom}(p)$, $X \sim \text{NBin}(r,p) \Rightarrow X = X_1 + \cdots + X_r$
- sum of Norm: $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \Rightarrow Z_1 + Z_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$
- comparable Expo: $X_{1,2} \sim \text{Expo}(\lambda_{1,2}) \Rightarrow P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$
- min of Expo: $X_j \sim \text{Expo}(\lambda_j)$, $L = min(X_1, X_2, \cdots, X_n) \Rightarrow L \sim \text{Expo}(\lambda_1+\lambda_2+\cdots+\lambda_n)$
- Expo max - min: $M - L \sim \text{Expo}(\lambda)$, independent of L
- Bin $\Rightarrow$ HGeom: $X \sim \text{Bin}(n,p)$, $Y \sim \text{Bin}(m,p) \Rightarrow X|X+Y = r \sim \text{HGeom}(n,m,r)$
- HGeom $\Rightarrow$ Bin: $X \sim \text{HGeom}(w,b,n)$, $w+b \to \infty$ s.t. $\frac{w}{w+b}$ fixed $\Rightarrow$ X converges to $\text{Bin}(n, \frac{w}{w+b})$
- Bin $\Rightarrow$ Pois: $X \sim \text{Bin}(n,p)$, $n \to \infty, p \to 0$ and $\lambda = np$ fixed $\Rightarrow X$ converges to $\text{Pois}(\lambda)$
- Pois $\Rightarrow$ Bin: $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \Rightarrow X|X+Y = n \sim \text{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$
- Chicken-Egg: $X+Y = N \sim \text{Pois}(\lambda)$, $X|X+Y = n \sim \text{Bin}(n,p) \Rightarrow X \sim \text{Pois}(\lambda p)$
- shift-scale of Norm: $Z \sim \mathcal{N}(0,1)$, $X = \mu + \sigma Z \Rightarrow X \sim \mathcal{N}(\mu, \sigma^2)$

Covariance and Correlation

- $Cov(X,Y) = E((X-E(X))(Y-E(Y))) = E(XY) - E(X)E(Y)$
- $Cov(X,X) = Var(X)$; $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,c) = 0$; $Cov(aX,Y) = aCov(X,Y)$
- $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$
- $Var(X_1+\cdots+X_n) = Var(X_1)+\cdots+Var(X_n)+\mathbf{2}\sum_{i<j} Cov(X_i,X_j)$
- indepenent $\Rightarrow$ uncorrelated, uncorrelated $\not\Rightarrow$ independent ($Y = X^2$)

Universality of Unif

- $U \sim \text{Unif}(0,1)$ has $f(x) = 1$ and $F(x) = x$, $0 < x < 1$
- Let F be a continuous, strictly increasing CDF:
  1. $U \sim \text{Unif}(0,1)$ and $X = F^{-1}(U) \Rightarrow X$ is an r.v. with CDF F
  2. X is an r.v. with CDF $F \Rightarrow F(X) \sim \text{Unif}(0,1)$

Generating Functions

- PGF: $g_X(t) = E(t^X) = \sum_{k=0}^{\infty} p_k t^k$
  - $X = X_1 + X_2 + \cdots + X_n$, then $g_X(t) = g_{X_1}(t) g_{X_2}(t) \cdots g_{X_n}(t)$
- MGF: $M_X(t) = E(e^{tX}) = \sum_k e^{tk} P(X = k)$ or $\int e^{tx} f(x)dx$ which determines the distribution.
  - $X = X_1+X_2+\cdots+X_n$, then $M_X(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t)$
  - $M(a+bX) = e^{at}M(bt)$

Change of Variables

- 1 dim: $Y = g(X)$ where g is differentiable and strictly increasing(decreasing) $\Rightarrow f_Y(y) = f_X(x)|\frac{dx}{dy}|$
- multi dim: $\mathbf{Y} = g(\mathbf{X})$ where g is invertible $\Rightarrow f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(x)||\frac{\partial \mathbf{x}}{\partial \mathbf{y}}||$, $|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}| = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$

Order Statistics

- joint PDF: $f_{X_{(1)},\ldots,X_{()}}(x_1,\ldots,x_n) = n! f(x_1)\cdots f(x_n)$
- PDF: $f_{X_{(j)}}(x) = n\binom{n-1}{j-1} f(x) F^{j-1}(x)(1-F(x))^{n-j}$
- CDF: $P(X_{(j)} \leq x) = \sum_{k=j}^{n} \binom{n}{k} F^k(x)(1-F(x))^{n-k}$

Beta

- $\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} = \frac{(a-1)!(b-1)!}{(a+b-1)!}$
- Beta-Bin Conjungacy: prior $p \sim \text{Beta}(a,b)$, $X|p \sim \text{Bin}(n,p)$, then $p|X = k \sim \text{Beta}(a+k, b+n-k)$, $E(p|X=k) = \frac{a+k}{a+b+n}$

Gamma

- $\Gamma(a) = \int_0^{\infty} x^a e^{-x} \frac{1}{x} dx = (a-1)!$
- sum of Expo: $X_j \sim \text{Expo}(\lambda) \Rightarrow X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda)$
- Bank-Post Office: $X \sim \text{Gamma}(a,\lambda)$, $Y \sim \text{Gamma}(b,\lambda)$, $T = X+Y, W = \frac{X}{X+Y} \Rightarrow T \sim \text{Gamma}(a+b,\lambda)$, $W \sim \text{Beta}(a,b)$

Conditional Expectation and Variance

- X, Y are independent $\Rightarrow E(Y|X) = E(Y)$
- $E(h(X)Y|X) = h(X)E(Y|X)$
- *linearity in front!!*
- Adam's Law: $E(E(Y|X)) = E(Y)$ / $E(E(Y|X,Z)|Z) = E(Y|Z)$
- Eve's Law: $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$

Inference

- Bayesian Inference: r.v. θ with prior distribution, X with $f_{X|\theta}$ and observed $X = k \Rightarrow$ posterior $f_{\theta|X=k}$
  1. MAP: $\hat{\theta} = arg\max_{\theta} f(\theta|X=k)$
  2. LSE/MSE: $\hat{\theta} = E(\theta|X=k)$
- Classical Inference: constant θ, observation $P(X=k)$ is function of θ
  1. MLE: $\hat{\theta} = arg\max_{\theta} P(X=k)$

Sampling and Limits

- Sample Mean: $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$, $E(\bar{X}_n) = \mu$, $Var(\bar{X}_n) = \frac{\sigma^2}{n}$
- Sample Variance: $S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$, $E(S_n^2) = \sigma^2$
- Law of Large Numbers
  1. As $n \to \infty$, $P(\bar{X}_n \to \mu) = 1$
  2. As $n \to \infty$, $\forall \epsilon > 0, P(|\bar{X}_n - \mu| > \epsilon) \to 0$
- CLT: As $n \to \infty$, $\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \to \mathcal{N}(0,1)$
  1. $Y \sim \text{Pois}(n) \Rightarrow Y \sim \mathcal{N}(n,n)$
  2. $Y \sim \text{Gamma}(n,\lambda) \Rightarrow Y \sim \mathcal{N}(\frac{n}{\lambda}, \frac{n}{\lambda^2})$
  3. $Y \sim \text{Bin}(n,p) \Rightarrow Y \sim \mathcal{N}(np, npq)$

Inequalities

- Cauchy-Schwartz: $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
- Second Moment: $P(X=0) \leq \frac{Var(X)}{E(X^2)}$
- Jensen's: $g(x)$ is convex, $E(g(X)) \geq g(E(X))$; concave, $\leq$
- Markov's: $P(|X| \geq a) \leq \frac{E|X|}{a}$
- Chebtshev's: $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$
- Chernoff's: $P(X \geq a) \leq \min_{t>0} \frac{E(e^{tX})}{e^{ta}}$

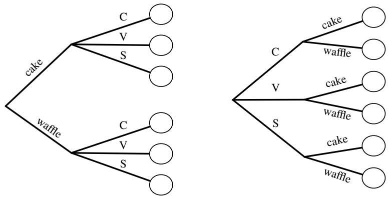
Probability Cheatsheet v2.0

Compiled by William Chen (http://wzchen.com) and Joe Blitzstein, with contributions from Sebastian Chiu, Yuan Jiang, Yuqi Hou, and Jessy Hwang. Material based on Joe Blitzstein's (@stat110) lectures (http://stat110.net) and Blitzstein/Hwang's Introduction to Probability textbook (http://bit.ly/introprobability). Licensed under CC BY-NC-SA 4.0. Please share comments, suggestions, and errors at http://github.com/wzchen/probability_cheatsheet.

Last Updated December 12, 2015

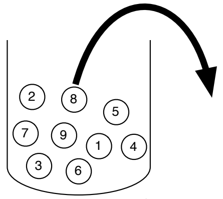
Counting

Multiplication Rule



Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, . . . , and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment.

Sampling Table



The sampling table gives the number of possible samples of size k out of a population of size n, under various assumptions about how the sample is collected.

	Order Matters	Not Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Thinking Conditionally

Independence

Independent Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More formally, A and B (which have nonzero probability) are independent if and only if one of the following equivalent statements holds:

$$P(A \cap B) = P(A)P(B)$$
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Conditional Independence A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws A useful identity that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. Analogous results hold with more than two sets.

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Joint, Marginal, and Conditional

Joint Probability $P(A \cap B)$ or $P(A, B)$ – Probability of A and B.

Marginal (Unconditional) Probability $P(A)$ – Probability of A.

Conditional Probability $P(A|B) = P(A, B)/P(B)$ – Probability of A, given that B occurred.

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Probability of an Intersection or Union

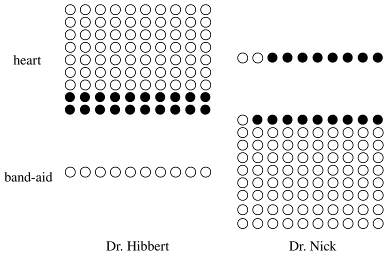
Intersections via Conditioning

$$P(A, B) = P(A)P(B|A)$$
$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

Unions via Inclusion-Exclusion

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C). \end{aligned}$$

Simpson's Paradox



It is possible to have

$$P(A \mid B, C) < P(A \mid B^c, C) \text{ and } P(A \mid B, C^c) < P(A \mid B^c, C^c)$$
$$\text{yet also } P(A \mid B) > P(A \mid B^c).$$

Law of Total Probability (LOTP)

Let $B_1, B_2, B_3, \dots B_n$ be a *partition* of the sample space (i.e., they are disjoint and their union is the entire sample space).

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n)$$

For **LOTP with extra conditioning**, just add in another event C!

$$P(A|C) = P(A|B_1, C)P(B_1|C) + \dots + P(A|B_n, C)P(B_n|C)$$
$$P(A|C) = P(A \cap B_1|C) + P(A \cap B_2|C) + \dots + P(A \cap B_n|C)$$

Special case of LOTP with B and B^c as partition:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Bayes' Rule

Bayes' Rule, and with extra conditioning (just add in C!)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}$$

We can also write

$$P(A|B, C) = \frac{P(A, B, C)}{P(B, C)} = \frac{P(B, C|A)P(A)}{P(B, C)}$$

Odds Form of Bayes' Rule

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \frac{P(A)}{P(A^c)}$$

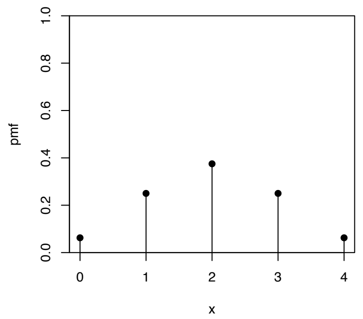
The *posterior odds* of A are the *likelihood ratio* times the *prior odds*.

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a *discrete* random variable takes on the value x.

$$p_X(x) = P(X = x)$$

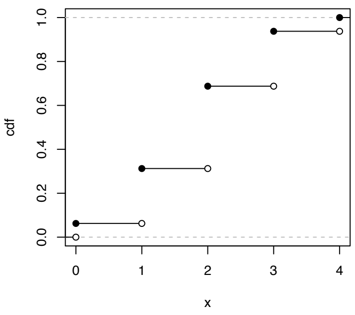


The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Cumulative Distribution Function (CDF) Gives the probability that a random variable is less than or equal to x.

$$F_X(x) = P(X \leq x)$$



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Intuitively, two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s X and Y are independent if for *all* values of x and y

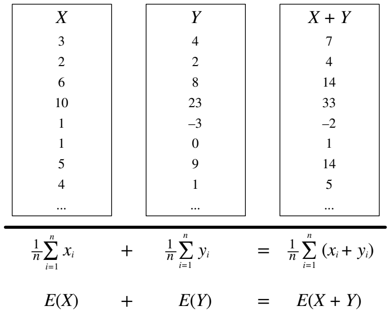
$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Expected Value and Indicators

Expected Value and Linearity

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \dots$ are all of the distinct possible values that X can take, the expected value of X is

$$E(X) = \sum_i x_i P(X = x_i)$$



X	Y	$X + Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	–3	–2
1	0	1
5	9	14
4	1	5
...	...	...

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad + \quad \frac{1}{n}\sum_{i=1}^{n} y_i \quad = \quad \frac{1}{n}\sum_{i=1}^{n} (x_i + y_i)$$

$$E(X) \quad + \quad E(Y) \quad = \quad E(X + Y)$$

Linearity For any r.v.s X and Y, and constants $a, b, c,$

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

Conditional Expected Value is defined like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x xP(X = x|A)$$

Indicator Random Variables

Indicator Random Variable is a random variable that takes on the value 1 or 0. It is always an indicator of some event: if the event occurs, the indicator is 1; otherwise it is 0. They are useful for many problems about counting how many events of some kind occur. Write

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Note that $I_A^2 = I_A, I_A I_B = I_{A \cap B}$, and $I_{A \cup B} = I_A + I_B - I_A I_B$.

Distribution $I_A \sim \text{Bern}(p)$ where $p = P(A)$.

Fundamental Bridge The expectation of the indicator for event A is the probability of event A: $E(I_A) = P(A)$.

Variance and Standard Deviation

$$\text{Var}(X) = E\left(X - E(X)\right)^2 = E(X^2) - (E(X))^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Continuous RVs, LOTUS, UoU

Continuous Random Variables (CRVs)

What's the probability that a CRV is in an interval? Take the difference in CDF values (or use the PDF as described later).

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

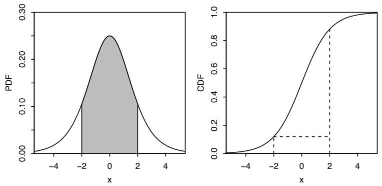
$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

What is the Probability Density Function (PDF)? The PDF f is the derivative of the CDF F.

$$F'(x) = f(x)$$

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



To find the probability that a CRV takes on a value in an interval, integrate the PDF over that interval.

$$F(b) - F(a) = \int_a^b f(x)dx$$

How do I find the expected value of a CRV? Analogous to the discrete case, where you sum x times the PMF, for CRVs you integrate x times the PDF.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

LOTUS

Expected value of a function of an r.v. The expected value of X is defined this way:

$$E(X) = \sum_x xP(X = x) \text{ (for discrete } X\text{)}$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \text{ (for continuous } X\text{)}$$

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_x g(x)P(X = x) \text{ (for discrete } X\text{)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ (for continuous } X\text{)}$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ is the number of bike wheels you see in that hour and $h(X) = \binom{X}{2} = \frac{X(X-1)}{2}$ is the number of *pairs* of bikes such that you see both of those bikes in that hour.

What's the point? You don't need to know the PMF/PDF of $g(X)$ to find its expected value. All you need is the PMF/PDF of X.

Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a Uniform(0,1) random variable. When you plug a Uniform(0,1) r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable X has CDF

$$F(x) = 1 - e^{-x}, \text{ for } x > 0$$

By UoU, if we plug X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0, 1)$$

Similarly, if $U \sim \text{Unif}(0, 1)$ then $F^{-1}(U)$ has CDF F. The key point is that for any continuous random variable X, we can transform it into a Uniform random variable and back by using its CDF.

Moments and MGFs

Moments

Moments describe the shape of a distribution. Let X have mean μ and standard deviation σ, and $Z = (X - \mu)/\sigma$ be the *standardized* version of X. The kth moment of X is $\mu_k = E(X^k)$ and the kth standardized moment of X is $m_k = E(Z^k)$. The mean, variance, skewness, and kurtosis are important summaries of the shape of a distribution.

Mean $E(X) = \mu_1$

Variance $\text{Var}(X) = \mu_2 - \mu_1^2$

Skewness $\text{Skew}(X) = m_3$

Kurtosis $\text{Kurt}(X) = m_4 - 3$

Moment Generating Functions

MGF For any random variable X, the function

$$M_X(t) = E(e^{tX})$$

is the **moment generating function (MGF)** of X, if it exists for all t in some open interval containing 0. The variable t could just as well have been called u or v. It's a bookkeeping device that lets us work with the *function* M_X rather than the *sequence* of moments.

Why is it called the Moment Generating Function? Because the kth derivative of the moment generating function, evaluated at 0, is the kth moment of X.

$$\mu_k = E(X^k) = M_X^{(k)}(0)$$

This is true by Taylor expansion of e^{tX} since

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}$$

MGF of linear functions If we have $Y = aX + b$, then

$$M_Y(t) = E(e^{t(aX+b)}) = e^{bt}E(e^{(at)X}) = e^{bt}M_X(at)$$

Uniqueness *If it exists, the MGF uniquely determines the distribution.* This means that for any two random variables X and Y, they are distributed the same (their PMFs/PDFs are equal) if and only if their MGFs are equal.

Summing Independent RVs by Multiplying MGFs. If X and Y are independent, then

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t) \cdot M_Y(t)$$

The MGF of the sum of two random variables is the product of the MGFs of those two random variables.

Joint PDFs and CDFs

Joint Distributions

The **joint CDF** of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

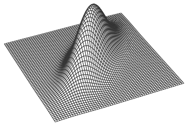
In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

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

Conditioning and Bayes' rule for continuous r.v.s

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Hybrid Bayes' rule

$$f_X(x|A) = \frac{P(A|X = x)f_X(x)}{P(A)}$$

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

$$P(X = x) = \sum_y P(X = x, Y = y)$$

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Independence of Random Variables

Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. For discrete random variables:

$$E(g(X, Y)) = \sum_x \sum_y g(x, y)P(X = x, Y = y)$$

For continuous random variables:

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$$

Covariance and Transformations

Covariance and Correlation

Covariance is the analog of variance for two random variables.

$$\text{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance and Independence If two random variables are independent, then they are uncorrelated. The converse is not necessarily true (e.g., consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$).

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0 \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance The variance of a sum can be found by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$$

If X and Y are independent then they have covariance 0, so

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

If $X_1, X_2, \dots, X_n$ are identically distributed and have the same covariance relationships (often by **symmetry**), then

$$\text{Var}(X_1 + X_2 + \dots + X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

Covariance Properties For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(Y, X) \\
\text{Cov}(X + a, Y + b) &= \text{Cov}(X, Y) \\
\text{Cov}(aX, bY) &= ab\text{Cov}(X, Y) \\
\text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) \\
&\quad + \text{Cov}(X, Z)
\end{aligned}$$

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

Transformations

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

The derivative of the inverse transformation is called the **Jacobian**.

Two Variable Transformations Similarly, let's say we know the joint PDF of U and V but are also interested in the random vector (X, Y) defined by $(X, Y) = g(U, V)$. Let

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}$$

be the **Jacobian matrix.** If the entries in this matrix exist and are continuous, and the determinant of the matrix is never 0, then

$$f_{X,Y}(x, y) = f_{U,V}(u, v)\left|\left|\frac{\partial(u, v)}{\partial(x, y)}\right|\right|$$

The inner bars tells us to take the matrix's determinant, and the outer bars tell us to take the absolute value. In a 2×2 matrix,

$$\left|\left|\begin{matrix} a & b \\ c & d \end{matrix}\right|\right| = |ad - bc|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$$

Example Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. Then for each fixed t,

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(t-x)^2/2}dx$$

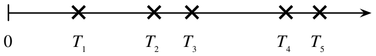
By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0, 2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is $\text{Pois}(\lambda t)$.
2. Numbers of arrivals in disjoint time intervals are independent.

For example, the numbers of arrivals in the time intervals $[0,5]$, $(5,12)$, and $[13,23)$ are independent with $\text{Pois}(5\lambda), \text{Pois}(7\lambda), \text{Pois}(10\lambda)$ distributions, respectively.



Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0,t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \le t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim \text{Expo}(\lambda)$. By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. $\text{Expo}(\lambda)$, i.e., the differences $T_n - T_{n-1}$ are i.i.d. $\text{Expo}(\lambda)$.

Order Statistics

Definition Let's say you have n i.i.d. r.v.s $X_1, X_2, \dots, X_n$. If you arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. So $X_{(1)}$ is the smallest in the list and $X_{(n)}$ is the largest in the list.

Note that the order statistics are *dependent*, e.g., learning $X_{(4)} = 42$ gives us the information that $X_{(1)}, X_{(2)}, X_{(3)}$ are ≤ 42 and $X_{(5)}, X_{(6)}, \dots, X_{(n)}$ are ≥ 42.

Distribution Taking n i.i.d. random variables $X_1, X_2, \dots, X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are:

$$F_{X_{(i)}}(x) = P(X_{(i)} \le x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n \binom{n-1}{i-1} F(x)^{i-1} (1 - F(x))^{n-i} f(x)$$

Uniform Order Statistics The jth order statistic of i.i.d. $U_1, \dots, U_n \sim \text{Unif}(0,1)$ is $U_{(j)} \sim \text{Beta}(j, n-j+1)$.

Conditional Expectation

Conditioning on an Event We can find $E(Y|A)$, the expected value of Y given that event A occurred. A very important case is when A is the event $X = x$. Note that $E(Y|A)$ is a *number*. For example:

- The expected value of a fair die roll, given that it is prime, is $\frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 5 = \frac{10}{3}$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success. Let A be the event that the first 3 trials are all successes. Then

  $$E(Y|A) = 3 + 7p$$

  since the number of successes among the last 7 trials is $\text{Bin}(7,p)$.
- Let $T \sim \text{Expo}(1/10)$ be how long you have to wait until the shuttle comes. Given that you have already waited t minutes, the expected additional waiting time is 10 more minutes, by the memoryless property. That is, $E(T|T > t) = t + 10$.

Discrete Y	**Continuous Y**
$E(Y) = \sum_y yP(Y = y)$	$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$
$E(Y\|A) = \sum_y yP(Y = y\|A)$	$E(Y\|A) = \int_{-\infty}^{\infty} y f(y\|A) dy$

Conditioning on a Random Variable We can also find $E(Y|X)$, the expected value of Y given the random variable X. This is *a function of the random variable X*. It is *not* a number except in certain special cases such as if $X \perp\!\!\!\perp Y$. To find $E(Y|X)$, find $E(Y|X = x)$ and then plug in X for x. For example:

- If $E(Y|X = x) = x^3 + 5x$, then $E(Y|X) = X^3 + 5X$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success and X be the number of successes among the first 3 trials. Then $E(Y|X) = X + 7p$.
- Let $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. Then $E(Y|X = x) = x^2$ since if we know $X = x$ then we know $Y = x^2$. And $E(X|Y = y) = 0$ since if we know $Y = y$ then we know $X = \pm\sqrt{y}$, with equal probabilities (by symmetry). So $E(Y|X) = X^2, E(X|Y) = 0$.

