Probability and Stochastic Processes

after Erhan Cinlar and Sheldon M. Ross, notes by Billy Fang

1 Introduction to Probability Theory

1.1 Introduction

1.2 Sample Space and Events

sample space (S): set of all possible outcomes of an experiment

event (E): any subset of the sample space; $E \subset S$

union of two events E and F ($E \cup F$): either E or F occurs

intersection of two events E and F ($E \cap F$ or EF): both E and F occur

E and F are **mutually exclusive** if $EF = \emptyset$

complement of a set E (E^c): all outcomes in S that are not in E

1.3 Probabilities Defined on Events

probability of an event E ($P(E)$)

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. For any sequence of events $E_1, E_2, \ldots$ that are mutually exclusive,
$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c)$

$P(E \cup F) = P(E) + P(F) - P(EF)$

1.4 Conditional Probabilities

conditional probability that E occurs given that F has occurred: $P(E|F) = \frac{P(EF)}{P(F)}$ where $P(F) > 0$

1.5 Independent Events

events E and F are **independent** if

$$P(EF) = P(E)P(F)$$

$\Rightarrow P(E|F) = P(E)$ and $P(F|E) = F(E)$

events E and F are **dependent** if they are not independent

events $E_1, E_2, \ldots, E_n$ are independent if for every subset $E_{1'}, E_{2'}, \ldots, E_{r'}, r \leq n$ of those events,

$$P(E_{1'}E_{2'}\cdots E_{r'}) = P(E_{1'})P(E_{2'})\cdots P(E_{r'})$$

pairwise independent events are not necessarily jointly independent

A sequence of success-fail experiments consists of **independent trials** if for all $i_1, i_2, \ldots, i_n$,

$$P(E_{i_1}E_{i_2}\cdots E_{i_n}) = \prod_{j=1}^{n} P(E_{i_j})$$

where $E_i, i \geq 1$ denotes the event that the ith experiment was a success.

1.6 Bayes' Formula

let $F_1, F_2, \ldots, F_n$ are mutually exclusive events s.t. $\bigcup_{i=1}^{n} F_i = S$.

$\Rightarrow E = \bigcup_{i=1}^{n} EF_i$ and EF_i are mutually exclusive

$\Rightarrow P(E) = \sum_{i=1}^{n} P(EF_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

2 Random Variables

2.1 Random Variables

random variables: real-valued functions defined on the sample space

indicator random variable: "categorical," example:

$$I = \begin{cases} 1, & \text{if the lifetime of the battery is 2 or more years} \\ 0, & \text{otherwise} \end{cases}$$

See Example 2.5 on page 23 for a good example

discrete random variables take on either a finite our countable number of possible values

continuous random variables take on a continuum of possible values

cumulative distribution function (cdf) for random variable X is defined for any real number b, $-\infty < b < \infty$ by

$$F(b) = P\{X \leq b\}$$

1. $F(b)$ is a nondecreasing function of b
2. $\lim_{b\to\infty} F(b) = F(\infty) = 1$
3. $\lim_{b\to-\infty} F(b) = F(-\infty) = 0$

Strict inequality: $\mathbb{P}\{X < b\} = \lim_{h\to 0^+} \mathbb{P}\{X \le b - h\} = \lim_{h\to 0^+} F(b-h)$

2.2 Discrete Random Variables

probability mass function (pmf) of a discrete random variable X:

$$p(a) = \mathbb{P}\{X = a\}$$

$p(a)$ is positive for at most a countable number of values of a; if X must assume one of the values $x_1, x_2, \ldots$, then

$$\begin{aligned} &p(x_i) > 0, \quad i = 1, 2, \ldots \\ &p(x) = 0, \quad \text{all other values of } x \end{aligned}$$

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

$$F(a) = \sum_{\text{all } x_i \le a} p(x_i)$$

2.2.1 The Bernoulli Random Variable

success-failure; $X = 1$ for success, $X = 0$ for failure

$$p(0) = \mathbb{P}\{X = 0\} = 1 - p$$

$$p(1) = \mathbb{P}\{X = 1\} = p$$

where p is the probability the trial is a success, $0 \le p \le 1$

Bernoulli random variable X has the pmf above for some $p \in (0, 1)$

2.2.2 The Binomial Random Variable

n independent trials, success with probability p, failure $1 - p$

binomial random variable X with parameters (n, p): represents the number of successes in the n trials

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}$$

for $i = 0, 1, \ldots, n$.

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p + (1-p))^n = 1$$

Terminology: if X is a binomial random variable with parameters (n, p), "X has a binomial distribution with parameters (n, p)."

2.2.3 The Geometric Random Variable

independent trials, each with probability p of success

geometric random variable X with parameter p is the number of trials until first success

$$p(n) = \mathbb{P}\{X = n\} = (1-p)^{n-1}p$$

for $n = 1, 2, \ldots$

$$\sum_{n=1}^{\infty} p(n) = p \sum_{n=1}^{\infty} (1-p)^{n-1} = 1$$

2.2.4 The Poisson Random Variable

$X = x \in \{0, 1, 2, \ldots\}$ is a **Poisson** random variable with parameter λ if for some $\lambda > 0$,

$$p(i) = \mathbb{P}\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}$$

for $i = 0, 1, \ldots$

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$$

(power series)

Poisson random variable may approximate a binomial random variable when n is large and p is small. Suppose X is a binomial random variable with parameters (n, p) and let $\lambda = np$. Then

$$\mathbb{P}\{X = i\} = \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} = \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} = \frac{n(n-1)\cdots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$$

For n large and p small,

- $$\left(1-\frac{\lambda}{n}\right)^n = \left(1+\left(-\frac{\lambda}{n}\right)\right)^{(-n/\lambda)(-\lambda)} \approx e^{-\lambda}$$
- $$\frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1$$
- $$(1-\frac{\lambda}{n})^i \approx 1$$

So, $\mathbb{P}\{X = i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$

The Multinomial distribution (page 88)

an experiment has r possible outcomes, the ith outcome has probability p_i. If n of these experiments are performed and the outcome of each experiment does not affect any of the other experiments, the probability that the ith outcome appears x_i times for $i = 1, \ldots, r$ is

$$\frac{n!}{x_1!x_2!\cdots x_r!}p_1^{x_1}p_2^{x_2}\cdots p_r^{x_r}$$

where $\sum_{i=1}^{r} x_i = n$. The multinomial distribution is a generalization of the binomial distribution.

2.3 Continuous Random Variables

X is a **continuous** random variable if there exists a nonnegative function $f(x)$ defined for all $x \in \mathbb{R}$, such that for any set B of real numbers,

$$\mathbb{P}\{X \in B\} = \int_B f(x)dx$$

and $f(x)$ must satisfy

$$1 = \mathbb{P}\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

$f(x)$ is the **probability density function (pdf)** of the random variable X

$\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)dx$

but $\mathbb{P}\{X = a\} = \int_a^a f(x)dx = 0$

cdf:

$$F(a) = \mathbb{P}\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)dx$$

$$\frac{d}{da}F(a) = f(a)$$

$$\mathbb{P}\{a - \varepsilon/2 \le X \le a + \varepsilon/2\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)dx \approx \varepsilon f(a)$$

($f(a)$ is a measure of how likely it is that the random variable will be near a)

2.3.1 The Uniform Random Variable

uniformly distributed over $(0, 1)$:

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

density function because $f(x) \ge 0$ and $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 dx = 1$

$X = x \in (0, 1)$

$$\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)dx = b - a$$

X is a **uniform** random variable on (α, β) if its pdf is

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta \\ 0, & \text{otherwise} \end{cases}$$

cdf:

$$F(a) = \int_{-\infty}^{a} f(x)dx = \begin{cases} 0, & a \le \alpha \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha < a < \beta \\ 1, & a \ge \beta \end{cases}$$

2.3.2 Exponential Random Variables

exponential random variable with parameter $\lambda > 0$ has pdf:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

cdf:

$$F(a) = \int_0^a \lambda e^{-\lambda x} dx = 1 - e^{-\lambda a}, a \geq 0$$

2.3.3 Gamma Random Variables

a **gamma random variable** with shape parameter α and rate parameter λ has pdf:

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where the gamma function is defined by

$$\Gamma(\alpha) \int_0^\infty e^{-x} x^{\alpha-1} dx$$

for $n \in \mathbb{Z}$,

$$\Gamma(n) = (n-1)!$$

2.3.4 Normal Random Variables

if X is a **normal random variable** with parameters μ and σ^2, pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, x \in \mathbb{R}$$

transformation: if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Y = \alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$

Proof:

Suppose $\alpha > 0$

$$F_Y(a) = \mathbb{P}\{Y \leq a\}$$
$$= \mathbb{P}\{\alpha X + \beta \leq a\}$$
$$= \mathbb{P}\{X \leq \frac{a-\beta}{\alpha}\}$$
$$= F_X(\frac{a-\beta}{\alpha})$$
$$\int_{-\infty}^{(a-\beta)/\alpha} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$
$$= \int_{-\infty}^a \frac{1}{\alpha\sigma\sqrt{2\pi}} \exp\left\{\frac{-(v-(\alpha\mu+\beta))^2}{2\alpha^2\sigma^2}\right\} dv$$
$$= \int_{-\infty}^a f_Y(v) dv$$

(change of variables $v = \alpha x + \beta$)

Similar if $\alpha < 0$

standard normal distribution: $\mathcal{N}(0,1)$

standardizing: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$

2.4 Expectation of a Random Variable

2.4.1 The Discrete Case

expected value of X:

$$\mathbb{E}[X] = \sum_{x:p(x)>0} xp(x)$$

Expectation of a Bernoulli random variable with parameter p:

$$\mathbb{E}[X] = 0(1-p) + 1(p) = p$$

Expectation of a binomial random variable with parameters n and p:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{n} ip(i) \\
&= \sum_{i=0}^{n} i\binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} \frac{n!}{(n-i)!(i-1)!} p^i (1-p)^{n-i} \text{ (note: when } i=0\text{, the whole addend is 0)} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(n-1)!(i-1)!} p^{i-1}(1-p)^{n-i} \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\
&= np(p+(1-p))^{n-1} \\
&= np
\end{aligned}$$

where $k = i-1$.

