Introduction

Sample Space

Data/observation: $\tilde{x}=(x_1,\ldots,x_n)'$.
Sample space: $S=\Omega=\{\tilde{x}\}$.
Functions: $X:\Omega\to\mathbb{R}$ rv (random variable).
Functions: $\vec{X}:\Omega\to\mathbb{R}^k$ rvec (random vector).

Stochastic Process, Probability Space

Suppose you observe continuously in time function $t\geq 0$. Then $\Omega=\{\tilde{x}(t)\}$.
Let $X_t(\vec{x}(t))=x_t$.
$\{x_t\}$: stochastic process.
$\{\Omega, \text{collection of events}, P\}$: probability space.
Sample space: $\Omega=\{\omega\}$.
Sample point: ω .
$X:\Omega\to\mathbb{R}$, so $X(\omega)$ is a number.
Events: Subsets of Ω ; A is a typical lable.

Expectation, Probability

We will assign expectation to rv's. Notation: $E(X)$.
Then we define $P(A)=E(I_A)$ where $I_A=\{\begin{matrix}1 \text{ if } A \text{ occurs}\\ 0 \text{ otherwise}\end{matrix}$.

Random Variables

For rv's, $X=X^+-X^-$.
Note: $|X|=X^++X^-$ (this avoids having $\infty-\infty$).
You can write $X^+=\dfrac{X+|X|}{2}$ and $X^-=\dfrac{X-|X|}{2}$.

Expectation

Axioms of E

1. $X\geq 0\Rightarrow E(X)\geq 0$.
2. $E(1)=1$; in general $E(\text{constant})=\text{constant}$.
3. $E(cX+dY)=cE(X)+dE(Y)$ if $E(X)$ and $E(Y)$ exist.
4. $X_n\uparrow X\Rightarrow E(X)=\lim_{n\to\infty} E(X_n)$ Called Monotone Convergence Theorem (MCT).

   Note: $X(\omega)$ is an increasing sequence of numbers ($X_1(\omega)\leq X_2(\omega)\leq\cdots$) and $X(\omega)=\lim_{n\to\infty} X_n(\omega)$ (meaning of $X_n\uparrow X$).
   Note: $\text{LHS}=\text{RHS}$ only if RHS exists.
   Note: $\omega=\text{lub}$.
5. $\left.\begin{matrix}X_n\geq 0\\ X_n\uparrow X\end{matrix}\right\}\Rightarrow E(X)=\lim_{n\to\infty} E(X_n)$.

6. If $X_1, \ldots, X_n \geq 0$, then $E\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} E(X_i)$.

Note: 1→4 is equivalent to 1→3 + 5 is equivalent to 1→3 + 6.

Consequences

1. $X \leq Y \Leftrightarrow E(X) \leq E(Y)$.
2. If $A_1 \Rightarrow A_2$ (occurrence of A_1 implies occurrence of A_2), then $P(A_1) \leq P(A_2)$.
3. $|E(X)| \leq E(|X|)$, $-E(|X|) \leq E(X) \leq E(|X|)$
4. Assume $E(X^+)$ and $E(X^-)$ aren't ∞. Then $E(X) = E(X^+) - E(X^+)$ and $E(|X|) = E(X^+) + E(X^+)$

Note: Often convenient to assume $E(|X|) < \infty$, and we will assume it as needed.

Theorem: Kolmogorov Axioms of P

1. $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. If A_1 and A_2 are disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.
4. If $A_1, A_2, \ldots$ have no overlap (ie disjoint), then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

   Or, equivalently, if $A_n \uparrow A_\infty$, then $\lim_{n \to \infty} P(A_n) = P\left(\lim_{n \to \infty} A_n\right) = A_\infty$

Markov's Inequality

$P(|X| \geq c) \leq \frac{E(|X|)}{c}$.

Definition

$X = Y$ in mean square if $E\left[(X-Y)^2\right] = 0$. We write $X \overset{\text{ms}}{=} Y$.

Proposition

If $X \overset{\text{ms}}{=} Y$, then $X = Y$ with probability 1 in the sense that $P(X = Y) = 1$. We will write $X \overset{\text{wp1}}{=} Y$ ("with probability 1") or $X \overset{\text{as}}{=} Y$ ("almost surely").

Note: $X \overset{\text{as}}{=} Y$ does not imply $X \overset{\text{ms}}{=} Y$ as the second moment may not exist (they may be infinite). If $E(X^2)$ and $E(Y^2)$ and $X \overset{\text{as}}{=} Y$, then $X \overset{\text{ms}}{=} Y$.

Proposition

Let $Y \geq 0$ and $E(Y) = 0$. then $Y \overset{\text{wp1}}{=} 0$.

Jensen's Inequality

If *g* is convex (ie for any point x_0 there exists a constant *c* such that $g(x_0) + c(x - x_0) \leq g(x)$), then $E(g(X)) \geq g(E(X))$.

Moments of Random Variables

Definition

$E(X)$, $E(X^2)$,.. are the moments of *X*.

Variance and Standard Deviation

Variance: $\sigma^2 = \text{var}(X) = E\left[(X-\mu)^2\right] = E(X^2) - (E(X))^2$, where $\mu = E(X)$.

Standard Deviation: $\sigma = \text{SD}(x) = \sqrt{\text{var}(X)}$.

Lemma

1. $E(aX+b) = aE(X)+b$.
2. $\text{var}(aX+b) = a^2 \text{var}(X)$.

Note: If X is a rv, then $\dfrac{X-\mu}{\sigma}$ has mean 0 and variance 1.

Moments of Random Vectors

Definitions

Random vector: $\underbrace{\vec{X}}_{n\times 1} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$.

Expectation: $E(\vec{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix} = \vec{\mu}$.