Properties of Conditional Expectation

1. $E(Y|X) = E(Y)$ if $X \perp\!\!\!\perp Y$
2. $E(h(X)W|X) = h(X)E(W|X)$ (**taking out what's known**) In particular, $E(h(X)|X) = h(X)$.
3. $E(E(Y|X)) = E(Y)$ (**Adam's Law**, a.k.a. Law of Total Expectation)

Adam's Law (a.k.a. Law of Total Expectation) can also be written in a way that looks analogous to LOTP. For any events $A_1, A_2, \dots, A_n$ that partition the sample space,

$$E(Y) = E(Y|A_1)P(A_1) + \dots + E(Y|A_n)P(A_n)$$

For the special case where the partition is A, A^c, this says

$$E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$$

Eve's Law (a.k.a. Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

MVN, LLN, CLT

Law of Large Numbers (LLN)

Let $X_1, X_2, X_3 \dots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$$

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \dots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. The CLT says

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

If the X_i are i.i.d. with mean μ_X and variance σ_X^2, then $\mu_Y = n\mu_X$ and $\sigma_Y^2 = n\sigma_X^2$. For the sample mean $\bar{X}_n$, the CLT says

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \sigma_X^2/n)$$

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$. The CLT says that if we standardize the sum $X_1 + \dots + X_n$ then the distribution of the sum converges to $\mathcal{N}(0,1)$ as $n \to \infty$:

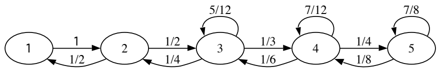
$$\frac{1}{\sigma\sqrt{n}}(X_1 + \dots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0,1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF, Φ. In terms of the sample mean, the CLT says

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{D} \mathcal{N}(0,1)$$

Markov Chains

Definition



A Markov chain is a random walk in a **state space**, which we will assume is finite, say $\{1, 2, \dots, M\}$. We let X_t denote which element of the state space the walk is visiting at time t. The Markov chain is the sequence of random variables tracking where the walk is at all points in time, $X_0, X_1, X_2, \dots$. By definition, a Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent.* In symbols,

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

State Properties

A state is either recurrent or transient.

- If you start at a **recurrent state**, then you will always return back to that state at some point in the future. ♪*You can check-out any time you like, but you can never leave.* ♪
- Otherwise you are at a **transient state**. There is some positive probability that once you leave you will never return. ♪*You don't have to go home, but you can't stay here.* ♪

A state is either periodic or aperiodic.

- If you start at a **periodic state** of period k, then the GCD of the possible numbers of steps it would take to return back is $k > 1$.
- Otherwise you are at an **aperiodic state**. The GCD of the possible numbers of steps it would take to return back is 1.

Transition Matrix

Let the state space be $\{1, 2, \dots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

$$q_{ij} = P(X_{n+1} = j | X_n = i)$$

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i, j) element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

If X_0 is distributed according to the row vector PMF $\vec{p}$, i.e., $p_j = P(X_0 = j)$, then the PMF of X_n is $\vec{p}Q^n$.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. If a chain (on a finite state space) is irreducible, then all of its states are recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.

A chain is **reversible** with respect to $\vec{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

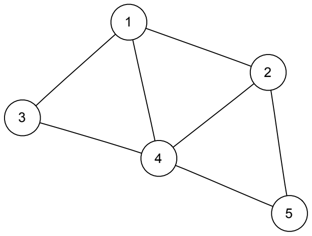
Let us say that the vector $\vec{s} = (s_1, s_2, \dots, s_M)$ be a PMF (written as a row vector). We will call $\vec{s}$ the **stationary distribution** for the chain if $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \dots$ also have the stationary distribution.

For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\vec{s}\,' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Reversibility Condition Implies Stationarity If you have a PMF $\vec{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\vec{s}$ is stationary.

Random Walk on an Undirected Network



If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{3}{14}, \frac{4}{14}, \frac{2}{14})$.

Continuous Distributions

Uniform Distribution

Let us say that U is distributed $\textrm{Unif}(a, b)$. We know the following:

Properties of the Uniform For a Uniform distribution, the probability of a draw from any interval within the support is proportional to the length of the interval. See *Universality of Uniform* and *Order Statistics* for other properties.

Example William throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. William's darts have a Uniform distribution on the surface of the room. The Uniform is the only distribution where the probability of hitting in any specific region is proportional to the length/area/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal Distribution

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the Central Limit Theorem, which states that the sample mean of i.i.d. r.v.s will approach a Normal distribution as the sample size grows, regardless of the initial distribution.

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0, 1)$ by the following transformation:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0, 1)$, has mean 0 and variance 1. Its CDF is denoted by Φ.

Exponential Distribution

Let us say that X is distributed $\textrm{Expo}(\lambda)$. We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but a shooting star is not "due" to come just because you've waited so long. Your waiting time is memoryless; the additional time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed $\textrm{Expo}(4)$ hours. Here $\lambda = 4$ is the **rate parameter**, since shooting stars arrive at a rate of 1 per 1/4 hour on average. The expected time until the next shooting star is $1/\lambda = 1/4$ hour.

Expos as a rescaled Expo(1)

$$Y \sim \textrm{Expo}(\lambda) \rightarrow X = \lambda Y \sim \textrm{Expo}(1)$$

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \textrm{Expo}(\lambda)$ and any positive numbers s and t,

$$P(X > s + t | X > s) = P(X > t)$$

Equivalently,

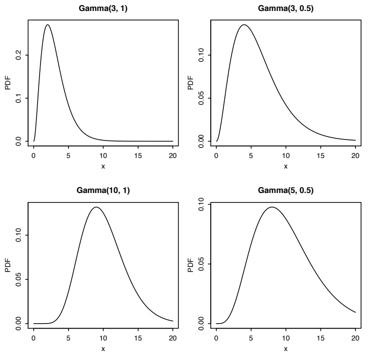
$$X - a | (X > a) \sim \textrm{Expo}(\lambda)$$

For example, a product with an $\textrm{Expo}(\lambda)$ lifetime is always "as good as new" (it doesn't experience wear and tear). Given that the product has survived a years, the additional time that it will last is still $\textrm{Expo}(\lambda)$.

Min of Expos If we have independent $X_i \sim \textrm{Expo}(\lambda_i)$, then $\min(X_1, \dots, X_k) \sim \textrm{Expo}(\lambda_1 + \lambda_2 + \dots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \textrm{Expo}(\lambda)$, then $\max(X_1, \dots, X_k)$ has the same distribution as $Y_1 + Y_2 + \dots + Y_k$, where $Y_j \sim \textrm{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution

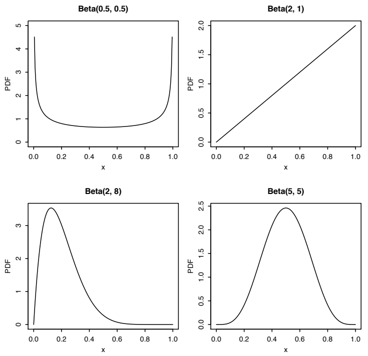


Let us say that X is distributed $\textrm{Gamma}(a, \lambda)$. We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed $\textrm{Expo}(\lambda)$. You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is $\textrm{Gamma}(n, \lambda)$.

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is Exponential with mean 2 minutes. Only one person at a time can be served. The distribution of your waiting time until it's your turn to be served is $\textrm{Gamma}(3, \frac{1}{2})$.

Beta Distribution



Conjugate Prior of the Binomial In the Bayesian approach to statistics, parameters are viewed as random variables, to reflect our uncertainty. The *prior* for a parameter is its distribution before observing data. The *posterior* is the distribution for the parameter after observing data. Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p in a Binomial, then the posterior distribution on p given the Binomial data is also Beta-distributed. Consider the following two-level model:

$$X|p \sim \textrm{Bin}(n, p)$$
$$p \sim \textrm{Beta}(a, b)$$

Then after observing $X = x$, we get the posterior distribution

$$p|(X = x) \sim \textrm{Beta}(a + x, b + n - x)$$

Order statistics of the Uniform See *Order Statistics*.

Beta-Gamma relationship If $X \sim \textrm{Gamma}(a, \lambda)$, $Y \sim \textrm{Gamma}(b, \lambda)$, with $X \perp\!\!\!\perp Y$ then

- $\frac{X}{X+Y} \sim \text{Beta}(a,b)$
- $X+Y \perp\!\!\!\perp \frac{X}{X+Y}$

This is known as the **bank–post office result**.

χ^2 (Chi-Square) Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0,1)$$

$$X \sim \text{Gamma}(n/2, 1/2)$$

Discrete Distributions

Distributions for four sampling schemes

	Replace	No Replace
Fixed # trials (n)	Binomial (Bern if $n=1$)	HGeom
Draw until r success	NBin (Geom if $r=1$)	NHGeom

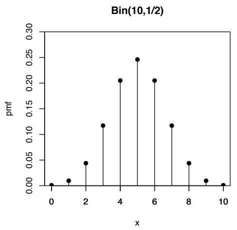
Bernoulli Distribution

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial ($n=1$). Let us say that X is distributed Bern(p). We know the following:

Story A trial is performed with probability p of "success", and X is the indicator of success: 1 means success, 0 means failure.

Example Let X be the indicator of Heads for a fair coin toss. Then $X \sim \text{Bern}(\frac{1}{2})$. Also, $1 - X \sim \text{Bern}(\frac{1}{2})$ is the indicator of Tails.

Binomial Distribution



Let us say that X is distributed Bin(n,p). We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial is either a success or a failure, each with the same probability p of success. We can also write X as a sum of multiple independent Bern(p) random variables. Let $X \sim \text{Bin}(n,p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. Then

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed Bin($10, \frac{3}{4}$).

Properties Let $X \sim \text{Bin}(n,p), Y \sim \text{Bin}(m,p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1-p)$
- **Sum** $X + Y \sim \text{Bin}(n+m, p)$
- **Conditional** $X|(X+Y=r) \sim \text{HGeom}(n,m,r)$
- **Binomial-Poisson Relationship** Bin(n,p) is approximately Pois(λ) if p is small.
- **Binomial-Normal Relationship** Bin(n,p) is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Geometric Distribution

Let us say that X is distributed Geom(p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has probability $\frac{1}{10}$ to catch Mew, the number of failed pokeballs will be distributed Geom($\frac{1}{10}$).

First Success Distribution

Equivalent to the Geometric distribution, except that it includes the first success in the count. This is 1 more than the number of failures. If $X \sim \text{FS}(p)$ then $E(X) = 1/p$.

Negative Binomial Distribution

Let us say that X is distributed NBin(r,p). We know the following:

Story X is the number of "failures" that we will have before we achieve our rth success. Our successes have probability p.

Example Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed NBin($3, 0.6$).

Hypergeometric Distribution

Let us say that X is distributed HGeom(w,b,n). We know the following:

Story In a population of w desired objects and b undesired objects, X is the number of "successes" we will have in a draw of n objects, without replacement. The draw of n objects is assumed to be a **simple random sample** (all sets of n objects are equally likely).

Examples Here are some HGeom examples.

- Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n Pokemon in the forest, and X is the number of Pikachus in our encounters.
- The number of Aces in a 5 card hand.
- You have w white balls and b black balls, and you draw n balls. You will draw X white balls.
- You have w white balls and b black balls, and you draw n balls without replacement. The number of white balls in your sample is HGeom(w,b,n); the number of black balls is HGeom(b,w,n).
- **Capture-recapture** A forest has N elk, you capture n of them, tag them, and release them. Then you recapture a new sample of size m. How many tagged elk are now in the new sample? HGeom($n, N-n, m$)

Poisson Distribution

Let us say that X is distributed Pois(λ). We know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, it is reasonable to model the number of accidents in a month at that intersection as Pois(2). Then the number of accidents that happen in two months at that intersection is distributed Pois(4).

Properties Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

1. **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. **Conditional** $X|(X+Y=n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
3. **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Multivariate Distributions

Multinomial Distribution

Let us say that the vector $\vec{X} = (X_1, X_2, X_3, \dots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \dots, p_k)$.

Story We have n items, which can fall into any one of the k buckets independently with the probabilities $\vec{p} = (p_1, p_2, \dots, p_k)$.

Example Let us assume that every year, 100 students in the Harry Potter Universe are randomly and independently sorted into one of four houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_4(100, \vec{p})$, where $\vec{p} = (0.25, 0.25, 0.25, 0.25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

Joint PMF For $n = n_1 + n_2 + \cdots + n_k$,

$$P(\vec{X} = \vec{n}) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

Marginal PMF, Lumping, and Conditionals Marginally, $X_i \sim \text{Bin}(n, p_i)$ since we can define "success" to mean category i. If you lump together multiple categories in a Multinomial, then it is still Multinomial. For example, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$ for $i \neq j$ since we can define "success" to mean being in category i or j. Similarly, if $k = 6$ and we lump categories 1-2 and lump categories 3-5, then

$$(X_1 + X_2, X_3 + X_4 + X_5, X_6) \sim \text{Mult}_3(n, (p_1 + p_2, p_3 + p_4 + p_5, p_6))$$

Conditioning on some X_j also still gives a Multinomial:

$$X_1, \dots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1-p_k}, \dots, \frac{p_{k-1}}{1-p_k}\right)\right)$$

Variances and Covariances We have $X_i \sim \text{Bin}(n, p_i)$ marginally, so $\text{Var}(X_i) = np_i(1-p_i)$. Also, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$.

Multivariate Uniform Distribution

See the univariate Uniform for stories and examples. For the 2D Uniform on some region, probability is proportional to area. Every point in the support has equal density, of value $\frac{1}{\text{area of region}}$. For the 3D Uniform, probability is proportional to volume.

Multivariate Normal (MVN) Distribution

A vector $\vec{X} = (X_1, X_2, \dots, X_k)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \dots, t_k$. The parameters of the Multivariate Normal are the **mean vector** $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$ and the **covariance matrix** where the (i,j) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Bivariate Normal (X,Y) with $\mathcal{N}(0,1)$ marginal distributions and correlation $\rho \in (-1,1)$ is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right),$$

with $\tau = \sqrt{1-\rho^2}$.

Distribution Properties

Important CDFs

Standard Normal Φ

Exponential(λ) $F(x) = 1 - e^{-\lambda x}$, for $x \in (0, \infty)$

Uniform(0,1) $F(x) = x$, for $x \in (0, 1)$

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. Bin$(n, p)$ can be thought of as a sum of i.i.d. Bern(p) r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. Gamma$(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. Expo(λ) r.v.s.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$ $\longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$. NBin$(r, p)$ can be thought of as a sum of i.i.d. Geom(p) r.v.s.
5. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Distributions

1. $\text{Bin}(1, p) \sim \text{Bern}(p)$
2. $\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$
3. $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
4. $\chi_n^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$
5. $\text{NBin}(1, p) \sim \text{Geom}(p)$

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \geq a) \leq \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$
4. **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave

Formulas

Geometric Series

$$1 + r + r^2 + \cdots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{1 - r^n}{1 - r}$$

$$1 + r + r^2 + \cdots = \frac{1}{1 - r} \text{ if } |r| < 1$$

Exponential Function (e^x)

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Gamma and Beta Integrals

You can sometimes solve complicated-looking integrals by pattern-matching to a gamma or beta integral:

$$\int_0^\infty x^{t-1} e^{-x} \, dx = \Gamma(t) \qquad \int_0^1 x^{a-1}(1 - x)^{b-1} \, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

Also, $\Gamma(a + 1) = a\Gamma(a)$, and $\Gamma(n) = (n - 1)!$ if n is a positive integer.

Euler's Approximation for Harmonic Sums

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.577\ldots$$

Stirling's Approximation for Factorials

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

Miscellaneous Definitions

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \geq a\}$; the median is the case $a = 1/2$.

log Statisticians generally use log to refer to natural log (i.e., base e).

i.i.d r.v.s Independent, identically-distributed random variables.

Example Problems

Contributions from Sebastian Chiu

Calculating Probability

A textbook has n typos, which are randomly scattered amongst its n pages, independently. You pick a random page. What is the probability that it has no typos? **Answer:** There is a $\left(1 - \frac{1}{n}\right)$ probability that any specific typo isn't on your page, and thus a $\boxed{\left(1 - \frac{1}{n}\right)^n}$ probability that there are no typos on your page. For n large, this is approximately $e^{-1} = 1/e$.

Linearity and Indicators (1)

In a group of n people, what is the expected number of distinct birthdays (month and day)? What is the expected number of birthday matches? **Answer:** Let X be the number of distinct birthdays and I_j be the indicator for the jth day being represented.

$$E(I_j) = 1 - P(\text{no one born on day } j) = 1 - (364/365)^n$$

By linearity, $\boxed{E(X) = 365\left(1 - (364/365)^n\right)}$. Now let Y be the number of birthday matches and J_i be the indicator that the ith pair of people have the same birthday. The probability that any two specific people share a birthday is $1/365$, so $\boxed{E(Y) = \binom{n}{2}/365}$.

Linearity and Indicators (2)

*This problem is commonly known as the **hat-matching problem**.* There are n people at a party, each with hat. At the end of the party, they each leave with a random hat. What is the expected number of people who leave with the right hat? **Answer:** Each hat has a $1/n$ chance of going to the right person. By linearity, the average number of hats that go to their owners is $\boxed{n(1/n) = 1}$.

Linearity and First Success

*This problem is commonly known as the **coupon collector problem**.* There are n coupon types. At each draw, you get a uniformly random coupon type. What is the expected number of coupons needed until you have a complete set? **Answer:** Let N be the number of coupons needed; we want $E(N)$. Let $N = N_1 + \cdots + N_n$, where N_1 is the draws to get our first new coupon, N_2 is the *additional* draws needed to draw our second new coupon and so on. By the story of the First Success, $N_2 \sim \text{FS}((n - 1)/n)$ (after collecting first coupon type, there's $(n - 1)/n$ chance you'll get something new). Similarly, $N_3 \sim \text{FS}((n - 2)/n)$, and $N_j \sim \text{FS}((n - j + 1)/n)$. By linearity,

$$E(N) = E(N_1) + \cdots + E(N_n) = \frac{n}{n} + \frac{n}{n - 1} + \cdots + \frac{n}{1} = \boxed{n \sum_{j=1}^{n} \frac{1}{j}}$$

This is approximately $n(\log(n) + 0.577)$ by Euler's approximation.

Orderings of i.i.d. random variables

I call 2 UberX's and 3 Lyfts at the same time. If the time it takes for the rides to reach me are i.i.d., what is the probability that all the Lyfts will arrive first? **Answer:** Since the arrival times of the five cars are i.i.d., all 5! orderings of the arrivals are equally likely. There are 3!2! orderings that involve the Lyfts arriving first, so the probability that the Lyfts arrive first is $\boxed{\frac{3!2!}{5!} = 1/10}$. Alternatively, there are $\binom{5}{3}$ ways to choose 3 of the 5 slots for the Lyfts to occupy, where each of the choices are equally likely. One of these choices has all 3 of the Lyfts arriving first, so the probability is $\boxed{1/\binom{5}{3} = 1/10}$.

Expectation of Negative Hypergeometric

What is the expected number of cards that you draw before you pick your first Ace in a shuffled deck (not counting the Ace)? **Answer:** Consider a non-Ace. Denote this to be card j. Let I_j be the indicator that card j will be drawn before the first Ace. Note that $I_j = 1$ says that j is before all 4 of the Aces in the deck. The probability that this occurs is $1/5$ by symmetry. Let X be the number of cards drawn before the first Ace. Then $X = I_1 + I_2 + \ldots + I_{48}$, where each indicator corresponds to one of the 48 non-Aces. Thus,

$$E(X) = E(I_1) + E(I_2) + \ldots + E(I_{48}) = 48/5 = \boxed{9.6}.$$

Minimum and Maximum of RVs

What is the CDF of the maximum of n independent Unif(0,1) random variables? **Answer:** Note that for r.v.s $X_1, X_2, \ldots, X_n$,

$$P(\min(X_1, X_2, \ldots, X_n) \geq a) = P(X_1 \geq a, X_2 \geq a, \ldots, X_n \geq a)$$

Similarly,

$$P(\max(X_1, X_2, \ldots, X_n) \leq a) = P(X_1 \leq a, X_2 \leq a, \ldots, X_n \leq a)$$

We will use this principle to find the CDF of $U_{(n)}$, where $U_{(n)} = \max(U_1, U_2, \ldots, U_n)$ and $U_i \sim \text{Unif}(0, 1)$ are i.i.d.

$$\begin{aligned} P(\max(U_1, U_2, \ldots, U_n) \leq a) &= P(U_1 \leq a, U_2 \leq a, \ldots, U_n \leq a) \\ &= P(U_1 \leq a)P(U_2 \leq a) \ldots P(U_n \leq a) \\ &= \boxed{a^n} \end{aligned}$$

for $0 < a < 1$ (and the CDF is 0 for $a \leq 0$ and 1 for $a \geq 1$).

Pattern-matching with e^x Taylor series

For $X \sim \text{Pois}(\lambda)$, find $E\left(\frac{1}{X + 1}\right)$. **Answer:** By LOTUS,

$$E\left(\frac{1}{X + 1}\right) = \sum_{k=0}^{\infty} \frac{1}{k + 1} \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k + 1)!} = \boxed{\frac{e^{-\lambda}}{\lambda}(e^{\lambda} - 1)}$$

Adam's Law and Eve's Law

William really likes speedsolving Rubik's Cubes. But he's pretty bad at it, so sometimes he fails. On any given day, William will attempt $N \sim \text{Geom}(s)$ Rubik's Cubes. Suppose each time, he has probability p of solving the cube, independently. Let T be the number of Rubik's Cubes he solves during a day. Find the mean and variance of T. **Answer:** Note that $T|N \sim \text{Bin}(N, p)$. So by Adam's Law,

$$E(T) = E(E(T|N)) = E(Np) = \boxed{\frac{p(1-s)}{s}}$$

Similarly, by Eve's Law, we have that

$$\text{Var}(T) = E(\text{Var}(T|N)) + \text{Var}(E(T|N)) = E(Np(1-p)) + \text{Var}(Np)$$

$$= \frac{p(1-p)(1-s)}{s} + \frac{p^2(1-s)}{s^2} = \boxed{\frac{p(1-s)(p+s(1-p))}{s^2}}$$

MGF – Finding Moments

Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X. **Answer:** The MGF of an Expo(λ) is $M(t) = \frac{\lambda}{\lambda - t}$. To get the third moment, we can take the third derivative of the MGF and evaluate at $t = 0$:

$$\boxed{E(X^3) = \frac{6}{\lambda^3}}$$

But a much nicer way to use the MGF here is via pattern recognition: note that $M(t)$ looks like it came from a geometric series:

$$\frac{1}{1 - \frac{t}{\lambda}} = \sum_{n=0}^{\infty} \left(\frac{t}{\lambda}\right)^n = \sum_{n=0}^{\infty} \frac{n!}{\lambda^n} \frac{t^n}{n!}$$

The coefficient of $\frac{t^n}{n!}$ here is the nth moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n.

Markov chains (1)

Suppose X_n is a two-state Markov chain with transition matrix

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{array}$$

Find the stationary distribution $\vec{s} = (s_0, s_1)$ of X_n by solving $\vec{s}Q = \vec{s}$, and show that the chain is reversible with respect to $\vec{s}$. **Answer:** The equation $\vec{s}Q = \vec{s}$ says that

$$s_0 = s_0(1-\alpha) + s_1\beta \text{ and } s_1 = s_0(\alpha) + s_0(1-\beta)$$

By solving this system of linear equations, we have

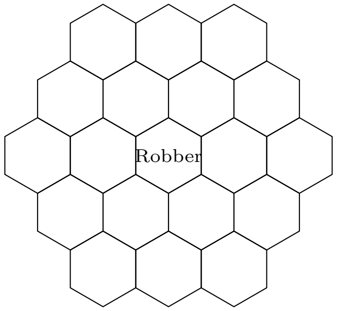
$$\boxed{\vec{s} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)}$$

To show that the chain is reversible with respect to $\vec{s}$, we must show $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is done if we can show $s_0 q_{01} = s_1 q_{10}$. And indeed,

$$s_0 q_{01} = \frac{\alpha\beta}{\alpha+\beta} = s_1 q_{10}$$

Markov chains (2)

William and Sebastian play a modified game of Settlers of Catan, where every turn they randomly move the robber (which starts on the center tile) to one of the adjacent hexagons.



