STAT 870 Outline

Course outline

Measure theoretic foundations of probability (3 weeks):

- σ-fields

- Measurability arguments

- Formal definition of expected value

- Fatou's lemma, monotone convergence theorem, dominated convergence theorem.

- Modes of convergence: in probability, in mean square, almost sure.

Course outline continued:

- Statements of some of the following famous theorems of probability:
  - Weak law of large numbers
  - Strong law of large numbers
  - Lindeberg central limit theorem
  - Martingale convergence theorems
  - Ergodic theorems
  - Renewal theorem

Course outline continued:

- One week introductions to each of:
  - Markov Chains
  - Poisson Processes
  - Point Processes
  - Birth and Death Processes
  - Queuing Theory
  - Brownian motion and diffusions
  - Simulation
- Student presentations (one week)

Models for coin tossing

Toss coin n times.

On trial k write down a 1 for heads and 0 for tails.

Typical outcome is $\omega = (\omega_1, \ldots, \omega_n)$ a sequence of zeros and ones.

Example: $n = 3$ gives 8 possible outcomes

$$\begin{aligned}\Omega = \{&(0,0,0),(0,0,1),(0,1,0),(0,1,1),\\ &(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}\,.\end{aligned}$$

General case: set of all possible outcomes is $\Omega = \{0,1\}^n$; $\text{card}(\Omega) = 2^n$.

Meaning of *random* not defined here. Interpretation of probability is usually long run limiting relative frequency (but then we deduce existence of long run limiting relative frequency from axioms of probability).

Probability measure: function P defined on the set of all subsets of Ω such that: with the following properties:

1. For each $A \subset \Omega$, $P(A) \in [0,1]$.

2. If $A_1, \ldots, A_k$ are *pairwise disjoint* (meaning that for $i \neq j$ the intersection $A_i \cap A_j$ which we usually write as $A_i A_j$ is the empty set $\emptyset$) then

$$P(\cup_1^k A_j) = \sum_1^k P(A_j)$$

3. $P(\Omega) = 1$.

Probability modelling: select family of possible probability measures.

Make best match between mathematics, real world.

interpretation of probability: long run limiting relative frequency

Coin tossing problem: many possible probability measures on Ω.

For $n = 3$, Ω has 8 elements and $2^8 = 256$ subsets.

To specify P: specify 256 numbers. Generally impractical.

Instead: *model* by listing some assumptions about P.

Then deduce what P is, or how to calculate $P(A)$

Three approaches to modelling coin tossing:

1. Counting model:
$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega} \qquad (1)$$
Disadvantage: no insight for other problems.

2. Equally likely elementary outcomes: if $A = \{\omega_1\}$ and $B = \{\omega_2\}$ are two singleton sets in Ω then $P(A) = P(B)$. If card$(\Omega) = m$, say $\Omega = (\omega_1, \ldots, \omega_m)$ then
$$\begin{aligned} P(\Omega) &= P(\cup_1^m \{\omega_j\}) \\ &= \sum_1^m P(\{\omega_j\}) \\ &= mP(\{(\omega_1\}) \end{aligned}$$
So $P(\{\omega_i\}) = 1/m$ and (1) holds.

Defect of models: infinite Ω not easily handled.

Toss coin till first head. Natural Ω is set of all sequences of k zeros followed by a one.

OR: $\Omega = \{0, 1, 2, \ldots\}$.

Can't assume all elements equally likely.

Third approach: model using **independence**:

Coin tossing example: $n = 3$.

Define $A = \{\omega : \omega_1 = 1, \omega_2 = 0, \omega_3 = 1\}$ and

$$\begin{aligned} A_1 &= \{\omega : \omega_1 = 1\} \\ A_2 &= \{\omega : \omega_2 = 0\} \\ A_3 &= \{\omega : \omega_3 = 1\}\,. \end{aligned}$$

Then $A = A_1 \cap A_2 \cap A_3$

Note $P(A) = 1/8$, $P(A_i) = 1/2$.

So: $P(A) = \prod P(A_i)$

General case: n tosses. $B_i \subset \{0, 1\}$; $i = 1, \ldots, n$

Define

$$A_i = \{\omega : \omega_i \in B_i\} \qquad A = \cap A_i.$$

It is possible to prove that

$$P(A) = \prod P(A_i)$$

Jargon to come later: random variables X_i defined by $X_i(\omega) = \omega_i$ are independent.

Basis of most common modelling tactic.

Assume

$$P(\{\omega : \omega_i = 1\}) = P(\{\omega : \omega_i = 0\}) = 1/2 \quad (2)$$

and for any set of events of form given above

$$P(A) = \prod P(A_i)\,. \quad (3)$$

Motivation: long run rel freq interpretation plus assume outcome of one toss of coin incapable of influencing outcome of another toss.

Advantages: generalizes to infinite Ω.

Toss coin infinite number of times:

$$\Omega = \{\omega = (\omega_1, \omega_2, \cdots)\}$$

is an uncountably infinite set. Model assumes for any n and any event of the form $A = \cap_1^n A_i$ with each $A_i = \{\omega : \omega_i \in B_i\}$ we have

$$P(A) = \prod_1^n P(A_i) \quad (4)$$

For a *fair* coin add the assumption that

$$P(\{\omega : \omega_i = 1\}) = 1/2\,. \quad (5)$$

Is $P(A)$ determined by these assumptions??

Consider $A = \{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B\}$ where $B \subset \Omega_n = \{0,1\}^n$. Our assumptions guarantee

$$P(A) = \frac{\text{number of elements in } B}{\text{number of elements in } \Omega_n}$$

In words, our model specifies that the first n of our infinite sequence of tosses behave like the equally likely outcomes model.

Define C_k to be the event *first head occurs after k consecutive tails*:

$$C_k = A_1^c \cap A_2^c \cdots \cap A_k^c \cap A_{k+1}$$

where $A_i = \{\omega : \omega_i = 1\}$; A^c means complement of A. Our assumption guarantees

$$\begin{aligned} P(C_k) &= P(A_1^c \cap A_2^c \cdots \cap A_k^c \cap A_{k+1}) \\ &= P(A_1^c) \cdots P(A_k^c) P(A_{k+1}) \\ &= 2^{-(k+1)} \end{aligned}$$

Complicated Events: examples

$$A_1 \equiv \{\omega : \lim_{n\to\infty} (\omega_1 + \cdots + \omega_n)/n \text{ exists } \}$$

$$A_2 \equiv \{\omega : \lim_{n\to\infty} (\omega_1 + \cdots + \omega_n)/n = 1/2\}$$

$$A_3 \equiv \{\omega : \lim_{n\to\infty} \sum_1^n (2\omega_k - 1)/k \text{ exists } \}$$

- Strong Law of Large Numbers: for our model $P(A_2) = 1$.

- In fact, $A_3 \subset A_2 \subset A_1$.

- If $P(A_2) = 1$ then $P(A_1) = 1$.

- In fact $P(A_3) = 1$ so $P(A_2) = P(A_1) = 1$.

Some mathematical questions to answer:

1. Do (4) and (5) determine $P(A)$ for every $A \subset \Omega$? [NO]

2. Do (4) and (5) determine $P(A_i)$ for $i = 1, 2, 3$? [YES]

3. Are (4) and (5) logically consistent? [YES]

Probability Definitions

Probability Space (or **Sample Space**): ordered triple $(\Omega, \mathcal{F}, P)$.

- Ω is a set (possible outcomes).

- $\mathcal{F}$ is a family of subsets (**events**) of Ω with the property that $\mathcal{F}$ is a σ-field (or Borel field or σ-algebra):

  1. The empty set $\emptyset$ and Ω are members of $\mathcal{F}$.

  2. $A \in \mathcal{F}$ implies $A^c = \{\omega \in \Omega : \omega \not\in A\} \in \mathcal{F}$

  3. $A_1, A_2, \cdots$ all in $\mathcal{F}$ implies

$$A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}\,.$$

- P a function, domain $\mathcal{F}$, range a subset of $[0, 1]$ satisfying:

  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.

  2. **Countable additivity**: $A_1, A_2, \cdots$ **pairwise disjoint** $(j \neq k \implies A_j A_k = \emptyset)$

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms guarantee can compute probabilities by usual rules, including approximation without contradiction.

Consequences:

1. **Finite additivity** $A_1, \cdots, A_n$ pairwise disjoint:
$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)\,.$$

2. For any event A $P(A^c) = 1 - P(A)$.

3. If $A_1 \subset A_2 \subset \cdots$ are events then
$$P(\bigcup_1^\infty A_i) = \lim_{n\to\infty} P(A_n)\,.$$

4. If $A_1 \supset A_2 \supset \cdots$ then
$$P(\bigcap_1^\infty A_i) = \lim_{n\to\infty} P(A_n)\,.$$

Most subtle point is σ-field, $\mathcal{F}$. Needed to avoid some contradictions which arise if you try to define $P(A)$ for every subset A of Ω when Ω is a set with uncountably many elements.

Events in Set Notation

Event that Y_n converges to 0 is

$$A \equiv \{\omega : \lim_{n\to\infty} Y_n(\omega) = 0\}$$

Not explicitly written in terms of simple events involving only a finite number of Ys.

Recall basic definition of limit: y_n converges to 0 if $\forall \epsilon > 0\ \exists N$ such that $\forall n \geq N$ we have $|y_n| \leq \epsilon$.

Convert the definition in A into set theory notation:

- replace y_n by $Y_n(\omega)$,

- replace each *for every* by an intersection

- replace each *there exists* with a union.

We get

$$A = \bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |Y_n(\omega)| \leq \epsilon\}$$

Not obvious A is event because intersection over $\epsilon > 0$ is uncountable.

However, the intersection is countable. Let

$$B_\epsilon \equiv \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |Y_n(\omega)| \le \epsilon\}$$

Notice that

$$\epsilon' < \epsilon \implies B_{\epsilon'} \subset B_\epsilon$$

This means that

$$\bigcap_{\epsilon>0} B_\epsilon = \bigcap_{m=1}^{\infty} B_{1/m}$$

A is countable intersection of countable unions of countable intersections of events, so A is an event.

Here are some other events:

Sequence S_n has a limit. Sequence s_n has a limit if $\exists s_\infty$ such that $\forall \epsilon > 0$ $\exists N$ such that $\forall n \ge N$ we have $|s_n - s_\infty| \le \epsilon$. Mechanically get event:

$$\bigcup_s \bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |S_n(\omega) - s| \le \epsilon\}$$

Intersection over ϵ can be made countable. Union over s, however, is not easy to make countable. Instead use theorem of analysis to describe existence of a limit.

A sequence s_n has a limit if and only if the sequence is Cauchy.

Cauchy sequence: $\forall\epsilon > 0\ \exists N$ such that $\forall n \geq N$ we have $|s_n - s_N| \leq \epsilon$. $\{S_n$ has a limit$\}$ is

$$\bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|S_n - S_N| \leq \epsilon\}$$

Intersection over all $\epsilon > 0$ is countable intersection over $\epsilon = 1/r$ for positive integers r.