Expectation of a geometric random variable with parameter p:

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} np(1-p)^{n-1} = p\sum_{n=1}^{\infty} nq^{n-1}$$

where $q = 1-p$,

$$= p\sum_{n=1}^{\infty} \frac{d}{dq}(q^n) = p\frac{d}{dq}\left(\sum_{n=1}^{\infty} q^n\right) = p\frac{d}{dq}\left(\frac{q}{1-q}\right) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

Expectation of a Poisson random variable with parameter λ:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} = \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^i}{(i-1)!} = \lambda e^{-\lambda}\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

2.4.2 The Continuous Case

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

Expectation of a uniform random variable distributed over (α, β):

$$\mathbb{E}[X] = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\beta + \alpha}{2}$$

Expectation of an exponential random variable with parameter λ:

$$\mathbb{E}[X] = \int_{0}^{\infty} x\lambda e^{-\lambda x} dx$$

integration by parts with $dv = \lambda e^{-\lambda x}, u = x$,

$$= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx$$

$$= 0 - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty$$

$$= \frac{1}{\lambda}$$

Expectation of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathbb{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2} dx$$

writing x as $(x - \mu) + \mu$,

$$\mathbb{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)e^{-(x-\mu)^2/2\sigma} dx + \frac{\mu}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx$$

let $y = x - \mu$

$$\mathbb{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} ye^{-y^2/2\sigma} dy + \mu \int_{-\infty}^{\infty} f(x)dx$$

by symmetry, the first integral is 0,

$$\mathbb{E}[X] = \mu \int_{-\infty}^{\infty} f(x)dx = \mu$$

Expectation of a gamma random variable X with parameters α, λ

(Proof omitted)

$$\mathbb{E}[X] = \alpha/\lambda$$

2.4.3 Expectation of a Function of a Random Variable

If X is a discrete random variable with pmf $p(x)$, then for any real-valued function g,

$$\mathbb{E}[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$$

Proof in the discrete case:

Let $Y = g(X)$ and let g be invertible. Then $\mathbb{P}\{Y = y\} = \mathbb{P}\{X = g^{-1}(y)\} = p(g^{-1}(y))$. Letting $x = g^{-1}(y)$,

$$\mathbb{E}[Y] = \sum_{y:\mathbb{P}\{Y=y\}>0} y\mathbb{P}\{Y = y\} = \sum_{x:p(x)>0} g(x)p(x)$$

If X is a continuous random variable with pdf $f(x)$, then for any real-valued function g,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Proof in the continuous case:

Assume g is an invertible, monotonically increasing function, and let $Y = g(X)$. Then the cdf of $g(X)$ is

$$F_Y(y) = \mathbb{P}\{Y < y\} = \mathbb{P}\{X < g^{-1}(y)\} = F(g^{-1}(y))$$

The pdf of $g(X)$ is

$$f_Y(y) = F_Y'(y) = [F'(g^{-1}(y))][(g^{-1})'(y)] = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{f(x)}{g'(x)}$$

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_{-\infty}^{\infty} g(x)\frac{f(x)}{g'(x)}g'(x)dx = \int_{-\infty}^{\infty} g(x)f(x)dx$$

through the change of variable $g^{-1}(y) = x$

If a and b are constants, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof in the discrete case:

$$\mathbb{E}[aX + b] = \sum_{x:p(x)>0} (ax + b)p(x)$$
$$= a\sum_{x:p(x)>0} xp(x) + b\sum_{x:p(x)>0} p(x)$$
$$= a\mathbb{E}[X] + b$$

Proof in the continuous case:

$$\mathbb{E}[aX + b] = \int_{-\infty}^{\infty} (ax + b)f(x)dx$$
$$= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx$$
$$= a\mathbb{E}[X] + b$$

$\mathbb{E}[X]$ is called **expected value**, the **mean**, or the first **moment** of X.

$\mathbb{E}[X^n]$, $n \geq 1$ is the nth moment of X;

$$\mathbb{E}[X^n] = \begin{cases} \sum_{x:p(x)>0} x^n p(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)dx, & X \text{ is continuous} \end{cases}$$

(see §2.6)

variance of random variable X:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Variance of the normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X-\mu)^2] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx\end{aligned}$$

substitute $y = (x-\mu)/\sigma$,

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy$$

integrate by parts with $u = y$, $dv = ye^{-y^2/2}dy$,

$$\frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-y^2/2} dy \right)$$

$$\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

$$= \sigma^2$$

Suppose x is continuous with pdf f, and let $\mathbb{E}[X] = \mu$, then

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$$

$$= \mathbb{E}[X^2 - 2\mu X + \mu^2]$$

$$= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x) dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx$$

$$= \mathbb{E}[X^2] - 2\mu^2 + \mu^2$$

$$= \mathbb{E}[X^2] - \mu^2$$

$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

which also works in the discrete case.

2.5 Jointly Distributed Random Variables

2.5.1 Joint Distribution Functions

joint cumulative distribution function (cdf) of X and Y:

$$F(a, b) = \mathbb{P}\{X \leq a, Y \leq b\}$$

for $-\infty < a, b < \infty$.

Note that the cdf of X is $F_X(a) = \mathbb{P}\{X \leq a\} = \mathbb{P}\{X \leq a, Y < \infty\} = F(a, \infty)$ and similarly, $F_Y(b) = \mathbb{P}\{Y \leq b\} = F(\infty, b)$

If X and Y are both discrete, the **joint probability mass function (pmf)** of X and Y is

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

pmf of X:

$$p_X(x) = \mathbb{P}\{X = x\} = \sum_{y:p(x,y)>0} p(x,y)$$

pmf of Y:

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x,y)$$

X and Y are **jointly continuous** if $\exists$ a nonnegative function $f(x,y)$ defined for all $x, y \in \mathbb{R}$ s.t. for all sets A and B of real numbers,

$$\mathbb{P}\{X \in A, Y \in B\} = \int_B \int_A f(x,y)dx\ dy$$

which is the **joint probability density function** of X and Y.

finding the pdf of X and Y:

$$\mathbb{P}\{X \in A\} = \mathbb{P}\{X \in A, Y \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} \int_A f(x,y)dx\ dy = \int_A f_X(x)dx$$

where

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

similarly,

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

Relating the cdf with the pdf:

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

$$\frac{d^2}{da\ db} F(a,b) = f(a,b)$$

If X and Y are random variables, and g is a function of 2 variables, then

$$\mathbb{E}[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$$

or

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dx\ dy$$

Example:

$$\mathbb{E}[X+Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)f(x,y)dx\ dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dx\ dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x,y)dx\ dy$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Generalized:

$$\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}[X_i]$$

for random variables X_i and constants a_i

Verifying the expectation of a binomial random variable X with parameters n and p:

$$X = \sum_{i=1}^{n} X_i$$

where

$$X_i = \begin{cases} 1, & i\text{th trial is a success} \\ 0, & i\text{th trial is a failure} \end{cases}$$

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$$

see book for interesting hat example and coupon example

2.5.2 Independent Random Variables

Random variables X and Y are **independent** if for all a, b,

$$\mathbb{P}\{X \leq a, Y \leq b\} = \mathbb{P}\{X \leq a\}\mathbb{P}\{Y \leq b\}$$

By this definition, if X and Y are independent, the following are true:

- For all a, b, the events $E_a = \{X \leq a\}$ and $F_b = \{Y \leq b\}$ are independent
- For all a, b, $F(a, b) = F_X(a)F_Y(b)$
- For discrete X and Y, $p(x, y) = p_X(x)p_Y(y)$
- For continuous X and Y, $f(x, y) = f_X(x)f_Y(y)$

Proof that $(p(x, y) = p_X(x)p_Y(y)) \Rightarrow (X$ and Y are independent):

$$\begin{aligned} \mathbb{P}\{X \leq a, Y \leq b\} &= \sum_{y \leq b} \sum_{x \leq a} p(x, y) \\ &= \sum_{y \leq b} \sum_{x \leq a} p_X(x)p_Y(y) \\ &= \sum_{y \leq b} p_Y(y) \sum_{x \leq a} p_X(x) \\ &= \mathbb{P}\{Y \leq b\}\mathbb{P}\{X \leq a\} \end{aligned}$$

Similar in the continuous case:

$$\iint f(x)g(y)dx\ dy = \left(\int f(x)dx\right)\left(\int g(y)dy\right)$$

If X and Y are independent, then for any functions h and g,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

Proof in the continuous case, assuming X and Y are jointly continuous:

$$\mathbb{E}[g(X)h(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x, y)dx\ dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dx\ dy$$
$$= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx$$
$$= \mathbb{E}[h(Y)]\mathbb{E}[g(X)]$$

2.5.3 Covariance and Variance of Sums of Random Variables

covariance of random variables X and Y:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$
$$= \mathbb{E}[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]]$$
$$= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$

Consider the case where X and Y are indicator variables for whether or not A and B occur:

$$X = \begin{cases} 1, & \text{A occurs} \\ 0, & \text{otherwise} \end{cases} \qquad Y = \begin{cases} 1, & \text{B occurs} \\ 0, & \text{otherwise} \end{cases}$$

Then

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{P}\{X=1, Y=1\} - \mathbb{P}\{X=1\}\mathbb{P}\{Y=1\}$$

$$\mathrm{Cov}(X,Y) > 0 \Leftrightarrow \mathbb{P}\{X=1, Y=1\} > \mathbb{P}\{X=1\}\mathbb{P}\{Y=1\}$$
$$\Leftrightarrow \frac{\mathbb{P}\{X=1, Y=1\}}{\mathbb{P}\{X=1\}} > \mathbb{P}\{Y=1\}$$
$$\Leftrightarrow \mathbb{P}\{Y=1|X=1\} > \mathbb{P}\{Y=1\}$$

which shows that $\mathrm{Cov}(X,Y)$ is positive if the outcome $X=1$ makes it more likely that $Y=1$ (as well as the reverse).

Positive covariance indicates the Y increases with X while negative covariance indicates Y decreases as X increases.

See excellent example on page 51

Properties of Covariance

1. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
2. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
3. $\mathrm{Cov}(cX,Y) = c\ \mathrm{Cov}(X,Y)$
4. $\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$

Proof of (4):

$$\mathrm{Cov}(X, Y+Z) = \mathbb{E}[X(Y+Z)] - \mathbb{E}[X]\mathbb{E}[Y+Z]$$
$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]$$
$$= \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$$

(4) generalizes to

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

Useful way to express variance of sum of random variables:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j\right)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$
$$= \sum_{i=1}^{n} \mathrm{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \mathrm{Cov}(X_i, X_j)$$
$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j<i} \mathrm{Cov}(X_i, X_j)$$

In the case that X_i are independent random variables for $i = 1, \ldots, n$,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

If $X_1, \ldots, X_n$ are *independent and identically distributed (i.i.d.)*, then the **sample mean** is

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

Suppose $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, then

- $\mathbb{E}[\overline{X}] = \mu$ because
$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{m} \mathbb{E}[X_i] = \mu$$

- $\mathrm{Var}(\overline{X}) = \sigma^2/n$ because
$$\mathrm{Var(X)} = \left(\frac{1}{n}\right)^2 = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

- $\mathrm{Cov}(\overline{X}, X_i - \overline{X}) = 0, i = 1, \ldots, n$ because
$$\mathrm{Cov}(\overline{X}, X_i - \overline{X}) = \mathrm{Cov}(\overline{X}, X_i) - \mathrm{Cov}(\overline{X}, \overline{X})$$
$$= \frac{1}{n}\mathrm{Cov}\left(X_i + \sum_{j\neq i} X_j, X_i\right) - \mathrm{Var}(\overline{X})$$

$$= \frac{1}{n}\text{Cov}(X_i, X_i) + \frac{1}{n}\text{Cov}\left(\sum_{j\neq i} X_j, X_i\right) - \frac{\sigma^2}{n}$$

$$\frac{\sigma^2}{n} + 0 - \frac{\sigma^2}{n} = 0$$

due to the fact that X_i and $\sum_{j\neq i} X_j$ are independent and have covariance 0.