(a) Is this Markov chain irreducible? Is it aperiodic? **Answer:** $\boxed{\text{Yes to both.}}$ The Markov chain is irreducible because it can get from anywhere to anywhere else. The Markov chain is aperiodic because the robber can return back to a square in $2, 3, 4, 5, \ldots$ moves, and the GCD of those numbers is 1.

(b) What is the stationary distribution of this Markov chain? **Answer:** Since this is a random walk on an undirected graph, the stationary distribution is proportional to the degree sequence. The degree for the corner pieces is 3, the degree for the edge pieces is 4, and the degree for the center pieces is 6. To normalize this degree sequence, we divide by its sum. The sum of the degrees is $6(3) + 6(4) + 7(6) = 84$. Thus the stationary probability of being on a corner is $3/84 = 1/28$, on an edge is $4/84 = 1/21$, and in the center is $6/84 = 1/14$.

(c) What fraction of the time will the robber be in the center tile in this game, in the long run? **Answer:** By the above, $\boxed{1/14}$.

(d) What is the expected amount of moves it will take for the robber to return to the center tile? **Answer:** Since this chain is irreducible and aperiodic, to get the expected time to return we can just invert the stationary probability. Thus on average it will take $\boxed{14}$ turns for the robber to return to the center tile.

Problem-Solving Strategies

Contributions from Jessy Hwang, Yuan Jiang, Yuqi Hou

1. **Getting started.** Start by *defining relevant events and random variables.* ("Let A be the event that I pick the fair coin"; "Let X be the number of successes.") Clear notion is important for clear thinking! Then decide what it is that you're supposed to be finding, in terms of your notation ("I want to find $P(X = 3|A)$"). Think about what type of object your answer should be (a number? A random variable? A PMF? A PDF?) and what it should be in terms of.

   Try simple and extreme cases. To make an abstract experiment more concrete, try *drawing a picture* or making up numbers that could have happened. Pattern recognition: does the structure of the problem resemble something we've seen before?

2. **Calculating probability of an event.** Use counting principles if the naive definition of probability applies. Is the probability of the complement easier to find? Look for symmetries. Look for something to condition on, then apply Bayes' Rule or the Law of Total Probability.

3. **Finding the distribution of a random variable.** First make sure you need the full distribution not just the mean (see next item). Check the *support* of the random variable: what values can it take on? Use this to rule out distributions that don't fit. Is there a *story* for one of the named distributions that fits the problem at hand? Can you write the random variable as a function of an r.v. with a known distribution, say $Y = g(X)$?

4. **Calculating expectation.** If it has a named distribution, check out the table of distributions. If it's a function of an r.v. with a named distribution, try LOTUS. If it's a count of something, try breaking it up into indicator r.v.s. If you can condition on something natural, consider using Adam's law.

5. **Calculating variance.** Consider independence, named distributions, and LOTUS. If it's a count of something, break it up into a sum of indicator r.v.s. If it's a sum, use properties of covariance. If you can condition on something natural, consider using Eve's Law.

6. **Calculating $E(X^2)$.** Do you already know $E(X)$ or $\text{Var}(X)$? Recall that $\text{Var}(X) = E(X^2) - (E(X))^2$. Otherwise try LOTUS.

7. **Calculating covariance.** Use the properties of covariance. If you're trying to find the covariance between two components of a Multinomial distribution, X_i, X_j, then the covariance is $-np_ip_j$ for $i \neq j$.

8. **Symmetry.** If $X_1, \ldots, X_n$ are i.i.d., consider using symmetry.

9. **Calculating probabilities of orderings.** Remember that all $n!$ ordering of i.i.d. continuous random variables $X_1, \ldots, X_n$ are equally likely.

10. **Determining independence.** There are several equivalent definitions. Think about simple and extreme cases to see if you can find a counterexample.

11. **Do a painful integral.** If your integral looks painful, see if you can write your integral in terms of a known PDF (like Gamma or Beta), and use the fact that PDFs integrate to 1?

12. **Before moving on.** Check some simple and extreme cases, check whether the answer seems plausible, check for biohazards.

Biohazards

Contributions from Jessy Hwang

1. **Don't misuse the naive definition of probability.** When answering "What is the probability that in a group of 3 people, no two have the same birth month?", it is *not* correct to treat the people as indistinguishable balls being placed into 12 boxes, since that assumes the list of birth months {January, January, January} is just as likely as the list {January, April, June}, even though the latter is six times more likely.

2. **Don't confuse unconditional, conditional, and joint probabilities.** In applying $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, it is *not* correct to say "$P(B) = 1$ because we know B happened"; $P(B)$ is the *prior* probability of B. Don't confuse $P(A|B)$ with $P(A, B)$.

3. **Don't assume independence without justification.** In the matching problem, the probability that card 1 is a match and card 2 is a match is not $1/n^2$. Binomial and Hypergeometric are often confused; the trials are independent in the Binomial story and dependent in the Hypergeometric story.

4. **Don't forget to do sanity checks.** Probabilities must be between 0 and 1. Variances must be ≥ 0. Supports must make sense. PMFs must sum to 1. PDFs must integrate to 1.

5. **Don't confuse random variables, numbers, and events.** Let X be an r.v. Then $g(X)$ is an r.v. for any function g. In particular, X^2, $|X|$, $F(X)$, and $I_{X>3}$ are r.v.s. $P(X^2 < X | X \geq 0)$, $E(X)$, $\text{Var}(X)$, and $g(E(X))$ are numbers. $X = 2$ and $F(X) \geq -1$ are events. It does not make sense to write $\int_{-\infty}^{\infty} F(X)dx$, because $F(X)$ is a random variable. It does not make sense to write $P(X)$, because X is not an event.

6. **Don't confuse a random variable with its distribution.** To get the PDF of X^2, you can't just square the PDF of X. The right way is to use transformations. To get the PDF of $X+Y$, you can't just add the PDF of X and the PDF of Y. The right way is to compute the convolution.

7. **Don't pull non-linear functions out of expectations.** $E(g(X))$ does not equal $g(E(X))$ in general. The St. Petersburg paradox is an extreme example. See also Jensen's inequality. The right way to find $E(g(X))$ is with LOTUS.

Distributions in R

Command	What it does
`help(distributions)`	shows documentation on distributions
`dbinom(k,n,p)`	PMF $P(X=k)$ for $X \sim \text{Bin}(n,p)$
`pbinom(x,n,p)`	CDF $P(X \leq x)$ for $X \sim \text{Bin}(n,p)$
`qbinom(a,n,p)`	ath quantile for $X \sim \text{Bin}(n,p)$
`rbinom(r,n,p)`	vector of r i.i.d. $\text{Bin}(n,p)$ r.v.s
`dgeom(k,p)`	PMF $P(X=k)$ for $X \sim \text{Geom}(p)$
`dhyper(k,w,b,n)`	PMF $P(X=k)$ for $X \sim \text{HGeom}(w,b,n)$
`dnbinom(k,r,p)`	PMF $P(X=k)$ for $X \sim \text{NBin}(r,p)$
`dpois(k,r)`	PMF $P(X=k)$ for $X \sim \text{Pois}(r)$
`dbeta(x,a,b)`	PDF $f(x)$ for $X \sim \text{Beta}(a,b)$
`dchisq(x,n)`	PDF $f(x)$ for $X \sim \chi_n^2$
`dexp(x,b)`	PDF $f(x)$ for $X \sim \text{Expo}(b)$
`dgamma(x,a,r)`	PDF $f(x)$ for $X \sim \text{Gamma}(a,r)$
`dlnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{LN}(m,s^2)$
`dnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{N}(m,s^2)$
`dt(x,n)`	PDF $f(x)$ for $X \sim t_n$
`dunif(x,a,b)`	PDF $f(x)$ for $X \sim \text{Unif}(a,b)$

The table above gives R commands for working with various named distributions. Commands analogous to `pbinom`, `qbinom`, and `rbinom` work for the other distributions in the table. For example, `pnorm`, `qnorm`, and `rnorm` can be used to get the CDF, quantiles, and random generation for the Normal. For the Multinomial, `dmultinom` can be used for calculating the joint PMF and `rmultinom` can be used for generating random vectors. For the Multivariate Normal, after installing and loading the `mvtnorm` package `dmvnorm` can be used for calculating the joint PDF and `rmvnorm` can be used for generating random vectors.

Recommended Resources

- Introduction to Probability Book (`http://bit.ly/introprobability`)
- Stat 110 Online (`http://stat110.net`)
- Stat 110 Quora Blog (`https://stat110.quora.com/`)
- Quora Probability FAQ (`http://bit.ly/probabilityfaq`)
- R Studio (`https://www.rstudio.com`)
- LaTeX File (`github.com/wzchen/probability_cheatsheet`)

Please share this cheatsheet with friends!
`http://wzchen.com/probability-cheatsheet`

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k \in \{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k \in \{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t},\ qe^t < 1$
Negative Binomial NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n \in \{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r,\ qe^t < 1$
Hypergeometric HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k \in \{0,1,2,\dots,n\}$	$\mu = \frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right) n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x\ -\ \mu)^2/(2\sigma^2)}$ $x \in (-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2 t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x \in (0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t},\ t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x \in (0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a,\ t<\lambda$
Beta Beta(a,b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x \in (0,1)$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x \in (0,\infty)$	$\theta = e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x \in (0,\infty)$	n	$2n$	$(1-2t)^{-n/2},\ t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x \in (-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist

Stat 110 Final Review, Fall 2011

Prof. Joe Blitzstein

1 General Information

The final will be on Thursday 12/15, from 2 PM to 5 PM. No books, notes, computers, cell phones, or calculators are allowed, except that you may bring four pages of standard-sized paper (8.5” x 11”) with anything you want written (or typed) on both sides. There will be approximately 8 problems, equally weighted. The material covered will be cumulative since probability *is* cumulative.

To study, I recommend solving lots and lots of practice problems! It’s a good idea to work through as many of the problems on this handout as possible without looking at solutions (and then discussing with others and looking at solutions to check your answers and for any problems where you were really stuck), and to take at least two of the practice finals under timed conditions using only four pages of notes. Carefully going through class notes, homeworks, and handouts (especially this handout and the midterm review handout) is also important, as long as it is done *actively* (intermixing reading, thinking, solving problems, and asking questions).

2 Topics

- Combinatorics: multiplication rule, tree diagrams, binomial coefficients, permutations and combinations, sampling with/without replacement when order does/doesn’t matter, inclusion-exclusion, story proofs.
- Basic Probability: sample spaces, events, axioms of probability, equally likely outcomes, inclusion-exclusion, unions, intersections, and complements.
- Conditional Probability: definition and meaning, writing $P(A_1 \cap A_2 \cap \cdots \cap A_n)$ as a product, Bayes’ Rule, Law of Total Probability, thinking conditionally, prior vs. posterior probability, independence vs. conditional independence.
- Random Variables: definition and interpretations, stories, discrete vs. continuous, distributions, CDFs, PMFs, PDFs, MGFs, functions of a r.v., indicator r.v.s, memorylessness of the Exponential, universality of the Uniform, Poisson approximation, Poisson processes, Beta as conjugate prior for the Binomial. sums (convolutions), location and scale.

- Expected Value: linearity, fundamental bridge, variance, standard deviation, covariance, correlation, using expectation to prove existence, LOTUS.

- Conditional Expectation: definition and meaning, taking out what's known, conditional variance, Adam's Law (iterated expectation), Eve's Law.

- Important Discrete Distributions: Bernoulli, Binomial, Geometric, Negative Binomial, Hypergeometric, Poisson.

- Important Continuous Distributions: Uniform, Normal, Exponential, Gamma, Beta, Chi-Square, Student-t.

- Jointly Distributed Random Variables: joint, conditional, and marginal distributions, independence, Multinomial, Multivariate Normal, change of variables, order statistics.

- Convergence: Law of Large Numbers, Central Limit Theorem.

- Inequalities: Cauchy-Schwarz, Markov, Chebyshev, Jensen.

- Markov chains: Markov property, transition matrix, irreducibility, stationary distributions, reversibility.

- Strategies: conditioning, symmetry, linearity, indicator r.v.s, stories, checking whether answers make sense (e.g., looking at simple and extreme cases and avoiding category errors).

- Some Important Examples: birthday problem, matching problem (de Montmort), Monty Hall, gambler's ruin, prosecutor's fallacy, testing for a disease, capture-recapture (elk problem), coupon (toy) collector, St. Petersburg paradox, Simpson's paradox, two envelope paradox, waiting time for HH vs. waiting time for HT, store with a random number of customers, bank-post office example, Bayes' billiards, random walk on a network, chicken and egg.

3 Important Distributions

3.1 Table of Distributions

The table below *will be provided on the final* (included as the last page). This is meant to help avoid having to memorize formulas for the distributions (or having to take up a lot of space on your pages of notes). Here $0 < p < 1$ and $q = 1 - p$. The parameters for Gamma and Beta are positive real numbers; n, r, and w are positive integers, as is b for the Hypergeometric.

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	p	$P(X=1) = p, P(X=0) = q$	p	pq
Binomial	n, p	$\binom{n}{k}p^k q^{n-k}$, for $k \in \{0, 1, \dots, n\}$	np	npq
Geometric	p	$q^k p$, for $k \in \{0, 1, 2, \dots\}$	q/p	q/p^2
NegBinom	r, p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0, 1, 2, \dots\}$	rq/p	rq/p^2
Hypergeom	w, b, n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$, for $k \in \{0, 1, \dots, n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1 - \frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$, for $k \in \{0, 1, 2, \dots\}$	λ	λ
Uniform	$a < b$	$\frac{1}{b-a}$, for $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ, σ^2	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Exponential	λ	$\lambda e^{-\lambda x}$, for $x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	a, λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x}x^{-1}$, for $x > 0$	a/λ	a/λ^2
Beta	a, b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$, for $0 < x < 1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
χ^2	n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$, for $x > 0$	n	$2n$
Student-t	n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1 + x^2/n)^{-(n+1)/2}$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$

3.2 Connections Between Distributions

The table above summarizes the PMFs/PDFs of the important distributions, and their means and variances, but it does not say where each distribution comes from (stories), or how the distributions interrelate. Some of these connections between distributions are listed below.

Also note that some of the important distributions are special cases of others. Bernoulli is a special case of Binomial; Geometric is a special case of Negative Binomial; Unif(0,1) is a special case of Beta; and Exponential and χ^2 are both special cases of Gamma.

1. **Binomial**: If $X_1, \dots, X_n$ are i.i.d. Bern(p), then $X_1 + \cdots + X_n \sim \text{Bin}(n, p)$.

2. **Neg. Binom.**: If $G_1, \dots, G_r$ are i.i.d. Geom(p), then $G_1 + \cdots + G_r \sim \text{NBin}(r, p)$.

3. **Location and Scale**: If $Z \sim \mathcal{N}(0, 1)$, then $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

   If $U \sim \text{Unif}(0, 1)$ and $a < b$, then $a + (b - a)U \sim \text{Unif}(a, b)$.

   If $X \sim \text{Expo}(1)$, then $\lambda^{-1}X \sim \text{Expo}(\lambda)$.

   If $Y \sim \text{Gamma}(a, \lambda)$, then $\lambda Y \sim \text{Gamma}(a, 1)$.

4. **Symmetry**: If $X \sim \text{Bin}(n, 1/2)$, then $n - X \sim \text{Bin}(n, 1/2)$.

   If $U \sim \text{Unif}(0, 1)$, then $1 - U \sim \text{Unif}(0, 1)$.

   If $Z \sim \mathcal{N}(0, 1)$, then $-Z \sim \mathcal{N}(0, 1)$.

5. **Universality of Uniform**: Let F be the CDF of a continuous r.v., such that F^{-1} exists. If $U \sim \text{Unif}(0, 1)$, then $F^{-1}(U)$ has CDF F. Conversely, if $X \sim F$, then $F(X) \sim \text{Unif}(0, 1)$.

6. **Uniform and Beta**: $\text{Unif}(0, 1)$ is the same distribution as $\text{Beta}(1, 1)$. The jth order statistic of n i.i.d. $\text{Unif}(0, 1)$ r.v.s is $\text{Beta}(j, n - j + 1)$.

7. **Beta and Binomial**: Beta is the conjugate prior to Binomial, in the sense that if $X|p \sim \text{Bin}(n, p)$ and the prior is $p \sim \text{Beta}(a, b)$, then the posterior is $p|X \sim \text{Beta}(a + X, b + n - X)$.

8. **Gamma**: If $X_1, \dots, X_n$ are i.i.d. Expo(λ), then $X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda)$.

9. **Gamma and Poisson**: In a Poisson process of rate λ, the number of arrivals in a time interval of length t is Pois(λt), while the time of the nth arrival is Gamma(n, λ).

10. **Gamma and Beta**: If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$ are independent, then $X/(X+Y) \sim \text{Beta}(a, b)$ is independent of $X + Y \sim \text{Gamma}(a + b, \lambda)$.

11. **Chi-Square**: χ^2_n is the same distribution as $\text{Gamma}(n/2, 1/2)$.

12. **Student-t**: If $Z \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2_n$ are independent, then $\frac{Z}{\sqrt{Y/n}}$ has the Student-t distribution with n degrees of freedom. For $n = 1$, this becomes the Cauchy distribution, which we can also think of as the distribution of Z_1/Z_2 with Z_1, Z_2 i.i.d. $\mathcal{N}(0, 1)$.

4 Sums of Independent Random Variables

Let $X_1, X_2, \ldots, X_n$ be *independent* random variables. The table below shows the distribution of their sum, $X_1 + X_2 + \cdots + X_n$, for various important cases depending on the distribution of X_i. The central limit theorem says that a sum of a large number of i.i.d. r.v.s will be *approximately* Normal, while these are exact distributions.

X_i	$\sum_{i=1}^n X_i$
Bernoulli(p)	Binomial(n, p)
Binomial(m_i, p)	Binomial($\sum_{i=1}^n m_i, p$)
Geometric(p)	NBin(n, p)
NBin(r_i, p)	NBin($\sum_{i=1}^n r_i, p$)
Poisson(λ_i)	Poisson($\sum_{i=1}^n \lambda_i$)
Unif(0,1)	Triangle(0,1,2) ($n = 2$)
$\mathcal{N}(\mu_i, \sigma_i^2)$	$\mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$
Exponential(λ)	Gamma(n, λ)
Gamma(α_i, λ)	Gamma($\sum_{i=1}^n \alpha_i, \lambda$)
Z_i^2, for $Z_i \sim \mathcal{N}(0, 1)$	χ^2_n

5 Review of Some Useful Results

5.1 De Morgan's Laws

$$(A_1 \cup A_2 \cdots \cup A_n)^c = A_1^c \cap A_2^c \cdots \cap A_n^c,$$
$$(A_1 \cap A_2 \cdots \cap A_n)^c = A_1^c \cup A_2^c \cdots \cup A_n^c.$$

5.2 Complements

$$P(A^c) = 1 - P(A).$$

5.3 Unions

$$P(A \cup B) = P(A) + P(B) - P(A \cap B);$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i), \text{ if the } A_i \text{ are disjoint;}$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i);$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \left((-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \right) \text{ (Inclusion-Exclusion).}$$

5.4 Intersections

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B),$$
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1}).$$

5.5 Law of Total Probability

If $E_1, E_2, \ldots, E_n$ are a partition of the sample space S (i.e., they are disjoint and their union is all of S) and $P(E_i) \neq 0$ for all i, then

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i).$$

An analogous formula holds for conditioning on a continuous r.v. X with PDF $f(x)$:

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x)f(x)dx.$$

Similarly, to go from a joint PDF $f(x, y)$ for (X, Y) to the marginal PDF of Y, integrate over all values of x:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

5.6 Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Often the denominator $P(B)$ is then expanded by the Law of Total Probability. For continuous r.v.s X and Y, Bayes' Rule becomes

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}.$$

5.7 Expected Value, Variance, and Covariance

Expected value is *linear*: for any random variables X and Y and constant c,

$$E(X + Y) = E(X) + E(Y),$$

$$E(cX) = cE(X).$$

Variance can be computed in two ways:

$$\text{Var}(X) = E(X - EX)^2 = E(X^2) - (EX)^2.$$

Constants come out from variance as the constant squared:

$$\text{Var}(cX) = c^2\text{Var}(X).$$

For the variance of the sum, there is a covariance term:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y),$$

where

$$\text{Cov}(X, Y) = E((X - EX)(Y - EY)) = E(XY) - (EX)(EY).$$

So if X and Y are uncorrelated, then the variance of the sum is the sum of the variances. Recall that *independent implies uncorrelated but not vice versa.* Covariance is symmetric:

$$\text{Cov}(Y, X) = \text{Cov}(X, Y),$$

and covariances of sums can be expanded as

$$\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W).$$

Note that for c a constant,

$$\text{Cov}(X, c) = 0,$$

$$\text{Cov}(cX, Y) = c\text{Cov}(X, Y).$$

The correlation of X and Y, which is between -1 and 1 , is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}.$$

This is also the covariance of the standardized versions of X and Y.

5.8 Law of the Unconscious Statistician (LOTUS)

Let X be a discrete random variable and h be a real-valued function. Then $Y = h(X)$ is a random variable. To compute EY using the definition of expected value, we would need to first find the PMF of Y and use $EY = \sum_y yP(Y = y)$. The Law of the Unconscious Statistician says we can use the PMF of X directly:

$$Eh(X) = \sum_x h(x)P(X = x).$$

Similarly, for X a continuous r.v. with PDF $f_X(x)$, we can find the expected value of $Y = h(X)$ by integrating $h(x)$ times the PDF of X, without first finding $f_Y(y)$:

$$Eh(X) = \int_{-\infty}^{\infty} h(x)f_X(x)dx.$$

5.9 Indicator Random Variables

Let A and B be events. Indicator r.v.s bridge between probability and expectation: $P(A) = E(I_A)$, where I_A is the indicator r.v. for A. It is often useful to think of a "counting" r.v. as a sum of indicator r.v.s. Indicator r.v.s have many pleasant properties. For example, $(I_A)^k = I_A$ for any positive number k, so it's easy to handle moments of indicator r.v.s. Also note that

$$I_{A\cap B} = I_A I_B,$$

$$I_{A\cup B} = I_A + I_B - I_A I_B.$$

5.10 Symmetry

There are many beautiful and useful forms of symmetry in statistics. For example:

1. If X and Y are i.i.d., then $P(X < Y) = P(Y < X)$. More generally, if $X_1, \dots, X_n$ are i.i.d., then $P(X_1 < X_2 < \dots X_n) = P(X_n < X_{n-1} < \dots < X_1)$, and likewise all $n!$ orderings are equally likely (in the continuous case it follows that $P(X_1 < X_2 < \dots X_n) = \frac{1}{n!}$, while in the discrete case we also have to consider ties).

2. If we shuffle a deck of cards and deal the first two cards, then the probability is 1/52 that the second card is the Ace of Spades, since by symmetry it's equally likely to be any card; it's not necessary to do a law of total probability calculation conditioning on the first card.

3. Consider the Hypergeometric, thought of as the distribution of the number of white balls, where we draw n balls from a jar with w white balls and b black balls (without replacement). By symmetry and linearity, we can immediately get that the expected value is $n\frac{w}{w+b}$, even though the trials are not independent, as the jth ball is equally likely to be any of the balls, and linearity still holds with dependent r.v.s.

4. By symmetry we can see immediately that if T is Cauchy, then $1/T$ is also Cauchy (since if we flip the ratio of two i.i.d. $\mathcal{N}(0,1)$ r.v.s, we still have the ratio of two i.i.d. $\mathcal{N}(0,1)$ r.v.s!).

5. $E(X_1|X_1+X_2) = E(X_2|X_1+X_2)$ by symmetry if X_1 and X_2 are i.i.d. So by linearity, $E(X_1|X_1+X_2) + E(X_2|X_1+X_2) = E(X_1+X_2|X_1+X_2) = X_1+X_2$, which gives $E(X_1|X_1+X_2) = (X_1+X_2)/2$.