Y_n **is summable**: the sequence of partial sums $S_n = \sum_1^n Y_i$ has a limit so

$$\{Y_n \text{ summable}\} = \bigcap_{r=1}^\infty \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty \{\left|\sum_{N+1}^n Y_j\right| \le 1/r\}$$

Event $S_n > 0$ **for infinitely many** n: $\forall N \exists n \ge N$ such that $S_n > 0$. is

$$\bigcap_{N=1}^\infty \bigcup_{n=N}^\infty \{S_n > 0\}$$

limit superior of S_n is 1 is intersection of two events, $\limsup S_n \le 1$ and $\limsup S_n \ge 1$. Former is $\forall \epsilon > 0 \exists N$ such that $\forall n \ge N$ $S_n \le 1 + \epsilon$. Latter is $\forall \epsilon > 0$ and $\forall N$ there is an $n \ge N$ such that $S_n \ge 1 - \epsilon$. Event is $A^* \cap A_*$ where

$$A^* = \bigcap_{r=1}^\infty \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty \{S_n \le 1 + 1/r\}$$

$$A_* = \bigcap_{r=1}^\infty \bigcap_{N=1}^\infty \bigcup_{n=N}^\infty \{S_n \ge 1 + 1/r\}$$

Random Variables:

Vector valued random variable: function X, domain Ω, range in $\mathbb{R}^p$ such that

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

is defined for any constants $(x_1, \ldots, x_p)$. Notation: $X = (X_1, \ldots, X_p)$ and

$$X_1 \le x_1, \ldots, X_p \le x_p$$

is shorthand for an event:

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\}$$

X function on Ω so X_1 function on Ω.

Formal definitions:

The **Borel** σ-field in $\mathbb{R}^p$ is the smallest σ-field in $\mathbb{R}^p$ containing every open ball

$$B_y(r) = \{x \in \mathbb{R}^p : |x - y| < r\}\,.$$

(To see that there is, in fact, such a "smallest" σ-field you prove the following assertions:

1. The intersection of an arbitrary family of σ-fields is a σ-field.
2. There is at least one σ-field of subsets of $\mathbb{R}^p$ containing every open ball.

Now define the Borel σ-field in $\mathbb{R}^p$ to be

$$\mathcal{B}(\mathbb{R}^p) = \cap \mathcal{F}$$

where the intersection runs over all σ-fields $\mathcal{F}$ which contain every open ball.)

Every common set is a Borel set, that is, in the Borel σ-field.

Example: If O is an open set then O is Borel.

Proof: For each x in O there is a point y all of whose co-ordinates are rational numbers and a rational number r such that

$$x \in B_y(r) \subset O$$

Now O is the union of all these $B_y(r)$.

(Every $x \in O$ is in one of the $B_y(r)$ and every point in any $B_y(r)$ is in O.)

But the union is countable because there are only countably many possible pairs (y, r) with all the co-ordinates rational numbers.

Independence

Events A and B **independent** if

$$P(AB) = P(A)P(B)\,.$$

Events A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any set of distinct indices $i_1, \ldots, i_r$ between 1 and p.

Example: $p = 3$

$$\begin{aligned} P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1A_2) &= P(A_1)P(A_2) \\ P(A_1A_3) &= P(A_1)P(A_3) \\ P(A_2A_3) &= P(A_2)P(A_3) \end{aligned}$$

Need all equations to be true for independence!

Example: Toss a coin twice. If A_1 is the event that the first toss is a Head, A_2 is the event that the second toss is a Head and A_3 is the event that the first toss and the second toss are different. then $P(A_i) = 1/2$ for each i and for $i \neq j$

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)\,.$$

Rvs $X_1, \ldots, X_p$ are **independent** if

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod P(X_i \in A_i)$$

for any choice of $A_1, \ldots, A_p$.

σ-fields $\mathcal{F}_1, \ldots, \mathcal{F}_p$ are **independent** if

$$P(A_1 \cap \cdots \cap A_p) = \prod P(A_i)$$

for any choice of events $A_1 \in \mathcal{F}_1, \ldots, A_p \in \mathcal{F}_p$.

Theorem 1 *1. If X and Y are independent and discrete then*

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

for all x, y

2. If X and Y are discrete and

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

for **all** *x, y then X and Y are independent.*

Theorem 2 *If $X_1,\ldots,X_p$ are independent and $Y_i = g_i(X_i)$ then $Y_1,\ldots,Y_p$ are independent. Moreover, $(X_1,\ldots,X_q)$ and $(X_{q+1},\ldots,X_p)$ are independent.*

Proof: The event

$$Y_1 \in A_1 \cap \cdots \cap Y_p \in A_p$$

is exactly the same as the event

$$X_1 \in g_1^{-1}(A_1) \cap \cdots \cap X_p \in g_p^{-1}(A_p)$$

so the first statement is easy.

The second statement is proved using a standard technique:

We must show

$$P\{(X_1,\ldots,X_q) \in A; (X_{q+1},\ldots,X_p) \in B\} = P\{(X_1,\ldots,X_q) \in A\}P\{(X_{q+1},\ldots,X_p) \in B\} \quad (6)$$

We study the collections of A, B pairs for which (6) holds.

Product Spaces

Suppose $\mathcal{X}_i, \mathcal{F}_i$ are pairs for $i = 1, \ldots, p$.

Each $\mathcal{X}_i$ a set; $\mathcal{F}_i$ a σ-field of subsets of $\mathcal{X}_i$.

The Cartesian product of the sets $\mathcal{X}_i$ is

$$\mathcal{X} = \{(x_1, \ldots, x_p) : x_i \in \mathcal{X}_i; i = 1, \ldots, p\}$$

We write

$$\mathcal{X} = \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_p$$

We define a subset A of $\mathcal{X}$ to be a measurable rectangle if

$$A = A_1 \otimes \cdots \otimes A_p$$

where each A_i is in $\mathcal{F}_i$.

We define the product σ-field on $\mathcal{X}$ as

$$\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$$

where $\mathcal{F}$ is the smallest σ-field containing all measurable rectangles.

Note: *smallest* means intersection of all σ-fields containing the family of measurable rectangles. This intersection is not empty (consider power set of $\mathcal{X}$). And any intersection of σ-fields is a σ-field.

How do we prove independence result:

Step 1: True for pairs A, B where both A and B are rectangles by definition of independence.

Step 2: Fix rectangle B. Consider collection of sets A for which (6) holds. It is closed under finite disjoint unions. It is closed under the action of taking complements. It contains the all measurable rectangles. So it contains the smallest field containing all measurable rectangles.

(Field is like σ field but with finite unions and intersections.)

The collection of sets A for which (6) holds is a monotone class. So it contains the smallest monotone class containing the field generated by the rectangles. So it contains the σ-field generated by the rectangles – the product σ-field.

Monotone Class: collection $\mathcal{C}$ of subsets of a given set, closed under increasing countable unions and decreasing countable intersections:

1. If $A_1 \subset A_2 \subset \cdots$ are in $\mathcal{C}$ then $\bigcup_1^\infty A_i \in \mathcal{C}$.

2. If $A_1 \supset A_2 \supset \cdots$ are in $\mathcal{C}$ then $\bigcap_1^\infty A_i \in \mathcal{C}$.

Lemma: The smallest monotone class containing a field $\mathcal{F}_o$ is a σ-field.

Proof of Lemma: Let $\mathcal{C}$ be smallest monotone class containing $\mathcal{F}_o$ (intersection of all monotone classes containing $\mathcal{F}_o$). Put $\mathcal{M} = \{A \in \mathcal{C} : A^c \in \mathcal{C}\}$.

- $\mathcal{M}$ is a monotone class.
- Since $\mathcal{F}_o$ is a field and a subset of $\mathcal{C}$, $\mathcal{M}$ contains $\mathcal{F}_o$.
- Since $\mathcal{C}$ is smallest monotone class containing $\mathcal{F}_o$ and $\mathcal{M}$ is another monotone class containing $\mathcal{F}_o$ we see $\mathcal{C} \subset \mathcal{M}$.
- This means every $A \in \mathcal{C}$ has the property $A^c \in \mathcal{C}$. In other words $\mathcal{C}$ is closed under the operation of taking complements, one of the defining properties of a σ-field.

- Similar argument for finite unions in notes.

- So $\mathcal{C}$ is a field. Since $\mathcal{C}$ is monotone class and field $\mathcal{C}$ is a σ-field.

Related mathematical problem. We use independence two ways: as a modelling tactic and as a computational tool.

We often model by assuming some random variables are independent.

Suppose $X_1, X_2, \cdots$ are an infinite sequence of independent coin tosees?

But can we suppose so? Does there exist $(\Omega, \mathcal{F}, \mathbf{P})$ and random variables $X_1, X_2, \cdots$ defined on Ω such that they are independent and 0, 1 valued?

Yes: use extension theorems.

Product measures:

Suppose P_i is a probability measure on $(\Omega_i, \mathcal{F}_i)$. We define a product measure on

$$\Omega = \Omega_1 \otimes \cdots \otimes \Omega_P$$

by

$$P(A_1 \otimes \cdots \otimes A_p) = \prod P_i(A_i)$$

This formula extends to the product σ-field by say the Caratheodory extension theorem.

Use: the maps $X_i : \Omega \mapsto \Omega_i$ given by

$$X_i(\omega_1, \ldots, \omega_p) = \omega_i$$

These rvs are independent and $P(X_i \in A_i) = P_i(A_i)$.

Can even take $p = \infty$ using Kolmogorov consistency theorem.

Conditional probability

Important modeling and computation technique:

Def'n: $P(A|B) = P(AB)/P(B)$ if $P(B) \neq 0$.

Def'n: For discrete rvs X, Y conditional pmf of Y given X is

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= f_{X,Y}(x, y)/f_X(x) \\ &= f_{X,Y}(x, y)/\sum_t f_{X,Y}(x, t) \end{aligned}$$

IDEA: used as both computational tool and modelling tactic.

Specify joint distribution by specifying "marginal" and "conditional".

Modelling:

Assume $X \sim$ Poisson(λ).

Assume $Y|X \sim$Binomial(X, p).

Let $Z = X - Y$. Joint law of Y, Z?

$$
\begin{aligned}
&P(Y = y, Z = z) \\
&\quad = P(Y = y, X - Y = z) \\
&\quad = P(Y = y, X = z + y) \\
&\quad = P(Y = y | X = y + z) P(X = y + z) \\
&\quad = \binom{z+y}{y} p^y (1-p)^z e^{-\lambda} \lambda^{z+y} / (z+y)! \\
&\quad = \exp\{-p\lambda\} \frac{(p\lambda)^y}{y!} \exp\{(1-p)\lambda\} \frac{\{(1-p)\lambda\}^z}{z!}
\end{aligned}
$$

So: Y, Z independent Poissons.

Expected Value

Undergraduate definition of E: integral for absolutely continuous X, sum for discrete. But: $\exists$ rvs which are neither absolutely continuous nor discrete.

General definition of E.

