Stat 110 Notesheet

Elvin Lo

1 Math

Taylor: $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$ $\quad \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n = e^x$

Vandermonde: $\sum_{k=0}^{r} \binom{m}{k}\binom{r}{r-k} = \binom{m+n}{r}$

Harmonic series: $\sum_{k=1}^{n} \frac{1}{k} \approx \log(n) + 0.577$

Bayes' Billiards: $\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}$ for $0 \le k \le n$

Geometric: $1 + \cdots + r^n = \frac{1-r^{n+1}}{1-r}$

Pattern matching integrals: $\int_0^\infty x^{t-1}e^{-x}dx = \Gamma(t)$ $\quad \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

2 Properties and Stories of Probability

Bayes' with two pieces of information: Suppose we want to calculate with Bayes' the probability $P(A \mid B \cap C)$.

- Updating sequentially, we may consider that we have first observed B, and then everything following will be conditioned on B.
$$P(A \mid B, C) = \frac{P(C \mid A, B)P(A \mid B)}{P(C \mid B)}.$$
- Updating simultaneously, we simply have
$$P(A \mid B \cap C) = \frac{P(B \cap C \mid A)P(A)}{P(B \cap C)}.$$

Remarks on independence of events:

- If A and B are independent, then all the following are independent: A and B^c, A^c and B, A^c and B^c.
- Independence is not transitive.
- **Conditional independence:** Events A and B are conditionally independent given E if
$$P(A \cap B|E) = P(A|E)P(B|E).$$
But conditional independence does not imply independence, and vice versa. Two events independent given E also may not be independent given E^c.
- **Functions of independent r.v.s:** if X and Y independent r.v.s, then any function of X is independent of any function of Y. (See section 8 for definition of independence of r.v.s, and conditional independence is defined analogously)

Miscellaneous classical probability:

- **PMF of $g(X)$:** if g is one-to-one, then we simply have $P(g(X) = g(x)) = P(X = x)$

Simpson's: This is a statistical phenomenon where an association between two variables in a population emerges, disappears, or reverses when the population is divided into subpopulations.

General form. Given events A, B, C, we say that we have a Simpson's paradox if both

$$P(A|B,C) < P(A|B^c,C) \qquad P(A|B,C^c) < P(A|B^c,C^c).$$

but

$$P(A|B) > P(A|B^c).$$

Explanation. The LOTP shows why this can happen:

$$P(A|B) = P(A|C,B)P(C|B) + P(A|C^c,B)P(C^c|B),$$
$$P(A|B^c) = P(A|C,B^c)P(C|B^c) + P(A|C^c,B^c)P(C^c|B^c).$$

The above equations express $P(A|B)$ as a weighted average of $P(A|C,B)$ and $P(A|C^c,B)$, and $P(A|B^c)$ as a weighted average of $P(A|C,B^c)$ and $P(A|C^c,B^c)$. If the corresponding weights were the same in both of these weighted averages, then Simpson's paradox could not occur. But the weights may be very different, e.g.,

$$P(C|B) < P(C|B^c) \text{ and } P(C^c|B) > P(C^c|B^c),$$

and if the differences in weights are sufficiently drastic, then Simpson's may occur.

Case study. To build intuition, consider the classical example: we have two doctors, Dr. Hibbert and Dr. Nick, each performing two types of surgeries: heart surgery and Band-Aid removal. Let A be the event of a successful surgery, B be the event that Dr. Nick is the surgeon, and C be the event that the surgery is a heart surgery.

Simpson's paradox may arise here if the probability of a successful surgery is lower under Dr. Nick than under Dr. Hibbert whether we condition on heart surgery or on Band-Aid removal, but the overall probability of success is higher for Dr. Nick.

Monty Hall: Consider the Monty Hall problem and assume WLOG that we begin by choosing door 1. Suppose we decide preemptively that we want to switch doors. Now let us calculate the unconditional and conditional probabilities of getting the car. Define events

- W that we get the car;
- C_i that the car is behind door i;
- M_j that Monty opens door j to reveal a goat (for $j = 2, 3$).