5.11 Change of Variables

Let $\mathbf{Y} = g(\mathbf{X})$, where g is a differentiable function from $\mathbb{R}^n$ to itself whose inverse exists, and $\mathbf{X} = (X_1, \ldots, X_n)$ is a continuous random vector with PDF $f_{\mathbf{X}}$. The PDF of $\mathbf{Y}$ can be found using the Jacobian of the transformation g:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|,$$

where $\mathbf{x} = g^{-1}(\mathbf{y})$ and $|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}|$ is the absolute value of the Jacobian determinant of g^{-1} (here $|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}|$ can either be found directly or by using the reciprocal of $|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}|$).

In the case $n = 1$, this says that if $Y = g(X)$ where g is differentiable with $g'(x) > 0$ everywhere, then

$$f_Y(y) = f_X(x)\frac{dx}{dy},$$

which is easily remembered if written in the form

$$f_Y(y)dy = f_X(x)dx.$$

Remember when using this that $f_Y(y)$ is a function of y (found by solving for x in terms of y), and the bounds for y should be specified. For example, if $y = e^x$ and x ranges over $\mathbb{R}$, then y ranges over $(0, \infty)$.

5.12 Order Statistics

Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with PDF f and CDF F. The order statistics are obtained by sorting the X_i's, with $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The marginal PDF of the jth order statistic is

$$f_{X_{(j)}}(x) = n\binom{n-1}{j-1} f(x)F(x)^{j-1}(1 - F(x))^{n-j}.$$

5.13 Moment Generating Functions

The moment generating function of X is the function

$$M_X(t) = E(e^{tX}),$$

if this exists for all t in some open interval containing 0. For $X_1, \ldots, X_n$ independent, the MGF of the sum $S_n = X_1 + \cdots + X_n$ is

$$M_{S_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t),$$

which is often much easier to deal with than a convolution. The name "moment generating function" comes from the fact that the derivatives of M_X at 0 give the moments of X:

$$M'_X(0) = E(X), M''_X(0) = E(X^2), M'''_X(0) = E(X^3), \ldots.$$

Sometimes these can be computed "all at once" without explicitly taking derivatives, by finding the Taylor series for $M_X(t)$, e.g., the MGF of $X \sim \text{Expo}(1)$ is $\frac{1}{1-t}$ for $t < 1$, which is the geometric series $\sum_{j=0}^{\infty} t^n = \sum_{j=0}^{\infty} n! \frac{t^n}{n!}$ for $|t| < 1$. So the nth moment of X is $n!$ (the coefficient of $\frac{t^n}{n!}$).

5.14 Conditional Expectation

The conditional expected value $E(Y|X = x)$ is a number (for each x) which is the average value of Y, given the information that $X = x$. The definition is analogous to the definition of EY: just replace the PMF or PDF by the conditional PMF or conditional PDF.

It is often very convenient to just directly condition on X to obtain $E(Y|X)$, which is a random variable (it is a function of X). This intuitively says to average Y, treating X as if it were a known constant: $E(Y|X = x)$ is a function of x, and $E(Y|X)$ is obtained from $E(Y|X = x)$ by "changing x to X". For example, if $E(Y|X = x) = x^3$, then $E(Y|X) = X^3$.

Important properties of conditional expectation:

$$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X) \text{ (Linearity)};$$

$$E(Y|X) = E(Y) \text{ if } X \text{ and } Y \text{ are independent};$$

$$E(h(X)Y|X) = h(X)E(Y|X) \text{ (Taking out what's known)};$$

$$E(Y) = E(E(Y|X)) \text{ (Iterated Expectation/Adam's Law)};$$

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X)) \text{ (Eve's Law)}.$$

The latter two identities are often useful for finding the mean and variance of Y: first condition on some choice of X where the conditional distribution of Y given X is easier to work with than the unconditional distribution of Y, and then account for the randomness of X.

5.15 Convergence

Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ and variance σ^2. The sample mean is defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The Strong Law of Large Numbers says that with probability 1, the sample mean converges to the true mean:

$$\bar{X}_n \to \mu \text{ with probability 1.}$$

The Weak Law of Large Numbers (which follows from Chebyshev's Inequality) says that $\bar{X}_n$ will be very close to μ with very high probability: for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty.$$

The Central Limit Theorem says that the sum of a large number of i.i.d. random variables is approximately Normal in distribution. More precisely, standardize the sum $X_1 + \cdots + X_n$ (by subtracting its mean and dividing by its standard deviation); then the standardized sum approaches $\mathcal{N}(0,1)$ in distribution (i.e., the CDF of the standardized sum converges to Φ). So

$$\frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1) \text{ in distribution.}$$

In terms of the sample mean,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \to \mathcal{N}(0,1) \text{ in distribution.}$$

5.16 Inequalities

When probabilities and expected values are hard to compute exactly, it is useful to have inequalities. One simple but handy inequality is Markov's Inequality:

$$P(X > a) \le \frac{E|X|}{a},$$

for any $a > 0$. Let X have mean μ and variance σ^2. Using Markov's Inequality with $(X-\mu)^2$ in place of X gives Chebyshev's Inequality:

$$P(|X - \mu| > a) \le \sigma^2/a^2.$$

For convex functions g (convexity of g is equivalent to $g''(x) \geq 0$ for all x, assuming this exists), there is Jensen's Inequality (the reverse inequality holds for concave g):

$$E(g(X)) \geq g(E(X)) \text{ for } g \text{ convex.}$$

The Cauchy-Schwarz inequality bounds the expected product of X and Y:

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}.$$

If X and Y have mean 0 and variance 1, this reduces to saying that the correlation is between -1 and 1. It follows that correlation is *always* between -1 and 1.

5.17 Markov Chains

Consider a Markov chain $X_0, X_1, \ldots$ with transition matrix $Q = (q_{ij})$, and let $\mathbf{v}$ be a row vector listing the initial probabilities of being in each state. Then $\mathbf{v}Q^n$ is the row vector listing the probabilities of being in each state after n steps, i.e., the jth component is $P(X_n = j)$.

A vector $\mathbf{s}$ of probabilities (adding to 1) is *stationary* for the chain if $\mathbf{s}Q = \mathbf{s}$; by the above, if a chain starts out with a stationary distribution then the distribution stays the same forever. Any irreducible Markov chain has a unique stationary distribution $\mathbf{s}$, and the chain converges to it: $P(X_n = i) \to s_i$ as $n \to \infty$.

If $\mathbf{s}$ is a vector of probabilities (adding to 1) that satisfies the reversibility condition $s_i q_{ij} = s_j q_{ji}$ for all states i, j, then it automatically follows that $\mathbf{s}$ is a stationary distribution for the chain; not all chains have this condition hold, but for those that do it is often easier to show that $\mathbf{s}$ is stationary using the reversibility condition than by showing $\mathbf{s}Q = \mathbf{s}$.

6 Common Mistakes in Probability

6.1 Category errors

A category error is a mistake that not only happens to be wrong, but also it is wrong in *every possible universe*. If someone answers the question "How many students are in Stat 110?" with "10, since it's one ten," that is wrong (and a very bad approximation to the truth); but there is no *logical* reason the enrollment couldn't be 10, aside from the logical necessity of learning probability for reasoning about uncertainty in the world. But answering the question with "-42" or "π" or "pink elephants" would be a category error. To help avoid being categorically wrong, always think about what type an answer should have. Should it be an integer? A nonnegative integer? A number between 0 and 1? A random variable? A distribution?

- Probabilities must be between 0 and 1.

  Example: When asked for an approximation to $P(X > 5)$ for a certain r.v. X with mean 7, writing "$P(X > 5) \approx E(X)/5$." This makes two mistakes: Markov's inequality gives $P(X > 5) \leq E(X)/5$, but this is an *upper bound*, not an approximation; and here $E(X)/5 = 1.4$, which is silly as an approximation to a probability since $1.4 > 1$.

- Variances must be nonnegative.

  Example: For X and Y independent r.v.s, writing that "$\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$", which can immediately be seen to be wrong from the fact that it becomes negative if $\text{Var}(Y) > \text{Var}(X)$ (and 0 if X and Y are i.i.d.). The correct formula is $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(-Y) - 2\text{Cov}(X, Y)$, which is $\text{Var}(X) + \text{Var}(Y)$ if X and Y are uncorrelated.

- Correlations must be between -1 and 1.

  Example: It is common to confuse covariance and correlation; they are related by $\text{Corr}(X, Y) = \text{Cov}(X, Y)/(\text{SD}(X)\text{SD}(Y))$, which is between -1 and 1.

- The range of possible values must make sense.

  Example: Two people each have 100 friends, and we are interested in the distribution of $X =$ (number of mutual friends). Then writing "$X \sim \mathcal{N}(\mu, \sigma^2)$" doesn't make sense since X is an *integer* (sometimes we use the Normal as an *approximation* to, say, Binomials, but exact answers should be given unless an approximation is specifically asked for); "$X \sim \text{Pois}(\lambda)$" or "$X \sim \text{Bin}(500, 1/2)$" don't make sense since X has possible values $0, 1, \dots, 100$.

- Units should make sense.

  Example: A common careless mistake is to divide by the variance rather than the standard deviation when standardizing. Thinking of X as having units makes it clear whether to divide by variance or standard deviation, e.g., if X is measured in light years, then $E(X)$ and $\mathrm{SD}(X)$ are also measured in light years (whereas $\mathrm{Var}(X)$ is measured in squared light years), so the standardized r.v. $\frac{X-E(X)}{\mathrm{SD}(X)}$ is unitless (as desired).

  Thinking about units also helps explain the change of variables formula,

  $$f_X(x)dx = f_Y(y)dy.$$

  If, for example, X is measured in nanoseconds and $Y = X^3$, then the units of $f_X(x)$ inverse nanoseconds and the units of $f_Y(y)$ are inverse cubed nanoseconds, and we need the dx and dy to make both sides be unitless (remember that we can think of $f_X(x)dx$ as the probability that X is in a tiny interval of length dx, centered at x).

- A number can't equal a random variable (unless the r.v. is actually a constant). Quantities such as $E(X), P(X > 1), F_X(1), \mathrm{Cov}(X, Y)$ are *numbers*. We often use the notation "$X = x$", but this is shorthand for an *event* (it is the set of all possible outcomes of the experiment where X takes the value x).

  Example: A store has $N \sim \mathrm{Pois}(\lambda)$ customers on a certain day, each of whom spends an average of μ dollars. Let X be the total amount spent by the customers. Then "$E(X) = N\mu$" doesn't make sense, since $E(X)$ is a number, while the righthand side is an r.v.

  Example: Writing something like "$\mathrm{Cov}(X, Y) = 3$ if $Z = 0$ and $\mathrm{Cov}(X, Y) = 1$ if $Z = 1$" doesn't make sense, as $\mathrm{Cov}(X, Y)$ is just one number. Similarly, students sometimes write "$E(Y) = 3$ when $X = 1$" when they mean $E(Y|X = 1) = 3$. This is both conceptually wrong since $E(Y)$ is a number, the overall average of Y, and careless notation that could lead, e.g., to getting "$E(X) = 1$ if $X = 1$, and $E(X) = 0$ if $X = 0$" rather than $EX = p$ for $X \sim \mathrm{Bern}(p)$.

- Don't replace a r.v. by its mean, or confuse $E(g(X))$ with $g(EX)$.

  Example: On the bidding for an unknown asset problem (#6 on the final from 2008), a common mistake is to replace the random asset value V by its mean, which completely ignores the *variability* of V.

Example: If $X - 1 \sim \text{Geom}(1/2)$, then $2^{E(X)} = 4$, but $E(2^X)$ is infinite (as in the St. Petersburg Paradox), so confusing the two is infinitely wrong. In general, if g is convex then Jensen's inequality says that $E(g(X)) \geq g(EX)$.

- An event is not a random variable.

  Example: If A is an event and X is an r.v., it does not make sense to write "$E(A)$" or "$P(X)$". There is of course a deep connection between events and r.v.s, in that for any event A there is a corresponding indicator r.v. I_A, and given an r.v. X and a number x, we have events $X = x$ and $X \leq x$.

- Dummy variables in an integral can't make their way out of the integral.

  Example: In LOTUS for a r.v. X with PDF f, the letter x in $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$ is a dummy variable; we could just as well write $\int_{-\infty}^{\infty} g(t)f(t)dt$ or $\int_{-\infty}^{\infty} g(u)f(u)du$ or even $\int_{-\infty}^{\infty} g(\Box)f(\Box)d\Box$, but the x (or whatever this dummy variable is called) can't migrate out of the integral.

- A random variable is not the same thing as its distribution! See Section 6.4.

- The conditional expectation $E(Y|X)$ must be a function of X (possibly a constant function, but it must be computable just in terms of X). See Section 6.5.

6.2 Notational paralysis

Another common mistake is a reluctance to introduce notation. This can be both a symptom and a cause of not seeing the structure of a problem. Be sure to define your notation clearly, carefully distinguishing between constants, random variables, and events.

- Give objects names if you want to work with them.

  Example: Suppose that we are interested in a LogNormal r.v. X (so $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$ for some μ, σ^2). Then $\log(X)$ is clearly an important object, so we should give it a name, say $Y = \log(X)$. Then, $X = e^Y$ and, for example, we can easily obtain the moments of X using the MGF of Y.

  Example: Suppose that we want to show that

$$E(\cos^4(X^2+1)) \geq (E(\cos^2(X^2+1)))^2.$$

  The essential pattern is that there is a r.v. on the right and its square on the left; so let $Y = \cos^2(X^2+1)$, which turns the desired inequality into the statement $E(Y^2) \geq (EY)^2$, which we know is true because variance is nonnegative.

- Introduce clear notation for events and r.v.s of interest.

  Example: In the Calvin and Hobbes problem (from HW 3 and the final from 2010), clearly the event "Calvin wins the match" is important (so give it a name) and the r.v. "how many of the first two games Calvin wins" is important (so give it a name). Make sure that events you define really are events (they are subsets of the sample space, and it must make sense to talk about whether the event occurs) and that r.v.s you define really are r.v.s (they are functions mapping the sample space to the real line, and it must make sense to talk about their distributions and talk about them as a numerical summary of some aspect of the random experiment).

- Think about location and scale when applicable.

  Example: If $Y_j \sim \text{Expo}(\lambda)$, it may be very convenient to work with $X_j = \lambda Y_j$, which is Expo(1). In studying $X \sim \mathcal{N}(\mu, \sigma^2)$, it may be very convenient to write $X = \mu + \sigma Z$ where $Z \sim \mathcal{N}(0,1)$ is the standardized version of X.

6.3 Common sense and checking answers

Whenever possible (i.e., when not under severe time pressure), look for simple ways to check your answers, or at least to check that they are plausible. This can be done in various ways, such as using the following methods.

1. *Miracle checks.* Does your answer seem intuitively plausible? Is there a category error? Did asymmetry appear out of nowhere when there should be symmetry?

2. *Checking simple and extreme cases.* What is the answer to a simpler version of the problem? What happens if $n = 1$ or $n = 2$, or as $n \to \infty$, if the problem involves showing something for all n?

3. *Looking for alternative approaches and connections with other problems.* Is there another natural way to think about the problem? Does the problem relate to other problems we've seen?

- Probability is full of counterintuitive results, but not impossible results!

  Example: Suppose that we have $P(\text{snow Saturday}) = P(\text{snow Sunday}) = 1/2$. Then we can't say "$P(\text{snow over the weekend}) = 1$"; clearly there is *some* chance of no snow, and of course the mistake is to ignore the need for disjointness.

Example: In finding $E(e^X)$ for $X \sim \text{Pois}(\lambda)$, obtaining an answer that can be negative, or an answer that isn't an increasing function of λ (intuitively, it is clear that larger λ should give larger average values of e^X).

- Check simple and extreme cases whenever possible.

  Example: Suppose we want to derive the mean and variance of a Hypergeometric, which is the distribution of the number of white balls if we draw n balls without replacement from a bag containing w white balls and b black balls. Suppose that using indicator r.v.s, we (correctly) obtain that the mean is $\mu = \frac{nw}{w+b}$ and the variance is $(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1-\frac{\mu}{n})$.

  Let's check that this makes sense for the simple case $n = 1$: then the mean and variance reduce to those of a $\text{Bern}(w/(w+b))$, which makes sense since with only 1 draw, it doesn't matter whether sampling is with replacement.

  Now let's consider an extreme case where the total number of balls $(w+b)$ is extremely large compared with n. Then it shouldn't matter much whether the sampling is with or without replacement, so the mean and variance should be very close to those of a $\text{Bin}(n, w/(b+w))$, and indeed this is the case. If we had an answer that did not make sense in simple and extreme cases, we could then look harder for a mistake or explanation.

  Example: Let $X_1, X_2, \dots, X_{1000}$ be i.i.d. with a continuous distribution, and consider the question of whether the event $X_1 < X_2$ is independent of the event $X_1 < X_3$. Many students guess intuitively that they are independent. But now consider the more extreme question of whether $P(X_1 < X_2 | X_1 < X_3, X_1 < X_4, \dots, X_1 < X_{1000})$ is $P(X_1 < X_2)$. Here most students guess intuitively (and correctly) that

$$P(X_1 < X_2 | X_1 < X_3, X_1 < X_4, \dots, X_1 < X_{1000}) > P(X_1 < X_2),$$

  since the evidence that X_1 is less than all of $X_3, \dots, X_{1000}$ suggests that X_1 is very small. Yet this more extreme case is the same in principle, just different in degree. Similarly, the Monty Hall problem is easier to understand with 1000 doors than with 3 doors. To show algebraically that $X_1 < X_2$ is not independent of $X_1 < X_3$, note that $P(X_1 < X_2) = 1/2$, while

$$P(X_1 < X_2 | X_1 < X_3) = \frac{P(X_1 < X_2, X_1 < X_3)}{P(X_1 < X_3)} = \frac{1/3}{1/2} = \frac{2}{3},$$

  where the numerator is 1/3 since the smallest of X_1, X_2, X_3 is equally likely to be any of them.

- Check that PMFs are nonnegative and sum to 1, and PDFs are nonnegative and integrate to 1 (or that it is at least plausible), when it is not too messy.

  Example: Writing that the PDF of X is "$f(x) = \frac{1}{5}e^{-5x}$ for all $x > 0$ (and 0 otherwise)" is immediately seen to be wrong by integrating (the constant in front should be 5, which can also be seen by recognizing this as an Expo(5). Writing that the PDF is "$f(x) = \frac{1+e^{-x}}{1+x}$ for all $x > 0$ (and 0 otherwise)" doesn't make sense since even though the integral is hard to do directly, clearly $\frac{1+e^{-x}}{1+x} > \frac{1}{1+x}$, and $\int_0^\infty \frac{1}{1+x}dx$ is infinite.

  Example: Consider the following problem: "You are invited to attend 6 weddings next year, independently with all months of the year equally likely. What is the probability that no two weddings are in the same month?" A common mistake is to treat the weddings as indistinguishable. But no matter how generic and cliched weddings can be sometimes, there must be *some* way to distinguish two weddings!

  It often helps to make up concrete names, e.g., saying "ok, we need to look at the possible schedulings of the weddings of Daenerys and Drogo, of Cersei and Robert, ...". There are 12^6 equally likely possibilities and, for example, it is much more likely to have 1 wedding per month in January through June than to have all 6 weddings in January (whereas treating weddings as indistinguishable would suggest having these be equal).

6.4 Random variables vs. distributions

A random variable is not the same thing as its distribution! We call this confusion *sympathetic magic*, and the consequences of this confusion are often disastrous. Every random variable has a distribution (which can always be expressed using a CDF, which can be expressed by a PMF in the discrete case, and which can be expressed by a PDF in the continuous case).

Every distribution can be used as a blueprint for generating r.v.s (for example, one way to do this is using Universality of the Uniform). But that doesn't mean that doing something to a r.v. corresponds to doing it to the distribution of the r.v. Confusing a distribution with a r.v. with that distribution is like confusing a map of a city with the city itself, or a blueprint of a house with the house itself. *The word is not the thing, the map is not the territory.*

- A function of a r.v. is a r.v.

Example: Let X be discrete with possible values $0, 1, 2, \dots$ and PMF $p_j = P(X = j)$, and let $Y = X + 3$. Then Y is discrete with possible values $3, 4, 5, \dots$, and its PMF is given by $P(Y = k) = P(X = k - 3) = p_{k-3}$ for $k \in \{3, 4, 5, \dots\}$. In the continuous case, if $Y = g(X)$ with g differentiable and strictly increasing, then we can use the change of variables formula to find the PDF of Y from the PDF of X. If we only need $E(Y)$ and not the distribution of Y, we can use LOTUS. A common mistake is not seeing why these transformations of X are themselves r.v.s and how to handle them.

Example: For X, Y i.i.d., writing that "$E \max(X, Y) = EX$ since $\max(X, Y)$ is either X or Y, both of which have mean EX"; this misunderstands how and why $\max(X, Y)$ is a r.v. Of course, we should have $E \max(X, Y) \geq EX$ since $\max(X, Y) \geq X$.

- Avoid sympathetic magic.

  Example: Is it possible to have two r.v.s X, Y which have the same distribution but are *never* equal, i.e., the event $X = Y$ never occurs?

  Example: In finding the PDF of XY, writing something like "$f_X(x)f_Y(y)$." This is a category error since if we let $W = XY$, we want a function $f_W(w)$, not a function of two variables x and y. The mistake is in thinking that the PDF of the product is the product of the PDFs, which comes from not understanding well what a distribution really is.

  Example: For r.v.s X and Y with PDFs f_X and f_Y respectively, the event $\{X < Y\}$ is very different conceptually from the inequality $f_X < f_Y$. In fact, it is impossible that for all t, $f_X(t) < f_Y(t)$, since both sides integrate to 1.

- A CDF $F(x) = P(X \leq x)$ is a way to specify the distribution of X, and is a function defined for all real values of x. Here X is the r.v., and x is any number; we could just as well have written $F(t) = P(X \leq t)$.

  Example: Why must a CDF $F(x)$ be defined for all x and increasing everywhere, and why is it *not* true that a CDF integrates to 1?

6.5 Conditioning

It is easy to make mistakes with conditional probability so it is important to think carefully about what to condition on and how to carry that out. *Conditioning is the soul of statistics.*

- Condition on *all* the evidence!

  Example: In the Monty Hall problem, if Monty opens door 2 then we can't just use "$P(\text{door 1 has car}|\text{door 2 doesn't have car}) = 1/2$," since this does not condition on all the evidence: we know not just that door 2 does not have the car, but also that Monty opened door 2. *How the information was collected is itself information.* Why is this additional information relevant? To see this, contrast the problem as stated with the variant where Monty randomly chooses to open one of the 2 doors not picked by the contestant (so there is a chance of revealing the car and spoiling the game): different information is obtained in the two scenarios. This is another example where looking at an extreme case helps (consider the analogue of the Monty Hall problem with a billion doors).

  Example: In the murder problem, a common mistake (often made by defense attorneys, intentionally or otherwise) is to focus attention on $P(\text{murder}|\text{abuse})$, which is irrelevant since we know the woman has been murdered, and we are interested in the probability of guilt given all the evidence (including the fact that the murder occurred).

- Don't destroy information.

  Example: Let $X \sim \text{Bern}(1/2)$ and $Y = 1 + W$ with $W \sim \text{Bern}(1/2)$ independent of X. Then writing "$E(X^2|X = Y) = E(Y^2) = 2.5$" is wrong (in fact, $E(X^2|X = Y) = 1$ since if $X = Y$, then $X = 1$), where the mistake is destroying the information $X = Y$, thinking we're done with that information once we have plugged in Y for X. A similar mistake is easy to make in the two envelope paradox.