Variance: $\Sigma = \text{var}(\vec{X}) = E\left[(\vec{X}-\vec{\mu})(\vec{X}-\vec{\mu})^T\right] = \{E[(X_i-\mu_i)(X_j-\mu_j)]\}_{i,j=0}^{n} = \{\text{cov}(X_i, X_j)\}_{i,j=0}^{n}$. Notice the diagonal elements of Σ are the variances of the components.

Lemma

Let A denote a constant matrix.

1. $E(A\vec{X}+\vec{b}) = A\,E(\vec{X})+\vec{b}$.
2. $\text{var}(A\vec{X}+\vec{b}) = A\,\text{var}(\vec{X})A^T$.

Note: If $A = \vec{1}^T$, then $\text{var}(\vec{1}^T\vec{X}) = \text{var}(X_1+\cdots+X_n) = \vec{1}^T\Sigma\,\vec{1} = \sum_{i,j}\text{cov}(X_i, X_j)$.

Note: If $\Sigma = \text{var}(\vec{X})$, then $\Sigma = \Sigma^T$. Also, $\vec{c}^T\Sigma\,\vec{c} = \text{var}(\vec{c}^T\vec{X}) \geq 0$, thus Σ is positive semi-definite (write $\Sigma \geq 0$). If for any $\vec{c} \neq \vec{0}$, $\text{var}(\vec{c}^T\vec{X}) > 0$, then Σ is positive definite (write $\Sigma > 0$) and Σ^{-1} exists.

Facts

1. $\Sigma > 0$ iff Σ^{-1} exists.
2. $\Sigma = Q^T D Q$ where $Q^T Q = I = QQ^T$ (so Q is orthogonal matrix) and D is a diagonal matrix with non-negative diagonal elements (which are the eigenvalues of Σ)..
3. Σ^{-1} exists iff $\det(\Sigma) > 0$. Note $\det(\Sigma) = \det(Q^T D Q) = \det(DQQ^T) = \det(D) = \prod$ diagonal elements .

Linear Prediction

Let $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$. Want to predict Y using $\vec{X}$. We assume 0 means and only use linear predictor $\hat{Y} = \vec{a}^T\vec{X}$.

Choose $\vec{a}$ to minimize the mean square error $D(\vec{a}) = E\left[(Y-\hat{Y})^2\right] = \Sigma_{YY} + \vec{a}^T\Sigma_{\vec{X}\vec{X}}\vec{a} - 2\vec{a}^T\Sigma_{Y\vec{X}}$. Since

$\frac{dD}{d\vec{a}} = 2\Sigma_{\vec{X}\vec{X}}\vec{a} - 2\Sigma_{Y\vec{X}} = \vec{0} \Rightarrow \Sigma_{\vec{X}\vec{X}}\vec{a} = \Sigma_{Y\vec{X}}$ (*).

If $\Sigma_{\vec{X}\vec{X}}^{-1}$ exists, then (*) yields $\vec{a} = \Sigma_{\vec{X}\vec{X}}^{-1}\Sigma_{Y\vec{X}}$. More generally, if $\vec{a}$ solves (*), then it minimizes $D(\vec{a})$, and $\hat{Y} = \vec{a}^T\vec{X}$ is the best linear least squares predictor.

Distributions

Let $\Omega = S = \{\omega\}$, $X : S \to \mathbb{R}$, $\vec{X} : S \to \mathbb{R}^n$.

Definitions: Continuous and Discrete Random Variables

1. If $E[h(X)] = \int h(x) f(x) dx, '\forall' h$, then X is an (absolutely) continuous random variable and f is the probability density function.
2. If $E[h(X)] = \sum h(x) f(x) dx, '\forall' h$, then X is an discrete rv and f is the probability function ($f(x) = P(X = x)$).

Similar definitions apply to random vectors.

Definition: Independence

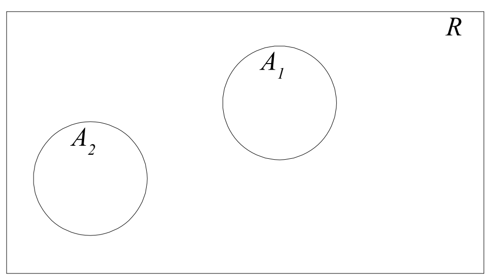
$X_1, X_2, \dots$ are independent if $E[h_1(X)h_2(X)\cdots] = E[h_1(X)]E[h_2(X)]\cdots$.

Definition: Independent Events

A_1 and A_2 are independent events if I_{A_1} and I_{A_2} are independent.

Note: A_1 and A_2 independent implies $P(A_1 A_2) = P(A_1)P(A_2)$.

Spatial Poisson Processes



- Throw N particles onto $R \subset \mathbb{R}^{(.)}$ independently and in a uniform way.
- Let $N(A_i)$ be the number of points in A_i. $N(A_i) \sim \text{multinomial}(N, \vec{P})$.
- The density of points is $\frac{N}{|R|} = \rho$.
- Let $R \to \mathbb{R}^{(.)}$ and $N \to \infty$ such that ρ is fixed.
- Then $N(A_i) \xrightarrow{d} \text{Poisson}(\rho|A_i|)$. Also, $N(A_1), N(A_2), \dots$ become independent in the limit.

This yields a Poisson Point Process on $\mathbb{R}^{(.)}$

Probability Generating Functions

Definitions: Probability Generating Function

1. $G(s) = E[s^X]$, where $X = 1, 2, \dots$ is a counting rv.
2. $G(\vec{s}) = E[s_1^{X_1} \cdots s_n^{X_n}]$, where $X_i = 1, 2, \dots$ is a counting rv.

Note:
1. If X and Y are independent, then $G_{X+Y}(s)=G_X(s)G_Y(s)$. This extends to sums of finite number of rv's.
2. $G_X(s)=E(s^X)=\sum P(X=k)s^k=P(X=0)+P(X=1)s+\cdots$ a power series.
3. If $|s|<1$, then $|G_X(s)|=|E(s^X)|\le E(|s^X|)=E(|s|^X)<\infty$. So the radius of convergence is at least 1.
4. $\frac{d^k}{ds^k}G(s)=E[X(X-1)(X-2)\cdots(X-(k-1))s^{X-k}]\Rightarrow G^{(k)}(1)=E[X(X-1)(X-2)\cdots(X-(k-1))]$.
5. $P(X=k)=\frac{G_X^{(k)}(0)}{k!}$.