Then we may calculate the unconditional probability by LOTP:

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) = 0 + \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Supposing M_j happens for $j = 2, 3$, we may calculate the conditional probability using Bayes',

$$P(C_1|M_j) = \frac{P(M_j|C_1)P(C_1)}{P(M_j)} = \frac{(1/2)(1/3)}{1/2} = \frac{1}{3},$$

so then the conditional probability of winning is again $2/3$ if we switch.

Gambler's ruin: Two gamblers, A and B, make a sequence of \$1 bets. In each bet, gambler A has probability p of winning, and gambler B has probability $q = 1 - p$ of winning. Gambler A starts with i dollars and gambler B starts with $N - i$ dollars.

We may visualize this game as a random walk which terminates with A's victory at position N or terminates with B's victory at position 0. To solve this problem, we may use states to derive what we call a difference equation, and then construct and solve the characteristic polynomial of the difference equation. Solving, we have that the probability of A winning with a starting wealth of i is

$$p_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N} & \text{if } p \neq 1/2 \\ \frac{i}{N} & \text{if } p = 1/2 \end{cases}.$$

3 Random Variables

Universality of the Uniform: Consider X continuous with CDF F strictly increasing on its support (so F^{-1} exists).

- $F(X) \sim \text{Unif}(0, 1)$
- $F^{-1}(U)$ has CDF F, where $U \sim \text{Unif}(0, 1)$.

Transformation/change of variables: Consider r.v. X and let $Y = g(X)$, where g is differentiable and strictly increasing or strictly decreasing. Then

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

where $x = g^{-1}(y)$, and the support of Y is all $g(x)$ with x in the support of X.

Order statistics: The order statistic $X_{(i)}$ is the i^{th} smallest of i.i.d. $X_1, \ldots, X_n$. We have

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

Special uniform case: if $X_i \sim \text{Unif}(0,1)$, we get $X_{(j)} \sim \text{Beta}(j, n-j+1)$ and $E\left(U_{(j)}\right) = {}^{j}/_{n+1}$.

MGFs: We have $M_X(t) = E(X^{tX})$, and this helps us determine the n^{th} moments by:

$$E\left(X^n\right) = M^{(n)}(0).$$

If it exists (i.e., is finite on some interval containing 0), the MGF uniquely determines the distribution. Also if X has MGF $M(t)$, then $a+bX$ has MGF $E\left(e^{t(a+bX)}\right) = e^{at}E\left(e^{btX}\right) = e^{at}M(bt)$.

Moments: skewness is the third standardized moment, $\text{Skew}(X) = E\left({}^{X-\mu}/_{\sigma}\right)^3$. Kurtosis is a shifted fourth standardized moment, $\text{Kurt}(X) = E\left({}^{X-\mu}/_{\sigma}\right)^4 - 3$, and measures tailedness of a distribution relative to the normal distribution.

Symmetry: X is symmetric about μ, which must be the mean and median should it exist, if $X-\mu$ and $\mu-X$ have the same distribution (i.e., if for all x we have $f(x) = f(2\mu - x)$)

4 Expected Value

Covariance: Covariance is a linear measure of whether X, Y tend to move in the same direction,

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

But r.v.s can be dependent in nonlinear ways and still have zero covariance. We have the properties:

- $\text{Cov}(X+Y, Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$
- $\text{Cov}(cX, Y) = c\,\text{Cov}(X,Y)$
- $\text{Var}(X) = \text{Cov}(X,X)$
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$, or more generally
$$\text{Var}\left(X_1 + \cdots + X_n\right) = \text{Var}\left(X_1\right) + \cdots + \text{Var}\left(X_n\right) + 2\sum_{i<j}\text{Cov}\left(X_i, X_j\right)$$
For X_i identically distributed and with symmetric covariances, the RHS becomes $n\,\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$.
- Intuitively, if X, Y independent, then $\text{Cov}(X,Y) = 0$.

Correlation: correlation is invariant under shifting/scaling of X, Y and is always in $[-1, 1]$,

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

LOTUS: Given r.v. X and $g : \mathbb{R} \to \mathbb{R}$, then summing over all possible values of X we have

$$E(g(X)) = \sum_x g(x)P(X=x) \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x).$$

In two-dimensions, with $g : \mathbb{R}^2 \to \mathbb{R}$, we have

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

Monotonicity of expectation: let X and Y be r.v.s such that $X \geq Y$ with probability 1. Then $E(X) \geq E(Y)$, with equality holding iff $X = Y$ with probability 1.