  Example: On the bidding for an unknown asset problem (#6 on the final from 2008), a very common mistake is to forget to condition on the bid being accepted. In fact, should have $E(V|\text{bid accepted}) < E(V)$ since if the bid is accepted, it restricts how much the asset could be worth (intuitively, this is similar to"buyer's remorse": it is common (though not necessarily rational) for someone to regret making an offer if the offer is accepted immediately, thinking that is a sign that a lower offer would have sufficed).

- Independence shouldn't be assumed without justification, and it is important to be careful not to implicitly assume independence without justification.

  Example: For $X_1, \dots, X_n$ i.i.d., we have $\text{Var}(X_1 + ... + X_n) = n\text{Var}(X_1)$, but this is not equal to $\text{Var}(X_1 + ... + X_1) = \text{Var}(nX_1) = n^2\text{Var}(X_1)$. For example, if X and Y are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $X + Y \sim \mathcal{N}(2\mu, 2\sigma^2)$, while $X + X = 2X \sim \mathcal{N}(2\mu, 4\sigma^2)$.

Example: Is it always true that if $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda)$, then $X + Y \sim \text{Pois}(2\lambda)$? What is an example of a sum of Bern(p)'s which is not Binomial?

Example: In the two envelope paradox, it is not true that the amount of money in the first envelope is independent of the indicator of which envelope has more money.

- Independence is completely different from disjointness!

  Example: Sometimes students try to visualize independent events A and B with two non-overlapping ovals in a Venn diagram. Such events in fact *can't* be independent (unless one has probability 0), since learning that A happened gives a great deal of information about B: it implies that B did not occur.

- Independence is a symmetric property: if A is independent of B, then B is independent of A. *There's no such thing as unrequited independence.*

  Example: If it is non-obvious whether A provides information about B but obvious that B provides information about A, then A and B can't be independent.

- The marginal distributions can be extracted from the joint distribution, but knowing the marginal distributions does not determine the joint distribution.

  Example: Calculations that are purely based on the marginal CDFs F_X and F_Y of dependent r.v.s X and Y may not shed much light on events such as $\{X < Y\}$ which involve X and Y jointly.

- Keep the distinction between prior and posterior probabilities clear.

  Example: Suppose that we observe evidence E. Then writing "$P(E) = 1$ since we know for sure that E happened" is careless; we have $P(E|E) = 1$, but $P(E)$ is the *prior probability* (the probability before E was observed).

- Don't confuse $P(A|B)$ with $P(B|A)$.

  Example: This mistake is also known as the *prosecutor's fallacy* since it is often made in legal cases (but not always by the prosecutor!). For example, the prosecutor may argue that the probability of guilt given the evidence is very high by attempting to show that the probability of the evidence given innocence is very low, but in and of itself this is insufficient since it does not use the prior probability of guilt. Bayes' rule thus becomes *Bayes' ruler*, measuring the weight of the evidence by relating $P(A|B)$ to $P(B|A)$ and showing us how to update our beliefs based on evidence.

- Don't confuse $P(A|B)$ with $P(A, B)$.

  Example: The law of total probability is often wrongly written without the weights as "$P(A) = P(A|B) + P(A|B^c)$" rather than $P(A) = P(A, B) + P(A, B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

- The expression $Y|X$ does not denote a r.v.; it is notation indicating that in working with Y, we should use the conditional distribution of Y given X (i.e., treat X as a known constant). The expression $E(Y|X)$ *is* a r.v., and is a function of X (we have summed or integrated over the possible values of Y).

  Example: Writing "$E(Y|X) = Y$" is wrong, except if Y is a function of X, e.g., $E(X^3|X) = X^3$; by definition, $E(Y|X)$ must be $g(X)$ for some function g, so any answer for $E(Y|X)$ that is not of this form is a category error.

7 Stat 110 Final from 2006

1. The number of fish in a certain lake is a Pois(λ) random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a Pois(λ) random variable).

(a) Find $E(Y^2)$ (simplify).

(b) Find $E(1/Y)$ (in terms of λ; do not simplify yet).

(c) Find a simplified expression for $E(1/Y)$. Hint: $k!(k+1) = (k+1)!$.

2. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

In (c) through (f), X and Y are i.i.d. (independent identically distributed) positive random variables. Assume that the various expected values exist.

(a) (probability that a roll of 2 fair dice totals 9) ____ (probability that a roll of 2 fair dice totals 10)

(b) (probability that 65% of 20 children born are girls) ____ (probability that 65% of 2000 children born are girls)

(c) $E(\sqrt{X})$ ____ $\sqrt{E(X)}$

(d) $E(\sin X)$ ____ $\sin(EX)$

(e) $P(X+Y>4)$ ____ $P(X>2)P(Y>2)$

(f) $E\left((X+Y)^2\right)$ ____ $2E(X^2)+2(EX)^2$

3. A fair die is rolled twice, with outcomes X for the 1st roll and Y for the 2nd roll.

(a) Compute the covariance of $X+Y$ and $X-Y$ (simplify).

(b) Are $X+Y$ and $X-Y$ independent? Justify your answer clearly.

(c) Find the moment generating function $M_{X+Y}(t)$ of $X+Y$ (your answer should be a function of t and can contain unsimplified finite sums).

4. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? (Simplify.) Justify your answer. Hint: no integrals are needed.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

(c) What is the expected total time that Alice needs to spend at the post office?

5. Bob enters a casino with $X_0 = 1$ dollar and repeatedly plays the following game: with probability 1/3, the amount of money he has increases by a factor of 3; with probability 2/3, the amount of money he has decreases by a factor of 3. Let X_n be the amount of money he has after playing this game n times. Thus, X_{n+1} is $3X_n$ with probability 1/3 and is $3^{-1}X_n$ with probability 2/3.

(a) Compute $E(X_1)$, $E(X_2)$ and, in general, $E(X_n)$. (Simplify.)

(b) What happens to $E(X_n)$ as $n \to \infty$? Let Y_n be the number of times out of the first n games that Bob triples his money. What happens to Y_n/n as $n \to \infty$?

(c) Does X_n converge to some number c as $n \to \infty$ (with probability 1) and if so, what is c? Explain.

6. Let X and Y be independent standard Normal r.v.s and let $R^2 = X^2 + Y^2$ (where $R > 0$ is the distance from (X, Y) to the origin).

(a) The distribution of R^2 is an example of three of the "important distributions" listed on the last page. State which three of these distributions R^2 is an instance of, specifying the parameter values.

(b) Find the PDF of R. (Simplify.) Hint: start with the PDF $f_W(w)$ of $W = R^2$.

(c) Find $P(X > 2Y + 3)$ in terms of the standard Normal CDF Φ. (Simplify.)

(d) Compute $\text{Cov}(R^2, X)$. Are R^2 and X independent?

7. Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. Unif(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$.

(a) Which important distribution is the distribution of X very close to? Specify what the parameters are, and state which theorem justifies your choice.

(b) Give a simple but accurate approximation for $P(X > 17)$. Justify briefly.

(c) Find the moment generating function (MGF) of X.

8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $E(X_1) = 3$, and consider the sum $S_n = X_1 + X_2 + \cdots + X_n$.

(a) What is $E(X_1 X_2 X_3 | X_1)$? (Simplify. Your answer should be a function of X_1.)

(b) What is $E(X_1|S_n) + E(X_2|S_n) + \cdots + E(X_n|S_n)$? (Simplify.)

(c) What is $E(X_1|S_n)$? (Simplify.) Hint: use (b) and symmetry.

9. An urn contains red, green, and blue balls. Balls are chosen randomly with replacement (each time, the color is noted and then the ball is put back.) Let r, g, b be the probabilities of drawing a red, green, blue ball respectively ($r + g + b = 1$).

(a) Find the expected number of balls chosen before obtaining the first red ball, not including the red ball itself. (Simplify.)

(b) Find the expected number of different *colors* of balls obtained before getting the first red ball. (Simplify.)

(c) Find the probability that at least 2 of n balls drawn are red, given that at least 1 is red. (Simplify; avoid sums of large numbers of terms, and $\sum$ or $\cdots$ notation.)

10. Let $X_0, X_1, X_2, \dots$ be an irreducible Markov chain with state space $\{1, 2, \dots, M\}$, $M \geq 3$, transition matrix $Q = (q_{ij})$, and stationary distribution $\mathbf{s} = (s_1, \dots, s_M)$. The initial state X_0 is given the stationary distribution, i.e., $P(X_0 = i) = s_i$.

(a) On average, how many of $X_0, X_1, \dots, X_9$ equal 3? (In terms of $\mathbf{s}$; simplify.)

(b) Let $Y_n = (X_n - 1)(X_n - 2)$. For $M = 3$, find an example of Q (the transition matrix for the *original* chain $X_0, X_1, \dots$) where $Y_0, Y_1, \dots$ is Markov, and another example of Q where $Y_0, Y_1, \dots$ is not Markov. Mark which is which and briefly explain. In your examples, make $q_{ii} > 0$ for at least one i and make sure it is possible to get from any state to any other state eventually.

(c) If each column of Q sums to 1, what is $\mathbf{s}$? Verify using the definition of *stationary*.

8 Stat 110 Final from 2007

1. Consider the birthdays of 100 people. Assume people's birthdays are independent, and the 365 days of the year (exclude the possibility of February 29) are equally likely.

(a) Find the expected number of birthdays represented among the 100 people, i.e., the expected number of days that at least 1 of the people has as his or her birthday (your answer can involve unsimplified fractions but should not involve messy sums).

(b) Find the covariance between how many of the people were born on January 1 and how many were born on January 2.

2. Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $(E(XY))^2$ ____ $E(X^2)E(Y^2)$

(b) $P(|X+Y|>2)$ ____ $\frac{1}{10}E((X+Y)^4)$

(c) $E(\ln(X+3))$ ____ $\ln(E(X+3))$

(d) $E(X^2e^X)$ ____ $E(X^2)E(e^X)$

(e) $P(X+Y=2)$ ____ $P(X=1)P(Y=1)$

(f) $P(X+Y=2)$ ____ $P\left(\{X \geq 1\} \cup \{Y \geq 1\}\right)$

3. Let X and Y be independent Pois(λ) random variables. Recall that the moment generating function (MGF) of X is $M(t) = e^{\lambda(e^t-1)}$.

(a) Find the MGF of $X + 2Y$ (simplify).

(b) Is $X + 2Y$ also Poisson? Show that it is, or that it isn't (whichever is true).

(c) Let $g(t) = \ln M(t)$ be the log of the MGF of X. Expanding $g(t)$ as a Taylor series

$$g(t) = \sum_{j=1}^{\infty} \frac{c_j}{j!} t^j$$

(the sum starts at $j = 1$ because $g(0) = 0$), the coefficient c_j is called the jth *cumulant* of X. Find c_j in terms of λ, for all $j \geq 1$ (simplify).

4. Consider the following conversation from an episode of *The Simpsons*:

> Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
>
> Homer: *Lisa, a guy who's got lots of ivory is less likely to hurt Stampy than a guy whose ivory supplies are low.*

Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

(a) Define clear notation for the various events of interest here.

(b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

(c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

5. Empirically, it is known that 49% of children born in the U.S. are girls (and 51% are boys). Let N be the number of children who will be born in the U.S. in March 2009, and assume that N is a Pois(λ) random variable, where λ is known. Assume that births are independent (e.g., don't worry about identical twins).

Let X be the number of girls who will be born in the U.S. in March 2009, and let Y be the number of boys who will be born then (note the importance of choosing good notation: boys have a Y chromosome).

(a) Find the joint distribution of X and Y. (Give the joint PMF.)

(b) Find $E(N|X)$ and $E(N^2|X)$.

6. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

(b) Find $P\left(X_1 = \min(X_1, X_2, X_3)\right)$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

7. Let $X_1, X_2, \ldots$ be i.i.d. random variables with CDF $F(x)$. For every number x, let $R_n(x)$ count how many of $X_1, \ldots, X_n$ are less than or equal to x.

(a) Find the mean and variance of $R_n(x)$ (in terms of n and $F(x)$).

(b) Assume (for this part only) that $X_1, \ldots, X_4$ are known constants. Sketch an example showing what the graph of the function $\frac{R_4(x)}{4}$ might look like. Is the function $\frac{R_4(x)}{4}$ necessarily a CDF? Explain briefly.

(c) Show that $\frac{R_n(x)}{n} \to F(x)$ as $n \to \infty$ (with probability 1).

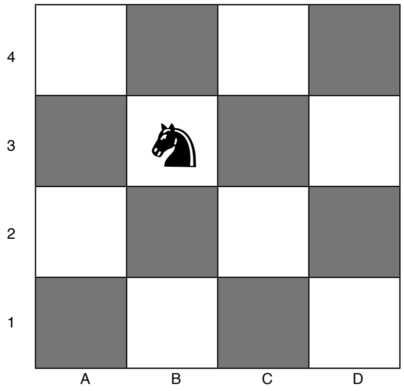
8. (a) Let T be a Student-t r.v. with 1 degree of freedom, and let $W = 1/T$. Find the PDF of W (simplify). Is this one of the "important distributions"?

Hint: no calculus is needed for this (though it can be used to check your answer).

(b) Let $W_n \sim \chi^2_n$ (the Chi-Square distribution with n degrees of freedom), for each $n \geq 1$. Do there exist a_n and b_n such that $a_n(W_n - b_n) \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$? If so, find them; if not, explain why not.

(c) Let $Z \sim \mathcal{N}(0,1)$ and $Y = |Z|$. Find the PDF of Y, and approximate $P(Y < 2)$.

9. Consider a knight randomly moving around on a 4 by 4 chessboard:



The 16 squares are labeled in a grid, e.g., the knight is currently at the square B3, and the upper left square is A4. Each move of the knight is an L-shape: two squares horizontally followed by one square vertically, or vice versa. For example, from B3 the knight can move to A1, C1, D2, or D4; from A4 it can move to B2 or C3. Note that from a white square, the knight always moves to a gray square and vice versa.

At each step, the knight moves randomly, each possibility equally likely. Consider the stationary distribution of this Markov chain, where the states are the 16 squares.

(a) Which squares have the highest stationary probability? Explain very briefly.

(b) Compute the stationary distribution (simplify). Hint: random walk on a graph.

9 Stat 110 Final from 2008

1. Joe's iPod has 500 different songs, consisting of 50 albums of 10 songs each. He listens to 11 random songs on his iPod, with all songs equally likely and chosen independently (so repetitions may occur).

(a) What is the PMF of how many of the 11 songs are from his favorite album?

(b) What is the probability that there are 2 (or more) songs from the same album among the 11 songs he listens to? (Do not simplify.)

(c) A pair of songs is a "match" if they are from the same album. If, say, the 1st, 3rd, and 7th songs are all from the same album, this counts as 3 matches. Among the 11 songs he listens to, how many matches are there on average? (Simplify.)

2. Let X and Y be *positive* random variables, *not necessarily independent.* Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $P(X+Y>2)$ ____ $\frac{EX+EY}{2}$

(b) $P(X+Y>3)$ ____ $P(X>3)$

(c) $E(\cos(X))$ ____ $\cos(EX)$

(d) $E(X^{1/3})$ ____ $(EX)^{1/3}$

(e) $E(X^Y)$ ____ $(EX)^{EY}$

(f) $E\left(E(X|Y)+E(Y|X)\right)$ ____ $EX+EY$

3. (a) A woman is pregnant with twin boys. Twins may be either identical or fraternal (non-identical). In general, 1/3 of twins born are identical. Obviously, identical twins must be of the same sex; fraternal twins may or may not be. Assume that identical twins are equally likely to be both boys or both girls, while for fraternal twins all possibilities are equally likely. Given the above information, what is the probability that the woman's twins are identical?

(b) A certain genetic characteristic is of interest. For a random person, this has a numerical value given by a $\mathcal{N}(0, \sigma^2)$ r.v. Let X_1 and X_2 be the values of the genetic characteristic for the twin boys from (a). If they are identical, then $X_1 = X_2$; if they are fraternal, then X_1 and X_2 have correlation ρ. Find $\text{Cov}(X_1, X_2)$ in terms of ρ, σ^2.

4. (a) Consider i.i.d. Pois(λ) r.v.s $X_1, X_2, \dots$. The MGF of X_j is $M(t) = e^{\lambda(e^t-1)}$. Find the MGF $M_n(t)$ of the sample mean $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. (Hint: it may help to do the $n = 2$ case first, which itself is worth a lot of partial credit, and then generalize.)

(b) Find the limit of $M_n(t)$ as $n \to \infty$. (You can do this with almost no calculation using a relevant theorem; or you can use (a) and that $e^x \approx 1 + x$ if x is very small.)

5. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? Justify your answer. Hint: no integrals are needed.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

(c) What is the expected total time that Alice needs to spend at the post office?

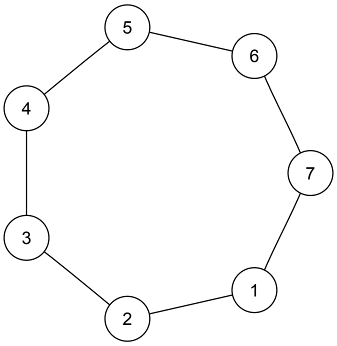
6. You are given an amazing opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown, except that it is worth at least nothing, and at most a million dollars. So the true value V of the prize is considered to be Uniform on [0,1] (measured in millions of dollars).

You can choose to bid any amount b (in millions of dollars). You have the chance to get the prize for considerably less than it is worth, but you could also lose money if you bid too much. Specifically, if $b < \frac{2}{3}V$, then the bid is rejected and nothing is gained or lost. If $b \geq \frac{2}{3}V$, then the bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V). What is your optimal bid b (to maximize the expected payoff)?

7. (a) Let $Y = e^X$, with $X \sim \text{Expo}(3)$. Find the mean and variance of Y (simplify).

(b) For $Y_1, \dots, Y_n$ i.i.d. with the same distribution as Y from (a), what is the approximate distribution of the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$ when n is large? (Simplify, and specify all parameters.)

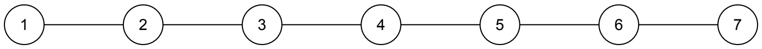
8.



(a) Consider a Markov chain on the state space $\{1, 2, \ldots, 7\}$ with the states arranged in a "circle" as shown above, and transitions given by moving one step clockwise or counterclockwise with equal probabilities. For example, from state 6, the chain moves to state 7 or state 5 with probability 1/2 each; from state 7, the chain moves to state 1 or state 6 with probability 1/2 each. The chain starts at state 1.

Find the stationary distribution of this chain.

(b) Consider a new chain obtained by "unfolding the circle." Now the states are arranged as shown below. From state 1 the chain always goes to state 2, and from state 7 the chain always goes to state 6. Find the new stationary distribution.



10 Stat 110 Final from 2009

1. A group of n people play "Secret Santa" as follows: each puts his or her name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's own name, so some may have to buy gifts for themselves (on the bright side, some may like self-selected gifts better). Assume $n \geq 2$.

(a) Find the expected number of people who pick their own names (simplify).

(b) Find the expected number of pairs of people, A and B, such that A picks B's name and B picks A's name (where $A \neq B$ and order doesn't matter; simplify).

(c) Let X be the number of people who pick their own names. Which of the "important distributions" are conceivable as the distribution of X, just based on the possible values X takes (you do not need to list parameter values for this part)?

(d) What is the *approximate* distribution of X if n is large (specify the parameter value or values)? What does $P(X = 0)$ converge to as $n \to \infty$?

2. Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(X^3)$ ____ $\sqrt{E(X^2)E(X^4)}$

(b) $P(|X+Y|>2)$ ____ $\frac{1}{16}E((X+Y)^4)$

(c) $E(\sqrt{X+3})$ ____ $\sqrt{E(X+3)}$

(d) $E(\sin^2(X))+E(\cos^2(X))$ ____ 1

(e) $E(Y|X+3)$ ____ $E(Y|X)$

(f) $E(E(Y^2|X))$ ____ $(EY)^2$

3. Let $Z \sim \mathcal{N}(0,1)$. Find the 4th moment $E(Z^4)$ in the following two different ways:

(a) using what you know about how certain powers of Z are related to other distributions, along with information from the table of distributions.

(b) using the MGF $M(t) = e^{t^2/2}$, by writing down its Taylor series and using how the coefficients relate to moments of Z, *not* by tediously taking derivatives of $M(t)$. Hint: you can get this series immediately from the Taylor series for e^x.

4. A chicken lays n eggs. Each egg independently does or doesn't hatch, with probability p of hatching. For each egg that hatches, the chick does or doesn't survive (independently of the other eggs), with probability s of survival. Let $N \sim \text{Bin}(n, p)$ be the number of eggs which hatch, X be the number of chicks which survive, and Y be the number of chicks which hatch but don't survive (so $X + Y = N$).

(a) Find the distribution of X, preferably with a clear explanation in words rather than with a computation. If X has one of the "important distributions," say which (including its parameters).

(b) Find the joint PMF of X and Y (simplify).

(c) Are X and Y independent? Give a clear explanation in words (of course it makes sense to see if your answer is consistent with your answer to (b), but you can get full credit on this part even without doing (b); conversely, it's not enough to just say "by (b), ..." without further explanation).

5. Suppose we wish to approximate the following integral (denoted by b):

$$b = \int_{-\infty}^{\infty} (-1)^{\lfloor x \rfloor} e^{-x^2/2} dx,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x (e.g., $\lfloor 3.14 \rfloor = 3$).

(a) Write down a function $g(x)$ such that $E(g(X)) = b$ for $X \sim \mathcal{N}(0,1)$ (your function should *not* be in terms of b, and should handle normalizing constants carefully).

(b) Write down a function $h(u)$ such that $E(h(U)) = b$ for $U \sim \text{Unif}(0,1)$ (your function should *not* be in terms of b, and can be in terms of the function g from (a) and the standard Normal CDF Φ).

(c) Let $X_1, X_2, \dots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ with n large, and let g be as in (a). What is the approximate distribution of $\frac{1}{n}(g(X_1) + \cdots + g(X_n))$? Simplify the parameters fully (in terms of b and n), and mention which theorems you are using.

6. Let X_1 be the number of emails received by a certain person today and let X_2 be the number of emails received by that person tomorrow, with X_1 and X_2 i.i.d.

(a) Find $E(X_1|X_1 + X_2)$ (simplify).

(b) For the case $X_j \sim \text{Pois}(\lambda)$, find the conditional distribution of X_1 given X_1+X_2, i.e., $P(X_1 = k|X_1+X_2 = n)$ (simplify). Is this one of the "important distributions"?

7. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

8. Let X_n be the price of a certain stock at the start of the nth day, and assume that $X_0, X_1, X_2, \dots$ follows a Markov chain with transition matrix Q (assume for simplicity that the stock price can never go below 0 or above a certain upper bound, and that it is always rounded to the nearest dollar).

(a) A lazy investor only looks at the stock once a year, observing the values on days $0, 365, 2 \cdot 365, 3 \cdot 365, \dots$. So the investor observes $Y_0, Y_1, \dots,$ where Y_n is the price after n years (which is $365n$ days; you can ignore leap years). Is $Y_0, Y_1, \dots$ also a Markov chain? Explain why or why not; if so, what is its transition matrix?

(b) The stock price is always an integer between \$0 and \$28. From each day to the next, the stock goes up or down by \$1 or \$2, all with equal probabilities (except for days when the stock is at or near a boundary, i.e., at \$0, \$1, \$27, or \$28).

If the stock is at \$0, it goes up to \$1 or \$2 on the next day (after receiving government bailout money). If the stock is at \$28, it goes down to \$27 or \$26 the next day. If the stock is at \$1, it either goes up to \$2 or \$3, or down to \$0 (with equal probabilities); similarly, if the stock is at \$27 it either goes up to \$28, or down to \$26 or \$25. Find the stationary distribution of the chain (simplify).

11 Stat 110 Final from 2010

1. Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match.

(a) What is the probability that Calvin wins the match (in terms of p)?

Hint: condition on the results of the first k games (for some choice of k).

(b) Find the expected number of games played.