Variance of a binomial random variable X with parameters n and p:

$$X = \sum_{i=1}^{n} X_i$$

where

$$X_i = \begin{cases} 1, & i\text{th trial is a success} \\ 0, & i\text{th trial is a failure} \end{cases}$$

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i)$$

$$\text{Var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \mathbb{E}[X_i] - (\mathbb{E}[X_i])^2 = p - p^2$$

because $X_i^2 = X_i$ (because $1^2 = 1$ and $0^2 = 0$). Thus,

$$\text{Var}(X) = np(1-p)$$

Sampling from a finite population: the hypergeometric

population of N individuals, Np in favor, $N - Np$ opposed, p is unknown; how to estimate p by randomly choosing and determining the positions of n members of the population

let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th person chosen is in favor} \\ 0, & \text{otherwise} \end{cases}$$

The usual estimator of p is $\sum_{i=1}^{n} X_i/n$

$$\mathbb{E}\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$$

because $\mathbb{E}[X_i] = Np/N = p$ for all i.

$$\text{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j}\sum \text{Cov}(X_i, X_j)\right)$$

Note that $\text{Var}(X_i) = p(1-p)$ because X_i is Bernoulli

Because $i \neq j$,

$$\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \mathbb{P}\{X_i = 1, X_j = 1\} - p^2$$

$$= \mathbb{P}\{X_i = 1\}\mathbb{P}\{X_j = 1 | X_i = 1\} - p^2$$
$$= \frac{Np}{N}\frac{Np-1}{N-1} - p^2$$

So,

$$\text{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{i<j}\sum \text{Cov}(X_i, X_j)\right)$$
$$= \frac{1}{n^2}\left(np(1-p) + 2\binom{n}{2}\left(\frac{Np}{N}\frac{Np-1}{N-1} - p^2\right)\right)$$
$$= \frac{1}{n^2}\left(np(1-p) + n(n-1)\left(\frac{p(p-1)}{N-1}\right)\right)$$
$$= \frac{p(1-p)}{n} - \frac{(n-1)p(1-p)}{n(N-1)}$$

Variance of the estimator increases as N increases; as $N \to \infty$, variance approaches $p(1-p)/n$. Makes sense because for N large, each X_i will be approx. independent, so $\sum_{i=1}^{n} X_i$ will be approx. binomial distribution with parameters n and p.

Think of $\sum_{i=1}^{n} X_i$ as the number of white balls obtained when n balls are randomly selected from a population consisting of Np white and $N - Np$ black balls; this random variable is **hypergeometric** and has pmf

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i = k\right\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$$

Let X and Y be continuous and independent, and let pdf of X and Y be f and g respectively; let $F_{X+Y}(a)$ be the cdf of $X+Y$. Then

$$F_{X+Y}(a) = \mathbb{P}\{X + Y \leq a\}$$
$$= \iint_{x+y\leq a} f(x)g(y)dx\ dy$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f(x)g(y)dx\ dy$$
$$= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f(x)dx\right)g(y)dy$$
$$= \int_{-\infty}^{\infty} F_X(a-y)g(y)dy$$

The cdf F_{X+Y} is the **convolution** of the distributions F_X and F_Y

To find the pdf of $X+Y$,

$$f_{X+Y}(a) = \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)g(y)dy$$
$$= \int_{-\infty}^{\infty}\frac{d}{da}(F_X(a-y))g(y)dy$$

$$= \int_{-\infty}^{\infty} f(a-y)g(y)dy$$

Sum of two independent random variables X and Y uniformly distributed on $(0,1)$

$$f(a) = g(a) = \begin{cases} 1, & 0 < a < 1 \\ 0, & \text{otherwise} \end{cases}$$

since $\mathbb{P}\{X \leq a\} = \int_{-\infty}^{a} f(x)dx = \int_0^a dx = a$

Then

$$f_{X+Y}(a) = \int_0^1 f(a-y)dy$$

since $g(y) = 0$ for other values of y.

Case 1: $0 \leq a \leq 1$

$$f_{X+Y}(a) = \int_0^a dy = a$$

because for $y > a$, $f(a-y) = 0$

Case 2: $1 < a < 2$

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

because $f(a-y) = 1$ only for y s.t. $a - 1 < y < 1$

$$f_{X+Y}(a) = \begin{cases} a, & 0 \leq a \leq 1 \\ 2-a, & 1 < a < 2 \\ 0, & \text{otherwise} \end{cases}$$

Sum of independent Poisson random variables X and Y with means λ_1 and λ_2

$$\mathbb{P}\{X+Y=n\} = \sum_{k=0}^{n} \mathbb{P}\{X=k, Y=n-k\}$$

$$= \sum_{k=0}^{n} \mathbb{P}\{X=k\}\mathbb{P}\{Y=n-k\}$$

$$= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!}$$

$$= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\lambda_1^k \lambda_2^{n-k}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n$$

$X_1 + X_2$ follows a Poisson distribution with mean $\lambda_1 + \lambda_2$

n random variables $X_1, X_2, \ldots, X_n$ are independent if, for all values $a_1, a_2, \ldots, a_n$,

$$\mathbb{P}\{X_1 \leq a_1, \ldots, X_n \leq a_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \leq a_i\}$$

see pg. 58 for example on **order statistics**

2.5.4 Joint Probability Distribution of Functions of Random Variables

X_1 and X_2 are jointly continuous random variables with joint pdf $f(x_1, x_2)$

suppose $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2 satisfying the following conditions:

1. $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 in terms of y_1 and y_2 with solutions given by, say, $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$

2. g_1 and g_2 have continuous partial derivatives at all points (x_1, x_2) and are such that

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} \equiv \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \neq 0$$

   at all poitns (x_1, x_2).

then Y_1 and Y_2 are jointly continuous by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}|J(x_1, x_2)|^{-1}$$

where $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$. This comes from differentiating both sides of the following equation w.r.t. y_1 and y_2:

$$\mathbb{P}\{Y_1 \leq y_1, Y_2 \leq y_2\} = \iint_{(x_1,x_2):g_1(x_1,x_2)\leq y_1; g_2(x_1,x_2)\leq y_2} f_{X_1,X_2}(x_1, x_2)dx_1\, dx_2$$

see pg. 61 for generalizing to more variables.

2.6 Moment Generating Functions

(see §2.4.3) **moment generating function** $\phi(t)$ of random variable X is defined for all values t by

$$\phi(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum_x e^{tx}p(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx}f(x)dx, & X \text{ is continuous} \end{cases}$$

all the moments of X can be obtained by successively differentiating $\phi(t)$.

$$\phi'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d}{dt}(e^{tX})\right] = \mathbb{E}[Xe^{tX}]$$

so $\phi'(0) = \mathbb{E}[X]$

in general, the nth derivative of $\phi(t)$ evaluated at $t = 0$ equals $\mathbb{E}[X^n]$ for $n \geq 1$

Example: Binomial Distribution with n and p

$$\phi(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} = (pe^t + 1 - p)^n$$

Hence,

$$\mathbb{E}[X] = \phi'(0) = \left[n(pe^t + 1 - p)^{n-1} pe^t\right]_{t=0} = np$$

$$\mathbb{E}[X^2] = \phi''(0) = n(n-1)p^2 + np$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$$

Example: Poisson Distribution with mean λ

$$\phi(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

$\mathbb{E}[X] = \phi'(0) = \lambda$

$\mathbb{E}[X^2] = \phi''(0) = \lambda^2 + \lambda$

$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda$

See book for more examples as well as a table of moment generating functions.

If X and Y are independent,

$$\phi_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = \phi_X(t)\phi_Y(t)$$

There is a one-to-one correspondence between the moment generating function and the distribution function of a random variable.

See book for Poisson paradigm, Laplace transform, multivariate normal distribution

2.6.1 Joint Distribution of Sample Mean and Sample Variance from a Normal Population

$X_1, \ldots, X_n$ i.i.d. random variables, each mean μ and variance σ^2. **Sample variance**:

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Note that

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (X_i - \mu + \mu - \overline{X})^2$$

$$= \left(\sum_{i=1}^{n} (X_i - \mu)^2\right) + n(\mu - \overline{X})^2 + 2(\mu - \overline{X}) \sum_{i=1}^{n} (X_i - \mu)$$

$$= \left(\sum_{i=1}^{n} (X_i - \mu)^2\right) + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})(n\overline{X} - n\mu)$$

$$= \left(\sum_{i=1}^{n} (X_i - \mu)^2 \right) + n(\mu - \overline{X})^2 - 2n(\mu - \overline{X})^2$$
$$= \left(\sum_{i=1}^{n} (X_i - \mu)^2 \right) - n(\mu - \overline{X})^2$$

So,

$$\begin{aligned} \mathbb{E}[(n-1)S^2] &= \sum_{i=1}^{n} \mathbb{E}[(X_i - \mu)^2] - n\mathbb{E}[(\overline{X} - \mu)^2] \\ &= n\sigma^2 - n\text{Var}(\overline{X}) \\ &= n\sigma^2 - n(\sigma^2/n) \\ &= (n-1)\sigma^2 \end{aligned}$$

So $\mathbb{E}[S^2] = \sigma^2$

If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then $\sum_{i=1}^{n} Z_i^2$ is a **chi-squared random variable** with n degrees of freedom

see book for more details

2.7 The Distribution of the Number of Events that Occur

2.8 Limit Theorems

Markov's Inequality: If X is a random variable that takes only nonnegative values, then for any $a > 0$,

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}$$

Proof for the case that X is continuous with density f:

$$\mathbb{E}[X] = \int_0^\infty xf(x)dx \geq \int_a^\infty xf(x)dx \geq \int_a^\infty af(x)dx = a\mathbb{P}\{X \geq a\}$$

Chebyshev's Inequality: If X is a random variable with mean μ and variance σ^2, then for any $k > 0$,

$$\mathbb{P}\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

Proof:

$(X - \mu)^2$ is a nonnegative random variable, apply Markov's inequality with $a = k^2$:

$$\mathbb{P}\{(X - \mu)^2 \geq k^2\} \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2}$$

Because $(X - \mu)^2 \geq k^2 \Leftrightarrow |X - \mu| \geq k$, then

$$\mathbb{P}\{|X - \mu| \geq k\} \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

These inequalities allow us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known.

Strong Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common distribution, let $\mathbb{E}[X_i] = \mu$. Then, with probability 1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty$$

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each with mean μ and variance σ^2. Then the distribution of $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ tends to the standard normal as $n \to \infty$:

$$\mathbb{P}\left\{\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

This holds for all any distribution of the X_is.

see text for proof of CLT

2.9 Stochastic Processes

A **stochastic process** $\{X(t), t \in T\}$ is a collection of random variables $X(t)$ for each $t \in T$. Refer to $X(t)$ as the *state* of the process at time t, and T as the *index* set of the process. The process is *discrete-time* if T is countable, and *continuous-time* if T is an interval of the real line. The *state space* of a stochastic process is the set of all possible values that $X(t)$ can assume.