A random variable X is **simple** if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Def'n: For a simple rv X we define

$$E(X) = \sum a_i P(A_i)$$

For positive random variables which are not simple we extend our definition by approximation:

Def'n: If $X \geq 0$ (almost surely, $P(X \geq 0) = 1$) then

$$E(X) = \text{sup}\{E(Y) : 0 \leq Y \leq X, Y \text{ simple}\}$$

Def'n: We call X **integrable** if

$$E(|X|) < \infty\,.$$

In this case we define

$$E(X) = E(\text{max}(X,0)) - E(\text{max}(-X,0))$$

Facts: E is a linear, monotone, positive operator:

1. **Linear**: $E(aX+bY) = aE(X)+bE(Y)$ provided X and Y are integrable.

2. **Positive**: $P(X \ge 0) = 1$ implies $E(X) \ge 0$.

3. **Monotone**: $P(X \ge Y) = 1$ and X, Y integrable implies $E(X) \ge E(Y)$.

Major technical theorems:

Monotone Convergence: If $0 \le X_1 \le X_2 \le \cdots$ a.s. and $X = \lim X_n$ (which exists a.s.) then

$$E(X) = \lim_{n \to \infty} E(X_n)$$

Dominated Convergence: If $|X_n| \le Y_n$ and $\exists$ rv X st $X_n \to X$ a.s. and rv Y st $Y_n \to Y$ with $E(Y_n) \to E(Y) < \infty$ then

$$E(X_n) \to E(X)$$

Often used with all Y_n the same rv Y.

Fatou's Lemma: If $X_n \ge 0$ then

$$E(\liminf X_n) \le \liminf E(X_n)$$

Theorem: With this definition of E if X has density $f(x)$ (even in $\mathbb{R}^p$ say) and $Y = g(X)$ then

$$E(Y) = \int g(x)f(x)dx\,.$$

(This could be a multiple integral.) If X has pmf f then

$$E(Y) = \sum_x g(x)f(x)\,.$$

Works even if X has density but Y doesn't.

Def'n: r^{th} moment (about origin) of a real rv X is $\mu_r' = E(X^r)$ (provided it exists). Generally use μ for $E(X)$. The r^{th} central moment is

$$\mu_r = E[(X-\mu)^r]$$

Call $\sigma^2 = \mu_2$ the variance.

Def'n: For an $\mathbb{R}^p$ valued rv X $\mu_X = E(X)$ is the vector whose i^{th} entry is $E(X_i)$ (provided all entries exist).

Def'n: The $(p \times p)$ variance covariance matrix of X is

$$Var(X) = E\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment. More generally if $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ both have all components with finite second moments then

$$\text{Cov}(X,Y) = \text{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right]$$

We have

$$\text{Cov}(AX+b, CY+d) = A\text{Cov}(X,Y)B^T$$

for general (conforming) matrices A, C and vectors b and d.

Moments and probabilities of rare events are closely connected as will be seen in a number of important probability theorems. Here is one version of Markov's inequality (one case is Chebyshev's inequality):

$$\begin{aligned} P(|X-\mu| \geq t) &= E[1(|X-\mu| \geq t)] \\ &\leq E\left[\frac{|X-\mu|^r}{t^r} 1(|X-\mu| \geq t)\right] \\ &\leq \frac{E[|X-\mu|^r]}{t^r} \end{aligned}$$

The intuition is that if moments are small then large deviations from average are unlikely.

Moments and independence

Theorem: If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$E(X_1 \cdots X_p) = E(X_1) \cdots E(X_p)$$

Proof: Usual order: simple Xs first, then positive, then integrable.

Suppose each X_i is simple:

$$X_i = \sum_j x_{ij} 1(X_i = x_{ij})$$

where the x_{ij} are the possible values of X_i.

Then

$$\begin{aligned}
&E(X_1 \cdots X_p) \\
&\qquad = \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \times \\
&\qquad\qquad E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p}) \\
&\qquad = \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \times \\
&\qquad\qquad P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&\qquad = \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \times \\
&\qquad\qquad P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&\qquad = \left[\sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1})\right] \times \cdots \times \\
&\qquad\qquad \left[\sum_{j_p} x_{pj_p} P(X_p = x_{pj_p})\right] \\
&\qquad = \prod E(X_i)
\end{aligned}$$

General $X_i > 0$: $X_{i,n}$ is X_i rounded down to the nearest multiple of 2^{-n} (to a maximum of n). Each $X_{i,n}$ is simple and $X_{1,n}, \ldots, X_{p,n}$ are independent. Thus

$$\mathrm{E}(\prod X_{j,n}) = \prod \mathrm{E}(X_{j,n})$$

for each n. If

$$X_n^* = \prod X_{j,n}$$

then

$$0 \le X_1^* \le X_2^* \le \cdots$$

and X_n^* converges to $X^* = \prod X_i$ so that

$$\mathrm{E}(X^*) = \lim \mathrm{E}(X_n^*)$$

by monotone convergence. Also by monotone convergence

$$\lim \prod \mathrm{E}(X_{j,n}) = \prod \mathrm{E}(X_j)) < \infty$$

This shows both that X^* is integrable and that

$$E(\prod X_j) = \prod \mathsf{E}(X_j)$$

The general case uses the fact that we can write each X_i as the difference of its positive and negative parts:

$$X_i = \mathsf{max}(X_i, 0) - \mathsf{max}(-X_i, 0)$$

Just expand out the product and use the previous case.

Lebesgue Integration

Lebesgue integral defined much the same way as E.

Borel function f *simple* if

$$f(x) = \sum_1^n a_i 1(x \in A_i)$$

for almost all $x \in \mathbb{R}^p$ and some constants a_i and Borel sets A_i with $\lambda(A_i) < \infty$). For such an f we define

$$\int f(x)dx = \sum a_i \lambda(A_i)$$

Again if

$$\sum a_i 1_{A_i} = \sum b_j 1_{B_j}$$

almost everywhere and all A_i and B_j have finite Lebesgue measure you must check that

$$\sum a_i \lambda(A_i) = \sum b_j \lambda(B_j)$$

If $f \geq 0$ almost everywhere and f is Borel define

$$\int f(x)dx = \text{sup}\{\int g(y)dy\}$$

where the sup ranges over all simple functions g such that $0 \leq g(x) \leq f(x)$ for almost all x. Call $f \geq 0$ integrable if $\int f(x)dx < \infty$.

Call a general f integrable if $|f|$ is integrable and define for integrable f

$$\begin{aligned}\int f(x)dx = \int & \text{max}(f(x), 0)dx \\ & - \int \text{max}(-f(x), 0)dx\end{aligned}$$

Remark: Again you must check that you have not changed the definition of f for either of the previous categories of f.

Facts: $\int$ is a linear, monotone, positive operator:

1. **Linear**: provided f and g are integrable
$$\int af(x)+bg(x)dx = a\int f(x)dx+b\,\text{inf}\,g(x)dx$$
2. **Positive**: If $f(x) \geq 0$ almost everywhere then $\int f(x)dx \geq 0$.
3. **Monotone**: If $f(x) > g(x)$ almost everywhere and f and g are integrable then
$$\int f(x)dx \geq \int g(x)dx.$$

Each of these facts is proved first for simple functions then for positive functions then for general integrable functions.

Major technical theorems:

Monotone Convergence: If $0 \le f_1 \le f_2 \le \cdots$ almost everywhere and $f = \lim f_n$ (which has to exist almost everywhere) then

$$\int f(x)dx = \lim_{n\to\infty} f_n(x)dx$$

Dominated Convergence: If:

1) $|f_n| \le g_n$

2) there is a Borel function f such that $f_n(x) \to f(x)$ for almost all x

3) there is a Borel function g such that $g_n(x) \to g(x)$ with $\int g_n(x)dx \to \int g(x)dx < \infty$

Then f is integrable and

$$\int f_n(x)dx \to \int f(x)dx$$

Fatou's Lemma: If $f_n \ge 0$ almost everywhere then

$$\int \liminf f_n(x)dx \le \liminf \int f_n(x)dx.$$

Notice frequent use of almost all or almost everywhere in hypotheses. In def' of E wherever we require a property of the function $X(\omega)$ we can require it to hold only for a set of ω whose complement has probability 0. In this case we say the property holds **almost surely**. For instance the dominated convergence theorem is usually written:

Dominated Convergence: If $|X_n| \le Y_n$ almost surely (often abbreviated a.s.) and there is a random variable X such that $X_n \to X$ a.s. and a random variable Y such that $Y_n \to Y$ almost surely with $E(Y_n) \to E(Y) < \infty$ then

$$E(X_n) \to E(X)$$

Hypothesis of almost sure convergence can be weakened.

Multiple Integration: Lebesgue integrals over $\mathbb{R}^p$ defined using Lebesgue measure on $\mathbb{R}^p$.

Iterated integrals wrt Lebesgue measure on $\mathbb{R}^1$ give same answer.

Theorem[Tonelli]: If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and $f \geq 0$ almost everywhere then for almost every $x \in \mathbb{R}^p$ the integral

$$g(x) \equiv \int f(x,y)dy$$

exists and

$$\int g(x)dx = \int f(x,y)dxdy$$

RHS denotes $p+q$ dimensional integral defined previously.

Theorem[Fubini] If $f : \mathbb{R}^{p+q} \mapsto \mathbb{R}$ is Borel and integrable then for almost every $x \in \mathbb{R}^p$ the integral

$$g(x) \equiv \int f(x,y)dy$$

exists and is finite. Moreover g is integrable and

$$\int g(x)dx = \int f(x,y)dxdy\,.$$

Results true for measures other than Lebesgue.

Conditional distributions, expectations

When X and Y are discrete we have

$$\mathsf{E}(Y|X=x) = \sum_y yP(Y=y|X=x)$$

for any x for which $P(X=x)$ is positive.

Defines a function of x.

This function evaluated at X gives rv which is ftn of X denoted

$$\mathsf{E}(Y|X)\,.$$

Example: $Y|X = x \sim$ Binomial(x, p). Since mean of a Binomial(n, p) is np we find

$$\mathsf{E}(Y|X = x) = px$$

and

$$\mathsf{E}(Y|X) = pX$$

Notice you simply replace x by X.