Conditioning

Let Y be a random variable.
Let A be an event with $E(I_A)=P(A)>0$.

Definition: Conditional Expectation

$E(Y|A)=\frac{E(X I_A)}{E(I_A)}$ is the conditional expectation of Y given the event A .

Proposition

Let $A_{1,}\ldots,A_n$ partition Ω . Then $E(Y)=\sum_i E(Y|A_i)P(A_i)$ (*).
Note: This is true for $n=\infty$.

Definition: Conditional Probability

$P(B|A)=E(I_B|A)=\frac{P(AB)}{P(A)}$.
Note: By setting $Y=I_B$, (*) is just $P(B)=\sum_i P(B|A_i)P(A_i)$.

Special Case: Regression

If X and Y are rv's, we want to talk about $E(Y|X=x)$.
If X is discrete, then it would just be $E(Y|X=x)=\frac{E(Y I_{\{X=x\}})}{P(X=x)}$.
If Y is also discrete, then $E(Y|X=x)=\sum_y y f(y|x)$, where $f(y|x)=\frac{f(x,y)}{f(x)}$. If Y is continuous, then $E(Y|X=x)=\int_y y f(y|x)$, where $f(y|x)=\frac{f(x,y)}{f(x)}$.

Now, $r(x)=E(Y|X=x)$ is a possible value of $r(X)$, which is called the conditional expectation of $Y|X$ and is denoted by $E(Y|X)$.

Proposition

$E[(Y-E(Y|X))h(X)]=0$, or equivalently $E[Y]=E[E(Y|X)h(X)]$.
Note: This means $E[(Y-E(Y|X))]=0\Leftrightarrow E[Y]=E[E(Y|X)]$.

Propositiion

$\text{var}(Y) = E[\text{var}(Y|X) + \text{var}(E(Y|X))]$.

Example: Matching Problem

If you have n letters and n envelopes, what is the probability of j matches if they are matched at random?

Let $A_{n,j} = \{\text{j matches}\}$. Then $P(A_{n,j}) = \frac{1}{j!}\sum_{k=j}^{n}\frac{(-1)^{k-j}}{(k-j)!}$.

Let the number of matches be $X = I_1 + \cdots + I_n$, where $I_i = 1$ if the *i*-th letter matches and is $\text{Bernoulli}\left(\frac{1}{n}\right)$. Then $E(X) = E\left(\sum X_i\right) = 1$, $\text{cov}(I_i, I_j) = \left(\frac{1}{n-1}\right)\left(\frac{1}{n}\right) - \left(\frac{1}{n}\right)^2, i \neq j$, and $\text{var}(X) = \sum \text{var}(I_i) + \sum_{i \neq j}\text{cov}(I_i, I_j) = 1$.

Example

Consider a graph with nodes $\{1, \ldots, n\}$ and edges $\{1, \ldots, M\}$. Let B be a set of nodes and $c(B) = \text{number of edges with exactly one node in } B$. Then $\max_B c(B) \geq \frac{M}{2}$.

Stochastic Process

Definition: Stochastic Process

A collection of random variables $\{X_t : t \in T\}$ is called a stochastic process.

Special Cases

- $T = \{t : t \geq 0\}$ continuous time process.
- $T = \{0, 1, 2, \ldots\}$ discrete time process.

Definition: Process With Independent Increment

Suppose for every $t_0 < t_1 < t_2 < \cdots$, the increments $\Delta X_{t_i} = X_{t_i} - X_{t_{i-1}}$ are independent. Then $\{X_t : t \in T\}$ is called a process with independent increment.

Definition: Poisson Counting Process

Consider a Poisson process of rate λ on $t \geq 0$. If $N(t)$ denotes the number of points in $[0, t]$, then $\{N(t) : t \geq 0\}$ is called a Poisson (counting) process of rate λ on $t \geq 0$.

Definition: Poisson Process

$\{N(t) : t \geq 0\}$ is a Poisson process of rate λ if the number of points in non-overlapping sets are independent and the number of points in a set A is distributed $\text{Poisson}(\lambda|A|)$.

Another Definition: Poisson Process

- No points on origin.
- Independent increments.
- $P(1 \text{ point in } (t, t+h)) = \lambda^h + o(h)$, where $\lim_{h \to 0}\frac{o(h)}{h} = 0$.
- $P(> 1 \text{ point in } (t, t+h)) = o(h)$.

Example

Take a Poisson process of rate λ on $t \geq 0$. Let $T_r = '$time to the r^{th} point. Then $T_r \sim \text{gamma}(r, \lambda)$ with pdf $f_{T_r}(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}$.

1. Let X_1 denote the time to the first point. Then $X_1 \sim \text{exponential}(\lambda)$ since $P(X_1 > x) = P(N(x) = 0) = e^{-\lambda x}$. If the joint pdf is $f(x_1, x_2)$, then $f(x_1, x_2) dx_1 dx_2 \approx P(X_1 \in (x_1, x_1 + dx_1), X_2 \in (x_2, x_2 + dx_2)) \approx e^{-\lambda x_1} \lambda dx_1 e^{-\lambda x_2} \lambda dx_2 \approx f(x_1) dx_1 f(x_2) dx_2$, which "implies" X_1 and X_2 are iid exponential(λ); in fact, $X_1, X_2, \ldots$ are iid exponential(λ). So $T_r = \sum_{i=1}^{r} X_i$ gamma(r, λ).
2. Since, $P(N(x) < r) = P(T_r > x) = 1 - F_{T_r}(x)$, so $f_{T_r}(x)$ is the gamma pdf by differentiation.