2.10 Extra stuff

In a sequence of independent success-failure trials (success with probability p), the number of successes that appear before the rth failure follows the **negative binomial distribution**.

$$X \sim NB(r,p)$$
$$\mathbb{P}\{X = k\} = \binom{k+r-1}{k}(1-p)^r p^k$$
$$\mathbb{E}[X] = \frac{pr}{1-p}$$
$$\mathrm{Var}(X) = \frac{pr}{(1-p)^2}$$

Continued from §2.2.4:

If N is a binomial random variable with parameters n large and p small,

$$\mathbb{P}\{N = i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

with $\mathbb{E}[N] = np = \lambda$.

From §2.4.1,

$$\mathbb{E}[N] = \sum_{i=0}^{\infty} i\mathbb{P}\{N = i\} = \sum_{i=1}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = \lambda \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda = np$$

Let N_t be the number of arrivals in the interval $[0, t]$. Let

$$\mathbb{P}\{N_t = k\} = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

for $k \geq 0$, so that $\mathbb{E}[N_t] = \lambda t$ is directly related with time.

Assume the arrivals are evenly spaced in the time interval.

Let T_k be the time of the kth arrival.

For **some fixed time** t, the probability that the first arrival comes after t is equal to the probability that there are no arrivals in the time interval $[0, t]$:

$$\mathbb{P}\{T_1 > t\} = \mathbb{P}\{N_t = 0\} = e^{-\lambda t}$$

(because $k = 0$).

$$F_{T_1}(t) = \mathbb{P}\{T_1 \leq t\} = 1 - \mathbb{P}\{T_1 > t\} = 1 - e^{-\lambda t}$$

for $t \geq 0$

$$f_{T_1}(t) = \frac{d}{dt}F_{T_1}(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$$

T_1 is an exponential random variable.

For **some fixed** k (let t be variable), examine the time of the kth arrival, T_k:

$$F_{T_k} = \mathbb{P}\{T_k \leq t\} = \mathbb{P}\{N_t \geq k\} = 1 - \mathbb{P}\{N_t \leq (k-1)\} = 1 - \sum_{j=0}^{k-1} \mathbb{P}\{N_t = j\} = 1 - \sum_{j=0}^{k-1} e^{-\lambda t}\frac{(\lambda t)^j}{j!}$$

$$f_{T_k} = -\sum_{j=0}^{k-1}\left[\frac{d}{dt}\left(e^{-\lambda t}\frac{(\lambda t)^j}{j!}\right)\right] = -\sum_{j=0}^{k-1}\left[e^{-\lambda t}\left(\frac{\lambda^j t^{j-1}}{(j-1)!} - \frac{\lambda^{j+1}t^j}{j!}\right)\right] = -e^{-\lambda t}\left(0 - \frac{\lambda^k t^{k-1}}{(k-1)!}\right) = \frac{\lambda e^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!}$$

because of telescoping. (Be careful with the case $j = 0$; it is not notated well above.)

$$\mathbb{E}[T_k] = \int_0^{\infty} \mathbb{P}\{T_k > t\}dt = \int_0^{\infty} \mathbb{P}\{N_t \leq k\}dt$$

$$= \int_0^{\infty}\left(\sum_{j=0}^{k} e^{-\lambda t}\frac{(\lambda t)^j}{j!}\right)dt = \sum_{j=0}^{k}\int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^j}{j!}dt = \sum_{j=0}^{k}\frac{1}{\lambda}\int_0^{\infty}\frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!}dt = \sum_{j=0}^{k}\frac{1}{\lambda} = \frac{k}{\lambda}$$

(Notice the gamma distribution.)

$$\mathbb{E}[T_k^2] = \int_0^{\infty} t^2\frac{\lambda e^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!}dt = \frac{(k+1)k}{\lambda^2}\int_0^{\infty}\frac{\lambda e^{-\lambda t}(\lambda t)^{k+1}}{(k+1)!}dt = \frac{(k+1)k}{\lambda^2}\int_0^{\infty}\mathbb{P}\{T_k = k+1\}dt = \frac{(k+1)k}{\lambda^2}$$

$$\mathrm{Var}[T_k] = \frac{(k+1)k}{\lambda^2} - \frac{k^2}{\lambda} = \frac{k}{\lambda^2}$$

Theorem: The interval lengths T_1, $T_2 - T_1$, $T_3 - T_2$, ... are i.i.d exponential variables with parameter λ.

$$\mathbb{E}[T_k] = \mathbb{E}[T_1 + (T_2 - T_1) + (T_3 - T_2) + \dots (T_k - T_{k-1})] = \frac{k}{\lambda}$$

$$\text{Var}[T_k] = \text{Var}[T_1 + (T_2 - T_1) + (T_3 - T_2) + \dots (T_k - T_{k-1})] = \frac{k}{\lambda^2}$$

Gamma density function with shape index k and scale parameter λ:

$$\frac{\lambda e^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!}$$

Gamma density function with shape index α and scale parameter λ:

$$\frac{\lambda e^{-\lambda t}(\lambda t)^{\alpha-1}}{\Gamma(\alpha)}$$

where $\Gamma(\alpha)$ is what makes it a pdf, and

$$\Gamma(\alpha) \equiv \int_0^\infty \lambda e^{-\lambda t}(\lambda t)^{\alpha-1} dt = \int_0^\infty e^{-x} x^{\alpha-1} dx$$

where $x = \lambda t$ (so, $dx = \lambda dt$).

Properties of Γ:

- $\Gamma(1) = 1$
- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
- $\Gamma(k) = (k-1)!$ if k is an integer
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Consider $X \sim \text{gamma}(\alpha, 1)$. Since $\int_0^\infty \frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)} dx = 1$, then

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$$

$$= [-e^{-x}x^{\alpha-1}]_0^\infty - \int_0^\infty -e^{-x}(\alpha-1)x^{\alpha-2} dx$$

$$= 0 + (\alpha - 1)\Gamma(\alpha - 1)$$

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

Since $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha - 1)!$ if $\alpha \in \mathbb{N}$

Consider $Y = X^2$ where $X \sim \mathcal{N}(0, 1)$

$$\mathbb{P}\{Y \le t\} = \mathbb{P}\{X^2 \le t\} = \mathbb{P}\{-\sqrt{t} \le X \le \sqrt{t}\} = \int_{-\sqrt{t}}^{\sqrt{t}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 2\int_0^{\sqrt{t}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

find pdf:

$$\frac{d}{dt}\mathbb{P}\{Y \le t\} = \frac{d}{dt}\left(2\int_0^{\sqrt{t}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx\right) = 2\left(\frac{e^{-t^2/2}}{\sqrt{2\pi}}\right)\left(\frac{1}{2}t^{-1/2}\right) = \frac{e^{-t/2}t^{-1/2}}{\sqrt{2\pi}} = \frac{\frac{1}{2}e^{-t/2}(\frac{1}{2}t)^{-1/2}}{\sqrt{\pi}}$$

$Y \sim \text{gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

If $X_1, \dots, X_n$ each $\sim \mathcal{N}(0, 1)$, then $\left(\sum_{i=1}^n X_i^2\right) \sim \text{gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$, which is the chi-square distribution.

Alternative way to compute expected value of a nonnegative random variable, $X \geq 0$:

$$\mathbb{E}[X] = \int_0^\infty (f(x)x)dx = \int_0^\infty f(x)\left(\int_0^x dy\right)dx = \int_0^\infty \int_0^x f(x)\,dy\,dx = \int_0^\infty \int_y^\infty f(x)\,dx\,dy$$

$$= \int_0^\infty \mathbb{P}\{X > y\}dy = \int_0^\infty (1 - F(y))dy$$

Beta distribution

U follows the beta distribution with shape pair (α, β) if

$$\mathbb{P}\{U \in du\} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u^{\alpha-1}(1-u)^{\beta-1}$$

for $\alpha > 0, \beta > 0$ and $u \in [0, 1]$.

3 Conditional Probability and Conditional Expectation

3.1 Introduction

3.2 The Discrete Case

if $\mathbb{P}(F) > 0$,

$$\mathbb{P}(E|F) \equiv \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$$

pmf of X given that $Y = y$, for all y s.t. $\mathbb{P}\{Y = y\} > 0$:

$$p_{X|Y}(x|y) = \mathbb{P}\{X = x|Y = y\} = \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}} = \frac{p(x, y)}{p_Y(y)}$$

cdf of X given that $Y = y$, for all y s.t. $\mathbb{P}\{Y = y\} > 0$:

$$F_{X|Y}(x|y) = \mathbb{P}\{X \leq x|Y = y\} = \sum_{a \leq x} p_{X|Y}(a|y)$$

conditional expectation of X given that $Y = y$:

$$\mathbb{E}[X|Y = y] = \sum_x x\mathbb{P}\{X = x|Y = y\} = \sum_x p_{X|Y}(x|y)$$

Note: properties of expectations still hold, for example:

$$\mathbb{E}\left[\sum_{i=1}^n X_i|Y = y\right] = \sum_{i=1}^n \mathbb{E}[X_i|Y = y]$$

if X is independent of Y, then $p_{X|Y}(x|y) = \mathbb{P}\{X = x|Y = y\} = \mathbb{P}\{X = x\}$

Example If X_1 and X_2 are independent binomial with (n_1, p) and (n_2, p), let $q = 1 - p$, then

$$\mathbb{P}\{X_1 = k | X_1 + X_2 = m\} = \frac{\mathbb{P}\{X_1 = k, X_2 = m-k\}}{\mathbb{P}\{X_1 + X_2 = m\}} = \frac{\binom{n_1}{k} p^k q^{n_1-k} \binom{n_2}{m-k} p^{m-k} q^{n_2-m+k}}{\binom{n_1+n_2}{m} p^m q^{n_1+n_2-m}} = \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1+n_2}{m}}$$

where we use the fact that $X_1 + X_2 \sim Binomial(n_1 + n_2, p)$

This is the **hypergeometric distribution** (see §2.5.3), the distribution of the number of blue balls that are chosen when a sample of m balls is randomly chosen from n_1 blue and n_2 red balls.

see p 100 for example on Poisson

3.3 The Continuous Case

If X and Y have joint pdf $f(x, y)$, then cdf of X given that $Y = y$ is defined for all y s.t. $f_Y(y) > 0$ by

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

To motivate this definition, notice that

$$f_{X|Y}(x|y)dx = \frac{f(x, y)dx\,dy}{f_Y(y)dy} \approx \frac{\mathbb{P}\{x \le X \le x + dx, y \le Y \le y + dy\}}{\mathbb{P}\{y \le Y \le y + dy\}} = \mathbb{P}\{x \le X \le x + dx | y \le Y \le y + dy\}$$

conditional expectation of X given that $Y = y$, is defined for all y s.t. $f_Y(y) > 0$, by

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

see book for examples

3.4 Computing Expectations by Conditioning

Let $\mathbb{E}[X|Y]$ be a function of the random variable Y: when $Y = y$, then $\mathbb{E}[X|Y] = \mathbb{E}[X|Y = y]$. $\mathbb{E}[X|Y]$ itself is random variable.