Here are some properties of the function

$$\mathsf{E}(Y|X=x)$$

1) Suppose A is a function defined on the range of X. Then

$$\mathsf{E}(A(X)Y|X=x) = A(x)\mathsf{E}(Y|X=x)$$

and so

$$\mathsf{E}(A(X)Y|X) = A(X)\mathsf{E}(Y|X)$$

Second assertion follows from first. Note that if $Z = A(X)Y$ then Z is discrete and

$$P(Z=z) = \sum_{x,y} P(Y=y, X=x)1(z=A(x)y)$$

Also

$$\begin{aligned} &P(Z=z|X=x) \\ &\quad = \frac{\sum_y P(Y=y, X=x)1(z=A(x)y)}{P(X=x)} \\ &\quad = \sum_y P(Y=y|X=x)1(z=A(x)y) \end{aligned}$$

Thus

$$
\begin{aligned}
&\mathsf{E}(Z|X=x) \\
&\quad = \sum_z zP(Z=z|X=x) \\
&\quad = \sum_z \sum_y zP(Y=y|X=x)\mathsf{1}(z=A(x)y) \\
&\quad = \sum_z \sum_y A(x)yP(Y=y|X=x)\mathsf{1}(z=A(x)y) \\
&\quad = A(x)\sum_y yP(Y=y|X=x)\sum_z \mathsf{1}(z=A(x)y) \\
&\quad = A(x)\sum_y yP(Y=y|X=x)
\end{aligned}
$$

2) Repeated conditioning: if X, Y and Z discrete then

$$\begin{aligned} \mathsf{E}\{\mathsf{E}(Z|X,Y)|X\} &= \mathsf{E}(Z|X) \\ \mathsf{E}\{\mathsf{E}(Y|X)\} &= \mathsf{E}(Y) \end{aligned}$$

3) Additivity

$$\mathsf{E}(Y+Z|X) = \mathsf{E}(Y|X) + \mathsf{E}(Z|X)$$

4) Putting the first two items together gives

$$\begin{aligned} &\mathsf{E}\{\mathsf{E}(A(X)Y|X)\} = \\ &\mathsf{E}\{A(X)\mathsf{E}(Y|X)\} = \mathsf{E}(A(X)Y) \end{aligned} \tag{7}$$

Definition of $\mathsf{E}(Y|X)$ when X and Y are not assumed to discrete:

$\mathsf{E}(Y|X)$ is rv which is measurable function of X satisfying(7).

Existence is measure theory problem.

Aside on "measurable": what sorts of events can be defined in terms of a family $\{Y_i : i \in I\}$?

Natural: any event of form $(Y_{i_1},\ldots,Y_{i_k}) \in C$ is "defined in terms of the family" for any finite set $i_1,\ldots,i_k$ and any (Borel) set C in S^k.

For countable S: each singleton $(s_1,\ldots,s_k) \in S^k$ Borel. So every subset of S^k Borel.

Natural: if you can define each of a sequence of events A_n in terms of the Ys then the definition "there exists an n such that (definition of A_n) $\ldots$" defines $\cup A_n$.

Natural: if A is definable in terms of the Ys then A^c can be defined from the Ys by just inserting the phrase "It is not true that" in front of the definition of A.

So family of events definable in terms of the family $\{Y_i : i \in I\}$ is a σ-field which includes every event of the form $(Y_{i_1},\ldots,Y_{i_k}) \in C$. We call the smallest such σ-field, $\mathcal{F}(\{Y_i : i \in I\})$, the σ-field generated by the family $\{Y_i : i \in I\}$.

Suppose X is discrete and $X^* = g(X)$ is a one to one transformation of X. Since $X = x$ is the same event as $X^* = g(x)$ we find

$$\mathsf{E}(Y|X = x) = \mathsf{E}(Y|X^* = g(x))$$

Let $h^*(u)$ denote the function $\mathsf{E}(Y|X^* = u)$ and $h(u) = \mathsf{E}(Y|X = u)$. Then

$$h(x) = h^*(g(x))$$

Thus

$$h(X) = h^*(g(X)) = h^*(X^*)$$

This just means

$$\mathsf{E}(Y|X) = \mathsf{E}(Y|X^*)$$

Interpretation.

Formula is “obvious”.

Example: Toss coin $n = 20$ times. Y is indicator of first toss is a heads. X is number of heads and X^* number of tails. Formula says:

$$\mathsf{E}(Y|X = 17) = \mathsf{E}(Y|X^* = 3)$$

In fact for a general k and n

$$\mathsf{E}(Y|X = k) = \frac{k}{n}$$

so

$$\mathsf{E}(Y|X) = \frac{X}{n}$$

At the same time

$$\mathsf{E}(Y|X^* = j) = \frac{n-j}{n}$$

so

$$\mathsf{E}(Y|X^*) = \frac{n-X^*}{n}$$

But of course $X = n - X^*$ so these are just two ways of describing the same random variable.

Another interpretation: Rv X partitions Ω into countable set of events of the form $X = x$.

Other rv X^* partitions Ω into the same events.

Then values of $\mathsf{E}(Y|X^* = x^*)$ are same as values of $\mathsf{E}(Y|X = x)$ but labelled differently.

To form $\mathsf{E}(Y|X)$ take value ω, compute $X(\omega)$ to determine member A of the partition we being conditionsed on, then write down corresponding $\mathsf{E}(Y|A)$.

Hence conditional expectation depends only on partition of Ω.

X not discrete: replace partition with σ-field. Suppose X and X^* 2 rvs such that $\mathcal{F}(X) = \mathcal{F}(X^*)$. Then:

- There is g Borel,one to one with one to one Borel inverse s.t. $X^* = g(X)$.

- $\mathsf{E}(Y|X) = \mathsf{E}(Y|X^*)$ almost surely.

In other words $\mathsf{E}(Y|X)$ depends *only* on the σ-field generated by X. We write

$$\mathsf{E}(Y|\mathcal{F}(X)) = \mathsf{E}(Y|X)$$

Def'n: Suppose $\mathcal{G}$ is sub-σ-field of $\mathcal{F}$. X is $\mathcal{G}$ measurable if, for every Borel B

$$\{\omega : X(\omega) \in B\} \in \mathcal{G}\,.$$

Def'n: $\mathsf{E}(Y|\mathcal{G})$ is any $\mathcal{G}$ measurable rv s.t. for every $\mathcal{G}$ measurable rv variable A we have

$$\mathsf{E}(AY) = \mathsf{E}\left\{A\mathsf{E}(Y|\mathcal{G})\right\}\,.$$

Again existence is measure theory problem.

Markov Chains

Stochastic process: family $\{X_i; i \in I\}$ of rvs I the **index set**. Often $I \subset \mathbb{R}$, e.g. $[0, \infty)$, $[0, 1]$ $\mathbb{Z}$ or $\mathbb{N}$.

Continuous time: I is an interval

Discrete time: $I \subset \mathbb{Z}$.

Generally all X_n take values in **state space** S. In following S is a finite or countable set; each X_n is discrete.

Usually S is $\mathbb{Z}$, $\mathbb{N}$ or $\{0, \ldots, m\}$ for some finite m.

Markov Chain: stochastic process $X_n; n \in \mathbb{N}$. taking values in a finite or countable set S such that for every n and every event of the form

$$A = \{(X_0, \ldots, X_{n-1}) \in B \subset S^n\}$$

we have

$$P(X_{n+1} = j | X_n = i, A) = P(X_1 = j | X_0 = i) \tag{8}$$

Notation: $\mathbf{P}$ is the (possibly infinite) array with elements

$$P_{ij} = P(X_1 = j | X_0 = i)$$

indexed by $i, j \in S$.

$\mathbf{P}$ is the (one step) **transition matrix** of the Markov Chain.

WARNING: in (8) we require the condition to hold **only** when

$$P(X_n = i, A) > 0$$

Evidently the entries in $\mathbf{P}$ are non-negative and

$$\sum_j P_{ij} = 1$$

for all $i \in S$. Any such matrix is called **stochastic**.

We define powers of $\mathbf{P}$ by

$$(\mathbf{P}^n)_{ij} = \sum_k \left(\mathbf{P}^{n-1}\right)_{ik} P_{kj}$$

Notice that even if S is infinite these sums converge absolutely.

Chapman-Kolmogorov Equations

Condition on X_{l+n-1} to compute

$$P(X_{l+n} = j | X_l = i)$$

$$\begin{aligned}
&P(X_{l+n} = j | X_l = i) \\
&\quad = \sum_k P(X_{l+n} = j, X_{l+n-1} = k | X_l = i) \\
&\quad = \sum_k P(X_{l+n} = j | X_{l+n-1} = k, X_l = i) \\
&\qquad \times P(X_{l+n-1} = k | X_l = i) \\
&\quad = \sum_k P(X_1 = j | X_0 = k) \\
&\qquad \times P(X_{l+n-1} = k | X_l = i) \\
&\quad = \sum_k P(X_{l+n-1} = k | X_l = i) \mathbf{P}_{kj}
\end{aligned}$$

Now condition on X_{l+n-2} to get

$$\begin{aligned}
&P(X_{l+n} = j | X_l = i) = \\
&\qquad \sum_{k_1 k_2} \mathbf{P}_{k_1 k_2} \mathbf{P}_{k_2 j} P(X_{l+n-2} = k_1 | X_l = i)
\end{aligned}$$

Notice: sum over k_2 computes k_1, j entry in matrix $\mathbf{P}\mathbf{P} = \mathbf{P}^2$.

$$P(X_{l+n} = j | X_l = i) = \sum_{k_1} (\mathbf{P}^2)_{k_1,j} P(X_{l+n-2} = k_1 | X_l = i)$$

We may now prove by induction on n that

$$P(X_{l+n} = j | X_l = i) = (\mathbf{P}^n)_{ij} .$$

This proves Chapman-Kolmogorov equations:

$$P(X_{l+m+n} = j | X_l = i) = \sum_k P(X_{l+m} = k | X_l = i) \times P(X_{l+m+n} = j | X_{l+m} = k)$$

These are simply a restatement of the identity

$$\mathbf{P}^{n+m} = \mathbf{P}^n \mathbf{P}^m .$$

Remark: It is important to notice that these probabilities depend on m and n but **not** on l. We say the chain has **stationary** transition probabilities. A more general definition of Markov chain than (8) is

$$\begin{aligned} P(X_{n+1} = j|X_n = i, A) & \\ &= P(X_{n+1} = j|X_n = i)\,. \end{aligned}$$

Notice RHS now permitted to depend on n.

Define $\mathbf{P}^{n,m}$: matrix with i,jth entry

$$P(X_m = j|X_n = i)$$

for $m > n$. Then

$$\mathbf{P}^{r,s}\mathbf{P}^{s,t} = \mathbf{P}^{r,t}$$

Also called Chapman-Kolmogorov equations. This chain does not have stationary transitions.

Remark: The calculations above involve sums in which all terms are positive. They therefore apply even if the state space S is countably infinite.

Extensions of the Markov Property

Function $f(x_0, x_1, \ldots)$ defined on S^∞ = all infinite sequences of points in S.

Let B_n be the event

$$f(X_n, X_{n+1}, \ldots) \in C$$

for suitable C in range space of f. Then

$$P(B_n|X_n = x, A) = P(B_0|X_0 = x) \tag{9}$$

for any event A of the form

$$\{(X_0, \ldots, X_{n-1}) \in D\}$$

Also

$$P(AB_n|X_n = x) = P(A|X_n = x)P(B_n|X_n = x) \tag{10}$$

"Given the present the past and future are conditionally independent."

Proof of (9):

Special case:

$$B_n = \{(X_{n+1} = x_1, \cdots, X_{n+m} = x_m\}$$

LHS of (9) evaluated by repeated conditioning (cf. Chapman-Kolmogorov):

$$\mathbf{P}_{x,x_1}\mathbf{P}_{x_1,x_2}\cdots\mathbf{P}_{x_{m-1},x_m}$$

Same for RHS.