Definition: Compound Poisson Process

Let $\{N(t) : t \geq 0\}$ be a Poisson Process of rate λ. Let $X_1, X_2, \ldots$ be iid and independent of $\{N(t) : t \geq 0\}$. Let $X_0 = 0$. Then $Y(t) = \sum_{i=1}^{N(t)} X_i, t \geq 0$ is a compound Poisson process.

Order Statistics

Definition: Order Statistics

Let $X_1, \ldots, X_n$ be iid with pdf $f(x)$ and df $F(x)$. Then $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics.

Probability Density Function

1. The df of $X_{(n)}$ is $F_{(n)}(x) = P(X_{(n)} \leq x) = P(X_{(1)} \leq x, \ldots, X_{(n)} \leq x) = P(X_{(1)} \leq x) \cdots P(X_{(n)} \leq x) = [F(x)]^n$, and thus $f_{(n)}(x) = F_{(n)}'(x) = n[F(x)]^{n-1} f(x)$.
2. For $i < j$, the joint pdf of $X_{(i)}$ and $X_{(j)}$ is $f_{(i)(j)}(a, b) = \frac{n!}{(i-1)!(j-1)!(n-j)!} (F(a))^{i-1} (F(b) - F(a))^{j-i-1} (1 - F(b))^{n-j} f(a) f(b)$ for $a < b$, and $f_{(i)(j)}(a, b) = 0$ 0 otherwise.
3. In the uniform case, the joint pdf of $X_{(1)}, \ldots, X_{(n)}$ is $f(x_{(1)}, \ldots, x_{(n)}) = n! f(x_{(1)}) \cdots f(x_{(n)})$ for $x_{(1)} < \cdots < x_{(n)}$, and $f(x_{(1)}, \ldots, x_{(n)}) = 0$ otherwise.

The Multinomial

Let items be classified in one of k ways. Consider $\underset{k \times 1}{\vec{Z}}$, a multivariate Bernoulli. Let $\vec{Y} = \vec{Z}_1 + \cdots + \vec{Z}_n$ ($\vec{Z}_i$'s iid), then it is a multinomial.

Probability Generating Functions

The pgf of $\vec{Z}$ is $G(\vec{s}) = E(s_1^{Z_1} \cdots s_k^{Z_k}) = p_1 s_1 + \cdots p_k s_k$. Note that $G(s_1, 1, \ldots, 1) = E(s_1^{Z_1}) = p_1 s_1 + p_2 + \cdots p_k = q_1 + p_1 s_1$, which is the Bernoulli pgf.

The pgf of $\vec{Y}$ is $G_{\vec{Y}}(\vec{s}) = [G_{\vec{Z}}(\vec{s})]^n = (p_1 s_1 + \cdots p_k s_k)^n$. Expand it to get $P(\vec{Y} = \vec{y}) = \binom{n}{y_1, \ldots, y_k} p_1^{y_1} \cdots p_k^{y_k}$ where $y_1, \ldots, y_k \geq 0$ and $y_1 + \cdots + y_k = n$.

Conditional Expectation as a Random Variable

Background

- If $E(Y^2)=0$, then $Y \overset{\text{ms}}{=} 0$.
- $Y \overset{\text{ms}}{=} X$ means $E\left[(Y-X)^2\right]=0$.

Definition

$E(Y|\vec{X})$ is that function of $\vec{X}$ which satisfies $E\left[\left(Y-E(Y|\vec{X})\right)h(\vec{X})\right]=0, \forall \text{ real } h$, or
$E[Y\,h(\vec{X})]=E[E(Y|\vec{X})h(\vec{X})], \forall \text{ real } h$ (*).
Equivalently, if $E(Y^2)<\infty$, $E(Y|\vec{X})$ is that function of $\vec{X}$ which minimizes $E[Y-E(Y|\vec{X})]$ (**).

Proposition

A solution to (**) is unique and satisfies (*), and vice versa.

Remarks

1. We still have to show that (*) and (**) has a solution.
2. For (*), there exists such a function and it is unique with probability 1.
3. For (**), there is such a function and it is unique up to mean square equivalence.
4. From (*), if $h(\vec{X})=1$, then we see $E[Y]=E[E(Y|\vec{X})]$.
5. For fixed $\vec{X}$, $E(Y|\vec{X})$ satisfies the axioms of E.
6. $E(E(Y|X_{1,}X_2)|X_1)=E(Y|X_1)$, i.e. if $E(Y|X_{1,}X_2)$ is a function of X_1 , then it equals $E(Y|X_1)$.
7. $\text{cov}(\ ,\ )$ is an inner product. So with mean 0, $\|Y\|=\sqrt{E(Y^2)}$, and the L_2 space results.

Independence

X and Y are independent if $E(g(Y)h(X))=E(g(Y))E(h(X)) \forall g, h$. This is the same as saying
$E(g(Y)|X)=E(g(Y)) \forall g$ (*).

Example

Let $X_{1,}\ldots,X_n$ be iid $N(\mu,\sigma^2)$. Then $\bar{X}=\dfrac{X_1+\cdots+X_n}{n}$ and $S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2$ are independent.

Limits and Convergence

Complex Numbers

Definitions

1. Complex number: $z=a+ib$ where $a, b \in \mathbb{R}$ and $i^2=-1$.
2. Conjugate: $\overline{a+ib}=a-ib$.
3. Square-length: $|z|^2=a^2+b^2$. Notice $|z|^2=z\,\bar{z}$.

Definition

$e^{it}=\cos t+i\sin t$

Note:

1. $e^{z} = e^{a+ib} = e^{a} e^{ib}$.
2. $\overline{e^{it}} = \cos t - i \sin t = \cos(-t) + i \sin(-t) = e^{-it}$.
3. $|e^{it}|^2 = \cos^2 t + \sin^2 t = 1 \Rightarrow |e^{it}| = 1$.