For all random variables X and Y,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

If Y is discrete,

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]\mathbb{P}\{Y = y\}$$

and if Y is continuous with pdf $f_Y(y)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] f_Y(y)dy$$

Proof in the discrete case:

$$\sum_y \mathbb{E}[X|Y = y]\mathbb{P}\{Y = y\} = \sum_y \sum_x x\mathbb{P}\{X = x|Y = y\}\mathbb{P}\{Y = y\}$$

$$= \sum_y \sum_x x\mathbb{P}\{X = x, Y = y\} = \sum_x x \sum_y \mathbb{P}\{X = x, Y = y\} = \sum_x x\mathbb{P}\{X = x\} = \mathbb{E}[X]$$

Example Expectation of the sum of a random number of random variables

N denotes number of accidents per week, X_i denotes the number of injuries in the ith accident. Let $\mathbb{E}[N] = 4$ and $\mathbb{E}[X_i] = 2$ for all $i \in \{1, \ldots, N\}$. What is the expected number of injuries?

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} X_i | N\right]\right]$$

But

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i | N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i | N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mathbb{E}[X]$$

which uses the independence of X_i and N. Note that above, $\mathbb{E}[X] = \mathbb{E}[X_i]$ for all $i \in \{1, \ldots, N\}$ Thus,

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N\mathbb{E}[X]] = \mathbb{E}[N]\mathbb{E}[X]$$

compound random variable $\sum_{i=1}^{N} X_i$, the sum of a random number N of i.i.d. random variables that are also independent of N.

Example Mean of a geometric distribution Trial has probability p of success. N represents number of trials before first success. Let Y be an indicator random variable that equals 1 if the first trial is a success, and 0 otherwise.

$$\begin{aligned}\mathbb{E}[N] &= p\mathbb{E}[N|Y = 1] + (1-p)\mathbb{E}[N|Y = 0] \\ &= p + (1-p)(1 + \mathbb{E}[N]) = 1/p\end{aligned}$$

see p 111-4 for excellent examples

Example Quicksort

Let M_n denote the expected number of comparisons needed by quick sort to sort a set of n distinct values. Condition on the rank of the initial value selected:

$$M_n = \sum_{j=1}^{n} \mathbb{E}[\text{number of comparisons}|\text{value selected is the } j\text{th smallest}]\frac{1}{n}$$

If the initial value selected is the jth smallest, two sets of size $j-1$ and $n-j$, and you need $n-1$ comparisons with the initial value.

$$M_n = \sum_{j=1}^{n}(n - 1 + M_{j-1} + M_{n-j})\frac{1}{n} = n - 1 + \frac{2}{n}\sum_{k=1}^{n-1} M_k$$

where $k = j - 1$ and we note that $M_0 = 0$ Then,

$$nM_n = n(n-1) + 2\sum_{k=1}^{n-1} M_k$$

Replace n by $n+1$

$$(n+1)M_{n+1} = (n+1)n + 2\sum_{k=1}^{n} M_k$$

Subtract the two equations

$$(n+1)M_{n+1} - nM_n = 2n + 2M_n$$

$$(n+1)M_{n+1} = (n+2)M_n + 2n$$

Iterate:

$$\frac{M_{n+1}}{n+2} = \frac{2n}{(n+1)(n+2)} + \frac{M_n}{n+1}$$

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

$$= \cdots$$

$$= 2\sum_{k=0}^{n-1} \frac{n-k}{(n+1-k)(n+2-k)}$$

since $M_1 = 0$

$$M_{n+1} = 2(n+2)\sum_{k=0}^{n-1} \frac{n-k}{(n+1-k)(n+2-k)} = 2(n+2)\sum_{i=1}^{n} \frac{i}{(i+1)(i+2)}$$

for $n \geq 1$, and we let $i = n-k$.

$$M_{n+1} = 2(n+2)\left[\sum_{i=1}^{n} \frac{2}{i+2} - \sum_{i=1}^{n} \frac{1}{i+1}\right]$$

approximate for large n:

$$\sim 2(n+2)\left[\int_3^{n+2} \frac{2}{x}dx - \int_2^{n+1} \frac{1}{x}dx\right]$$

$$= 2(n+2)[2\log(n+2) - \log(n+1) + \log 2 - 2\log 3]$$

$$= 2(n+2)\left[\log(n+2) + \log\frac{n+2}{n+1} + \log 2 - 2\log 3\right]$$

$$\sim 2(n+2)\log(n+2)$$

3.4.1 Computing Variances by Conditioning

Example Variance of geometric random variable

Independent trails with success p. Let N be the trial number of the first success. Let $Y = 1$ if first trial is success, $Y = 0$ otherwise.

Find $\mathbb{E}[N^2]$:

$$\mathbb{E}[N^2] = \mathbb{E}[\mathbb{E}[N^2|Y]]$$

$$\mathbb{E}[N^2|Y=1] = 1$$

$$\mathbb{E}[N^2|Y=0] = \mathbb{E}[(1+N)^2]$$

$$\mathbb{E}[N^2] = p + \mathbb{E}[(1+N)^2](1-p) = 1 + (1-p)\mathbb{E}[2N + N^2]$$

Since we showed that $\mathbb{E}[N] = 1/p$ earlier, then

$$\mathbb{E}[N^2] = 1 + \frac{2(1-p)}{p} + (1-p)\mathbb{E}[N^2]$$

$$\mathbb{E}[N^2] = \frac{2-p}{p^2}$$

Thus,

$$\mathrm{Var}(N) = \mathbb{E}[N^2] - (\mathbb{E}[N])^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

By the definition of variance,

$$\mathrm{Var}(X|Y=y) = \mathbb{E}\left[(X-\mathbb{E}[X|Y=y])^2|Y=y\right]$$

$$= \mathbb{E}[X^2|Y=y] + \mathbb{E}[-2X\mathbb{E}[X|Y=y]|Y=y] + \mathbb{E}[(\mathbb{E}[X|Y=y])^2|Y=y]$$

$$= \mathbb{E}[X^2|Y=y] - 2\mathbb{E}[X\mathbb{E}[X|Y=y]|Y=y] + (\mathbb{E}[X|Y=y])^2$$

$$= \mathbb{E}[X^2|Y=y] - 2(\mathbb{E}[X|Y=y])^2 + (\mathbb{E}[X|Y=y])^2$$

$$= \mathbb{E}[X^2|Y=y] - (\mathbb{E}[X|Y=y])^2$$

If we let $\mathrm{Var}(X|Y)$ be a function of Y that takes the value $\mathrm{Var}(X|Y=y)$ when $Y=y$, then we have this:

Proposition The Conditional Variance Formula

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$$

Proof

$$\mathbb{E}[\mathrm{Var}(X|Y)] = \mathbb{E}[\mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2] = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2] = \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2]$$

and

$$\mathrm{Var}(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y)^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2$$

so,

$$\mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y]) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Example Variance of a compound random variable

Let $X_1, \ldots$ be i.i.d. random variables with distribution F having mean μ and variance σ^2, and let them be independent of the nonnegative integer valued random variable N. $S = \sum_{i=1}^{N} X_i$ is a compound random variable.

$$\mathrm{Var}(S|N=n) = \mathrm{Var}\left(\sum_{i=1}^{N} X_i|N=n\right) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i|N=n\right) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$$

(By the same reasoning, $\mathbb{E}[S|N=n] = n\mu$) So,

$$\mathrm{Var}(S|N) = N\sigma^2$$

and

$$\mathbb{E}[S|N] = N\mu$$

and the conditional variance formula gives

$$\mathrm{Var}(S) = \mathbb{E}[N\sigma^2] + \mathrm{Var}(N\mu) = \sigma^2\mathbb{E}[N] + \mu^2\mathrm{Var}(N)$$

In the case that N is Poisson, then S is a **compound Poisson** random variable. Since the variance of a Poisson is equal to its mean, if N is Poisson with $\mathbb{E}[N] = \lambda$,

$$\mathrm{Var}(S) = \lambda\sigma^2 + \lambda\mu^2 = \lambda\mathbb{E}[X^2]$$

where X has distribution F.

See p 120 for another example

3.5 Computing Probabilities by Conditioning

If we have X be an indicator random variable that takes value 1 if E occurs and 0 otherwise, then

$$\mathbb{E}[X] = \mathbb{P}(E)$$

$$\mathbb{E}[X|Y = y] = \mathbb{P}(E|Y = y)$$

So if Y is discrete,

$$\mathbb{P}(E) = \mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]P(Y = y) = \sum_y \mathbb{P}(E|Y = y)\mathbb{P}(Y = y)$$

and if Y is continuous,

$$\mathbb{P}(E) = \mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y]\mathbb{P}(Y = y)dy = \int_{-\infty}^{\infty} \mathbb{P}(E|Y = y)f_Y(y)dy$$

from §3.3

Example Suppose X and Y are independent with pdfs f_X and f_Y. Compute $\mathbb{P}\{X < Y\}$.

$$\begin{aligned}
\mathbb{P}\{X < Y\} &= \int_{-\infty}^{\infty} \mathbb{P}\{X < Y|Y = y\}f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \mathbb{P}\{X > y|Y = y\}f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \mathbb{P}\{X < y\}f_Y(y)dy \\
&= \int_{-\infty}^{\infty} F_X(y)f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f_X(x)dx\right) f_Y(y)dy
\end{aligned}$$

Example Distribution of the sum of independent Bernoulli random variables

Let $X_1, \dots X_n$ be independent Bernoulli random variables with X_i having parameter p_i. Specifically, $\mathbb{P}\{X_i = 1\} = p_i$ and $\mathbb{P}\{X_i = 0\} = q_i = 1 - p_i$.

Let

$$P_k(j) = \mathbb{P}\{X_1 + \cdots + X_k = j\}$$

and note that

$$P_k(k) = \prod_{i=1}^{k} p_i$$

and

$$P_k(0) = \prod_{i=1}^{k} q_i$$

For $0 < j < k$, condition on X_k:

$$P_k(j) = \mathbb{P}\{X_1 + \cdots + X_k = j|X_k = 1\}p_k + \mathbb{P}\{X_1 + \cdots + X_k = j|X_k = 0\}q_k$$

$$= \mathbb{P}\{X_1 + \cdots + X_{k-1} = j - 1|X_k = 1\}p_k + \mathbb{P}\{X_1 + \cdots + X_{k-1} = j|X_k = 0\}q_k$$

$$= \mathbb{P}\{X_1 + \cdots + X_{k-1} = j-1\}p_k + \mathbb{P}\{X_1 + \cdots + X_{k-1} = j\}q_k$$

$$= p_k P_{k-1}(j-1) + q_k P_{k-1}(j)$$

Start with $P_1(1) = p_1$ and $P_1(0) = q_1$.