Events defined from $X_n, \ldots, X_{n+m}$: sum over appropriate vectors $x, x_1, \ldots, x_m$.

General case: monotone class techniques.

To prove (10) write

$$\begin{aligned} & P(AB_n|X_n = x) \\ & \qquad = P(B_n|X_n = x, A)P(A|X_n = x) \\ & \qquad = P(B_n|X_n = x)P(A|X_n = x) \end{aligned}$$

using (9).

Classification of States

If an entry $\mathbf{P}_{ij}$ is 0 it is not possible to go from state i to state j in one step. It may be possible to make the transition in some larger number of steps, however. We say i **leads to** j (or j is accessible from i) if there is an integer $n \ge 0$ such that

$$P(X_n = j | X_0 = i) > 0\,.$$

We use the notation $i \rightsquigarrow j$. Define $\mathbf{P}^0$ to be identity matrix $\mathbf{I}$. Then $i \rightsquigarrow j$ if there is an $n \ge 0$ for which $(\mathbf{P}^n)_{ij} > 0$.

States i and j **communicate** if $i \rightsquigarrow j$ and $j \rightsquigarrow i$.

Write $i \leftrightarrow j$ if i and j communicate.

Communication is an equivalence relation: reflexive, symmetric, transitive relation on states of S.

More precisely:

Reflexive: for all i we have $i \leftrightarrow j$.

Symmetric: if $i \leftrightarrow j$ then $j \leftrightarrow i$.

Transitive: if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.

Proof:

Reflexive: follows from inclusion of $n = 0$ in definition of leads to.

Symmetry is obvious.

Transitivity: suffices to check that $i \rightsquigarrow j$ and $j \rightsquigarrow k$ imply that $i \rightsquigarrow k$. But if $(\mathbf{P}^m)_{ij} > 0$ and $(\mathbf{P}^n)_{jk} > 0$ then

$$\begin{aligned}(\mathbf{P}^{m+n})_{ik} &= \sum_l (\mathbf{P}^m)_{il}(\mathbf{P}^n)_{lk} \\ &\geq (\mathbf{P}^m)_{ij}(\mathbf{P}^n)_{jk} \\ &> 0\end{aligned}$$

Any equivalence relation on a set partitions the set into **equivalence classes**; two elements are in the same equivalence class if and only if they are equivalent.

Communication partitions S into equivalence classes called **communicating classes**.

Example:

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Find communicating classes: start with say state 1, see where it leads.

- $1 \rightsquigarrow 2$, $1 \rightsquigarrow 3$ and $1 \rightsquigarrow 4$ in row 1.

- Row 4: $4 \rightsquigarrow 1$. So: (transitivity) 1, 2, 3 and 4 all in the same communicating class.

- Claim: none of these leads to 5, 6, 7 or 8.

  Suppose $i \in \{1,2,3,4\}$ and $j \in \{5,6,7,8\}$. Then $(\mathbf{P}^n)_{ij}$ is sum of products of $\mathbf{P}_{kl}$. Cannot be positive unless there is a sequence $i_0 = i, i_1, \ldots, i_n = j$ with $\mathbf{P}_{i_{k-1},i_k} > 0$ for $k = 1, \ldots, n$.

  Consider first k for which $i_k \in \{5,6,7,8\}$ Then $i_{k-1} \in \{1,2,3,4\}$ and so $\mathbf{P}_{i_{k-1},i_k} = 0$.

So: $\{1,2,3,4\}$ is a communicating class.

- $5 \rightsquigarrow 1$, $5 \rightsquigarrow 2$, $5 \rightsquigarrow 3$ and $5 \rightsquigarrow 4$.

- None of these lead to any of $\{5,6,7,8\}$ so $\{5\}$ must be communicating class.

- Similarly $\{6\}$ and $\{7,8\}$ are communicating classes.

Note: states 5 and 6 have special property. Each time you are in either state you run a risk of going to one of the states 1, 2, 3 or 4. Eventually you will make such a transition and then never return to state 5 or 6.

States 5 and 6 are **transient**.

To make this precise define hitting times:

$$T_k = \min\{n > 0 : X_n = k\}$$

We define

$$f_k = P(T_k < \infty | X_0 = k)$$

State k is **transient** if $f_k < 1$ and **recurrent** if $f_k = 1$.

Let N_k be number of times chain is ever in state k.

Claims:

1. If $f_i < 1$ then N_k has a Geometric distribution:

   $$P(N_k = r | X_0 = k) = f_k^{r-1}(1 - f_k)$$

   for $r = 1, 2, \ldots$.

2. If $f_i = 1$ then

   $$P(N_k = \infty | X_0 = k) = 1$$

Proof using **Strong Markov Property**:

Stopping time for the Markov chain is a random variable T taking values in $\{0, 1, \cdots\} \cup \{\infty\}$ such that for each finite k there is a function f_k such that

$$1(T = k) = f_k(X_0, \ldots, X_k)$$

Notice that T_k in theorem is a stopping time.

Standard shorthand notation: by

$$P^x(A)$$

we mean

$$P(A|X_0 = x) .$$

Similarly we define

$$\mathsf{E}^x(Y) = \mathsf{E}(Y|X_0 = x) .$$

strong Markov property:

Stopping time for the Markov chain is a random variable T taking values in $\{0, 1, \cdots\} \cup \{\infty\}$ such that for each finite k there is a function f_k such that

$$1(T = k) = f_k(X_0, \ldots, X_k)$$

Notice that T_k in theorem is a stopping time.

Standard shorthand notation: by

$$P^x(A)$$

we mean

$$P(A|X_0 = x)\,.$$

Similarly we define

$$\mathsf{E}^x(Y) = \mathsf{E}(Y|X_0 = x)\,.$$

Goal: explain and prove

$$\mathsf{E}(f(X_T,\ldots)|X_T,\ldots,X_0) = \mathsf{E}^{X_T}(f(X_0,\ldots))$$

Simpler claim:

$$P(X_{T+1}=j|X_T=i) = \mathbf{P}_{ij} = P^i(X_1=j)\,.$$

Notation: $A_k = \{X_k = i, T = k\}$

Notice: $A_k = \{X_T = k, T = k\}$:

$$\begin{aligned}
P(X_{T+1}=j|X_T=i) &= \frac{P(X_{T+1}=j, X_T=i)}{P(X_T=i)} \\
&= \frac{\sum_k P(X_{T+1}=j, X_T=i, T=k)}{\sum_k P(X_T=i, T=k)} \\
&= \frac{\sum_k P(X_{k+1}=j, A_k)}{\sum_k P(A_k)} \\
&= \frac{\sum_k P(X_{k+1}=j|A_k)P(A_k)}{\sum_k P(A_k)} \\
&= \frac{\sum_k P(X_1=j|X_0=i)P(A_k)}{\sum_k P(A_k)} \\
&= \mathbf{P}_{i,j}
\end{aligned}$$

Notice use of fact that $T = k$ is event defined in terms of $X_0, \ldots, X_k$.

Technical problems with proof:

• It might be that $P(T = \infty) > 0$. What are X_T and X_{T+1} on the event $T = \infty$.

Answer: condition also on $T < \infty$.

• Prove formula only for stopping times where $\{T < \infty\} \cap \{X_T = i\}$ has positive probability.

We will now fix up these technical details.

Suppose $f(x_0, x_1, \ldots)$ is a (measurable) function on $S^{\mathbb{N}}$. Put

$$Y_n = f(X_n, X_{n+1}, \ldots)\,.$$

Assume $\mathsf{E}(\,|Y_0|\,|X_0 = x) < \infty$ for all x. Claim:

$$\mathsf{E}(Y_n|X_n, A) = \mathsf{E}^{X_n}(Y_0) \qquad (11)$$

whenever A is any event defined in terms of $X_0, \ldots, X_n$.

Proof:

1 Family of f for which claim holds includes all indicators; see extension of Markov Property in previous lecture.

2 family of f for which claim is true is vector space (so if f, g in family then so is $af + bg$ for any constants a and b.

- So family of f for which claim is true includes all simple functions.

- family of f for which claim true is closed under monotone increasing limits (of non-negative f_n) by the Monotone Convergence theorem.

- So claim true for every non-negative integrable f.

- Claim follows for integrable f by linearity.

Aside on “measurable”: what sorts of events can be defined in terms of a family $\{Y_i : i \in I\}$?

Natural: any event of form $(Y_{i_1},\ldots,Y_{i_k}) \in C$ is "defined in terms of the family" for any finite set $i_1,\ldots,i_k$ and any (Borel) set C in S^k.

For countable S: each singleton $(s_1,\ldots,s_k) \in S^k$ Borel. So every subset of S^k Borel.

Natural: if you can define each of a sequence of events A_n in terms of the Ys then the definition "there exists an n such that (definition of A_n) ..." defines $\cup A_n$.

Natural: if A is definable in terms of the Ys then A^c can be defined from the Ys by just inserting the phrase "It is not true that" in front of the definition of A.

So family of events definable in terms of the family $\{Y_i : i \in I\}$ is a σ-field which includes every event of the form $(Y_{i_1},\ldots,Y_{i_k}) \in C$. We call the smallest such σ-field, $\mathcal{F}(\{Y_i : i \in I\})$, the σ-field generated by the family $\{Y_i : i \in I\}$.

Using the Markov property:

Toss coin till I get a head. What is the expected number of tosses?

Define state to be 0 if toss is tail and 1 if toss is heads.

Define $X_0 = 0$.

Let $N = \min\{n > 0 : X_n = 1\}$. Want

$$\mathrm{E}(N) = \mathrm{E}^0(N)$$

Note: if $X_1 = 1$ then $N = 1$. If $X_1 = 0$ then $N = 1 + \min\{n > 0 : X_{n+1} = 1\}$.

In symbols:

$$N = \min\{n > 0 : X_n = 1\} = f(X_1, X_2, \cdots)$$

and

$$N = 1 + 1(X_1 = 0) f(X_2, X_3, \cdots)$$

Take expected values starting from 0:

$$\mathsf{E}^0(N) = 1 + \mathsf{E}^0\{1(X_1 = 0)f(X_2, X_3, \cdots)\}$$

Condition on X_1 and get

$$\mathsf{E}^0(N) = 1 + \mathsf{E}^0[\mathsf{E}\{1(X_1 = 0)f(X_2, \cdots)|X_1\}]$$

But

$$\begin{aligned}\mathsf{E}\{1(X_1 = 0)f(X_2, X_3, \cdots)|X_1\}& \\ &= 1(X_1 = 0)\mathsf{E}^{X_1}\{f(X_1, X_2, \cdots)\} \\ &= 1(X_1 = 0)\mathsf{E}^0\{f(X_1, X_2, \cdots)\} \\ &= 1(X_1 = 0)\mathsf{E}^0(N)\end{aligned}$$

so that

$$\mathsf{E}^0(N) = 1 + p\mathsf{E}^0\{N\}$$

where p is the probability of tails. Solve for $\mathsf{E}(N)$ to get

$$\mathsf{E}(N) = \frac{1}{1-p}$$

This is the formula for expected value of the sort of geometric which starts at 1 and has p being the probability of failure.