Results

1. $|z_1 + z_2| \le |z_1| + |z_2|$.
2. $|z_1 z_2| = |z_1||z_2|$.
3. $|e^{it_2 - it_1}| \le |t_2 - t_1|$ (since $|e^{it_2 - it_1}| = \left| \int_{t_1}^{t_2} \frac{d(e^{it_2 - it_1})}{dt} dt \right| \le \left| \int_{t_1}^{t_2} \left| \frac{d(e^{it_2 - it_1})}{dt} \right| dt \right| = \left| \int_{t_1}^{t_2} |i\, e^{it}| dt \right| = \left| \int_{t_1}^{t_2} dt \right| = |t_2 - t_1|$).
4. $|e^{it_2 - it_1}| \le 2$.
5. $|\cos t_2 - \cos t_1| \le |t_2 - t_1|$.
6. $|\sin t_2 - \sin t_1| \le |t_2 - t_1|$.

Characteristic Functions

Definition: Characteristic Function

If X is a rv, its characteristic function is $c(t) = E(e^{itX})$.

Proposition

1. $c(0) = 1$, $|c(t)| \le 1$.
2. $\bar{c}(t) = c(-t)$.
3. $c_{aX+b} = e^{itb} c_X(at)$.

Inversion Theorem

1. If X has pdf $f(x)$, then $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-tx} c(t)\, dt$. Note $c(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$.
2. $F(y) - F(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ity} - e^{-itx}}{-it} dt$, where $F(y)$ and $F(x)$ are continuity points of F .

Notes

1. $c(t)$ determines the distribution.
2. If $\vec{X}$ is a rvec, $c(\vec{t}) = E\left(e^{i \vec{t}^T \vec{X}}\right)$.
3. From $c(\vec{t})$, we can get $F(\vec{x}) = P(\vec{X} \le \vec{x})$.

Facts

1. $F(\vec{x}) = F(x_1) \cdots F(x_n)$ if and only if $x_1, \ldots, x_n$ are independent.
2. $c(\vec{t}) = c(t_1) \cdots c(t_n)$ implies $F(\vec{x}) = F(x_1) \cdots F(x_n)$.

Proposition

If X and Y independent, then $c_{X+Y}(t) = c_X(t) c_Y(t)$.

Proof: $c_{X+Y}(t) = E\left(e^{it(X+Y)}\right) = E\left(e^{itX} e^{itY}\right) = E\left(e^{itX}\right) E\left(e^{itY}\right) = c_X(t) c_Y(t)$.

Note: $c_{X+Y}(t) = c_X(t) c_Y(t)$ does not imply independence.

Facts

1. If $m(\vec{t})$ (mgf of $\vec{X}$) exists in a neighbourhood of 0, then $c(\vec{t})=m(i\vec{t})$.
2. A version of Taylor's Theorem: If $g^{(n)}(0)$ exists, then $g(x)=\left(\sum_{j=0}^{n}\frac{g^{(j)}}{j!}x^j\right)+o(x^n)$ as $x\to 0$; here $o(x^n)$ is the remainder and has the property that $\lim_{x\to 0}\frac{o(x^n)}{n}=0$. Notice $1+x+o(x)=e^{x+o(x)}$ and $1+x+\frac{x^2}{2}+o(x)=e^{x+\frac{x^2}{2}+o(x)}$.

Notes

1. $e^z=\sum_{j=0}^{\infty}\frac{z^j}{j!}$.
2. Formally, $e^{itX}=\sum_{i=1}^{\infty}\frac{(it)^j}{j!}X^j$, and "so" $E(e^{itX})=\sum_{i=1}^{\infty}\frac{(it)^j}{j!}E(X^j)$, and "so" $c^{(j)}(0)=i^j E(X^j)$.
3. We "know" $m^{(j)}(0)=E(X^j)$.

Theorem

Let X be finite wp1. Then $c(t)$ is uniformly continuous. If $E(|X|^k)<\infty$, then $E(X^k e^{itX})$ is uniformly continuous (in t).

Consequences

1. If $E(|X|^k)<\infty$, then $c^{(k)}(0)$ exists. Note however that if $c^{(k)}(0)$ exists, $E(|X|^k)<\infty$ only if k is even.
2. $E(|X|^k)<\infty\Rightarrow c(t)=\sum_{i=1}^{k}\frac{(it)^j}{j!}E(X^j)+0(t^k)\Rightarrow c^{(k)}(0)=i^k E(X^k)$.

Convergence

Cauchy Property

If $a_n-a_m\to 0$ as $n,m\to\infty$, there exists an a such that $a_n\to a$.

Definitions: Types of Convergence

1. Mean Square: $X_n\overset{\text{ms}}{\to}X$ if $E\left[(X_n-X)^n\right]\to 0$ as $n\to\infty$.
2. Probability: $X_n\overset{\text{p}}{\to}X$ if for all $\varepsilon>0$ $P(|X_n-X|<\varepsilon)\to 1$ as $n\to\infty$.
3. Pointwise: $X_n\to X$ if $\lim_{n\to\infty}X_n(\omega)=X(\omega),\forall\omega\in\Omega$.
4. With Probability 1 or Almost Surely: $X_n\overset{\text{wp1}}{\to}X$ or $X_n\overset{\text{as}}{\to}X$ if $P\left(\lim_{n\to\infty}X_n=X\right)=1$. This is almost like conventional convergence, except for the set of ω s it may fall on has probability 0.

Fact: All these have the Cauchy Property in that $X_n\overset{\text{S}}{\to}X\Leftrightarrow X_m-X_n\overset{\text{S}}{\to}0$ $\forall m,n\to\infty$; the latter is called mutual convergence.

Theorem: Monotone Convergence Theorem (MCT)

If $0\le X_n$ and $X_n\uparrow X$ wp1, then $E(X_n)\to E(X)$.
Note: $0\le X_n$ can be replaced by $E(|X_1|)<\infty$.

Theorem: Dominate Convergence Theorem (DCT)

If $X_n\overset{\text{as}}{\to}X$ and $|X_n|<Y$ with $E(Y)<\infty$, then $E(X_n)\to E(X)$.