St. Petersburg paradox: Suppose a wealthy stranger offers to play the following game with you. You will flip a fair coin until it lands Heads for the first time, and you will receive \$2 if the game lasts for 1 round, \$4 if the game lasts for 2 rounds, \$8 if the game lasts for 3 rounds, and in general, \$$2n$ if the game lasts for n rounds. What is the fair value of this game (the expected payoff)? How much would you be willing to pay to play this game once?

Solution: Let X be your winnings from playing the game. By definition, $X = 2^N$ where N is the number of rounds that

the game lasts. We have

$$E(X) = E(2^N) = \tfrac{1}{2} \cdot 2 + \tfrac{1}{4} \cdot 4 + \cdots = \infty,$$

even though $E(N) = 2$ and $2^{E(N)} = 4$.

Lesson: The ∞ in this paradox is driven by an infinite "tail" of extremely rare events where you get extremely large payoffs. Cutting off this tail at some point, which makes sense in the real world, dramatically reduces the expected value of the game. This paradox illustrates the danger of confusing $E(g(X))$ with $g(E(X))$ when g is not linear.

5 Conditional Expectation

Conditional expectation: Define $E(X \mid A) = \sum_x xP(X = x \mid A)$ and $E(Y \mid X) = g(X)$ where $g(x) = E(Y \mid X = x)$.

- **Adam:** $E(E(Y \mid X)) = E(Y)$
- **Eve (law of total variance):** defining $\mathrm{Var}(Y \mid X) = E\left(Y^2 \mid X\right) - (E(Y \mid X))^2$, we have
$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X))$$
- **Dropping what's independent:** for independent X, Y, we have $E(Y \mid X) = E(Y)$
- **Taking out what's known:** for any function h, we have
$$E(h(X)Y \mid X) = h(X)E(Y \mid X)$$
- **LOTE:** For a partition $\{A_i\}$ of the sample space,
$$E(Y) = \sum_{i=1}^{n} E\left(Y \mid A_i\right) P\left(A_i\right)$$

6 Discrete Distributions

Hypergeometric: Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random without replacement. Let X be the number of white balls in the sample,

$$X \sim \mathrm{HGeom}(w, b, n) \qquad P(X = k) = \tbinom{w}{k}\tbinom{b}{n-k}/\tbinom{w+b}{n},$$

for k satisfying $0 \le k \le w$ and $0 \le n - k \le b$.

Binomial to hypergeometric: if $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ independent, $X \mid X + Y = r \sim \mathrm{HGeom}(n, m, r)$.
Elk capture-recapture problem: 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? Answer: $\mathrm{HGeom}(n, N - n, m)$

Negative Binomial: In a sequence of independent Bernoulli trials with success probability p, $X \sim \mathrm{NBin}(r, p)$ measures the number of failures before the r^{th} success. If X_i i.i.d. $\mathrm{Geom}(p)$, of course

$$X = X_1 + \cdots + X_r,$$

also implying that if $X \sim \mathrm{NBin}\left(r_1, p\right)$ and $Y \sim \mathrm{NBin}\left(r_2, p\right)$ independent, then $X + Y \sim \mathrm{NBin}\left(r_1 + r_2, p\right)$

Poisson: Poisson distributions are used when counting the number of successes in a time interval, where there are a large number of trials with each a small success probability. The parameter λ is the rate of occurrence; we would expect λ occurrences in this interval.