Initial Distributions

Meaning of unconditional expected values?

Markov property specifies only cond'l probs; no way to deduce marginal distributions.

For every dstbn π on S and transition matrix $\mathbf{P}$ there is a a stochastic process $X_0, X_1, \ldots$ with

$$P(X_0 = k) = \pi_k$$

and which is a Markov Chain with transition matrix $\mathbf{P}$.

Note Strong Markov Property proof used only conditional expectations.

Notation: π a probability on S. E^π and P^π are expected values and probabilities for chain with initial distribution π.

Summary of easily verified facts:

- For any sequence of states $i_0, \ldots, i_k$

  $$P(X_0 = i_0, \ldots, X_k = i_k) = \pi_{i_0} \mathbf{P}_{i_0 i_1} \cdots \mathbf{P}_{i_{k-1} i_k}$$

- For any event A:

  $$\mathbf{P}^\pi(A) = \sum_k \pi_k \mathbf{P}^k(A)$$

- For any bounded rv $Y = f(X_0, \ldots)$

  $$\mathsf{E}^\pi(Y) = \sum_k \pi_k \mathsf{E}^k(A)$$

Recurrence and Transience

Now consider a transient state k, that is, a state for which

$$f_k = P^k(T_k < \infty) < 1$$

Note that $T_k = \min\{n > 0 : X_n = k\}$ is a stopping time. Let N_k be the number of visits to state k. That is

$$N_k = \sum_{n=0}^{\infty} 1(X_n = k)$$

Notice that if we define the function

$$f(x_0, x_1, \ldots) = \sum_{n=0}^{\infty} 1(x_n = k)$$

then

$$N_k = f(X_0, X_1, \ldots)$$

Notice, also, that on the event $T_k < \infty$

$$N_k = 1 + f(X_{T_k}, X_{T_k+1}, \ldots)$$

and on the event $T_k = \infty$ we have

$$N_k = 1$$

In short:

$$N_k = 1 + f(X_{T_k}, X_{T_k+1}, \ldots)1(T_k < \infty)$$

Hence

$$\begin{aligned}
&\mathbf{P}^k(N_k = r) \\
&\quad = \mathsf{E}^k \{P(N_k = r|\mathcal{F}_T)\} \\
&\quad = \mathsf{E}^k \Big[P\Big\{1 + f(X_{T_k}, X_{T_k+1}, \ldots) \\
&\qquad\qquad \times 1(T_k < \infty) = r|\mathcal{F}_T\}] \\
&\quad = \mathsf{E}^k [1(T_k < \infty) \\
&\qquad\quad \times P^{X_{T_k}} \{f(X_0, X_1, \ldots) = r - 1\}\Big] \\
&\quad = \mathsf{E}^k \Big\{1(T_k < \infty) P^k(N_k = r - 1)\Big\} \\
&\quad = \mathsf{E}^k \{1(T_k < \infty)\} P^k(N_k = r - 1) \\
&\quad = f_k P^k(N_k = r - 1)
\end{aligned}$$

It is easily verified by induction, then, that

$$\mathbf{P}^k(N_k = r) = f_k^{r-1} P^k(N_k = 1)$$

But $N_k = 1$ if and only if $T_k = \infty$ so

$$\mathbf{P}^k(N_k = r) = f_k^{r-1}(1 - f_k)$$

so N_k has (chain starts from k) Geometric dist'n, mean $1/(1 - f_k)$. Argument also shows that if $f_k = 1$ then

$$P^k(N_k = 1) = P^k(N_k = 2) = \cdots$$

which can only happen if all these probabilities are 0. Thus if $f_k = 1$

$$P(N_k = \infty) = 1$$

Since

$$N_k = \sum_{n=0}^{\infty} 1(X_n = k)$$

$$\mathsf{E}^k(N_k) = \sum_{n=0}^{\infty} (\mathbf{P}^n)_{kk}$$

So: State k is transient if and only if

$$\sum_{n=0}^{\infty} (\mathbf{P}^n)_{kk} < \infty$$

and this sum is $1/(1 - f_k)$.

Proposition 1 *Recurrence (or transience) is a class property. That is, if i and j are in the same communicating class then i is recurrent (respectively transient) if and only if j is recurrent (respectively transient).*

Proof: Suppose i is recurrent and $i \leftrightarrow j$. There are integers m and n such that

$$(\mathbf{P}^m)_{ji} > 0 \quad \text{and} \quad (\mathbf{P}^n)_{ij} > 0$$

Then

$$\begin{aligned}\sum_k (\mathbf{P}^k)_{jj} &\ge \sum_{k\ge 0} (\mathbf{P}^{m+k+n})_{jj} \\ &\ge \sum_{k\ge 0} (\mathbf{P}^m)_{ji}(\mathbf{P}^k)_{ii}(\mathbf{P}^n)_{ij} \\ &= (\mathbf{P}^m)_{ji}\left\{\sum_{k\ge 0}(\mathbf{P}^k)_{ii}\right\}(\mathbf{P}^n)_{ij}\end{aligned}$$

The middle term is infinite and the two outside terms positive so

$$\sum_k (\mathbf{P}^k)_{jj} = \infty$$

which shows j is recurrent.

A finite state space chain has at least one recurrent state:

If all states we transient we would have for each k $P(N_k < \infty) = 1$. This would mean $P(\forall k\,.N_k < \infty) = 1$. But for any ω there must be at least one k for which $N_k = \infty$ (the total of a finite list of finite numbers is finite).

Infinite state space chain may have all states transient:

The chain X_n satisfying $X_{n+1} = X_n + 1$ on the integers has all states transient.

More interesting example:

- Toss a coin repeatedly.

- Let X_n be X_0 plus the number of heads minus the number of tails in the first n tosses.

- Let p denote the probability of heads on an individual trial.

$X_n - X_0$ is a sum of n iid random variables Y_i where $P(Y_i = 1) = p$ and $P(Y_i = -1) = 1 - p$.

SLLN shows X_n/n converges almost surely to $2p - 1$. If $p \neq 1/2$ this is not 0.

In order for X_n/n to have a positive limit we must have $X_n \to \infty$ almost surely so all states are visited only finitely many times. That is, all states are transient. Similarly for $p < 1/2$ $X_n \to -\infty$ almost surely and all states are transient.

Now look at $p = 1/2$. The law of large numbers argument no long shows anything. I will show that all states are recurrent.

Proof: We evaluate $\sum_n (\mathbf{P}^n)_{00}$ and show the sum is infinite. If n is odd then $(p_n)_{00} = 0$ so we evaluate

$$\sum_m (\mathbf{P}^{2m})_{00}$$

Now

$$(\mathbf{P}^{2m})_{00} = \binom{2m}{m} 2^{-2m}$$

According to Stirling's approximation

$$\lim_{m\to\infty} \frac{m!}{m^{m+1/2}e^{-m}\sqrt{2\pi}} = 1$$

Hence

$$\lim_{m\to\infty} \sqrt{m}(\mathbf{P}^{2m})_{00} = \frac{1}{\sqrt{\pi}}$$

Since

$$\sum \frac{1}{\sqrt{m}} = \infty$$

we are done.

Mean return times

Compute expected times to return. For $x \in S$ let T_x denote the hitting time for x.

Suppose x recurrent in **irreducible** chain (only one communicating class).

Derive equations for expected values of different T_x.

Each T_x is a certain function f_x applied to $X_1, \ldots$. Setting $\mu_{ij} = \mathsf{E}^i(T_j)$ we find

$$\mu_{ij} = \sum_k \mathsf{E}^i(T_j 1(X_1 = k))$$

Note that if $X_1 = x$ then $T_x = 1$ so

$$\mathsf{E}^i(T_j 1(X_1 = j)) = \mathbf{P}_{ij}$$

For $k \neq j$

$$T_x = 1 + f_x(X_2, X_3, \ldots)$$

and, by conditioning on $X_1 = k$ we find

$$\mathsf{E}^i(T_j 1(X_1 = k)) = \mathbf{P}_{ik}\left\{1 + \mathsf{E}^k(T_j)\right\}$$

This gives

$$\mu_{ij} = 1 + \sum_{k \neq j} \mathbf{P}_{ik}\mu_{kj} \tag{12}$$

Technically, I should check that the expectations in (12) are finite. All the random variables involved are non-negative, however, and the equation actually makes sense even if some terms are infinite. (To prove this you actually study

$$T_{x,n} = \min(T_x, n)$$

deriving an identity for a fixed n, letting $n \to \infty$ and applying the monotone convergence theorem.)

Here is a simple example:

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

The identity (12) becomes

$$\begin{aligned}
\mu_{1,1} &= 1 + \frac{1}{2}\mu_{2,1} + \frac{1}{2}\mu_{3,1} \\
\mu_{1,2} &= 1 + \frac{1}{2}\mu_{3,1} \\
\mu_{1,3} &= 1 + \frac{1}{2}\mu_{2,1} \\
\mu_{2,1} &= 1 + \frac{1}{2}\mu_{3,1} \\
\mu_{2,2} &= 1 + \frac{1}{2}\mu_{1,2} + \frac{1}{2}\mu_{3,2} \\
\mu_{2,3} &= 1 + \frac{1}{2}\mu_{1,3} \\
\mu_{3,1} &= 1 + \frac{1}{2}\mu_{2,1} \\
\mu_{3,2} &= 1 + \frac{1}{2}\mu_{1,2} \\
\mu_{3,3} &= 1 + \frac{1}{2}\mu_{1,3} + \frac{1}{2}\mu_{2,3}
\end{aligned}$$

Seventh and fourth show $\mu_{2,1} = \mu_{3,1}$. Similar calculations give $\mu_{ii} = 3$ and for $i \neq j$ $\mu_{i,j} = 2$.

Example: Coin tossing Markov Chain with $p = 1/2$ shows situation can be different when S is infinite. Equations above become:

$$\begin{aligned} m_{0,0} &= 1 + \frac{1}{2} m_{1,0} + \frac{1}{2} m_{-1,0} \\ m_{1,0} &= 1 + \frac{1}{2} m_{2,0} \end{aligned}$$

and many more.

Some observations:

Have to go through 1 to get to 0 from 2 so

$$m_{2,0} = m_{2,1} + m_{1,0}$$

Symmetry (switching H and T):

$$m_{1,0} = m_{-1,0}$$

Transition probabilities are **homogeneous**:

$$m_{2,1} = m_{1,0}$$

Conclusion:

$$\begin{aligned} m_{0,0} &= 1 + m_{1,0} \\ &= 1 + 1 + \frac{1}{2} m_{2,0} \\ &= 2 + m_{1,0} \end{aligned}$$

Notice that there are **no** finite solutions!

Summary of the situation:

Every state is recurrent.

All the expected hitting times m_{ij} are infinite.