Definition: Convergence in Distribution (Weak Convergence)
$X_n \xrightarrow{d} X$ (X_n converges to X weakly) if for all bounded continuous h we have $E(h(X_n)) \to E(h(X))$ (*).

Separating Class of *h*
The separation class of h s is the subset of all bounded continuous functions such that if (*) holds for them, then it holds for all bounded continuous functions.
One is $\{\sin tx, \cos tx \mid \forall t\} = \{e^{itX} \mid \forall t\}$, i.e. if $c_n(t) \to c(t)\ \forall t$, then $X_n \xrightarrow{d} X$.

Weak Law of Large Numbers (WLLN)

Let $X_{1,} \ldots, X_n$ be iid with mean μ . Then $\bar{X} \xrightarrow{d} \mu$.
Note: $Y_n \xrightarrow{d} c \Rightarrow Y_n \xrightarrow{p} c$.

Proof:
Let $c(t)$ be the cf of X_1 . We know $c(t) = 1 + i\mu t + o(t) = e^{i\mu t + o(t)}$. So

$$\begin{aligned} E\left(e^{it\bar{X}}\right) &= E\left(e^{i\frac{t}{n}(X_1 + \cdots + X_n)}\right) \\ &= E\left(e^{i\frac{t}{n}X_1}\right) \cdots E\left(e^{i\frac{t}{n}X_n}\right) \\ &= \left[c\left(\frac{t}{n}\right)\right]^n \\ &= \left[e^{i\mu\frac{t}{n} + o\left(\frac{t}{n}\right)}\right]^n \\ &= e^{i\mu t + n o\left(\frac{t}{n}\right)} \\ &= e^{i\mu t + \frac{o(t/n)}{1/n}} \\ &\to e^{i\mu t} \end{aligned}$$

which is the cf of the constant μ .

Central Limit Theorem

Let $X_{1,} X_{2,} \ldots$ be iid with mean μ and variance σ^2 . Then $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0,1)$.

Proof:
The cf of $X_1 - \mu$ is $E\left(e^{it(X_1 - \mu)}\right) = 1 - \dfrac{\sigma^2}{2} t^2 + o(t^2) = e^{\frac{-\sigma^2}{2} t^2 + o(t^2)}$. So the cf of $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ is

$$\begin{aligned} E\left(e^{it\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}\right) &= E\left(e^{\frac{it}{\sigma\sqrt{n}}[(X_1 - \mu) + \cdots + (X_n - \mu)]}\right) \\ &= \left[E\left(e^{\frac{it}{\sigma\sqrt{n}}(X_1 - \mu)}\right)\right]^n \\ &= \left[e^{\frac{-\sigma^2 t^2}{2\sigma^2 n} + o\left(\frac{t^2}{\sigma^2 n}\right)}\right]^n \\ &= e^{\frac{-t^2}{2} + n o\left(\frac{t^2}{\sigma^2 n}\right)} \\ &\to e^{\frac{-t^2}{2}} \end{aligned}$$

which is the cf of a $N(0,1)$.

Results In Convergence

$$\left.\begin{array}{c}\xrightarrow{\text{as}}\\ \xrightarrow{\text{ms}}\end{array}\right\}\Rightarrow\xrightarrow{\text{p}}\Rightarrow\xrightarrow{\text{d}}$$

Proposition

$X_n\xrightarrow{\text{d}}c\Rightarrow X_n\xrightarrow{\text{p}}c$.

Little Results

For any continuous function g:

1. $X_n\xrightarrow{\text{d}}X\Rightarrow g(X_n)\xrightarrow{\text{d}}g(X)$.
2. $X_n\xrightarrow{\text{p}}X\Rightarrow g(X_n)\xrightarrow{\text{p}}g(X)$.

Results are also true in the vector case.

Proposition

If for all $\varepsilon>0$, $\sum P(|Y_n|>\varepsilon)<\infty$, then $Y_n\xrightarrow{\text{as}}0$.

Corollaries

1. Let $\varepsilon_n\downarrow 0$. If $\sum P(|Y_n|>\varepsilon_n)<\infty$, then $Y_n\xrightarrow{\text{as}}0$.
2. If $X_n\xrightarrow{\text{p}}X$, then there exists a subsequence $\{X_{n_1},X_{n_2},\ldots\}$ such that $X_{n_k}\xrightarrow{\text{p}}X$.
3. $X_n\xrightarrow{\text{p}}X$ if and only if every subsequence has a further subsequence converging wp1 to X.

Proposition

Let $\varepsilon_n>0$. If $\sum\varepsilon_n<\infty$ and $\sum P(|X_{n+1}-X_n|>\varepsilon_n)<\infty$, then $X_n\xrightarrow{\text{as}}X$.

Application

Result

Let $R_i, i=0,1,\ldots$ be counting rv's with $E(R_i)<M$. Then $\sum_{i=0}^{\infty}R_i s^i$ converges wp1 on $|s|<1$.

Lemma

Let $A_1, A_2,\ldots$ be events with $P(A_i)=1$ $\forall i$. Then $P(\cap A_i)=1$.

Definition: Stopping Time

Let $X_0, X_1,\ldots$ be rv's and $N\in\{0,1,2,\ldots\}$ a counting rv. If the event $\{N=n\}$ (or $\{N\le n\}$) depends only on $X_0, X_1,\ldots,X_n$, then we call N a stopping time for the sequence.

Wald's Equation

Let $X_{0,} X_{1,} X_{2,} \dots$ be independent with $X_0=0$ and $X_{1,} X_{2,} \dots$ iid. Let N be a stopping time and $\mu=E(X_1)$. Then $E\left(\sum_{n=0}^{N} X_n\right)=\mu E(N)$.