See p 126 for example on best prize problem

See p 130 for example on ballot problem

3.6 Some applications

3.7 An Identity for Compound Random Variables

3.8 Extra stuff

Example

Let X be an exponential random variable with parameter λ.

$$\mathbb{E}[X|X > t] = \int_0^\infty \mathbb{P}\{X > x|X > t\}dx$$

Remember that for an exponential random variable,

$$\mathbb{P}\{X > t + u|X > t\} = \begin{cases} \frac{\mathbb{P}\{X>t+u\}}{\mathbb{P}\{X>t\}} = e^{-\lambda u}, & u \geq 0 \\ \frac{\mathbb{P}\{X>t\}}{\mathbb{P}\{X>t\}} = 1, & u < 0 \end{cases}$$

From the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > x\}dx$ for nonnegative random variable X, we have

$$\mathbb{E}[X|X > t] = \int_0^\infty \mathbb{P}\{X > x|X > t\}dx = \int_0^t dx + \int_t^\infty e^{-\lambda(x-t)}dx = t + \frac{1}{\lambda}$$

Example

Let $X_1, X_2, \ldots$ be i.i.d random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, $\forall i$.

Let N be a random variable, independent of the X_i, with $\mathbb{E}[N] = a$ and $\text{Var}(N) = b^2$.

Let random variable $S = \sum_1^N X_i$ for $N \geq 1$, and $S = 0$ for $N = 0$.

Compute expected value:

$$\mathbb{E}[S] = \sum_{n=0}^{\infty} \mathbb{E}[S|N = n]\mathbb{P}\{N = n\}$$

$$= \sum_{n=0}^{\infty} \mathbb{E}[X_1 + \cdots + X_n]\mathbb{P}\{N = n\}$$

$$= \sum_{n=0}^{\infty} n\mu\mathbb{P}\{N = n\} = \mu\mathbb{E}[N] = a\mu$$

Compute variance:

$$\mathbb{E}[S^2] = \sum_{n=0}^{\infty} \mathbb{E}[S^2|N=n]\mathbb{P}\{N=n\}$$

$$= \sum_{n=0}^{\infty} \mathbb{E}[(X_1 + \cdots + X_n)^2]\mathbb{P}\{N=n\}$$

$$= \sum_{n=0}^{\infty} \left[Var(X_1 + \cdots + X_n) + (\mathbb{E}[X_1 + \cdots + X_n])^2\right] \mathbb{P}\{N=n\}$$

$$= \sum_{n=0}^{\infty} (n\sigma^2 + n^2\mu^2)\mathbb{P}\{N=n\}$$

$$= \sigma^2\mathbb{E}[N] + \mu^2\mathbb{E}[N^2]$$

$$= \sigma^2\mathbb{E}[N] + \mu^2(Var(N) + (\mathbb{E}[N])^2)$$

$$= a\sigma^2 + \mu^2(b^2 + a^2)$$

$$Var(S) = \mathbb{E}[S^2] - (\mathbb{E}[S])^2 = a\sigma^2 + b^2\mu^2$$

Interesting instance where successes and failures are independent

N trials with success probability p, $N \sim \text{Poisson}(a)$

$L =$ number of successes

$M =$ number of failures

Let X_i be the number of successes at ith trial; i.e. $\mathbb{P}\{X_i = 1\} = p$, $\mathbb{P}\{X_i = 0\} = 1 - p = q$

$L = \sum_{i=1}^{N} X_i$

$$\mathbb{P}\{L=i, M=j\} = \mathbb{P}\{L=i, M=j, N=i+j\}$$

$$= \mathbb{P}\{L=i, M=j|N=i+j\}\mathbb{P}\{N=i+j\}$$

$$= \left(\binom{i+j}{i} p^i(1-p)^j\right)\left(e^{-a}\frac{a^{i+j}}{(i+j)!}\right)$$

$$= \frac{e^{-a}(ap)^i(aq)^j}{i!j!}$$

$$= \frac{e^{-ap}(ap)^i}{i!}\frac{e^{-aq}(aq)^j}{j!}$$

because $p+q=1$. Notice that this equals $\mathbb{P}\{L=i\}\mathbb{P}\{M=j\}$ where $L \sim \text{Poisson}(ap)$ and $M \sim \text{Poisson}(a(1-p))$ are independent.

Example

If the sun rose for n consecutive days, what is the probability that it will rise tomorrow, if we know that its rising or not is a Bernoulli random variable?

S_n: number of successes in first n trials

$X \sim \mathcal{U}(0,1)$: probability of success in one trial

$\mathbb{P}\{X \in dp\} = dp$, $p \in [0,1]$

$\mathbb{P}\{S_n = k | X = p\} = \binom{n}{k} p^k (1-p)^{n-k}$

$$\mathbb{P}\{S_{n+1} = n+1 | S_n = n\} = \frac{\mathbb{P}\{S_{n+1} = n+1\}}{\mathbb{P}\{S_n = n\}} = \frac{\int_0^1 p^{n+1} dp}{\int_0^1 p^n dp} = \frac{n+1}{n+2} \approx 1$$

for n large.

What is $\mathbb{P}\{X \in dp | S_n = k\}$?

$$\mathbb{P}\{X \in dp | S_n = k\} = \frac{\mathbb{P}\{X \in dp\}\mathbb{P}\{S_n = k | X = p\}}{\mathbb{P}\{S_n = k\}} = \frac{p^k (1-p)^{n-k}}{\int_0^1 u^k (1-u)^{n-k} du} dp = p^k (1-p)^{n-k} \frac{k!(n-k)!}{(n+1)!}$$

Where the denominator is similar to beta distribution:

$$\begin{aligned}
&\int_0^1 u^k (1-u)^{n-k} du \\
&= \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \int_0^1 \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} u^k (1-u)^{n-k} du \\
&= \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \\
&= \frac{k!(n-k)!}{(n+1)!}
\end{aligned}$$

Theorem

Let X be a random variable and ϕ be a positive bounded deterministic function.

$(X \text{ has pdf } f(x)) \Leftrightarrow (\mathbb{E}[\phi(X)] = \int_x \phi(x) f(x) dx)$

General strategy for finding pdfs for functions of random variables

$$\mathbb{E}[\phi(f(X))] = \int_a^b \phi(f(x))\mathbb{P}\{X \in dx\} = \cdots = \int_c^d \phi(y) g(y) dy$$

Then $g(y)dy = \mathbb{P}\{f(X) \in dy\}$

Example

Suppose $X \sim \text{expon}(\lambda)$

What is the pdf for λX?

$$\mathbb{E}[\lambda X] = \lambda \mathbb{E}[X] = 1$$

$$\text{Var}(\lambda X) = \lambda^2 \text{Var}(X) = 1$$

Let ϕ be an arbitrary positive, bounded, deterministic function.

Letting $y = \lambda x$,

$$\begin{aligned}
\mathbb{E}[\phi(\lambda X)] &= \int_0^\infty \phi(\lambda X)\lambda e^{-\lambda x}dx \\
&= \int_0^\infty \phi(y)e^{-y}dy \\
&= \mathbb{E}[\phi(Y)]
\end{aligned}$$

So pdf of Y is e^{-y}.

Conclusion: $(X \sim \text{expon}(\lambda)) \Leftrightarrow (\lambda X \sim \text{expon}(1))$

Example

Let $X \sim \text{gamma}(\alpha, \lambda)$ with shape parameter α and scale parameter λ.

Let $Y = \lambda X$

Let ϕ be an arbitrary positive, bounded, deterministic function.

$$\begin{aligned}
\mathbb{E}[\phi(Y)] = \mathbb{E}[\phi(\lambda X)] &= \int_0^\infty \phi(\lambda X)f(x)dx \\
&= \int_0^\infty \phi(y)\frac{\lambda e^{-y}y^{\alpha-1}}{\Gamma(\alpha)}\frac{1}{\lambda}dy
\end{aligned}$$

So $Y = \lambda X \sim \text{gamma}(\alpha, 1)$

Conclusion (concept of scale): $(X \sim \text{gamma}(\alpha, \lambda)) \Leftrightarrow (\lambda X \sim \text{gamma}(\alpha, 1))$

Example: Convolution

(Also see §2.5)

Let X and Y be independent, with pdfs f and g

What is pdf of $Z = X + Y$?

Let ϕ be an arbitrary positive bounded function.

Let $\hat{\phi}(X, Y) = \phi(X + Y)$

$$\begin{aligned}
&\mathbb{E}[\phi(Z)] = \mathbb{E}[\phi(X+Y)] = \mathbb{E}[\hat{\phi}(X,Y)] \\
&= \int_{-\infty}^\infty \int_{-\infty}^\infty \hat{\phi}(x,y)\mathbb{P}\{X \in dx, Y \in dy\} \\
&= \int_{-\infty}^\infty \int_{-\infty}^\infty \phi(x+y)f(x)g(y)dydx \\
&= \int_{-\infty}^\infty \int_{-\infty}^\infty \phi(z)f(x)g(z-x)dzdx \\
&= \int_{-\infty}^\infty \phi(z)\left(\int_{-\infty}^\infty f(x)g(z-x)dx\right)dz
\end{aligned}$$

$$= \int_{-\infty}^{\infty} \phi(z)h(z)dz$$

We call $h(z) = \int_{-\infty}^{\infty} f(z)g(z-x)dz$ the **convolution** of f and g, written $f * g$.

Example

Let $X \sim \text{gamma}(\alpha, \lambda)$

Let $Y \sim \text{gamma}(\beta, \lambda)$

Let X and Y be independent

Show that $Z = X + Y$ and $U = \frac{X}{X+Y}$ are independent.

Suffices to show that this is true for $\lambda = 1$; we can scale to any other value for λ.

Let ϕ be an arbitrary, positive, bounded function in two variables, and let $\hat{\phi}(x, y) = \phi(\frac{x}{x+y}, x+y)$.

(Goal: try to get $\iint \phi(u, z)h(u, z)dudz$)

Letting $y = z - x$, and $x = zu$,

$$\mathbb{E}[\phi(U, Z)] = \mathbb{E}[\hat{\phi}(X, Y)] = \int_0^{\infty} \int_0^{\infty} \hat{\phi}(x, y) f(x) g(y) dydx$$
$$= \int_0^{\infty} \int_0^{\infty} \frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)} \frac{e^{-y}y^{\beta-1}}{\Gamma(\beta)} \phi\left(\frac{x}{x+y}, x+y\right) dydx$$
$$= \int_0^{\infty} \int_x^{\infty} \frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)} \frac{e^{-(z-x)}(z-x)^{\beta-1}}{\Gamma(\beta)} \phi\left(\frac{x}{z}, z\right) dzdx$$
$$= \int_0^{\infty} \int_0^{z} \frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)} \frac{e^{-(z-x)}(z-x)^{\beta-1}}{\Gamma(\beta)} \phi\left(\frac{x}{z}, z\right) dxdz$$
$$= \int_0^{\infty} \int_0^{1} \frac{e^{-z}(zu)^{\alpha-1}(z(1-u))^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \phi(u, z)(z\, du)dz$$
$$= \int_0^{\infty} \int_0^{1} \left(\frac{e^{-z}z^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)}\right) \left(\Gamma(\alpha+\beta)\frac{u^{\alpha-1}(1-u)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}\right) \phi(u, z)dudz$$

$Z \sim \text{gamma}(\alpha + \beta, \lambda)$ and $U \sim \text{Beta}(\alpha, \beta)$ are independent.