All entries $\mathbf{P}^n_{ij}$ converge to 0.

Jargon: The states in this chain are null recurrent.

```
> p:= matrix(2,2,[[3/5,2/5],[1/5,4/5]]);
                        [3/5    2/5]
                   p := [          ]
                        [1/5    4/5]
> p2:=evalm(p*p):
> p4:=evalm(p2*p2):
> p8:=evalm(p4*p4):
> p16:=evalm(p8*p8):
```

This computes the powers (`evalm` understands matrix algebra).

Fact:

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

```
> evalf(evalm(p));
            [.6000000000    .4000000000]
            [                          ]
            [.2000000000    .8000000000]
> evalf(evalm(p2));
            [.4400000000    .5600000000]
            [                          ]
            [.2800000000    .7200000000]
> evalf(evalm(p4));
            [.3504000000    .6496000000]
            [                          ]
            [.3248000000    .6752000000]
> evalf(evalm(p8));
            [.3337702400    .6662297600]
            [                          ]
            [.3331148800    .6668851200]
> evalf(evalm(p16));
            [.3333336197    .6666663803]
            [                          ]
            [.3333331902    .6666668098]
```

Where did 1/3 and 2/3 come from?

Suppose we toss a coin $P(H) = \alpha_D$ and start the chain with Dry if we get heads and Wet if we get tails.

Then

$$P(X_0 = x) = \begin{cases} \alpha_D & x = \text{Dry} \\ \alpha_W = 1 - \alpha_D & x = \text{Wet} \end{cases}$$

and

$$\begin{aligned} P(X_1 = x) &= \sum_y P(X_1 = x | X_0 = y) P(X_0 = y) \\ &= \sum_y \alpha_y P_{y,x} \end{aligned}$$

Notice last line is a matrix multiplication of row vector α by matrix $\mathbf{P}$. A special α: if we put $\alpha_D = 1/3$ and $\alpha_W = 2/3$ then

$$\begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

So: if $P(X_0 = D) = 1/3$ then $P(X_1 = D) = 1/3$ and analogously for W. This means that X_0 and X_1 have the same distribution.

A probability vector α is called the initial distribution for the chain if

$$P(X_0 = i) = \alpha_i$$

A Markov Chain is **stationary** if

$$P(X_1 = i) = P(X_0 = i)$$

for all i

Finding stationary initial distributions. Consider $\mathbf{P}$ above. The equation

$$\alpha\mathbf{P} = \alpha$$

is really

$$\begin{aligned}\alpha_D &= 3\alpha_D/5 + \alpha_W/5 \\ \alpha_W &= 2\alpha_D/5 + 4\alpha_W/5\end{aligned}$$

The first can be rearranged to

$$\alpha_W = 2\alpha_D.$$

So can the second. If α is probability vector then

$$\alpha_W + \alpha_D = 1$$

so we get

$$1 - \alpha_D = 2\alpha_D$$

leading to

$$\alpha_D = 1/3$$

Some more examples:

$$\mathbf{P} = \begin{bmatrix} 0 & 1/3 & 0 & 2/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 2/3 & 0 & 1/3 & 0 \end{bmatrix}$$

Set $\alpha \mathbf{P} = \alpha$ and get

$$\begin{aligned} \alpha_1 &= \alpha_2/3 + 2\alpha_4/3 \\ \alpha_2 &= \alpha_1/3 + 2\alpha_3/3 \\ \alpha_3 &= 2\alpha_2/3 + \alpha_4/3 \\ \alpha_4 &= 2\alpha_1/3 + \alpha_3/3 \\ 1 &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \end{aligned}$$

First plus third gives

$$\alpha_1 + \alpha_3 = \alpha_2 + \alpha_4$$

so both sums 1/2. Continue algebra to get

$$(1/4, 1/4, 1/4, 1/4)\,.$$

```
p:=matrix([[0,1/3,0,2/3],[1/3,0,2/3,0],
          [0,2/3,0,1/3],[2/3,0,1/3,0]]);

               [ 0     1/3     0      2/3]
               [                         ]
               [1/3     0     2/3      0 ]
          p := [                         ]
               [ 0     2/3     0      1/3]
               [                         ]
               [2/3     0     1/3      0 ]
> p2:=evalm(p*p);
             [5/9     0     4/9      0 ]
             [                         ]
             [ 0     5/9     0      4/9]
        p2:= [                         ]
             [4/9     0     5/9      0 ]
             [                         ]
             [ 0     4/9     0      5/9]
> p4:=evalm(p2*p2):
> p8:=evalm(p4*p4):
> p16:=evalm(p8*p8):
> p17:=evalm(p8*p8*p):
```

```
> evalf(evalm(p16));
     [.5000000116 , 0 , .4999999884 , 0]
     [                                 ]
     [0 , .5000000116 , 0 , .4999999884]
     [                                 ]
     [.4999999884 , 0 , .5000000116 , 0]
     [                                 ]
     [0 , .4999999884 , 0 , .5000000116]
> evalf(evalm(p17));
     [0 , .4999999961 , 0 , .5000000039]
     [                                 ]
     [.4999999961 , 0 , .5000000039 , 0]
     [                                 ]
     [0 , .5000000039 , 0 , .4999999961]
     [                                 ]
     [.5000000039 , 0 , .4999999961 , 0]
```

```
> evalf(evalm((p16+p17)/2));
  [.2500, .2500, .2500, .2500]
  [                          ]
  [.2500, .2500, .2500, .2500]
  [                          ]
  [.2500, .2500, .2500, .2500]
  [                          ]
  [.2500, .2500, .2500, .2500]
```

$\mathbf{P}^n$ doesn't converges but$(\mathbf{P}^n + \mathbf{P}^{n+1})/2$ does.
Next example:

$$p = \begin{bmatrix} \frac{2}{5} & \frac{3}{5} & 0 & 0 \\ \frac{1}{5} & \frac{4}{5} & 0 & 0 \\ 0 & 0 & \frac{2}{5} & \frac{3}{5} \\ 0 & 0 & \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

Solve $\alpha\mathbf{P} = \alpha$:

$$\begin{aligned}
\alpha_1 &= \frac{2}{5}\alpha_1 + \frac{1}{5}\alpha_2 \\
\alpha_2 &= \frac{3}{5}\alpha_1 + \frac{4}{5}\alpha_2 \\
\alpha_3 &= \frac{2}{5}\alpha_3 + \frac{1}{5}\alpha_4 \\
\alpha_4 &= \frac{3}{5}\alpha_3 + \frac{4}{5}\alpha_4 \\
1 &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4
\end{aligned}$$

Second and fourth equations redundant. Get

$$\begin{aligned}
\alpha_2 &= 3\alpha_1 \\
3\alpha_3 &= \alpha_4 \\
1 &= 4\alpha_1 + 4\alpha_3
\end{aligned}$$

Pick α_1 in $[0, 1/4]$; put $\alpha_3 = 1/4 - \alpha_1$.

$$\alpha = (\alpha_1, 3\alpha_1, 1/4 - \alpha_1, 3(1/4 - \alpha_1))$$

solves $\alpha\mathbf{P} = \alpha$. So solution is not unique.

```
> p:=matrix([[2/5,3/5,0,0],[1/5,4/5,0,0],
      [0,0,2/5,3/5],[0,0,1/5,4/5]]);
               [2/5    3/5      0      0 ]
               [                         ]
               [1/5    4/5      0      0 ]
          p := [                         ]
               [ 0      0      2/5    3/5]
               [                         ]
               [ 0      0      1/5    4/5]
> p2:=evalm(p*p):
> p4:=evalm(p2*p2):
> p8:=evalm(p4*p4):
> evalf(evalm(p8*p8));
        [.2500000000 , .7500000000 , 0 , 0]
        [                                  ]
        [.2500000000 , .7500000000 , 0 , 0]
        [                                  ]
        [0 , 0 , .2500000000 , .7500000000]
        [                                  ]
        [0 , 0 , .2500000000 , .7500000000]
```

Notice that rows converge but to two different vectors:

$$\alpha^{(1)} = (1/4, 3/4, 0, 0)$$

and

$$\alpha^{(2)} = (0, 0, 1/4, 3/4)$$

Solutions of $\alpha\mathbf{P} = \alpha$ revisited? Check that

$$\alpha^{(1)}\mathbf{P} = \alpha^{(1)}$$

and

$$\alpha^{(2)}\mathbf{P} = \alpha^{(2)}$$

If $\alpha = \lambda\alpha^{(1)} + (1-\lambda)\alpha^{(2)}$ $(0 \le \lambda \le 1)$ then

$$\alpha\mathbf{P} = \alpha$$

so again solution is not unique.

Last example:

```
> p:=matrix([[2/5,3/5,0],[1/5,4/5,0],
             [1/2,0,1/2]]);

                   [2/5    3/5      0 ]
                   [                  ]
              p := [1/5    4/5      0 ]
                   [                  ]
                   [1/2     0     1/2]
> p2:=evalm(p*p):
> p4:=evalm(p2*p2):
> p8:=evalm(p4*p4):
> evalf(evalm(p8*p8));
  [.2500000000 .7500000000        0        ]
  [                                        ]
  [.2500000000 .7500000000        0        ]
  [                                        ]
  [.2500152588 .7499694824 .00001525878906]
```

Interpretation of examples

- For some $\mathbf{P}$ all rows converge to some α. In this case this α is a stationary initial distribution.

- For some $\mathbf{P}$ the locations of zeros flip flop. $\mathbf{P}^n$ does not converge. Observation: average
$$\frac{\mathbf{P}+\mathbf{P}^2+\cdots+\mathbf{P}^n}{n}$$
does converge.

- For some $\mathbf{P}$ some rows converge to one α and some to another. In this case the solution of $\alpha\mathbf{P}=\alpha$ is not unique.

Basic distinguishing features: pattern of 0s in matrix $\mathbf{P}$.

The ergodic theorem

Consider a finite state space chain. If x is a vector then the ith entry in $\mathbf{P}x$ is

$$\sum_j \mathbf{P}_{ij} x_j$$

Rows of $\mathbf{P}$ probability vectors, so a weighted average of the entries in x.

If the weights are strictly between 0 and 1 and the largest and smallest entries in x are not the same then $\sum_j \mathbf{P}_{ij} x_j$ is strictly between the largest and smallest entries in x. In fact

$$\sum_j \mathbf{P}_{ij} x_j - \text{min}(x_k) \\ \geq \min_j\{p_{ij}\}(\text{max}\{x_k\} - \text{min}\{x_k\})$$

and

$$\text{max}\{x_j\} - \sum_j \mathbf{P}_{ij} x_j \\ \geq \min_j\{p_{ij}\}(\text{max}\{x_k\} - \text{min}\{x_k\})$$

Now multiply $\mathbf{P}^r$ by $\mathbf{P}^m$.

ijth entry in $\mathbf{P}^{r+m}$ is a weighted average of the jth column of $\mathbf{P}^m$.