Renewal Process

Discrete Time Renewal Process

Let $X_{1,} X_{2,} \dots$ be iid with pgf $G(s)$ and let X_0-0 be renewal points. Assume $P(X_i=0)<1$

Let $S_k=X_0+X_1+\cdots+X_k$ be the time to the k -th renewal.

Let R_t be the number of renewals at time t (discrete, $R_0=1$ by convention) and let $u_t=E(R_t)$ ($u_0=1$).

Problem

From $G(s)$ we want the u_t 's. Set $U(s)=\sum_{t=0}^{\infty} u_t s^t$.

Theorem

1. $\sum_{t=0}^{\infty} R_t s^t=\sum_{r=0}^{\infty} s^{S_r}$.
2. $U(s)=\frac{1}{1-G(s)}$.

Note $G(s)=\frac{U(s)-1}{U(s)}$, so knowing the renewal probabilities yields that distribution of the inter-arrival times.

Definition: Regeneration Point

Take discrete stochastic process $\{X_t\}$. Suppose that whenever $\{X_t \in A\}$ the event a happens, the process regenerates in the sense that given a the present and future are independent of the past. Then a is a called a regeneration point.

In equations this is $E[g(\{X_s; s \ge t\}) | a, \{X_s; s<t\}]=E[g(\{X_s; s\ge t\})|a]$ where g is a function of the X_i 's.

Theorem: Elementary Renewal Theorem

Let $N(t)$ be the number of renewals up to t (not counting the initial item). Set $m(t)=E(N(t))$. This is the renewal function. Call $\mu \overset{\text{def}}{=} E(X_i)>0$ (since $P(X_i=0)<1$). Then $\frac{m(t)}{t} \to \frac{1}{\mu}$.

Proposition

$N(t)+1$ is a stopping time for $X_1, X_{2,} \dots$.

Corollary

$E(S_{N(t)+1})=E(X)E(N(t)+1)=\mu E(N(t)+1)$. This yields $E(N(t)+1)=\frac{E(S_{N(t)+1})}{\mu} \approx \frac{t}{\mu}$.

Theorem: Renewal Theorem

For the renewal process, $E(N(t)+1)=E\left(\sum_{k=0}^{t} R_k\right)=\sum_{k=0}^{t} u_t \Rightarrow \frac{1}{t}\sum_{k=0}^{t} u_t \to \frac{1}{\mu}$, which almost yields the stronger result $u_t \to \frac{1}{\mu}$.

Markov Process

When renewals happen the process begins over. This notional generalized to the at of a regeneration point for a stochastic process $\{X_t\}$. In particular, if for each t_0 , X_{t_0} is a regeneration point, then we have a Markov Process. This is defined in many different (equivalent) forms. We assume t to be "true".

Definition: Markov Process

$\{X_t\}$ is Markov if for each t , functions of the present and future $\{X_s, s\geq t\}$ are independent of functions of the past $\{X_s, s<t\}$ given the present X_t .

Definition: Markov Process

$\{X_t\}$ is Markov if for each "for all" functions h , $E\left[h(\{X_s, s\geq t\})|\{X_s, s<t\}\right]=E\left[h(\{X_s, s\geq t\})\right]$.

Definition: Markov Process

$\{X_t\}$ is Markov if $E\left[h\left(X_{t_{n+1}}\right)|X_{t_n}, X_{t_{n-1}}, \ldots, X_{t_1}\right]=E\left[h\left(X_{t_{n+1}}\right)|X_{t_n}\right]$ $\forall t_1<t_2<\cdots|X_t$.

Definition: Time Homogeneous

If the conditional structure does not change over time, then the process is said to be time homogeneous. There the conditional distribution of $X_{t+s}|X_t$ does not change with t .
Note: We will assume this is the case.
Note: This assumption of time homogeneity of the conditional distribution is different than that of stationary which one finds frequently in time series. There $\{X_t\}$ is stationary if the joint distribution of $X_{t_1+t}, \ldots, X_{t_n+t}$ do not change with t .

Definition: State Space

Let $\{X_t\}$ be Markov and time homogeneous. The state space is $\cup\{\text{all possible values of } X_t\}$.
Note: We mainly consider countable state spaces and talk of $X_t=i$ (the process is in state i at time t).

Note

Let $N(t)$ be the number of renewals up to t (not counting the initial item). The process $\{N(t), t\geq 0\}$ is a renewal counting process.
In continuous time if the inter-arrivals were iid $\text{exponential}(\lambda)$ then we would have a Poisson process of rate λ , and it would have the Markov property.
In discrete dime the Markov property would hold if the inter-arrivals were geometric.

Ageless Property (Lack of Memory)

$X\sim\text{exponential}(\lambda)$ satisfies $P(X>s+t|X>s)=P(X>t)$ $\forall s, t>0$, and it is the only continuous positive rv with this property.
$X\sim\text{geometric}(p)$ satisfies $P(X>s+t|X>s)=P(X>t)$ $\forall s, t>0; s, t\in\mathbb{Z}$, and it is the only counting rv with this property.

Definitions: Markov Chain States

1. States of which recurrence is uncertain is called transient.
2. States with finite recurrence times is called recurrent
3. Recurrent states with finite recurrence times are called positive recurrent. Otherwise they are called null recurrent.
4. States that only recur at multiples of $d>1$ (an integer) are periodic. Otherwise they are aperiodic.

Poisson Process

Definition: Nonhomogeneous Poisson Process

Consider $\{N_t, t\geq 0\}$ where $N_0=0$ and N_t be the number of points in $[0,t]$. Let $\lambda(t)\geq 0$ and set $m(t)=\int_0^t \lambda(u)\,du$.
$\{N_t, t\geq 0\}$ is a nonhomogeneous Poisson (counting) process of rate $\lambda(t)$ if it has independent increments and $N_t \sim \text{Poisson}(m(t))$.

Definition: Homogeneous Poisson Process

If $\lambda(t)=\lambda$ then the process is homogeneous and simply termed a Poisson Process.