$\mathbb{E}[X + Y] = \frac{\alpha+\beta}{\lambda}$

How to compute $\mathbb{E}\left[\frac{X}{X+Y}\right]$?

Method 1:

$$\int_0^1 u\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u^{\alpha-1}(1-u)^{\beta-1}du = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}u^{\alpha}(1-u)^{\beta-1}du = \frac{\alpha}{\alpha+\beta}$$

Method 2:

We know that $X + Y$ and $\frac{X}{X+Y}$ are independent.

$$\mathbb{E}\left[\frac{X}{X+Y}\right]\mathbb{E}[X+Y] = \mathbb{E}\left[\frac{X}{X+Y}(X+Y)\right]$$
$$\mathbb{E}\left[\frac{X}{X+Y}\right] = \frac{\mathbb{E}[X]}{\mathbb{E}[X+Y]} = \frac{\alpha/\lambda}{(\alpha+\beta)/\lambda} = \frac{\alpha}{\alpha+\beta}$$

Special case: $\alpha = \beta = \frac{1}{2}$

$\frac{X}{X+Y} \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

Because $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

pdf:

$$\frac{\Gamma(1)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})} u^{-\frac{1}{2}} (1-u)^{-\frac{1}{2}}$$
$$= \frac{1}{\pi\sqrt{u(1-u)}}$$

cdf (let $u = \sin^2 x$):

$$\mathbb{P}\left\{\frac{X}{X+Y} \le t\right\} = \int_0^t \frac{1}{\pi\sqrt{u(1-u)}} du$$
$$= \int_0^{\arcsin\sqrt{t}} \frac{1}{\pi \sin x \cos x} (2\sin x \cos x dx)$$
$$= \frac{2}{\pi} \arcsin\sqrt{t}$$

This is the **arcsine distribution**, which is a special case of the beta distribution (beta $\left(\frac{1}{2}, \frac{1}{2}\right)$)

4 Markov Chains

4.1 Introduction

Let $\{X_n\}_{n\in\{0,1,2,\ldots\}}$ be a stochastic process with state space D, which can be finite or countable, and transition matrix

$$P = [p_{ij}] = [\mathbb{P}\{X_{n+1} = j | X_n = i\}]$$

for $i, j \in D$.

X is a **Markov chain** if

$$\mathbb{P}\{X_{n+1} = j | X_0, \ldots, X_n\} = \mathbb{P}\{X_{n+1} = j | X_n\}$$

X is **time-homogeneous** if $\mathbb{P}\{X_{n+1} = j | X_n = i\}$ is free of n.

We will suppose that D is discrete (countable number of states).

The n-step transition probabilities

$$p_{ij}^{(n)} = \mathbb{P}\{X_{m+n} = j | X_m = i\}$$

are the i, j entries of the matrix P^n (see next section).

Note that

$$\sum_{j\in D} \mathbb{P}\{X_{n+1} = j | X_n = i\} = 1$$

(sum of rows in matrix is 1)

4.2 Chapman-Kolmogorov Equations

Let us denote $\mathbb{P}_i\{F\} = \mathbb{P}\{F|X_0 = i\}$ and $\mathbb{E}_i[F] = \mathbb{E}[F|X_0 = i]$ for any event F.

Ex. $\mathbb{P}_b\{X_1 = c, X_2 = a, X_3 = b\} = p_{bc}p_{ca}p_{ab}$ (step through each X_i).

$$\mathbb{P}_i\{X_2 = k\} = \sum_{j\in D} \mathbb{P}_i\{X_1 = j, X_2 = k\} = \sum_{j\in D} p_{ij}p_{jk} = p_{ik}^{(2)}$$

which is the i,k entry of P^2. Reasoning: we know $X_0 = i$ and $X_2 = k$; we add up the probabilities of all possible values of X_1, which turns out to be the matrix product of the ith row and kth column of P, which is the i,k entry of P^2, which equals $[p_{ij}^{(2)}]$.

In general,

$$\mathbb{P}_i\{X_{m+n} = k\} = \sum_{j\in D} \mathbb{P}_i\{X_n = j, X_{m+2n} = k\} = \sum_{j} p_{ij}^{(m)}p_{jk}^{(n)} = p_{ik}^{(m+n)}$$

which is the i,k entry of P^{m+n}. This is a **Chapman-Kolmogorov** equation.

4.3 Classification of States; Recurrency and Transiency

Definition There is a **path** from i to j if $p_{ij}^{(n)} > 0$ for some $n \geq 0$. We notate this $i \rightsquigarrow j$.

If $i \rightsquigarrow j$ and $j \rightsquigarrow i$, then they are in the same **class**. Note that this is an equivalence relation, so it partitions D into classes.

A Markov chain with only one class is said to be **irreducible**.

Let $N(j)$ be the number of visits to state j over all time $n = \{0, 1, 2, \ldots\}$.

Definition We say j is **recurrent** if $\mathbb{P}\{N(j) = +\infty\} = 1$ (that if the particle is at state j, it will eventually come back to state j). Otherwise, j is **transient**. Note that recurrency of state j also implies that the particle will be in state j infinitely many times, and that transiency of state j implies that it will visit state j only finitely many times.

Ross's definition: For any state i, let f_i denote the probability that, starting in state i, the process will ever reenter state i. If $f_i = 1$, we say i is **recurrent**; if $f_i < 1$, we say i is **transient**.

Theorem If i is recurrent and $i \rightsquigarrow j$, then j is recurrent and $j \rightsquigarrow i$.

> Define "success at trial m" to mean j is visited at least once between trials $m-1$ and m where a trial is the mth time the particle is in state i. The probability of such a success is strictly positive because $i \rightsquigarrow j$ (Murphy's law?). By the Strong Markov Property, test trials are independent Bernoulli trials. There are infinitely many trials because i is recurrent. Then the total number of successes is also infinite with probability 1. Since the number of visits to j is at least as large as the number of successes (can have multiple visits to j per success), then $\mathbb{P}\{N(j) = \infty\} = 1$. Further, if $i \rightsquigarrow j$, then $j \rightsquigarrow i$ since i is recurrent (you reach state j from state i, but must eventually get back to i). Or in other words, $\exists n : \mathbb{P}_j\{X_n = i\} > 0$.

Corollary If $i \rightsquigarrow j$ and $j \rightsquigarrow i$ (i.e., i and j are in the same class) and i is transient, then j is transient.

> Proof: Let i be transient and let i and j be in the same class. For sake of contradiction, let j be recurrent. There is a path from j to i, so by the theorem, i is recurrent, which is a contradiction.

Note that the condition for the corollary is stronger than for the theorem.

Theorem If D is finite and X is irreducible, then by the previous theorem, all states are recurrent. Having all states be transient is impossible, since each state will only be visited a finite amount of times, which would result in a time in which the particle is in no state, which is a contradiction.

However, if D is infinite, it is possible that all states are transient. Since there are an infinite number of states, each state can still be visited a finite number of times (transiency), and time can still continue on infinitely. See example below:

Example

Consider $D = \{0\} \cup \mathbb{N}$ and

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & 0 & \cdots \\ 1-p & 0 & p & 0 & 0 & \cdots \\ 0 & 1-p & 0 & p & 0 & \cdots \\ 0 & 0 & 1-p & 0 & p & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

Then

- $p > \frac{1}{2} \Rightarrow$ all states are transient
- $p < \frac{1}{2} \Rightarrow$ all states are recurrent (because state 0 is recurrent)
- $p = q \Rightarrow$ all states are recurrent (proof omitted)

Theorem More generally, if X is irreducible, either all states are recurrent or all states are transient. Thus, recurrency and transiency are **class properties**.

4.4 Stationary Distribution; Limiting Probabilities

Definition A state i is said to have period d if $p_{ii}^{(n)} = 0$ if n is not divisible by d, and d is the largest integer with this property. Or in other words, the period of a state i can be expressed as

$$d_i = GCF\{n \geq 0 : p_{ii}^{(n)} > 0\}$$

Periodicity is a **class property**, or more specifically, states in the same class have the same period.

Let π be a row vector whose entries are

$$\pi_i = \mathbb{P}\{X_0 = i\}$$

Thus π is the distribution of X_0.

Then

$$\mathbb{P}\{X_n = j\} = \sum_{i \in D} \mathbb{P}\{X_0 = i\}\mathbb{P}\{X_n = j | X_0 = i\} = \sum_{i \in D} \pi_i p_{ij}^{(n)}$$

which is the jth entry in the row vector πP^n.

Suppose $\pi = \pi P$. This implies $\pi = \pi P = \pi P^2 = \cdots$ so $\pi = \pi P^n \ \forall n$. Then

$$\mathbb{P}\{X_n = j\} = \sum_{i \in D} \pi_i p_{ij}^{(n)} = \pi_j$$

Such a π is called a **stationary distribution** for the process $X = \{X_n\}_{n \in \{0,1,\ldots\}}$. A stationary distribution exists if there is at least one recurrent state. Consequently, if a state space is finite, a stationary distribution exists because at least one state must be recurrent.

If D is recurrent and irreducible (for finite D, irreducibility suffices), then there is a *unique* π such that $\pi = \pi P$ and $\sum_{j \in D} \pi_i = 1$. This is because the solution space of $\pi = \pi P$ has dimension 1, that is, if π satisfies the equation, then so does $c\pi$, for any constant c. However, for π to represent a distribution, the sum of its entries must be 1, so normalizing any solution to $\pi = \pi P$ will give you the unique row vector.

Theorem Suppose that D is recurrent and irreducible and that all states are aperiodic (for finite D, irreducibility and aperiodicity suffice). Then

$$\pi_j = \lim_{n \to \infty} \mathbb{P}_i\{X_n = j\} = \lim_{n \to \infty} p_{ij}^{(n)}$$

exists for each j, the limit does not depend on the initial state i, and the row vector π of probabilities π_j is the unique stationary distribution. Note that each row of $\lim\limits_{n \to \infty} P^n$ is π.

Note: by definition, the limiting probability $\mathbb{P}_i\{X_n = j\}$ of a transient state j is 0. If a Markov chain has transient states, remove all rows and columns of the transition matrix that correspond to transient states, and find the stationary distribution of the smaller matrix that results (that is, just consider the recurrent states).

4.5 Rewards

Given $f : D \to \mathbb{R}$, we represent f as a column vector with entries $f(i)$ for each state i. $P^n f$ is a column vector, and let $P^n f(i)$ be the ith entry of $P^n f$. We can think of $f(j)$ as the reward for being in state j. Then the reward at time n is $f(X_n)$.

$$\begin{aligned}
\mathbb{E}_i[f(X_n)] &= \sum_{j \in D} f(j) \mathbb{P}_i\{X_n = j\} \\
&= \sum_{j \in D} f(j) p_{ij}^{(n)} \\
&= P^n f(i) \quad i\text{th entry of } P^n f, \text{a column vector}
\end{aligned}$$

Suppose reward is $f(X_n)$ at time n, and **discount factor** is α^n for $\alpha < 1$. This gives smaller rewards for larger n.