- **Sum of independent Poissons:** if $X \sim \mathrm{Pois}(\lambda_1)$ and $Y \sim \mathrm{Pois}(\lambda_2)$ independent, then $X + Y \sim \mathrm{Pois}(\lambda_1 + \lambda_2)$
- **Condition into Binomial:** we have $X \mid (X + Y = n) \sim \mathrm{Bin}(n, \lambda_1/\lambda_1+\lambda_2)$
- **Chicken-egg:** consider $N \sim \mathrm{Pois}(\lambda)$ items, and suppose we randomly and independently accept each item with probability p. Then the number of accepted items is $X \sim \mathrm{Pois}(\lambda p)$, the number of rejected items is $Y \sim \mathrm{Pois}(\lambda(1-p))$, and X and Y independent.
- **Binomial limit to Poisson:** suppose $X \sim \mathrm{Bin}(n, p)$ and $n \to \infty$ while $p \to 0$ such that $\lambda = np$ constant. Then the X PMF converges to $\mathrm{Pois}(\lambda)$.
- **Poisson paradigm:** consider events $A_1, \ldots, A_n$ with $p_j = P\left(A_j\right)$, where n large, p_j small, and A_j independent or weakly dependent. Then we have the Poisson approximation
$$X = \sum_{j=1}^{n} I\left(A_j\right) \dot{\sim} \mathrm{Pois}(E(X))$$
- **Poisson MGF:** given $X \sim \mathrm{Pois}(\lambda)$, we have $M_X(t) = E(e^{tX}) = e^{\lambda(e^t-1)}$

Poisson process of rate λ: The number of arrivals that occur in time t is distributed Pois(λt), and the numbers of arrivals in disjoint time intervals are independent.

- **Count-time duality:** let discrete r.v. N_t be the number of arrivals in time t, and continuous r.v. T_n the time of the n^{th} arrival. Then $T_n > t$ and $N_t < n$ are the same event.
- **Exponential waiting times:** we have $T_1 \sim \text{Expo}(\lambda), T_2 - T_1 \sim \text{Expo}(\lambda)$, and so on

7 Continuous Distributions

Normal: If $Z \sim \mathcal{N}(0,1)$:

- Symmetry of PDF: φ is even, $\varphi(z) = \varphi(-z)$
- Symmetry of CDF tails: $\Phi(z) + \Phi(-z) = 1$
- Symmetry of Z and $-Z$: $-Z \sim \mathcal{N}(0,1)$
- Normal MGF: for normal W, $E\left(e^W\right) = e^{E(W)+\frac{1}{2}\text{Var}(W)}$

Exponential: For $X \sim \text{Expo}(\lambda)$, we have support $x > 0$ and

$$f(x) = \lambda e^{-\lambda x} \qquad F(x) = 1 - e^{-\lambda x}$$

- **Memoryless property:** for $X \sim \text{Expo}(\lambda)$,
$$X - s \mid X \geq s \sim \text{Expo}(\lambda) \implies E(X \mid X \geq s) = s + E(X) = s + 1/\lambda$$
- **Minimum of independent Expos:** consider independent $X_i \sim \text{Expo}(\lambda_i)$, then
$$L = \min(X_1, \ldots, X_n) \implies L \sim \text{Expo}\left(\lambda_1 + \cdots + \lambda_n\right)$$
- **Maximum of independent Expos:** consider $T_1 = X_{(1)}$ the minimum of the Expos, then $T_2 = X_{(2)}$, and so on. For T_2, we know that all the other Expos are at least T_1, and so we may apply memoryless and then T_2 is again a minimum of independent Expos like T_1.
- **First arrival:** For independent $T_1 \sim \text{Expo}(\lambda_1)$ and $T_2 \sim \text{Expo}(\lambda_2)$, we have
$$P(T_1 < T_2) = \lambda_1/\lambda_1+\lambda_2$$
- **Scaling Expos:** $Y \sim \text{Expo}(\lambda) \implies X = \lambda Y \sim \text{Expo}(1)$.
- **Expo MGF:** for $X \sim \text{Expo}(\lambda)$, we have
$$M_X(t) = \lambda/\lambda-t$$

Beta: For $X \sim \text{Beta}(a,b)$, we have

$$f(x) \propto \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

- **Special uniform case:** Beta$(1,1)$ is the same as Unif$(0,1)$
- **Beta-Binomial conjugacy:** consider a series of Bernoulli trials with unknown success probability p with prior Beta$(1,1)$. We observe $X \mid p \sim \text{Bin}(n,p)$, the number of successes in the first n trials, and update to posterior
$$p \mid X = k \sim \text{Beta}(a+k, b+n-k)$$

Gamma:

- **Gamma is sum of Expos:** if X_i are i.i.d. Expo(λ), then
$$X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda).$$
This implies Gamma$(1,\lambda)$ is Expo(λ)
- **Sum of independent Gammas:** given independent $X \sim \text{Gamma}\left(a_1, \lambda\right)$ and $Y \sim \text{Gamma}\left(a_2, \lambda\right)$, we have $X + Y \sim \text{Gamma}\left(a_1 + a_2, \lambda\right)$
- **Gamma-Poisson conjugacy:** consider a Poisson process with unknown rate λ with prior Gamma(r_0, b_0). (Intuitively, r_0 is the number of prior arrivals and b_0 is total time taken for those prior arrivals.) We observe $Y_t \mid \lambda \sim \text{Pois}(\lambda t)$, the number of arrivals in time t, and update to posterior
$$\lambda \mid Y_t = k \sim \text{Gamma}(r_0 + y, b_0 + t)$$

- **Bank-Post Office story:** given independent $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$, we have independent r.v.s

$$T = X + Y \sim \text{Gamma}(a + b, \lambda) \qquad W = X/(X+Y) \sim \text{Beta}(a, b).$$

Chi-Square: We have $V \sim \chi_n^2$ has the Chi-Square distribution with n degrees of freedom if

$$V = Z_1^2 + \cdots + Z_n^2$$

for $Z_1, \ldots, Z_n$ i.i.d. $\mathcal{N}(0,1)$. Equivalently, χ_n^2 is the Gamma $(n/2, 1/2)$ distribution.

Remark: R.v.s with continuous distributions cannot have a probability mass, i.e., we cannot have $P(X = x) > 0$ for any x.

8 Joint Distributions

Define the joint PDF $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$. Bayes' and LOTP hold for both discrete and continuous r.v.s, though be careful in the hybrid case.

Independence criteria: Continuous r.v.s X, Y are independent if for all x and y,

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \iff f_{X,Y}(x,y) = f_X(x)f_Y(y) \iff f_{Y|X}(y \mid x) = f_Y(y).$$

For discrete X, Y, we have the analogous relations with the PMF in place of the PDF.

Factoring the joint PDF: If $f_{X,Y}$ factors as

$$f_{X,Y}(x,y) = g(x)h(y),$$

for some nonnegative functions g and h, then X and Y are independent. Also, if either g or h is a valid PDF, then the other one is valid too and g, h are the respective marginal PDFs. (The analogous result in the discrete case also holds.)

Multinomial:

- **Marginals are binomial:** if $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then $X_j \sim \text{Bin}(n, p_j)$
- **Multinomial lumping:** by combining categories and their corresponding probabilities in an Multinomial random vector, we get another Multinomial random vector
- **Multinomial conditioning:** if $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, then

$$(X_2, \ldots, X_k) \mid X_1 = n_1 \sim \text{Mult}_{k-1}\left(n - n_1, (p_2', \ldots, p_k')\right)$$

  where $p_j' = p_j / (p_2 + \cdots + p_k)$
- **Covariance in Multinomials:** if $(X_1, \ldots, X_k) \sim \text{Mult}_k(n, \mathbf{p})$, then for $i \neq j, \text{Cov}(X_i, X_j) = -np_ip_j$

Multivariate Normal: $\mathbf{X}$ is MVN if any linear combination of its components is Normally distributed. The parameters are the mean vector and covariance matrix. Of course each subvector and, in particular, each random vector component, is also MVN.

- **Concatenating independent MVNs:** if independent $\mathbf{X}$ and $\mathbf{Y}$ are MVN, then $X + Y$ is also MVN.
- **Uncorrelated implies independent for MVN:** if MVN $\mathbf{X}$ can be written $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with every component of $\mathbf{X}_1$ uncorrelated with every component of $\mathbf{X}_2$, then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent.

  In particular, if (X, Y) is Bivariate Normal and $\text{Corr}(X, Y) = 0$, then X and Y are independent.

9 Convergence

Given i.i.d. X_i with mean μ and variance σ^2, the sample mean $\bar{X}_n$ is an r.v. with mean μ and variance σ^2/n:

$$\bar{X}_n = (X_1 + \cdots + X_n)/n.$$

LLN: As n grows, the sample mean $\bar{X}_n$ converges to the true mean μ. The LLN comes in two versions with slightly different definitions of what it means for a sequence of r.v.s to converge to a number:

- **Strong LLN (pointwise):** $\bar{X}_n$ converges to the true mean μ pointwise with probability 1,

$$P\left(\bar{X}_n \to \mu\right) = 1.$$

- **Weak LLN (convergence in probability):** For all $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) = 0.$$

CLT: For large n, the distribution of $\bar{X}_n$ after standardization approaches a standard Normal. As $n \to \infty$,

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

In approximation form, this means that for large n, the distribution of $\bar{X}_n$ is approximately $\mathcal{N}\left(\mu, \sigma^2/n\right)$.