Hint: consider the first two games as a pair, then the next two as a pair, etc.

2. A DNA sequence can be represented as a sequence of letters, where the “alphabet” has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_1, p_2, p_3, p_4 respectively.

(a) In a DNA sequence of length 115, what is the expected number of occurrences of the expression “CATCAT” (in terms of the p_j)? (Note that, for example, the expression “CATCATCAT” counts as 2 occurrences.)

(b) What is the probability that the first A appears earlier than the first C appears, as letters are generated one by one (in terms of the p_j)?

(c) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a $\text{Unif}(0,1)$ r.v. before observing any data, and that then the first 3 letters observed are “CAT”. Given this information, what is the probability that the next letter is C?

3. Let X and Y be i.i.d. *positive* random variables. Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general). It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(e^{X+Y})$ ____ $e^{2E(X)}$

(b) $E(X^2e^X)$ ____ $\sqrt{E(X^4)E(e^{2X})}$

(c) $E(X|3X)$ ____ $E(X|2X)$

(d) $E(X^7Y)$ ____ $E(X^7E(Y|X))$

(e) $E(\frac{X}{Y}+\frac{Y}{X})$ ____ 2

(f) $P(|X-Y|>2)$ ____ $\frac{\text{Var}(X)}{2}$

4. Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \ldots,$ and let $p_k = P(X = x_k)$. The *entropy* of X is defined to be $H(X) = -\sum_{k=0}^{\infty} p_k \log_2(p_k)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

Hint: use properties of logs, and interpret part of the sum as an expected value.

(b) Find $H(X^3)$ for $X \sim \text{Geom}(p)$, in terms of $H(X)$.

(c) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

Hint: Consider $E(\log_2(W))$, where W is a r.v. taking value p_k with probability p_k.

5. Let $Z_1, \dots, Z_n \sim \mathcal{N}(0,1)$ be i.i.d.

(a) As a function of Z_1, create an Expo(1) r.v. X (your answer can also involve the standard Normal CDF Φ).

(b) Let $Y = e^{-R}$, where $R = \sqrt{Z_1^2 + \cdots + Z_n^2}$. Write down (but do not evaluate) an integral for $E(Y)$.

(c) Let $X_1 = 3Z_1 - 2Z_2$ and $X_2 = 4Z_1 + 6Z_2$. Determine whether X_1 and X_2 are independent (being sure to mention which results you're using).

6. Let $X_1, X_2, \dots$ be i.i.d. positive r.v.s. with mean μ, and let $W_n = \frac{X_1}{X_1 + \dots + X_n}$.

(a) Find $E(W_n)$.

Hint: consider $\frac{X_1}{X_1+\dots+X_n} + \frac{X_2}{X_1+\dots+X_n} + \dots + \frac{X_n}{X_1+\dots+X_n}$.

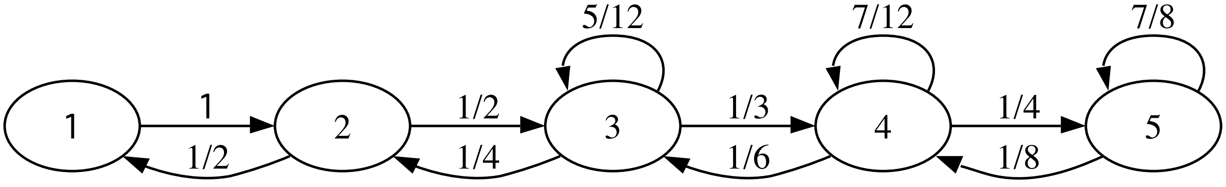
(b) What random variable does nW_n converge to as $n \to \infty$?

(c) For the case that $X_j \sim \text{Expo}(\lambda)$, find the distribution of W_n, preferably without using calculus. (If it is one of the "important distributions" state its name and specify the parameters; otherwise, give the PDF.)

7. A task is randomly assigned to one of two people (with probability 1/2 for each person). If assigned to the first person, the task takes an Expo(λ_1) length of time to complete (measured in hours), while if assigned to the second person it takes an Expo(λ_2) length of time to complete (independent of how long the first person would have taken). Let T be the time taken to complete the task.

(a) Find the mean and variance of T.

(b) Suppose instead that the task is assigned to *both* people, and let X be the time taken to complete it (by whoever completes it first, with the two people working independently). It is observed that after 24 hours, the task has not yet been completed. Conditional on this information, what is the expected value of X?



8. Find the stationary distribution of the Markov chain shown above, *without using matrices*. The number above each arrow is the corresponding transition probability.

Stat 110 Final Review Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Solutions to Stat 110 Final from 2006

1. The number of fish in a certain lake is a Pois(λ) random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a Pois(λ) random variable).

(a) Find $E(Y^2)$ (simplify).

We have $Y = X + 1$ with $X \sim \text{Pois}(\lambda)$, so $Y^2 = X^2 + 2X + 1$. So

$$E(Y^2) = E(X^2 + 2X + 1) = E(X^2) + 2E(X) + 1 = (\lambda + \lambda^2) + 2\lambda + 1 = \lambda^2 + 3\lambda + 1,$$

since $E(X^2) = \text{Var}(X) + (EX)^2 = \lambda + \lambda^2$.

(b) Find $E(1/Y)$ (in terms of λ; do not simplify yet).

By LOTUS,

$$E(\frac{1}{Y}) = E(\frac{1}{X+1}) = \sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda} \frac{\lambda^k}{k!}$$

(c) Find a simplified expression for $E(1/Y)$. Hint: $k!(k+1) = (k+1)!$.

$$\sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k+1)!} = \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} = \frac{e^{-\lambda}}{\lambda}(e^{\lambda} - 1) = \frac{1}{\lambda}(1 - e^{-\lambda}).$$

2. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

In (c) through (f), X and Y are i.i.d. (independent identically distributed) positive random variables. Assume that the various expected values exist.

(a) (probability that a roll of 2 fair dice totals 9) $\geq$ (probability that a roll of 2 fair dice totals 10)

The probability on the left is 4/36 and that on the right is 3/36 as there is only one way for both dice to show 5's.

(b) (probability that 65% of 20 children born are girls) $\geq$ (probability that 65% of 2000 children born are girls)

With a large number of births, by the LLN it becomes likely that the fraction that are girls is close to 1/2.

(c) $E(\sqrt{X}) \leq \sqrt{E(X)}$

By Jensen's inequality (or since $\text{Var}(\sqrt{X}) \geq 0$).

(d) $E(\sin X)$? $\sin(EX)$

The inequality can go in either direction. For example, let X be 0 or π with equal probabilities. Then $E(\sin X) = 0$, $\sin(EX) = 1$. But if we let X be $\pi/2$ or $5\pi/2$ with equal probabilities, then $E(\sin X) = 1$, $\sin(EX) = -1$.

(e) $P(X + Y > 4) \geq P(X > 2)P(Y > 2)$

The righthand side is $P(X > 2, Y > 2)$ by independence. The $\geq$ then holds since the event $X > 2, Y > 2$ is a subset of the event $X + Y > 4$.

(f) $E\left((X + Y)^2\right) = 2E(X^2) + 2(EX)^2$

The lefthand side is

$$E(X^2) + E(Y^2) + 2E(XY) = E(X^2) + E(Y^2) + 2E(X)E(Y) = 2E(X^2) + 2(EX)^2$$

since X and Y are i.i.d.

3. A fair die is rolled twice, with outcomes X for the 1st roll and Y for the 2nd roll.

(a) Compute the covariance of $X+Y$ and $X-Y$ (simplify).

$$\text{Cov}(X+Y, X-Y) = \text{Cov}(X,X) - \text{Cov}(X,Y) + \text{Cov}(Y,X) - \text{Cov}(Y,Y) = 0.$$

(b) Are $X+Y$ and $X-Y$ independent? Justify your answer clearly.

They are not independent: information about $X+Y$ may give information about $X-Y$. For example, if we know that $X+Y=12$, then $X=Y=6$, so $X-Y=0$.

(c) Find the moment generating function $M_{X+Y}(t)$ of $X+Y$ (your answer should be a function of t and can contain unsimplified finite sums).

Since X and Y are i.i.d., LOTUS gives

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = \left(\frac{1}{6}\sum_{k=1}^{6} e^{kt}\right)^2$$

4. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? (Simplify.) Justify your answer. Hint: no integrals are needed.

Alice begins to be served when either Bob or Claire leaves. By the memoryless property, the additional time needed to serve whichever of Bob or Claire is still there is Expo(λ). The time it takes to serve Alice is also Expo(λ), so by symmetry the probability is 1/2 that Alice is the last to be done being served.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

Use the order statistics results, or compute it directly:

$$P(\min(X,Y) > z) = P(X > z, Y > z) = P(X > z)P(Y > z) = e^{-2\lambda z},$$

so $\min(X, Y)$ has the Expo(2λ) distribution, with CDF $F(z) = 1 - e^{-2\lambda z}$.

(c) What is the expected total time that Alice needs to spend at the post office?

The expected time spent waiting in line is $\frac{1}{2\lambda}$ by (b). The expected time spent being served is $\frac{1}{\lambda}$. So the expected total time is

$$\frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}.$$

5. Bob enters a casino with $X_0 = 1$ dollar and repeatedly plays the following game: with probability 1/3, the amount of money he has increases by a factor of 3; with probability 2/3, the amount of money he has decreases by a factor of 3. Let X_n be the amount of money he has after playing this game n times. Thus, X_{n+1} is $3X_n$ with probability 1/3 and is $3^{-1}X_n$ with probability 2/3.

(a) Compute $E(X_1)$, $E(X_2)$ and, in general, $E(X_n)$. (Simplify.)

$$E(X_1) = \frac{1}{3} \cdot 3 + \frac{2}{3} \cdot 1/3 = \frac{11}{9}$$

$E(X_{n+1})$ can be found by conditioning on X_n.

$$E(X_{n+1}|X_n) = \frac{1}{3} \cdot 3X_n + \frac{2}{3} \cdot 3^{-1}X_n = \frac{11}{9}X_n,$$

so

$$E(X_{n+1}) = E(E(X_{n+1}|X_n)) = \frac{11}{9}E(X_n).$$

Then

$$E(X_2) = (\frac{11}{9})^2 = \frac{121}{81}$$

and in general,

$$E(X_n) = (\frac{11}{9})^n.$$

(b) What happens to $E(X_n)$ as $n \to \infty$? Let Y_n be the number of times out of the first n games that Bob triples his money. What happens to Y_n/n as $n \to \infty$?

By the above, $E(X_n) \to \infty$ as $n \to \infty$. By LLN, $Y_n/n \to \frac{1}{3}$ a.s. as $n \to \infty$.

(c) Does X_n converge to some number c as $n \to \infty$ and if so, what is c? Explain.

In the long run, Bob will win about 1/3 of the time and lose about 2/3 of the time. Note that a win and a loss cancel each other out (due to multiplying and dividing by 3), so X_n will get very close to 0. In terms of Y_n from (b), with probability 1

$$X_n = 3^{Y_n}3^{-(n-Y_n)} = 3^{n(2\frac{Y_n}{n}-1)} \to 0.$$

because Y_n/n approaches 1/3. So X_n converges to 0 (with probability 1).

6. Let X and Y be independent standard Normal r.v.s and let $R^2 = X^2 + Y^2$ (where $R > 0$ is the distance from (X, Y) to the origin).

(a) The distribution of R^2 is an example of three of the "important distributions" listed on the last page. State which three of these distributions R^2 is an instance of, specifying the parameter values. (For example, if it were Geometric with $p = 1/3$, the distribution would be Geom(1/3) and also NBin(1,1/3).)

It is χ^2_2, Expo(1/2), and Gamma(1,1/2).

(b) Find the PDF of R. (Simplify.) Hint: start with the PDF $f_W(w)$ of $W = R^2$.

$R = \sqrt{W}$ with $f_W(w) = \frac{1}{2}e^{-w/2}$ gives

$$f_R(r) = f_W(w)|dw/dr| = \frac{1}{2}e^{-w/2}2r = re^{-r^2/2}, \text{ for } r > 0.$$

(This is known as the *Rayleigh distribution*.)

(c) Find $P(X > 2Y + 3)$ in terms of the standard Normal CDF Φ. (Simplify.)

$$P(X > 2Y + 3) = P(X - 2Y > 3) = 1 - \Phi\left(\frac{3}{\sqrt{5}}\right)$$

since $X - 2Y \sim \mathcal{N}(0, 5)$.

(d) Compute $\text{Cov}(R^2, X)$. Are R^2 and X independent?

They are not independent since knowing X gives information about R^2, e.g., X^2 being large implies that R^2 is large. But R^2 and X are uncorrelated:

$\text{Cov}(R^2, X) = \text{Cov}(X^2+Y^2, X) = \text{Cov}(X^2, X)+\text{Cov}(Y^2, X) = E(X^3)-(EX^2)(EX)+0 = 0.$

7. Let $U_1, U_2, \ldots, U_{60}$ be i.i.d. Unif(0,1) and $X = U_1 + U_2 + \cdots + U_{60}$.

(a) Which important distribution is the distribution of X very close to? Specify what the parameters are, and state which theorem justifies your choice.

By the Central Limit Theorem, the distribution is approximately $\mathcal{N}(30, 5)$ since $E(X) = 30, \text{Var}(X) = 60/12 = 5$.

(b) Give a simple but accurate approximation for $P(X > 17)$. Justify briefly.

$$P(X > 17) = 1 - P(X \leq 17) = 1 - P\left(\frac{X - 30}{\sqrt{5}} \leq \frac{-13}{\sqrt{5}}\right) \approx 1 - \Phi\left(\frac{-13}{\sqrt{5}}\right) = \Phi\left(\frac{13}{\sqrt{5}}\right).$$

Since $13/\sqrt{5} > 5$, and we already have $\Phi(3) \approx 0.9985$ by the 68-95-99.7% rule, the value is extremely close to 1.

(c) Find the moment generating function (MGF) of X.

The MGF of U_1 is $E(e^{tU_1}) = \int_0^1 e^{tu} du = \frac{1}{t}(e^t - 1)$ for $t \neq 0$, and the MGF of U_1 is 1 for $t = 0$. Thus, the MGF of X is 1 for $t = 0$, and for $t \neq 0$ it is

$$E(e^{tX}) = E(e^{t(U_1 + \cdots + U_{60})}) = \left(E(e^{tU_1})\right)^{60} = \frac{(e^t - 1)^{60}}{t^{60}}.$$

8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $E(X_1) = 3$, and consider the sum $S_n = X_1 + X_2 + \cdots + X_n$.

(a) What is $E(X_1 X_2 X_3 | X_1)$? (Simplify. Your answer should be a function of X_1.)

$$E(X_1 X_2 X_3 | X_1) = X_1 E(X_2 X_3 | X_1) = X_1 E(X_2) E(X_3) = 9X_1.$$

(b) What is $E(X_1|S_n) + E(X_2|S_n) + \cdots + E(X_n|S_n)$? (Simplify.)

By linearity, it is $E(S_n|S_n)$, which is S_n.

(c) What is $E(X_1|S_n)$? (Simplify.) Hint: use (b) and symmetry.

By symmetry, $E(X_j|S_n) = E(X_1|S_n)$ for all j. Then by (b),

$$nE(X_1|S_n) = S_n,$$

so

$$E(X_1|S_n) = \frac{S_n}{n}.$$

9. An urn contains red, green, and blue balls. Balls are chosen randomly with replacement (each time, the color is noted and then the ball is put back.) Let r, g, b be the probabilities of drawing a red, green, blue ball respectively ($r + g + b = 1$).

(a) Find the expected number of balls chosen before obtaining the first red ball, not including the red ball itself. (Simplify.)

The distribution is Geom(r), so the expected value is $\frac{1-r}{r}$.

(b) Find the expected number of different *colors* of balls obtained before getting the first red ball. (Simplify.)

Use indicator random variables: let I_1 be 1 if green is obtained before red, and 0 otherwise, and define I_2 similarly for blue. Then

$$E(I_1) = P(\text{green before red}) = \frac{g}{g+r}$$

since "green before red" means that the first nonblue ball is green. Similarly, $E(I_2) = b/(b + r)$, so the expected number of colors obtained before getting red is

$$E(I_1 + I_2) = \frac{g}{g+r} + \frac{b}{b+r}.$$

(c) Find the probability that at least 2 of n balls drawn are red, given that at least 1 is red. (Simplify; avoid sums of large numbers of terms, and $\sum$ or $\cdots$ notation.)

$$P(\text{at least 2}|\text{ at least 1}) = \frac{P(\text{at least 2})}{P(\text{at least 1})} = \frac{1 - (1-r)^n - nr(1-r)^{n-1}}{1 - (1-r)^n}.$$

10. Let $X_0, X_1, X_2, \ldots$ be an irreducible Markov chain with state space $\{1, 2, \ldots, M\}$, $M \geq 3$, transition matrix $Q = (q_{ij})$, and stationary distribution $\mathbf{s} = (s_1, \ldots, s_M)$. The initial state X_0 is given the stationary distribution, i.e., $P(X_0 = i) = s_i$.

(a) On average, how many of $X_0, X_1, \ldots, X_9$ equal 3? (In terms of $\mathbf{s}$; simplify.)

Since X_0 has the stationary distribution, all of $X_0, X_1, \ldots$ have the stationary distribution. By indicator random variables, the expected value is $10s_3$.

(b) Let $Y_n = (X_n - 1)(X_n - 2)$. For $M = 3$, find an example of Q (the transition matrix for the *original* chain $X_0, X_1, \ldots$) where $Y_0, Y_1, \ldots$ is Markov, and another example of P where $Y_0, Y_1, \ldots$ is not Markov. Mark which is which and briefly explain. In your examples, make $q_{ii} > 0$ for at least one i and make sure it is possible to get from any state to any other state eventually.

Note that Y_n is 0 if X_n is 1 or 2, and Y_n is 2 otherwise. So the Y_n process can be viewed as merging states 1 and 2 of the X_n-chain into one state. Knowing the history of Y_n's means knowing when the X_n-chain is in State 3, without being able to distinguish State 1 from State 2.

If $q_{13} = q_{23}$, then Y_n is Markov since given Y_n, even knowing the past $X_0, \ldots, X_n$ does not affect the transition probabilities. But if $q_{13} \neq q_{23}$, then the Y_n past history can give useful information about X_n, affecting the transition probabilities. So one example (not the only possible example!) is

$$Q_1 = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \text{ (Markov)} \qquad Q_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \end{pmatrix} \text{ (not Markov).}$$

(c) If each column of Q sums to 1, what is $\mathbf{s}$? Verify using the definition of *stationary*.

The stationary distribution is uniform over all states:

$$\mathbf{s} = (1/M, 1/M, \ldots, 1/M).$$

This is because

$$\begin{pmatrix} 1/M & 1/M & \ldots & 1/M \end{pmatrix} Q = \frac{1}{M} \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} Q = \begin{pmatrix} 1/M & 1/M & \ldots & 1/M \end{pmatrix},$$

where the matrix multiplication was done by noting that multiplying a row vector of 1's times Q gives the column sums of Q.

2 Solutions to Stat 110 Final from 2007

1. Consider the birthdays of 100 people. Assume people's birthdays are independent, and the 365 days of the year (exclude the possibility of February 29) are equally likely.

(a) Find the expected number of birthdays represented among the 100 people, i.e., the expected number of days that at least 1 of the people has as his or her birthday (your answer can involve unsimplified fractions but should not involve messy sums).

Define indicator r.v.s I_j where $I_j = 1$ if the jth day of the year appears on the list of all the birthdays. Then $EI_j = P(I_j = 1) = 1 - (\frac{364}{365})^{100}$, so

$$E(\sum_{j=1}^{365} I_j) = 365 \left(1 - (\frac{364}{365})^{100}\right).$$

(b) Find the covariance between how many of the people were born on January 1 and how many were born on January 2.

Let X_j be the number of people born on January j. Then

$$\text{Cov}(X_1, X_2) = -\frac{100}{365^2}.$$

To see this, we can use the result about covariances in the Multinomial, or we can solve the problem directly as follows (or with various other methods). Let A_j be the indicator for the jth person having been born on January 1, and define B_j similarly for January 2. Then

$$E(X_1X_2) = E\left((\sum_i A_i)(\sum_j B_j)\right) = E(\sum_{i,j} A_iB_j) = 100 \cdot 99(\frac{1}{365})^2$$

since $A_iB_i = 0$, while A_i and B_j are independent for $i \neq j$. So

$$\text{Cov}(X_1, X_2) = 100 \cdot 99(\frac{1}{365})^2 - (\frac{100}{365})^2 = -\frac{100}{365^2}.$$

2. Let X and Y be positive random variables, *not necessarily independent*. Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $(E(XY))^2 \leq E(X^2)E(Y^2)$ (by Cauchy-Schwarz)

(b) $P(|X+Y|>2) \leq \frac{1}{10}E((X+Y)^4)$ (by Markov's Inequality)

(c) $E(\ln(X+3)) \leq \ln(E(X+3))$ (by Jensen)

(d) $E(X^2e^X) \geq E(X^2)E(e^X)$ (since X^2 and e^X are positively correlated)

(e) $P(X+Y=2)$? $P(X=1)P(Y=1)$ (What if X, Y are independent? What if $X \sim \text{Bern}(1/2)$ and $Y = 1 - X$?)

(f) $P(X+Y=2) \leq P(\{X \geq 1\} \cup \{Y \geq 1\})$ (left event is a subset of right event)

3. Let X and Y be independent Pois(λ) random variables. Recall that the moment generating function (MGF) of X is $M(t) = e^{\lambda(e^t-1)}$.

(a) Find the MGF of $X + 2Y$ (simplify).

$$E(e^{t(X+2Y)}) = E(e^{tX})E(e^{2tY}) = e^{\lambda(e^t-1)}e^{\lambda(e^{2t}-1)} = e^{\lambda(e^t+e^{2t}-2)}.$$

(b) Is $X + 2Y$ also Poisson? Show that it is, or that it isn't (whichever is true).

No, it is not Poisson. This can be seen by noting that the MGF from (a) is not of the form of a Poisson MGF, or by noting that $E(X + 2Y) = 3\lambda$, $\text{Var}(X + 2Y) = 5\lambda$ are not equal, whereas any Poisson random variable has mean equal to its variance.

(c) Let $g(t) = \ln M(t)$ be the log of the MGF of X. Expanding $g(t)$ as a Taylor series

$$g(t) = \sum_{j=1}^{\infty} \frac{c_j}{j!} t^j$$

(the sum starts at $j = 1$ because $g(0) = 0$), the coefficient c_j is called the jth *cumulant* of X. Find c_j in terms of λ, for all $j \geq 1$ (simplify).

Using the Taylor series for e^t,

$$g(t) = \lambda(e^t - 1) = \sum_{j=1}^{\infty} \lambda \frac{t^j}{j!},$$

so $c_j = \lambda$ for all $j \geq 1$.

4. Consider the following conversation from an episode of *The Simpsons*:

> Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
>
> Homer: *Lisa, a guy who's got lots of ivory is less likely to hurt Stampy than a guy whose ivory supplies are low.*

Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

(a) Define clear notation for the various events of interest here.

Let H be the event that the man will hurt Stampy, let L be the event that a man has lots of ivory, and let D be the event that the man is an ivory dealer.

(b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

Lisa observes that L is true. She suggests (reasonably) that this evidence makes D more likely, i.e., $P(D|L) > P(D)$. Implicitly, she suggests that this makes it likely that the man will hurt Stampy, i.e., $P(H|L) > P(H|L^c)$. Homer argues that $P(H|L) < P(H|L^c)$.

(c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

Homer does not realize that observing that Blackheart has so much ivory makes it much more likely that Blackheart is an ivory dealer, which in turn makes it more likely that the man will hurt Stampy. (This is an example of Simpson's Paradox.) It may be true that, *controlling for whether or not Blackheart is a dealer*, having high ivory supplies makes it less likely that he will harm Stampy: $P(H|L, D) < P(H|L^c, D)$ and $P(H|L, D^c) < P(H|L^c, D^c)$. However, this does not imply that $P(H|L) < P(H|L^c)$.