Notes

1. A Non homogeneous Poisson process is a Markov Chain. If $\lambda(t)=\lambda$ then it is also time homogeneous.
2. $N((t_1, t_2))$ is the number of points in (t_1, t_2) and is distributed $\text{Poisson}\left(\int_{t_1}^{t_2} \lambda(u)\,du\right)=\text{Poisson}(m(t_2)-m(t_1))$.
3. If $m(t)$ is strictly increasing (can be relaxed) then $\{N_{m^{-1}(t)}, t\geq 0\}$ is a homogeneous Poisson process of rate 1.

The Galton Watson Branching Process

We start with one individual (can be generalized to k) at generation/time 0. The individual has offspring according to an offspring distribution with pfg $G(z)$. These offspring live for one generation and each of them independently has offspring according to G . This generates a Markov Chain $\{X_t, t=0,1,2,\ldots\}$ where X_t is the number of people in the t -th generation. Since $X_0=1$, X_1 has pgf $G(z)=E\left[z^{X_1}\right]$. Set $\mu=E(X_1)$.
We are interested in calculating $\rho=\lim_{t\to\infty} P(X_t=0)$ which we may term the probability of ultimate extinction.

Theorem

Assume $0<P(X_1=0)<1$. Then

- $\mu<1\Rightarrow\rho=1$ (subcritical case).
- $\mu=1\Rightarrow\rho=1$ (critical case).
- $\mu>1\Rightarrow\rho=1$ (supercritical case).

Time Homogeneous Discrete Time Markov Chains

Consider $\{X_t, t=0,1,2,\ldots\}$. Set $\vec{p}(t)$ to be the pf of X_t ; $\vec{p}(0)$ represents the initial distribution. Define the n -step transition probabilities via $p_{ij}(n)=P(X_n=j\,|\,X_0=i)$. Set the matrix $P(n)=[p_{ij}(n)]_{i,j}$ and call $P=P(1)$ the transition matrix. Then $\vec{p}(t)^T=\vec{p}(0)^T P(t)=\vec{p}(t-1)^T P$.

Stationary/Equilibrium Distribution

If $\vec{p}(t)^T \to \vec{\pi}$ (a limiting pf), then we must have $\vec{\pi}^T = \vec{\pi}^T P$. Such a $\vec{\pi}$ is called a stationary or equilibrium distribution. In such a case $\vec{\pi}$ will not depend on the initial distribution and the process is then said to be ergodic.

From $\vec{p}(t)^T = \vec{p}(0)^T P(t)$, by taking $\vec{p}(0)^T = (1 \quad 0 \quad 0 \quad \cdots)$, $\vec{p}(0)^T = (0 \quad 1 \quad 0 \quad \cdots)$, etc, yields $\lim_{t\to\infty} P(t) = \begin{bmatrix} \vec{\pi} \\ \vec{\pi} \\ \vdots \end{bmatrix}$. Since $P(0) = I$ we conclude $P(t) = P^t$ and hence $\vec{p}(t) = \vec{p}(0) P^t$.

Chapman Kolmogorov Equations

Consider $t_1 < t_2 < t_3$. We have

$$\begin{aligned} P(X_{t_3} = j \mid X_{t_1} = i) &= \frac{P(X_{t_3} = j, X_{t_1} = i)}{P(X_{t_1} = i)} \\ &= \sum_k \frac{P(X_{t_3} = j, X_{t_2} = k, X_{t_1} = i)}{P(X_{t_1} = i)} \\ &= \sum_k P(X_{t_3} = j \mid X_{t_2} = k, X_{t_1} = i) \frac{P(X_{t_2} = k, X_{t_1} = i)}{P(X_{t_1} = i)} \\ &= \sum_k P(X_{t_3} = j \mid X_{t_2} = k) P(X_{t_2} = k \mid X_{t_1} = i) \end{aligned}$$

The equations $P(X_{t_3} = j \mid X_{t_1} = i) = \sum_k P(X_{t_3} = j \mid X_{t_2} = k) P(X_{t_2} = k \mid X_{t_1} = i)$ are called Chapman Kolmogorov equations (CKE). In the time homogeneous case they reduce to $P(s+t) = P(s) P(t)$ (in matrix form).

Definition: Recurrent State

State *i* is recurrent if starting from *i* one is certain to return.

- Notice that this is simply stating that the recurrence time is finite wp1.
- Note that if the recurrence time is finite then the number of recurrences must be infinite. The number of returns is $\sum_{t=1}^{\infty} I(X_t = i \mid X_0 = i)$ with mean $\sum_{t=1}^{\infty} E[I(X_t = i \mid X_0 = i)] = \sum_{t=1}^{\infty} p_{ii}(t)$, and it follows that state i is recurrent iff $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$.

Definitions

1. A recurrent state is positive recurrent if the mean return time is finite. Otherwise it is null recurrent.
2. A state which is not recurrent is called transient.
3. A state which can only return at multiples of $d > 1$ is called periodic. The smallest d is the period.

Definitions

1. j is accessible from i ($i \to j$) if $p_{i,j}(n) > 0$ for some $n \geq 0$.
2. i and j communicate ($i \leftrightarrow j$) if $i \to j$ and $j \to i$.
3. A Markov Chain is irreducible if all states communicate.

Theorem

If i and j communicate, then they are the same type (recurrent, transient, etc).

Remarks

1. States which are not periodic are termed aperiodic. The behavior of a periodic state can be determined by looking at times kd , $k = 0, 1, \ldots$.
2. For a finite Markov Chain which is irreducible all states must be positive recurrent.

3. A sketch of the one-step transition probabilities is often helpful.

Theorem

In a irreducible, aperiodic, positive recurrent Markov Chain,

$$\lim_{t\to\infty} P(X_t=i|X_0=i)=\lim_{t\to\infty} P(X_t=i)=\pi_i$$

where $\pi_i=\frac{1}{\text{mean recurrence time}}$, and $\vec{\pi}$ is a pf with $\vec{\pi}^T=\vec{\pi}^T P$.