Binomial convergence to Normal: $Y \sim \text{Bin}(n,p)$ is the sum of n i.i.d. Bern(p) r.v.s, and so for large n

$$Y \dot{\sim} \mathcal{N}(np, np(1-p)).$$

We may apply continuity correction to account for discreteness of Y:

$$P(Y=k) = P\left(k - \tfrac{1}{2} < Y < k + \tfrac{1}{2}\right) \approx \Phi\left(\frac{k+1/2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-1/2-np}{\sqrt{np(1-p)}}\right).$$

10 Inequalities

Cauchy-Schwarz (marginal bound on a joint expectation): for any r.v.s X and Y with finite variances,

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

Jensen (convexity): consider some r.v. X and function g. If g convex, then $E(g(X)) \geq g(E(X))$. If g concave, then $E(g(X)) \leq g(E(X))$. In either case, equality holds iff $g(X) = a + bX$ with probability 1 for some constants a, b.

Markov (tail probabilities): for any r.v. X and constant $a > 0$,

$$P(|X| \geq a) \leq {}^{E|X|}/_{a}.$$

Chebyshev (tail probabilities): let X have mean μ and variance σ^2. Then for any $a > 0$,

$$P(|X-\mu| \geq a) \leq {}^{\sigma^2}/_{a^2} \implies P(|X-\mu| \geq c\sigma) \leq {}^{1}/_{c^2},$$

upper bounding the probability of an r.v. being more than c standard deviations away from its mean.

11 Markov

Markov chain: A sequence of r.v.s $X_0, X_1, \ldots$ taking values in state space $\{1, 2, \ldots, M\}$, for which any particular transition probability depends only on the most recent state:

$$P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right) = P\left(X_{n+1} = j \mid X_n = i\right).$$

The $M \times M$ transition matrix $\mathbf{Q}$ has entries $q_{ij} = P\left(X_{n+1} = j \mid X_n = i\right)$ giving the transition probability from state i to state j. In Q, each row sums to 1. Similarly, $\mathbf{Q}^n$ has entries $q_{ij}^{(n)}$ giving the n-step transition probabilities.

Marginal distribution of X_n: letting row vector $\mathbf{t} = (t_1, t_2, \ldots, t_M)$ with $t_i = P\left(X_0 = i\right)$, then the marginal distribution of X_n is $\mathbf{t}Q^n$. That is, the j^{th} component of $\mathbf{t}Q^n$ is $P\left(X_n = j\right)$.

Classification of states:

- **Recurrent vs transient:** Starting at a recurrent state, you will always return back to that state eventually. Non-recurrent states are called transient, and there is some positive probability that after leaving we will never return.
- **Period:** The period k of state is the GCD of the possible numbers of steps it could take to return back to the state. A periodic state has $k > 1$, and an aperiodic state has $k = 1$. A chain is called periodic if any of its states are periodic, and aperiodic otherwise.
- **Irreducible chain:** a chain is irreducible if you can get from anywhere to anywhere. In an irreducible chain, all states are recurrent and have the same period.

Stationary distribution: A row vector $\mathbf{s} = (s_1, \ldots, s_M)$ of probabilities is a stationary distribution for a Markov chain with transition matrix Q if for all j

$$\sum_i s_i q_{ij} = s_j \iff \mathbf{s}Q = \mathbf{s}.$$

Intuitively, this is the long-run behavior of the chain, regardless of its initial conditions. For any irreducible Markov chain, there exists a unique stationary distribution (in which every state must have positive probability).

Reversible distribution: a chain is reversible w.r.t. the vector $\mathbf{s} = (s_1, \ldots, s_M)$ of probabilities if for all i, j we have

$$s_i q_{ij} = s_j q_{ji}$$

Reversibility implies stationary: if the chain is reversible w.r.t. $\mathbf{s}$, then $\mathbf{s}$ is a stationary distribution.