5. Empirically, it is known that 49% of children born in the U.S. are girls (and 51% are boys). Let N be the number of children who will be born in the U.S. in March 2009, and assume that N is a Pois(λ) random variable, where λ is known. Assume that births are independent (e.g., don't worry about identical twins).

Let X be the number of girls who will be born in the U.S. in March 2009, and let Y be the number of boys who will be born then (note the importance of choosing good notation: boys have a Y chromosome).

(a) Find the joint distribution of X and Y. (Give the joint PMF.)

Note that the problem is equivalent to the chicken and egg problem (the structure is identical). So X and Y are independent with $X \sim \text{Pois}(0.49\lambda), Y \sim \text{Pois}(0.51\lambda)$. The joint PMF is

$$P(X = i, Y = j) = (e^{-0.49\lambda}(0.49\lambda)^i/i!)(e^{-0.51\lambda}(0.51\lambda)^j/j!).$$

(b) Find $E(N|X)$ and $E(N^2|X)$.

Since X and Y are independent,

$$E(N|X) = E(X + Y|X) = X + E(Y|X) = X + EY = X + 0.51\lambda,$$

$$E(N^2|X) = E(X^2+2XY+Y^2|X) = X^2+2XE(Y)+E(Y^2) = (X+0.51\lambda)^2+0.51\lambda.$$

6. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (independent Exponentials with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

By linearity, independence, and the memoryless property, we get

$$E(X_1|X_1 > 1) + E(X_2|X_2 > 2) + E(X_3|X_3 > 3) = \lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1} + 6.$$

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

The desired probability is $P(X_1 \leq \min(X_2, X_3))$. Noting that $\min(X_2, X_3) \sim \text{Expo}(\lambda_2 + \lambda_3)$ is independent of X_1, we have

$$P(X_1 \leq \min(X_2, X_3)) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}.$$

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

Let $M = \max(X_1, X_2, X_3)$. Using the order statistics results from class or by directly computing the CDF and taking the derivative, for $x > 0$ we have

$$f_M(x) = 3(1 - e^{-x})^2 e^{-x}.$$

This is not one of the "important distributions". (The form is reminiscent of a Beta, but a Beta takes values between 0 and 1, while M can take any positive real value; in fact, $B \sim \text{Beta}(1, 3)$ if we make the transformation $B = e^{-M}$.)

7. Let $X_1, X_2, \ldots$ be i.i.d. random variables with CDF $F(x)$. For every number x, let $R_n(x)$ count how many of $X_1, \ldots, X_n$ are less than or equal to x.

(a) Find the mean and variance of $R_n(x)$ (in terms of n and $F(x)$).

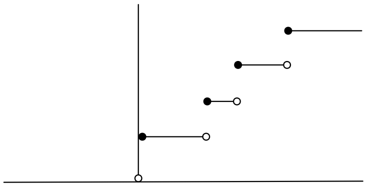
Let $I_j(x)$ be 1 if $X_j \leq x$ and 0 otherwise. Then

$$R_n(x) = \sum_{j=1}^{n} I_j(x) \sim \text{Bin}(n, F(x)),$$

so $ER_n(x) = nF(x)$ and $\text{Var}(R_n(x)) = nF(x)(1 - F(x))$.

(b) Assume (for this part only) that $X_1, \ldots, X_4$ are known constants. Sketch an example showing what the graph of the function $\frac{R_4(x)}{4}$ might look like. Is the function $\frac{R_4(x)}{4}$ necessarily a CDF? Explain briefly.

For $X_1, \ldots, X_4$ distinct, the graph of $\frac{R_4(x)}{4}$ starts at 0 and then has 4 jumps, each of size 0.25 (it jumps every time one of the X_i's is reached).



The $\frac{R_4(x)}{4}$ is the CDF of a discrete random variable with possible values X_1, X_2, X_3, X_4.

(c) Show that $\frac{R_n(x)}{n} \to F(x)$ as $n \to \infty$ (with probability 1).

As in (a), $R_n(x)$ is the sum of n i.i.d. Bern(p) r.v.s, where $p = F(x)$. So by the Law of Large Numbers, $\frac{R_n(x)}{n} \to F(x)$ as $n \to \infty$ (with probability 1).

8. (a) Let T be a Student-t r.v. with 1 degree of freedom, and let $W = 1/T$. Find the PDF of W (simplify). Is this one of the "important distributions"?

Hint: no calculus is needed for this (though it can be used to check your answer).

Recall that a Student-t with 1 degree of freedom (also known as a Cauchy) can be represented as a ratio X/Y with X and Y are i.i.d. $\mathcal{N}(0,1)$. But then the reciprocal Y/X is of the same form! So W is also Student-t with 1 degree of freedom, and PDF $f_W(w) = \frac{1}{\pi(1+w^2)}$.

(b) Let $W_n \sim \chi^2_n$ (the Chi-squared distribution with n degrees of freedom), for each $n \geq 1$. Do there exist a_n and b_n such that $a_n(W_n - b_n) \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$? If so, find them; if not, explain why not.

Write $W_n = \sum_{i=1}^n Z_i^2$ with the Z_i i.i.d. $\mathcal{N}(0,1)$. By the CLT, the claim is true with

$$b_n = E(W_n) = n \text{ and } a_n = \frac{1}{\sqrt{\text{Var}(W_n)}} = \frac{1}{\sqrt{2n}}.$$

(c) Let $Z \sim \mathcal{N}(0,1)$ and $Y = |Z|$. Find the PDF of Y, and approximate $P(Y < 2)$.

For $y \geq 0$, the CDF of Y is

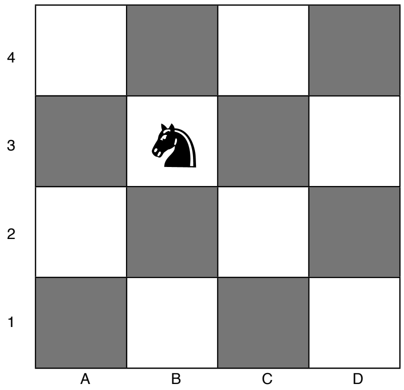
$$P(Y \leq y) = P(|Z| \leq y) = P(-y \leq Z \leq y) = \Phi(y) - \Phi(-y),$$

so the PDF of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2} + \frac{1}{\sqrt{2\pi}}e^{-y^2/2} = 2\frac{1}{\sqrt{2\pi}}e^{-y^2/2}.$$

By the 68-95-99.7% Rule, $P(Y < 2) \approx 0.95$.

9. Consider a knight randomly moving around on a 4 by 4 chessboard:



The 16 squares are labeled in a grid, e.g., the knight is currently at the square B3, and the upper left square is A4. Each move of the knight is an L-shape: two squares horizontally followed by one square vertically, or vice versa. For example, from B3 the knight can move to A1, C1, D2, or D4; from A4 it can move to B2 or C3. Note that from a white square, the knight always moves to a gray square and vice versa.

At each step, the knight moves randomly, each possibility equally likely. Consider the stationary distribution of this Markov chain, where the states are the 16 squares.

(a) Which squares have the highest stationary probability? Explain very briefly.

The four center squares (B2, B3, C2, C3) have the highest stationary probability since they are the most highly connected squares: for each of these squares, the number of possible moves to/from the square is maximized.

(b) Compute the stationary distribution (simplify). Hint: random walk on a graph.

Use symmetry to note that there are only three "types" of square: there are 4 center squares, 4 corner squares (such as A4), and 8 edge squares (such as B4; exclude corner squares from being considered edge squares). Recall from the Markov chain handout that the stationary probability of a state for random walk on an undirected network is *proportional to its degree.*

A center square here has degree 4, a corner square has degree 2, and an edge square has degree 3. So these have probabilities $4a, 2a, 3a$ respectively for some a. To find a, count the number of squares of each type to get $4a(4) + 2a(4) + 3a(8) = 1$, giving $a = 1/48$. Thus, each center square has stationary probability $4/48 = 1/12$; each corner square has stationary probability $2/48 = 1/24$; and each edge square has stationary probability $3/48 = 1/16$.

3 Solutions to Stat 110 Final from 2008

1. Joe's iPod has 500 different songs, consisting of 50 albums of 10 songs each. He listens to 11 random songs on his iPod, with all songs equally likely and chosen independently (so repetitions may occur).

(a) What is the PMF of how many of the 11 songs are from his favorite album?

The distribution is $\text{Bin}(n, p)$ with $n = 11, p = \frac{1}{50}$ (thinking of getting a song from the favorite album as a "success"). So the PMF is

$$\binom{11}{k}\left(\frac{1}{50}\right)^k\left(\frac{49}{50}\right)^{11-k}, \text{ for } 0 \leq k \leq 11.$$

(b) What is the probability that there are 2 (or more) songs from the same album among the 11 songs he listens to? (Do not simplify.)

This is a form of the birthday problem.

$$P(\text{at least 1 match}) = 1 - P(\text{no matches}) = 1 - \frac{50 \cdot 49 \cdot \dots \cdot 40}{50^{11}} = 1 - \frac{49!}{39! \cdot 50^{10}}.$$

(c) A pair of songs is a "match" if they are from the same album. If, say, the 1st, 3rd, and 7th songs are all from the same album, this counts as 3 matches. Among the 11 songs he listens to, how many matches are there on average? (Simplify.)

Defining an indicator r.v. I_{jk} for the event that the jth and kth songs match, we have $E(I_{jk}) = P(I_{jk} = 1) = 1/50$, so the expected number of matches is

$$\binom{11}{2}\frac{1}{50} = \frac{11 \cdot 10}{2 \cdot 50} = \frac{110}{100} = 1.1.$$

2. Let X and Y be *positive* random variables, *not necessarily independent.* Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $P(X + Y > 2) \leq \frac{EX+EY}{2}$ (by Markov and linearity)

(b) $P(X + Y > 3) \geq P(X > 3)$ (since $X > 3$ implies $X + Y > 3$ since $Y > 0$)

(c) $E(\cos(X))? \cos(EX)$ (e.g., let $W \sim$ Bern(1/2) and $X = aW + b$ for various a, b)

(d) $E(X^{1/3}) \leq (EX)^{1/3}$ (by Jensen)

(e) $E(X^Y)?(EX)^{EY}$ (take X constant or Y constant as examples)

(f) $E\left(E(X|Y) + E(Y|X)\right) = EX + EY$ (by linearity and Adam's Law)

3. (a) A woman is pregnant with twin boys. Twins may be either identical or fraternal (non-identical). In general, 1/3 of twins born are identical. Obviously, identical twins must be of the same sex; fraternal twins may or may not be. Assume that identical twins are equally likely to be both boys or both girls, while for fraternal twins all possibilities are equally likely. Given the above information, what is the probability that the woman's twins are identical?

By Bayes' Rule,

$$P(\text{identical}|BB) = \frac{P(BB|\text{identical})P(\text{identical})}{P(BB)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{4}\cdot\frac{2}{3}} = 1/2.$$

(b) A certain genetic characteristic is of interest. For a random person, this has a numerical value given by a $\mathcal{N}(0,\sigma^2)$ r.v. Let X_1 and X_2 be the values of the genetic characteristic for the twin boys from (a). If they are identical, then $X_1 = X_2$; if they are fraternal, then X_1 and X_2 have correlation ρ. Find $\text{Cov}(X_1, X_2)$ in terms of ρ, σ^2.

Since the means are 0, $\text{Cov}(X_1, X_2) = E(X_1X_2) - (EX_1)(EX_2) = E(X_1X_2)$. We find this by conditioning on whether the twins are identical or fraternal:

$$E(X_1X_2) = E(X_1X_2|\text{identical})\frac{1}{2}+E(X_1X_2|\text{fraternal})\frac{1}{2} = E(X_1^2)\frac{1}{2}+\rho\sigma^2\frac{1}{2} = \frac{\sigma^2}{2}(1+\rho).$$

4. (a) Consider i.i.d. Pois(λ) r.v.s $X_1, X_2, \ldots$. The MGF of X_j is $M(t) = e^{\lambda(e^t-1)}$. Find the MGF $M_n(t)$ of the sample mean $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. (Hint: it may help to do the $n = 2$ case first, which itself is worth a lot of partial credit, and then generalize.)

The MGF is

$$E(e^{\frac{t}{n}(X_1+\cdots+X_n)}) = \left(E(e^{\frac{t}{n}X_1})\right)^n = e^{n\lambda(e^{t/n}-1)},$$

since the X_j are i.i.d. and $E(e^{\frac{t}{n}X_1})$ is the MGF of X_1 evaluated at t/n.

(b) Find the limit of $M_n(t)$ as $n \to \infty$. (You can do this with almost no calculation using a relevant theorem; or you can use (a) and that $e^x \approx 1 + x$ if x is very small.)

By the Law of Large Numbers, $\bar{X}_n \to \lambda$ with probability 1. The MGF of the constant λ (viewed as a r.v. that always equals λ) is $e^{t\lambda}$. Thus, $M_n(t) \to e^{t\lambda}$ as $n \to \infty$.

5. A post office has 2 clerks. Alice enters the post office while 2 other customers, Bob and Claire, are being served by the 2 clerks. She is next in line. Assume that the time a clerk spends serving a customer has the Expo(λ) distribution.

(a) What is the probability that Alice is the last of the 3 customers to be done being served? Justify your answer. Hint: no integrals are needed.

Alice begins to be served when either Bob or Claire leaves. By the memoryless property, the additional time needed to serve whichever of Bob or Claire is still there is Expo(λ). The time it takes to serve Alice is also Expo(λ), so by symmetry the probability is 1/2 that Alice is the last to be done being served.

(b) Let X and Y be independent Expo(λ) r.v.s. Find the CDF of $\min(X, Y)$.

Use the order statistics results, or compute it directly:

$$P(\min(X, Y) > z) = P(X > z, Y > z) = P(X > z)P(Y > z) = e^{-2\lambda z},$$

so $\min(X, Y)$ has the Expo(2λ) distribution, with CDF $F(z) = 1 - e^{-2\lambda z}$.

(c) What is the expected total time that Alice needs to spend at the post office?

The expected time spent waiting in line is $\frac{1}{2\lambda}$ by (b). The expected time spent being served is $\frac{1}{\lambda}$. So the expected total time is

$$\frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}.$$

6. You are given an amazing opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown, except that it is worth at least nothing, and at most a million dollars. So the true value V of the prize is considered to be Uniform on [0,1] (measured in millions of dollars).

You can choose to bid any amount b (in millions of dollars). You have the chance to get the prize for considerably less than it is worth, but you could also lose money if you bid too much. Specifically, if $b < \frac{2}{3}V$, then the bid is rejected and nothing is gained or lost. If $b \geq \frac{2}{3}V$, then the bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V). What is your optimal bid b (to maximize the expected payoff)?

We choose a bid $b \geq 0$, which cannot be defined in terms of the unknown V. The expected payoff can be found by conditioning on whether the bid is accepted. The term where the bid is rejected is 0, so the expected payoff is

$$E(V - b|b \geq \frac{2}{3}V)P(b \geq \frac{2}{3}V) = \left(E(V|V \leq \frac{3}{2}b) - b\right) P(V \leq \frac{3}{2}b).$$

For $b \geq 2/3$, the bid is definitely accepted but we lose money on average, so assume $b < 2/3$. Then

$$\left(E(V|V \leq \frac{3}{2}b) - b\right) P(V \leq \frac{3}{2}b) = (\frac{3}{4}b - b)\frac{3}{2}b = -\frac{3}{8}b^2,$$

since given that $V \leq \frac{3}{2}b$, the conditional distribution of V is Uniform on $[0, \frac{3}{2}b]$.

The above expression is negative except at $b = 0$, so the optimal bid is 0: one should not play this game! What's the moral of this story? First, investing in an asset without any information about its value is a bad idea. Second, *condition on all the information*. It is crucial in the above calculation to use $E(V|V \leq \frac{3}{2}b)$ rather than $E(V) = 1/2$; knowing that the bid was accepted gives information about how much the mystery prize is worth!

7. (a) Let $Y = e^X$, with $X \sim \text{Expo}(3)$. Find the mean and variance of Y (simplify).

By LOTUS,

$$E(Y) = \int_0^\infty e^x (3e^{-3x})dx = \frac{3}{2},$$

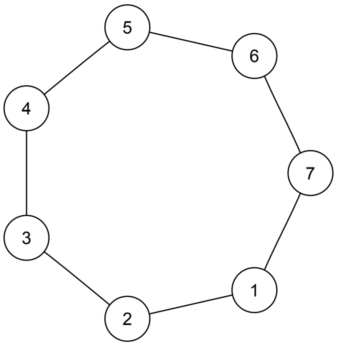
$$E(Y^2) = \int_0^\infty e^{2x} (3e^{-3x})dx = 3.$$

So $E(Y) = 3/2, \text{Var}(Y) = 3 - 9/4 = 3/4$.

(b) For $Y_1, \dots, Y_n$ i.i.d. with the same distribution as Y from (a), what is the approximate distribution of the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$ when n is large? (Simplify, and specify all parameters.)

By the CLT, $\bar{Y}_n$ is approximately $\mathcal{N}(\frac{3}{2}, \frac{3}{4n})$ for large n.

8.

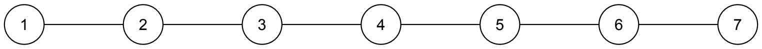


(a) Consider a Markov chain on the state space $\{1, 2, \ldots, 7\}$ with the states arranged in a "circle" as shown above, and transitions given by moving one step clockwise or counterclockwise with equal probabilities. For example, from state 6, the chain moves to state 7 or state 5 with probability 1/2 each; from state 7, the chain moves to state 1 or state 6 with probability 1/2 each. The chain starts at state 1.

Find the stationary distribution of this chain.

The symmetry of the chain suggests that the stationary distribution should be uniform over all the states. To verify this, note that the reversibility condition is satisfied. So the stationary distribution is $(1/7, 1/7, \ldots, 1/7)$.

(b) Consider a new chain obtained by "unfolding the circle." Now the states are arranged as shown below. From state 1 the chain always goes to state 2, and from state 7 the chain always goes to state 6. Find the new stationary distribution.



By the results from class for random walk on an undirected network, the stationary probabilities are proportional to the degrees. So we just need to normalize $(1, 2, 2, 2, 2, 2, 1)$, obtaining $(1/12, 1/6, 1/6, 1/6, 1/6, 1/6, 1/12)$.

4 Solutions to Stat 110 Final from 2009

1. A group of n people play "Secret Santa" as follows: each puts his or her name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's own name, so some may have to buy gifts for themselves (on the bright side, some may like self-selected gifts better). Assume $n \geq 2$.

(a) Find the expected number of people who pick their own names (simplify).

Let I_j be the indicator r.v. for the jth person picking his or her own name. Then $E(I_j) = P(I_j = 1) = \frac{1}{n}$. By linearity, the expected number is $n \cdot E(I_j) = 1$.

(b) Find the expected number of pairs of people, A and B, such that A picks B's name and B picks A's name (where $A \neq B$ and order doesn't matter; simplify).

Let I_{ij} the the indicator r.v. for the ith and jth persons having such a "swap" (for $i < j$). Then $E(I_{ij}) = P(i \text{ picks } j)P(j \text{ picks } i | i \text{ picks } j) = \frac{1}{n(n-1)}$.

Alternatively, we can get this by counting: there are $n!$ permutations for who picks whom, of which $(n-2)!$ have i pick j and j pick i, giving $\frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. So by linearity, the expected number is $\binom{n}{2} \cdot \frac{1}{n(n-1)} = \frac{1}{2}$.

(c) Let X be the number of people who pick their own names. Which of the "important distributions" are conceivable as the distribution of X, just based on the possible values X takes (you do not need to list parameter values for this part)?

Since X is an integer between 0 and n, the only conceivable "important distributions" are Binomial and Hypergeometric. Going further (which was not required), note that X actually can't equal $n - 1$, since if $n - 1$ people pick their own names then the remaining person must too. So the possible values are the integers from 0 to n except for $n - 1$, which rules out all of the "important distributions".

(d) What is the *approximate* distribution of X if n is large (specify the parameter value or values)? What does $P(X = 0)$ converge to as $n \to \infty$?

By the Poisson Paradigm, X is approximately Pois(1) for large n. As $n \to \infty$, $P(X = 0) \to 1/e$, which is the probability of a Pois(1) r.v. being 0.

2. Let X and Y be positive random variables, *not necessarily independent.* Assume that the various expected values below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general.) It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(X^3) \leq \sqrt{E(X^2)E(X^4)}$ (by Cauchy-Schwarz)

(b) $P(|X+Y|>2) \leq \frac{1}{16}E((X+Y)^4)$ (by Markov, taking 4th powers first)

(c) $E(\sqrt{X+3}) \leq \sqrt{E(X+3)}$ (by Jensen with a concave function)

(d) $E(\sin^2(X)) + E(\cos^2(X)) = 1$ (by linearity)

(e) $E(Y|X+3) = E(Y|X)$ (knowing $X+3$ is equivalent to knowing X)

(f) $E(E(Y^2|X)) \geq (EY)^2$ (by Adam's Law and Jensen)

3. Let $Z \sim \mathcal{N}(0,1)$. Find the 4th moment $E(Z^4)$ in the following two different ways:

(a) using what you know about how certain powers of Z are related to other distributions, along with information from the table of distributions.

Let $W = Z^2$, which we know is χ_1^2. By the table, $E(W) = 1, \text{Var}(W) = 2$. So $E(Z^4) = E(W^2) = \text{Var}(W) + (EW)^2 = 2 + 1 = 3$.

(b) using the MGF $M(t) = e^{t^2/2}$, by writing down its Taylor series and using how the coefficients relate to moments of Z, *not* by tediously taking derivatives of $M(t)$. Hint: you can get this series immediately from the Taylor series for e^x.

Plugging $t^2/2$ in for x in $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, we have

$$e^{t^2/2} = \sum_{n=0}^{\infty} \frac{t^{2n}}{n! \cdot 2^n}.$$

The t^4 term is

$$\frac{1}{2! \cdot 2^2} t^4 = \frac{1}{8} t^4 = \frac{3}{4!} t^4,$$

so the 4th moment is 3, which agrees with (a).

4. A chicken lays n eggs. Each egg independently does or doesn't hatch, with probability p of hatching. For each egg that hatches, the chick does or doesn't survive (independently of the other eggs), with probability s of survival. Let $N \sim \text{Bin}(n, p)$ be the number of eggs which hatch, X be the number of chicks which survive, and Y be the number of chicks which hatch but don't survive (so $X + Y = N$).

(a) Find the distribution of X, preferably with a clear explanation in words rather than with a computation. If X has one of the "important distributions," say which (including its parameters).

We will give a story proof that $X \sim \text{Bin}(n, ps)$. Consider any one of the n eggs. With probability p, it hatches. Given that it hatches, with probability s the chick survives. So the probability is ps of the egg hatching a chick which survives. Thus, $X \sim \text{Bin}(n, ps)$.

(b) Find the joint PMF of X and Y (simplify).

As in the chicken-egg problem from class, condition on N and note that only the $N = i + j$ term is nonzero: for any nonnegative integers i, j with $i + j \leq n$,

$$\begin{aligned} P(X = i, Y = j) &= P(X = i, Y = j | N = i + j) P(N = i + j) \\ &= P(X = i | N = i + j) P(N = i + j) \\ &= \binom{i+j}{i} s^i (1-s)^j \binom{n}{i+j} p^{i+j} (1-p)^{n-i-j} \\ &= \frac{n!}{i! j! (n-i-j)!} (ps)^i (p(1-s))^j (1-p)^{n-i-j}. \end{aligned}$$

(c) Are X and Y independent? Give a clear explanation in words (of course it makes sense to see if your answer is consistent with your answer to (b), but you can get full credit on this part even without doing (b); conversely, it's not enough to just say "by (b), ..." without further explanation).