So, if all the entries in row i of $\mathbf{P}^r$ are positive and the jth column of $\mathbf{P}^m$ is not constant, the ith entry in the jth column of $\mathbf{P}^{r+m}$ must be strictly between the minimum and maximum entries of the jth column of $\mathbf{P}^m$.

In fact, fix a j.

$\overline{x}_m =$ maximum entry in column j of $\mathbf{P}^m$

$\underline{x}_m$ the minimum entry.

Suppose all entries of $\mathbf{P}^r$ are positive.

Let $\delta > 0$ be the smallest entry in $\mathbf{P}^r$. Our argument above shows that

$$\overline{x}_{m+r} \leq \overline{x}_m - \delta(\overline{x}_m - \underline{x}_m)$$

and

$$\underline{x}_{m+r} \geq \underline{x}_m + \delta(\overline{x}_m - \underline{x}_m)$$

Putting these together gives

$$(\overline{x}_{m+r} - \underline{x}_{m+r}) \leq (1 - 2\delta)(\overline{x}_m - \underline{x}_m)$$

In summary the column maximum decreases, the column minimum increases and the gap between the two decreases exponentially along the sequence $m, m+r, m+2r, \ldots$.

This idea can be used to prove

Proposition 2 *Suppose X_n finite state space Markov Chain with stationary transition matrix $\mathbf{P}$. Assume that there is a power r such that all entries in $\mathbf{P}^r$ are positive. Then for $\mathbf{P}^k$ has all entries positive for all $k \geq r$ and $\mathbf{P}^n$ converges, as $n \to \infty$ to a matrix $\mathbf{P}^\infty$. Moreover,*

$$(\mathbf{P}^\infty)_{ij} = \pi_j$$

where π is the unique row vector satisfying

$$\pi = \pi\mathbf{P}$$

whose entries sum to 1.

Proof: First for $k > r$

$$(\mathbf{P}^k)_{ij} = \sum_k (\mathbf{P}^{k-r})_{ik}(\mathbf{P}^r)_{kj}$$

For each i there is a k for which $(\mathbf{P}^{k-r})_{ik} > 0$ and since $(\mathbf{P}^r)_{kj} > 0$ we see $(\mathbf{P}^k)_{ij} > 0$.

The argument before the proposition shows that

$$\lim_{j\to\infty} \mathbf{P}^{m+jk}$$

exists for each m and $k \geq r$. This proves $\mathbf{P}^n$ has a limit which we call $\mathbf{P}^\infty$. Since $\mathbf{P}^{n-1}$ also converges to $\mathbf{P}^\infty$ we find

$$\mathbf{P}^\infty = \mathbf{P}^\infty \mathbf{P}$$

Hence each row of $\mathbf{P}^\infty$ is a solution of $x\mathbf{P} = x$. The argument before the statement of the proposition shows all rows of $\mathbf{P}^\infty$ are equal. Let π be this common row.

Now if α is any vector whose entries sum to 1 then $\alpha\mathbf{P}^n$ converges to

$$\alpha\mathbf{P}^\infty = \pi$$

If α is any solution of $x = x\mathbf{P}$ we have by induction $\alpha\mathbf{P}^n = \alpha$ so $\alpha\mathbf{P}^\infty = \alpha$ so $\alpha = \pi$. That is exactly one vector whose entries sum to 1 satisfies $x = x\mathbf{P}$. •

Note conditions:

There is an r for which all entries in $\mathbf{P}^r$ are positive.

The chain has a finite state space.

Consider finite state space case: $\mathbf{P}^n$ need not have limit. Example:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Note $\mathbf{P}^{2n}$ is the identity while $\mathbf{P}^{2n+1} = \mathbf{P}$. Note, too, that

$$\frac{\mathbf{P}^0 + \cdots + \mathbf{P}^n}{n+1} \to \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{array}\right]$$

Consider the equations $\pi = \pi\mathbf{P}$ with $\pi_1 + \pi_2 = 1$. We get

$$\pi_1 = \frac{1}{2}\pi_1 + \frac{1}{2}(1 - \pi_1) = \frac{1}{2}$$

so that the solution to $\pi = \pi\mathbf{P}$ is again unique.

Def'n: The period d of a state i is the greatest common divisor of

$$\{n : (\mathbf{P}^n)_{ii} > 0\}$$

Lemma 1 *If $i \leftrightarrow j$ then i and j have the same period.*

Def'n: A state is **aperiodic** if its period is 1.

Proof: I do the case $d = 1$. Fix i. Let

$$G = \{k : (\mathbf{P}^k)_{ii} > 0\}$$

If $k_1, k_2 \in G$ then $k_1 + k_2 \in G$.

This (and aperiodic) implies (number theory argument) that there is an r such that $k \geq r$ implies $k \in G$.

Now find m and n so that

$$(\mathbf{P}^m)_{ij} > 0 \text{ and } (\mathbf{P}^n)_{ji} > 0$$

For $k > r+m+n$ we see $(\mathbf{P}^k)_{jj} > 0$ so the gcd of the set of k such that $(\mathbf{P}^k)_{jj} > 0$ is 1. $\bullet$

The case of period $d > 1$ can be dealt with by considering $\mathbf{P}^d$.

$$\mathbf{P} = \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

For this example $\{1,2,3\}$ is a class of period 3 states and $\{4,5\}$ a class of period 2 states.

$$\mathbf{P} = \left[\begin{array}{ccc} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{array}\right]$$

has a single communicating class of period 2.

A chain is **aperiodic** if all its states are aperiodic.

Hitting Times

Start irreducible recurrent chain X_n in state i. Let T_j be first $n > 0$ such that $X_n = j$. Define

$$m_{ij} = \mathsf{E}(T_j | X_0 = i)$$

First step analysis:

$$\begin{aligned} m_{ij} &= 1 \cdot P(X_1 = j | X_0 = i) \\ &\qquad + \sum_{k \neq j} (1 + \mathsf{E}(T_j | X_0 = k)) P_{ik} \\ &= \sum_j P_{ij} + \sum_{k \neq j} P_{ik} m_{kj} \\ &= 1 + \sum_{k \neq j} P_{ik} m_{kj} \end{aligned}$$

Example

$$\mathbf{P} = \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

The equations are

$$\begin{aligned}
m_{11} &= 1 + \frac{2}{5} m_{21} \\
m_{12} &= 1 + \frac{3}{5} m_{12} \\
m_{21} &= 1 + \frac{4}{5} m_{21} \\
m_{22} &= 1 + \frac{1}{5} m_{12}
\end{aligned}$$

The second and third equations give immediately

$$\begin{aligned}
m_{12} &= \frac{5}{2} \\
m_{21} &= 5
\end{aligned}$$

Then plug in to the others to get

$$\begin{aligned}
m_{11} &= 3 \\
m_{22} &= \frac{3}{2}
\end{aligned}$$

Notice stationary initial distribution is

$$\left(\frac{1}{m_{11}}, \frac{1}{m_{22}}\right)$$

Consider fraction of time spent in state j:

$$\frac{1(X_0 = j) + \cdots + 1(X_n = j)}{n+1}$$

Imagine chain starts in chain i; take expected value.

$$\frac{\sum_{r=1}^{n} \mathbf{P}_{ij}^{r} + 1(i = j)}{n+1}$$

If rows of $\mathbf{P}$ converge to π then fraction converges to π_j; i.e. limiting fraction of time in state j is π_j.

Heuristic: start chain in i. Expect to return to i every m_{ii} time units. So are in state i about once every m_{ii} time units; i.e. limiting fraction of time in state i is $1/m_{ii}$.

Conclusion: for an irreducible recurrent finite state space Markov chain

$$\pi_i = \frac{1}{m_{ii}}.$$

Real proof: Renewal theorem or variant.

Idea: $S_1 < S_2 < ...$ are times of visits to i. Segment i:

$$X_{S_{i-1}+1}, \ldots, X_{S_i}.$$

Segments are iid by Strong Markov.

Number of visits to i by time S_k is exactly k.

Total elapsed time is $S_k = T_1 + \cdots + T_k$ where T_i are iid.

Fraction of time in state i by time S_k is

$$\frac{k}{S_k} \to \frac{1}{m_{ii}}$$

by SLLN. So if fraction converges to π_i must have

$$\pi_i = \frac{1}{m_{ii}}.$$

Summary of Theoretical Results:

For an irreducible aperiodic positive recurrent Markov Chain:

1. $\mathbf{P}^n$ converges to a stochastic matrix $\mathbf{P}^\infty$.

2. Each row of $\mathbf{P}^\infty$ is π the unique stationary initial distribution.

3. The stationary initial distribution is given by
$$\pi_i = 1/m_i$$
where m_i is the mean return time to state i from state i.

If the state space is finite an irreducible chain is positive recurrent.

Ergodic Theorem

Notice slight of hand: I showed

$$\frac{\mathsf{E}\left\{\sum_{i=0}^{n} 1(X_i = k)\right\}}{n} \to \pi_i$$

but claimed

$$\frac{\sum_{i=0}^{n} 1(X_i = k)}{n} \to \pi_i$$

almost surely which is also true. This is a step in the proof of the ergodic theorem. For an irreducible positive recurrent Markov chain and any f on S such that $\mathsf{E}^{\pi}(f(X_0)) < \infty$:

$$\frac{\sum_0^n f(X_i)}{n} \to \sum \pi_j f(j)$$

almost surely. The limit works in other senses, too. You also get

$$\frac{\sum_0^n f(X_i, \ldots, X_{i+k})}{n} \to \mathsf{E}^{\pi}\left\{f(X_0, \ldots, X_k)\right\}$$

E.g. fraction of transitions from i to j goes to

$$\pi_i \mathbf{P}^{ij}$$

For an irreducible positive recurrent chain of period d:

1. $\mathbf{P}^d$ has d communicating classes each of which forms an irreducible aperiodic positive recurrent chain.

2. $(\mathbf{P}^{n+1} + \cdots + \mathbf{P}^{n+d})/d$ has a limit $\mathbf{P}^\infty$.

3. Each row of $\mathbf{P}^\infty$ is π the unique stationary initial distribution.

4. Stationary initial distribution places probability $1/d$ on each of the communicating classes in 1.

For an irreducible null recurrent chain:

1. $\mathbf{P}^n$ converges to 0 (pointwise).

2. there is no stationary initial distribution.

For an irreducible transient chain:

1. $\mathbf{P}^n$ converges to 0 (pointwise).

2. there is no stationary initial distribution.

For a chain with more than 1 communicating class:

1. If $\mathcal{C}$ is a recurrent class the submatrix $\mathbf{P}_{\mathcal{C}}$ of $\mathbf{P}$ made by picking out rows i and columns j for which $i, j \in \mathcal{C}$ is a stochastic matrix. The corresponding entries in $\mathbf{P}^n$ are just $(\mathbf{P}_{\mathcal{C}})^n$ so you can apply the conclusions above.

2. For any transient or null recurrent class the corresponding columns in $\mathbf{P}^n$ converge to 0.

3. If there are multiple positive recurrent communicating classes then the stationary initial distribution is not unique.