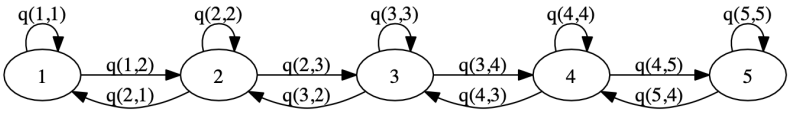
Other special stationary cases: If each column of the transition matrix Q sums to 1, then the uniform distribution over all states, $(1/M, 1/M, ..., 1/M)$, is a stationary distribution. A special case of this is Q symmetric.

Expected time to return: Consider an irreducible Markov chain with stationary distribution $\mathbf{s}$. Letting r_i be the expected time it takes the chain to return to i, given that it starts at i. Then $r_i = \frac{1}{s_i}$.

Random walk on an undirected network: this is a collection of nodes with bidirectionally-traversable edges such that from node i, the probabilities of traversing any edge at i are equal. Self-loops are allowed.

The degree of a node is the number of edges attached to it, and the reversible (and thus stationary) distribution is proportional to the degree sequence $(d_1, \ldots, d_n)$.

Birth-death chain: A birth-death chain on states $\{1, 2, \ldots, M\}$ is a Markov chain with transition matrix $Q = (q_{ij})$ such that $q_{ij} > 0$ if $|i - j| = 1$ and $q_{ij} = 0$ if $|i - j| \geq 2$. This says it's possible to go one step to the left and possible to go one step to the right (except at boundaries) but impossible to jump further in one step. For example, the chain shown below is a birth-death chain if the labeled transitions have positive probabilities, except for the loops from a state to itself, which are allowed to have 0 probability.



Solution. We will now show that any birth-death chain is reversible, and construct the stationary distribution. Let s_1 be a positive number, to be specified later. Since we want $s_1 q_{12} = s_2 q_{21}$, let

$$s_2 = s_1 q_{12}/q_{21}.$$

Then since we want $s_2 q_{23} = s_3 q_{32}$, let

$$s_3 = s_2 q_{23}/q_{32} = s_1 q_{12} q_{23}/(q_{32} q_{21})$$

Continuing in this way, let

$$s_j = \frac{s_1 q_{12} q_{23} \cdots q_{j-1,j}}{q_{j,j-1} q_{j-1,j-2} \cdots q_{21}}$$

for all states j with $2 \leq j \leq M$. Choose s_1 so that the s_j sum to 1. Then the chain is reversible with respect to $\mathbf{s}$, since $q_{ij} = q_{ji} = 0$ if $|i - j| \geq 2$ and by construction $s_i q_{ij} = s_j q_{ji}$ if $|i - j| = 1$. Thus, $\mathbf{s}$ is the stationary distribution.

Metropolis-Hasting: Let $\mathbf{s} = (s_1, \ldots, s_M)$ be a desired stationary distribution on state space $\{1, \ldots, M\}$. Assume that $s_i > 0$ for all i (if not, just delete any states i with $s_i = 0$ from the state space). Suppose that $P = (p_{ij})$ is the transition matrix for a Markov chain on state space $\{1, \ldots, M\}$. Intuitively, P is a Markov chain that we know how to run but that doesn't have the desired stationary distribution.

Our goal is to modify P to construct a Markov chain $X_0, X_1, \ldots$ with stationary distribution s. We will give a Metropolis-Hastings algorithm for this. Start at any state X_0 (chosen randomly or deterministically), and suppose that the new chain is currently at X_n. To make one move of the new chain, do the following.

1. If $X_n = i$, propose a new state j using the transition probabilities in the i^{th} row of the original transition matrix P.
2. Compute the acceptance probability
$$a_{ij} = \min\left({}^{s_j p_{ji}}/_{s_i p_{ij}}, 1\right).$$
3. Flip a coin that lands Heads with probability a_{ij}.
4. If the coin lands Heads, accept the proposal (i.e., go to j), setting $X_{n+1} = j$. Otherwise, reject the proposal (i.e., stay at i), setting $X_{n+1} = i$.

That is, the Metropolis-Hastings chain uses the original transition probabilities p_{ij} to propose where to go next, then accepts the proposal with probability a_{ij}, staying in its current state in the event of a rejection.