They are *not* independent, unlike in the chicken-egg problem from class (where N was Poisson). To see this, consider extreme cases: if $X = n$, then clearly $Y = 0$. This shows that X can yield information about Y.

5. Suppose we wish to approximate the following integral (denoted by b):

$$b = \int_{-\infty}^{\infty} (-1)^{\lfloor x \rfloor} e^{-x^2/2} dx,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x (e.g., $\lfloor 3.14 \rfloor = 3$).

(a) Write down a function $g(x)$ such that $E(g(X)) = b$ for $X \sim \mathcal{N}(0,1)$ (your function should *not* be in terms of b, and should handle normalizing constants carefully).

There are many possible solutions. By LOTUS, we can take $g(x) = \sqrt{2\pi}(-1)^{\lfloor x \rfloor}$. We can also just calculate b: by symmetry, $b = 0$ since $\lfloor -x \rfloor = -\lfloor x \rfloor - 1$ except when x is an integer, so the integral from $-\infty$ to 0 cancels that from 0 to ∞, so we can simply take $g(x) = 0$.

(b) Write down a function $h(u)$ such that $E(h(U)) = b$ for $U \sim \text{Unif}(0,1)$ (your function should *not* be in terms of b, and can be in terms of the function g from (a) and the standard Normal CDF Φ).

By Universality of the Uniform, $\Phi^{-1}(U) \sim \mathcal{N}(0,1)$, so define $X = \Phi^{-1}(U)$. Then $E(g(\Phi^{-1}(U))) = b$, so we can take $h(u) = g(\Phi^{-1}(u))$.

(c) Let $X_1, X_2, \dots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ with n large, and let g be as in (a). What is the approximate distribution of $\frac{1}{n}(g(X_1) + \cdots + g(X_n))$? Simplify the parameters fully (in terms of b and n), and mention which theorems you are using.

For the choice of g obtained from LOTUS, we have $E(g(X)) = b$ and $\text{Var}(g(X)) = 2\pi - b^2$ (since $g(x)^2 = 2\pi$), so by the CLT, the approximate distribution is $\mathcal{N}(b, (2\pi - b^2)/n)$.

For the choice $g(x) = 0$, the distribution is degenerate, giving probability 1 to the value 0.

6. Let X_1 be the number of emails received by a certain person today and let X_2 be the number of emails received by that person tomorrow, with X_1 and X_2 i.i.d.

(a) Find $E(X_1|X_1 + X_2)$ (simplify).

By symmetry, $E(X_1|X_1 + X_2) = E(X_2|X_1 + X_2)$. By linearity,

$$E(X_1|X_1 + X_2) + E(X_2|X_1 + X_2) = E(X_1 + X_2|X_1 + X_2) = X_1 + X_2.$$

So

$$E(X_1|X_1 + X_2) = (X_1 + X_2)/2.$$

(b) For the case $X_j \sim \text{Pois}(\lambda)$, find the conditional distribution of X_1 given $X_1 + X_2$, i.e., $P(X_1 = k|X_1 + X_2 = n)$ (simplify). Is this one of the "important distributions"?

By Bayes' Rule and the fact that $X_1 + X_2 \sim \text{Pois}(2\lambda)$,

$$\begin{aligned} P(X_1 = k|X_1 + X_2 = n) &= P(X_1 + X_2 = n|X_1 = k)P(X_1 = k)/P(X_1 + X_2 = n) \\ &= P(X_2 = n - k)P(X_1 = k)/P(X_1 + X_2 = n) \\ &= \frac{e^{-\lambda}\lambda^{n-k}}{(n-k)!}\frac{e^{-\lambda}\lambda^{k}}{k!}e^{2\lambda}(2\lambda)^{-n}n! \\ &= \binom{n}{k}\left(\frac{1}{2}\right)^{n}. \end{aligned}$$

Thus, the conditional distribution is $\text{Bin}(n, 1/2)$. Note that the λ disappeared! This is not a coincidence; there is an important statistical reason for this, but that is a story for another day and another course.

7. Let X_1, X_2, X_3 be independent with $X_i \sim \text{Expo}(\lambda_i)$ (so with possibly different rates). A useful fact (which you may use) is that $P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

(a) Find $E(X_1 + X_2 + X_3 | X_1 > 1, X_2 > 2, X_3 > 3)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

By linearity, independence, and the memoryless property, we get

$$E(X_1|X_1 > 1) + E(X_2|X_2 > 2) + E(X_3|X_3 > 3) = \lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1} + 6.$$

(b) Find $P(X_1 = \min(X_1, X_2, X_3))$, the probability that the first of the three Exponentials is the smallest. Hint: re-state this in terms of X_1 and $\min(X_2, X_3)$.

The desired probability is $P(X_1 \leq \min(X_2, X_3))$. Noting that $\min(X_2, X_3) \sim \text{Expo}(\lambda_2 + \lambda_3)$ is independent of X_1, we have

$$P(X_1 \leq \min(X_2, X_3)) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}.$$

(c) For the case $\lambda_1 = \lambda_2 = \lambda_3 = 1$, find the PDF of $\max(X_1, X_2, X_3)$. Is this one of the "important distributions"?

Let $M = \max(X_1, X_2, X_3)$. Using the order statistics results from class or by directly computing the CDF and taking the derivative, for $x > 0$ we have

$$f_M(x) = 3(1 - e^{-x})^2 e^{-x}.$$

This is not one of the "important distributions". (The form is reminiscent of a Beta, but a Beta takes values between 0 and 1, while M can take any positive real value; in fact, $B \sim \text{Beta}(1, 3)$ if we make the transformation $B = e^{-M}$.)

8. Let X_n be the price of a certain stock at the start of the nth day, and assume that $X_0, X_1, X_2, \dots$ follows a Markov chain with transition matrix Q (assume for simplicity that the stock price can never go below 0 or above a certain upper bound, and that it is always rounded to the nearest dollar).

(a) A lazy investor only looks at the stock once a year, observing the values on days $0, 365, 2 \cdot 365, 3 \cdot 365, \dots$. So the investor observes $Y_0, Y_1, \dots$, where Y_n is the price after n years (which is $365n$ days; you can ignore leap years). Is $Y_0, Y_1, \dots$ also a Markov chain? Explain why or why not; if so, what is its transition matrix?

Yes, it is a Markov chain: given the whole past history $Y_0, Y_1, \dots, Y_n$, only the most recent information Y_n matters for predicting Y_{n+1}, because $X_0, X_1, \dots$ is Markov. The transition matrix of $Y_0, Y_1, \dots$ is Q^{365}, since the kth power of Q gives the k-step transition probabilities.

(b) The stock price is always an integer between \$0 and \$28. From each day to the next, the stock goes up or down by \$1 or \$2, all with equal probabilities (except for days when the stock is at or near a boundary, i.e., at \$0, \$1, \$27, or \$28).
If the stock is at \$0, it goes up to \$1 or \$2 on the next day (after receiving government bailout money). If the stock is at \$28, it goes down to \$27 or \$26 the next day. If the stock is at \$1, it either goes up to \$2 or \$3, or down to \$0 (with equal probabilities); similarly, if the stock is at \$27 it either goes up to \$28, or down to \$26 or \$25. Find the stationary distribution of the chain (simplify).

This is an example of random walk on an undirected network, so we know the stationary probability of each node is proportional to its degree. The degrees are $(2, 3, 4, 4, \dots, 4, 4, 3, 2)$, where there are $29 - 4 = 25$ 4's. The sum of these degrees is 110 (coincidentally?). Thus, the stationary distribution is

$$\left(\frac{2}{110}, \frac{3}{110}, \frac{4}{110}, \frac{4}{110}, \dots, \frac{4}{110}, \frac{4}{110}, \frac{3}{110}, \frac{2}{110}\right),$$

with 25 $\frac{4}{110}$'s.

5 Solutions to Stat 110 Final from 2010

1. Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match.

(a) What is the probability that Calvin wins the match (in terms of p)?

Hint: condition on the results of the first k games (for some choice of k).

Let C be the event that Calvin wins the match, $X \sim \text{Bin}(2, p)$ be how many of the first 2 games he wins, and $q = 1 - p$. Then

$$P(C) = P(C|X=0)q^2 + P(C|X=1)(2pq) + P(C|X=2)p^2 = 2pqP(C) + p^2,$$

so $P(C) = \frac{p^2}{1-2pq}$. This can also be written as $\frac{p^2}{p^2+q^2}$ since $p + q = 1$. (Also, the problem can be thought of as gambler's ruin where each player starts out with \$2.)

Miracle check: Note that this should (and does) reduce to 1 for $p = 1$, 0 for $p = 0$, and $\frac{1}{2}$ for $p = \frac{1}{2}$. Also, it makes sense that the probability of Hobbes winning, which is $1 - P(C) = \frac{q^2}{p^2+q^2}$, can also be obtained by swapping p and q.

(b) Find the expected number of games played.

Hint: consider the first two games as a pair, then the next two as a pair, etc.

Think of the first 2 games, the 3rd and 4th, the 5th and 6th, etc. as "mini-matches." The match ends right after the first mini-match which isn't a tie. The probability of a mini-match not being a tie is $p^2 + q^2$, so the number of mini-matches needed is 1 plus a $\text{Geom}(p^2 + q^2)$ r.v. Thus, the expected number of games is $\frac{2}{p^2+q^2}$.

Miracle check: For $p = 0$ or $p = 1$, this reduces to 2. The expected number of games is maximized when $p = \frac{1}{2}$, which makes sense intuitively. Also, it makes sense that the result is symmetric in p and q.

2. A DNA sequence can be represented as a sequence of letters, where the "alphabet" has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_1, p_2, p_3, p_4 respectively.

(a) In a DNA sequence of length 115, what is the expected number of occurrences of the expression "CATCAT" (in terms of the p_j)? (Note that, for example, the expression "CATCATCAT" counts as 2 occurrences.)

Let I_j be the indicator r.v. of "CATCAT" appearing starting at position j, for $1 \leq j \leq 110$. Then $E(I_j) = (p_1p_2p_3)^2$, so the expected number is $110(p_1p_2p_3)^2$.

Miracle check: Stat 115 is the bioinformatics course here and Stat 110 is this course, so $109(p_1p_2p_3)^2$ would have been a much less aesthetically pleasing result (this kind of "off by one" error is extremely common in programming, but is not hard to avoid by doing a quick check). The number of occurrences is between 0 and 110, so the expected value should also be between 0 and 110.

(b) What is the probability that the first A appears earlier than the first C appears, as letters are generated one by one (in terms of the p_j)?

Consider the first letter which is an A or a C (call it X; alternatively, condition on the first letter of the sequence). This gives

$$P(\text{A before C}) = P(X \text{ is A}|X \text{ is A or C}) = \frac{P(X \text{ is A})}{P(X \text{ is A or C})} = \frac{p_1}{p_1 + p_2}.$$

Miracle check: The answer should be 1/2 for $p_1 = p_2$, should go to 0 as $p_1 \to 0$, should be increasing in p_1 and decreasing in p_2, and finding $P(\text{A before C})$ by $1 - P(\text{A before C})$ should agree with finding it by swapping p_1, p_2.

(c) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a $\text{Unif}(0, 1)$ r.v. before observing any data, and that then the first 3 letters observed are "CAT". Given this information, what is the probability that the next letter is C?

Let X be the number of C's in the data (so $X = 1$ is observed here). The prior is $p_2 \sim \text{Beta}(1, 1)$, so the posterior is $p_2|X = 1 \sim \text{Beta}(2, 3)$ (by the connection between Beta and Binomial, or by Bayes' Rule). Given p_2, the indicator of the next letter being C is $\text{Bern}(p_2)$. So given X (but not given p_2), the probability of the next letter being C is $E(p_2|X) = \frac{2}{5}$.

Miracle check: It makes sense that the answer should be strictly in between 1/2 (the mean of the prior distribution) and 1/3 (the observed frequency of C's in the data).

3. Let X and Y be i.i.d. *positive* random variables. Assume that the various expressions below exist. Write the most appropriate of $\leq$, $\geq$, $=$, or ? in the blank for each part (where "?" means that no relation holds in general). It is *not* necessary to justify your answers for full credit; some partial credit is available for justified answers that are flawed but on the right track.

(a) $E(e^{X+Y}) \geq e^{2E(X)}$ (write $E(e^{X+Y}) = E(e^X e^Y) = E(e^X)E(e^Y) = E(e^X)E(e^X)$ using the fact that X, Y are i.i.d., and then apply Jensen)

(b) $E(X^2 e^X) \leq \sqrt{E(X^4)E(e^{2X})}$ (by Cauchy-Schwarz)

(c) $E(X|3X) = E(X|2X)$ (knowing $2X$ is equivalent to knowing $3X$)

(d) $E(X^7 Y) = E(X^7 E(Y|X))$ (by Adam's law and taking out what's known)

(e) $E(\frac{X}{Y} + \frac{Y}{X}) \geq 2$ (since $E(\frac{X}{Y}) = E(X)E(\frac{1}{Y}) \geq \frac{EX}{EY} = 1$, and similarly $E(\frac{Y}{X}) \geq 1$)

(f) $P(|X - Y| > 2) \leq \frac{\text{Var}(X)}{2}$ (by Chebyshev, applied to the r.v. $W = X - Y$, which has variance $2\text{Var}(X)$: $P(|W - E(W)| > 2) \leq \text{Var}(W)/4 = \text{Var}(X)/2$)

4. Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \dots,$ and let $p_k = P(X = x_k)$. The *entropy* of X is defined to be $H(X) = -\sum_{k=0}^{\infty} p_k \log_2(p_k)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

Hint: use properties of logs, and interpret part of the sum as an expected value.

$$\begin{aligned} H(X) &= -\sum_{k=0}^{\infty}(pq^k)\log_2(pq^k) \\ &= -\log_2(p)\sum_{k=0}^{\infty} pq^k - \log_2(q)\sum_{k=0}^{\infty} kpq^k \\ &= -\log_2(p) - \frac{q}{p}\log_2(q), \end{aligned}$$

with $q = 1 - p$, since the first series is the sum of a Geom(p) PMF and the second series is the expected value of a Geom(p) r.v.

Miracle check: entropy must be positive (unless X is a constant), since it is the average "surprise" (where the "surprise" of observing $X = x_k$ is $-\log_2(p_k) = \log_2(\frac{1}{p_k})$).

(b) Find $H(X^3)$ for $X \sim \text{Geom}(p)$, in terms of $H(X)$.

Let $p_k = pq^k$. Since X^3 takes values $0^3, 1^3, 2^3, \dots$ with probabilities $p_0, p_1, \dots$ respectively, we have $H(X^3) = H(X)$.

Miracle check: The definition of entropy depends on the *probabilities* p_k of the values x_k, not on the values x_k themselves, so taking a one-to-one function of X should not change the entropy.

(c) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

Hint: Consider $E(\log_2(W))$, where W is a r.v. taking value p_k with probability p_k.

Let W be as in the hint. By Jensen, $E(\log_2(W)) \leq \log_2(EW)$. But

$$E(\log_2(W)) = \sum_k p_k \log_2(p_k) = -H(X),$$

$$EW = \sum_k p_k^2 = P(X = Y),$$

so $-H(X) \leq \log_2 P(X = Y)$. Thus, $P(X = Y) \geq 2^{-H(X)}$.

5. Let $Z_1, \dots, Z_n \sim \mathcal{N}(0,1)$ be i.i.d.

(a) As a function of Z_1, create an Expo(1) r.v. X (your answer can also involve the standard Normal CDF Φ).

Use Z_1 to get a Uniform and then the Uniform to get X: we have $\Phi(Z_1) \sim \text{Unif}(0,1)$, and we can then take $X = -\ln(1-\Phi(Z_1))$. By symmetry, we can also use $-\ln(\Phi(Z_1))$.

Miracle check: $0 < \Phi(Z_1) < 1$, so $-\ln(\Phi(Z_1))$ is well-defined and positive.

(b) Let $Y = e^{-R}$, where $R = \sqrt{Z_1^2 + \cdots + Z_n^2}$. Write down (but do not evaluate) an integral for $E(Y)$.

Let $W = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$, so $Y = e^{-\sqrt{W}}$. We will use LOTUS to write $E(Y)$ using the PDF of W (there are other possible ways to use LOTUS here, but this is simplest since we get a single integral and we know the χ_n^2 PDF). This gives

$$E(Y) = \int_0^\infty e^{-\sqrt{w}} \frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-w/2} dw.$$

(c) Let $X_1 = 3Z_1 - 2Z_2$ and $X_2 = 4Z_1 + 6Z_2$. Determine whether X_1 and X_2 are independent (being sure to mention which results you're using).

There are uncorrelated:

$$\text{Cov}(X_1, X_2) = 12\text{Var}(Z_1) + 10\text{Cov}(Z_1, Z_2) - 12\text{Var}(Z_2) = 0.$$

Also, (X_1, X_2) is Multivariate Normal since any linear combination of X_1, X_2 can be written as a linear combination of Z_1, Z_2 (and thus is Normal since the sum of two independent Normals is Normal). So X_1 and X_2 are independent.

6. Let $X_1, X_2, \dots$ be i.i.d. positive r.v.s. with mean μ, and let $W_n = \frac{X_1}{X_1+\cdots+X_n}$.

(a) Find $E(W_n)$.

Hint: consider $\frac{X_1}{X_1+\cdots+X_n} + \frac{X_2}{X_1+\cdots+X_n} + \cdots + \frac{X_n}{X_1+\cdots+X_n}$.

The expression in the hint equals 1, and by linearity and symmetry its expected value is $nE(W_n)$. So $E(W_n) = 1/n$.

Miracle check: in the case that the X_j are actually constants, $\frac{X_1}{X_1+\cdots+X_n}$ reduces to $\frac{1}{n}$. Also in the case $X_j \sim \text{Expo}(\lambda)$, part (c) shows that the answer should reduce to the mean of a $\text{Beta}(1, n-1)$ (which is $\frac{1}{n}$).

(b) What random variable does nW_n converge to as $n \to \infty$?

By the Law of Large Numbers, with probability 1 we have

$$nW_n = \frac{X_1}{(X_1 + \cdots + X_n)/n} \to \frac{X_1}{\mu} \text{ as } n \to \infty.$$

Miracle check: the answer should be a random variable since it's asked what r.v. nW_n converges to. It should *not* depend on n since we let $n \to \infty$.

(c) For the case that $X_j \sim \text{Expo}(\lambda)$, find the distribution of W_n, preferably without using calculus. (If it is one of the "important distributions" state its name and specify the parameters; otherwise, give the PDF.)

Recall that $X_1 \sim \text{Gamma}(1)$ and $X_2 + \cdots + X_n \sim \text{Gamma}(n-1)$. By the connection between Beta and Gamma (i.e., the bank-post office story), $W_n \sim \text{Beta}(1, n-1)$.

Miracle check: the distribution clearly always takes values between 0 and 1, and the mean should agree with the answer from (a).

7. A task is randomly assigned to one of two people (with probability 1/2 for each person). If assigned to the first person, the task takes an Expo(λ_1) length of time to complete (measured in hours), while if assigned to the second person it takes an Expo(λ_2) length of time to complete (independent of how long the first person would have taken). Let T be the time taken to complete the task.

(a) Find the mean and variance of T.

Write $T = IX_1 + (1-I)X_2$, with $I \sim \text{Bern}(1/2), X_1 \sim \text{Expo}(\lambda_1), X_2 \sim \text{Expo}(\lambda_2)$ independent. Then

$$ET = \frac{1}{2}(\lambda_1^{-1} + \lambda_2^{-1}),$$

$$\begin{aligned}
\text{Var}(T) &= E(\text{Var}(T|I)) + \text{Var}(E(T|I)) \\
&= E(I^2\frac{1}{\lambda_1^2} + (1-I)^2\frac{1}{\lambda_2^2}) + \text{Var}\left(\frac{I}{\lambda_1} + \frac{1-I}{\lambda_2}\right) \\
&= E(I\frac{1}{\lambda_1^2} + (1-I)\frac{1}{\lambda_2^2}) + \text{Var}\left(I(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})\right) \\
&= \frac{1}{2}(\frac{1}{\lambda_1^2} + \frac{1}{\lambda_2^2}) + \frac{1}{4}(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})^2.
\end{aligned}$$

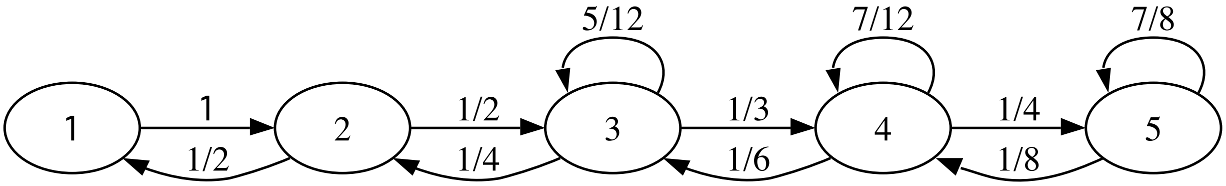
Miracle check: for $\lambda_1 = \lambda_2$, the two people have the same distribution so randomly assigning the task to one of the two should be equivalent to just assigning it to the first person (so the mean and variance should agree with those of an Expo(λ_1) r.v.). It makes sense that the mean is the average of the two means, as we can condition on whether $I = 1$ (though the variance is *greater* than the average of the two variances, by Eve's Law). Also, the results should be (and are) the same if we swap λ_1 and λ_2.

(b) Suppose instead that the task is assigned to *both* people, and let X be the time taken to complete it (by whoever completes it first, with the two people working independently). It is observed that after 24 hours, the task has not yet been completed. Conditional on this information, what is the expected value of X?

Here $X = \min(X_1, X_2)$ with $X_1 \sim \text{Expo}(\lambda_1), X_2 \sim \text{Expo}(\lambda_2)$ independent. Then $X \sim \text{Expo}(\lambda_1 + \lambda_2)$ (since $P(X > x) = P(X_1 > x)P(X_2 > x) = e^{-(\lambda_1+\lambda_2)x}$, or by results on order statistics). By the memoryless property,

$$E(X|X > 24) = 24 + \frac{1}{\lambda_1 + \lambda_2}.$$

Miracle check: the answer should be greater than 24 and should be very close to 24 if λ_1 or λ_2 is very large. Considering a Poisson process also helps make this intuitive.



8. Find the stationary distribution of the Markov chain shown above, *without using matrices*. The number above each arrow is the corresponding transition probability.

We will show that this chain is reversible by solving for $\mathbf{s}$ (which will work out nicely since this is a birth-death chain). Let q_{ij} be the transition probability from i to j, and solve for $\mathbf{s}$ in terms of s_1. Noting that $q_{ij} = 2q_{ji}$ for $j = i + 1$ (when $1 \leq i \leq 4$), we have that

$$s_1 q_{12} = s_2 q_{21} \text{ gives } s_2 = 2s_1.$$

$$s_2 q_{23} = s_3 q_{32} \text{ gives } s_3 = 2s_2 = 4s_1.$$

$$s_3 q_{34} = s_4 q_{43} \text{ gives } s_4 = 2s_3 = 8s_1.$$

$$s_4 q_{45} = s_5 q_{54} \text{ gives } s_5 = 2s_4 = 16s_1.$$

The other reversibility equations are automatically satisfied since here $q_{ij} = 0$ unless $|i - j| \leq 1$. Normalizing, the stationary distribution is

$$\left(\frac{1}{31}, \frac{2}{31}, \frac{4}{31}, \frac{8}{31}, \frac{16}{31}\right).$$

Miracle check: this chain "likes" going from left to right more than from right to left, so the stationary probabilities should be increasing from left to right. We also know that $s_j = \sum_i s_i q_{ij}$ (since if the chain is in the stationary distribution at time n, then it is also in the stationary distribution at time $n + 1$), so we can check, for example, that $s_1 = \sum_i s_i q_{i1} = \frac{1}{2}s_2$.