Total discounted reward is $\sum_{n=0}^{\infty} \alpha^n f(X_n)$

$$\begin{aligned}
\mathbb{E}_i\left[\sum_{n=0}^{\infty} \alpha^n f(X_n)\right] &= \sum_{n=0}^{\infty} \alpha^n \mathbb{E}_i[f(X_n)] \\
&= \sum_{n=0}^{\infty} \alpha^n P^n f(i)
\end{aligned}$$

Lettting $g(i) = \mathbb{E}_i\left[\sum_{n=0}^{\infty} \alpha^n f(X_n)\right]$ and g being a column vector with entries $g(i)$,

$$\begin{aligned}
g &= f + \alpha P f + \alpha^2 P^2 f + \cdots \\
&= f + \alpha P g
\end{aligned}$$

Then the problem is a system of equations:

$$\begin{bmatrix} g(a) \\ g(b) \\ g(c) \end{bmatrix} = \begin{bmatrix} f(a) \\ f(b) \\ f(c) \end{bmatrix} + \alpha \begin{bmatrix} p_{aa} & p_{ab} & p_{ac} \\ p_{ba} & p_{bb} & p_{bc} \\ p_{ca} & p_{cb} & p_{cc} \end{bmatrix} \begin{bmatrix} g(a) \\ g(b) \\ g(c) \end{bmatrix}$$

4.6 Time Averages

Fix a state j. Let $N_m(j)$ be the number of visits to state j during $[0, m]$. This can be interpreted as a reward function $f(x) = 1$.

$$\mathbb{E}_i[N_m(j)] = \sum_{n=0}^{m} p_{ij}^{(n)}$$

Theorem If X is irreducible recurrent and aperiodic, then $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ where π_j is part of the stationary distribution. Then from the equation above,

$$\lim_{m\to\infty} \frac{1}{m}\mathbb{E}_i[N_m(j)] = \lim_{m\to\infty} \frac{1}{m}\sum_{n=0}^{m} p_{ij}^{(n)} = \lim_{m\to\infty} p_{ij}^{(n)} = \pi_j$$

(See appendix for proof of the middle equality). In other words, the limiting probability π_j is also the long-term average of the expected number of visits to j per unit time.

Theorem Suppose X is irreducible recurrent (but not necessarily aperiodic). Let π be the stationary distribution. Then with probability one,

$$\lim_{m\to\infty} \frac{1}{m} N_m(j) = \pi_j$$

The limiting probability π_j is also the long-time average of the random number of visits to j per unit time. This is a basically a strong law of large numbers.

Proof:

> Use strong law of large numbers: if $L_1, L_2, \ldots$ are i.i.d. with mean μ, then with probability one,
>
> $$\lim_{n\to\infty} \frac{1}{n}(L_1 + \cdots + L_n) = \mu$$
>
> Let L_i be the lengths between successive visits to the fixed state j (i.e., L_i is the length between the ith and $(i+1)$st visit to j).
>
> Since the past and future become independent at each time of visit to j, the lengths $L_0, L_1, \cdots$ are independent, and $L_1, L_2, \cdots$ are i.i.d. and thus have the same mean μ. By the strong law,
>
> $$\lim_{n\to\infty} \frac{1}{n+1}(L_0 + L_1 + \cdots + L_n) = \mu$$
>
> with probability one. In other words, the visits to j occur once every μ time units on the average in the long run. Thus the number of visits to j per unit time is equal to $\frac{1}{\mu}$ in the long run:
>
> $$\lim_{m\to\infty} \frac{1}{m} N_m(j) = \frac{1}{\mu}$$
>
> with probability one. To show that $\frac{1}{\mu} = p_{ij}$, take the expectations of both sides and use the result from the previous theorem.

Corollary

$$\lim_{m\to\infty} \frac{1}{m}\sum_{n=0}^{m} f(X_n) = \sum_{j\in D} \pi_j f(j) = \pi f$$

Proof:

$\sum_{n=0}^{m} f(X_n)$ is the reward accumulated until time m if we receive $f(j)$ dollars each time we are in state j, for all j. So,

$$\sum_{n=0}^{m} f(X_n) = \sum_{j \in D} N_m(j)\ f(j).$$

Applying the earlier theorem, we get the result in the corollary.

The LHS is an average over time, whereas the RHS is an average of f over the state space D. The equality of these two averages is called **the ergodic principle**.

4.7 Transitions

Let $N_m(i,j)$ be the number of transitions from i to j during $[0,m]$. Every time the chain is in state i, there is the probability p_{ij} that the next jump is to state j. At each visit to i, interpret as a Bernoulli trial where "success" means "jumping to state j." Then long-run number of successes per trial is

$$\lim_{m \to \infty} \frac{N_m(i,j)}{N_m(i)} = p_{ij}$$

Theorem With this along with the results from the previous section, we have

$$\lim_{m \to \infty} \frac{1}{m} N_m(i,j) = \lim_{m \to \infty} \frac{1}{m} N_m(i) p_{ij} = \pi_i p_{ij}$$

This is useful if we have a reward $f(i,j)$ that depends on both the present and preceding state (jump from i to j).

Cumulative reward during $[0,m]$:

$$\sum_{n=1}^{m} f(X_{n-1}, X_n) = \sum_i \sum_j N_m(i,j) f(i,j)$$

Long-term reward per unit time:

$$\lim_{m \to \infty} \frac{1}{m} \sum_{n=1}^{m} f(X_{n-1}, X_n) = \sum_i \sum_j \pi_i p_{ij} f(i,j)$$

4.8 Rewards up to a random time

Let a particle move from state to state in D according to a Markov chain. It receives a reward of $f(j)$ each time it is in state j. Each dollar of time n is worth α^n dollars today, at time 0. The particle "dies" (stops receiving rewards) as soon as it enters A, where A is some fixed subset of D. The expected value of the present worth of all rewards we are to receive, given the initial state is i, is

$$g(i) = \mathbb{E}_i \sum_{n=0}^{T-1} \alpha^n f(X_n)$$

where $T = T_A$ is the time of the first visit to the set A, i.e.

$$T = \min\{n \geq 0 : X_n \in A\}$$

If $i \in A$, then $g(i) = 0$, otherwise:

Theorem Let A be a subset of D, let B be the complement of A. Suppose B is finite. Then for $i \in B$,

$$g(i) = f(i) + \sum_{j \in B} \alpha p_{ij} g(j).$$

Proof:

> Suppose $i \in B$. Then $T \geq 1$ and we receive a reward of $f(i)$ dollars at time 0 and the discounted expected value of all rewards to be received from time 1 on is equal to $g(X_1)$ dollars in time 1 dollars. Since $g(j) = 0$ for $j \in A$, the expected value of $g(X_1)$ dollars of time 1 is equal to $\sum_{j \in B} \alpha p_{ij} g(j)$ at time 0.

5 The Exponential Distribution and the Poisson Process

5.1 Introduction

5.2 The Exponential Distribution

5.2.1 Definition

pdf: $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

cdf: $F(x) = \int_{-\infty}^{x} f(y)dy = 1 - e^{-\lambda x}$, $x \geq 0$

Using integration by parts, where $u = x$ and $dv = \lambda e^{-\lambda x}$:

$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^\infty \lambda x e^{-\lambda x}dx = -xe^{-\lambda x}\big|_0^\infty + \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}$

$\phi(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \lambda \int_0^\infty e^{-(\lambda - t)x}dx = \frac{\lambda}{\lambda - t}$ for $t < \lambda$

$\mathbb{E}[X^2] = \frac{d^2}{dt^2}\phi(t)\Big|_{t=0} = \frac{2\lambda}{(\lambda-t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}$

$\text{Var}(X) = \frac{1}{\lambda^2}$

Example Expected discounted return is equal to reward earned up to an exponentially distributed random time (see book)

5.2.2 Properties of Exponential Distribution

Definition A random variable X is **memoryless** if

$$\mathbb{P}\{X > s + t | X > t\} = \mathbb{P}\{X > s\}, (\forall s, t \geq 0)$$

$$\Leftrightarrow \mathbb{P}\{X > s + t, X > t\} = \mathbb{P}\{X > s\}\mathbb{P}\{X > t\}$$

Since $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$, exponentially distributed random variables are memoryless.

See books for example problems on memorylessness of exponential

Definition: $a^+ = a$ if $a > 0$, and $a^+ = 0$ if $a \leq 0$.

Claim: The only *right continuous* function g that satisfies $g(s+t) = g(s)g(t)$ is $g(x) = e^{-\lambda x}$

Proof:

> Suppose $g(s+t) = g(s)g(t)$.
>
> $g(2/n) = g(1/n + 1/n) = (g(1/n))^2$
>
> Repeating yields $g(m/n) = (g(1/n))^m$
>
> Also, $g(1) = g(1/n + \cdots + 1/n) = (g(1/n))^n$

$g(1/n) = (g(1))^{1/n}$

$g(m/n) = (g(1/n))^m = (g(1))^{m/n}$

By the right continuity of g, we then have $g(x) = (g(1))^x$

$g(1) = (g(1/n))^2 \geq 0$

$g(x) = e^{-(-\log(g(1)))x}$

Definition: failure/hazard rate function is $r(t) = \frac{f(t)}{1-F(t)}$

Suppose lifetime X has survived for t; what is the probability it does not survive for additional time dt?

$\mathbb{P}\{X \in (t, t+dt) | X > t\} = \frac{\mathbb{P}\{X \in (t,t+dt), X>t\}}{\mathbb{P}\{X>t\}} = \frac{\mathbb{P}\{X \in (t,t+dt)\}}{\mathbb{P}\{X>t\}} \approx \frac{f(t)dt}{1-F(t)} = r(t)dt$

If $X \sim \text{expon}(\lambda)$, then $r(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$

$r(t)$ uniquely determines distribution F:

$r(t) = \frac{\frac{d}{dt}F(t)}{1-F(t)}$

Integrate both sides:

$\log(1-F(t)) = -\int_0^t r(t)dt + k$

$1 - F(t) = e^k \exp\left\{-\int_0^t r(t)dt\right\}$

Setting $t = 0$ shows that $k = 0$

$F(t) = 1 - \exp\left\{-\int_0^t r(t)dt\right\}$

Claim: exponential random variables are the only ones that are memoryless

Proof: We showed above that memoryless is equivalent to having a constant failure rate function, and that exponential random variables have a constant failure rate function. If a failure rate function is constant, then by the equation above, $1 - F(t) = e^{-ct}$, which shows that it must be exponential.

See book for example on **hyperexponential** random variable.

5.2.3 Further Properties of the Exponential Distribution

Sum of i.i.d. exponential random variables is gamma

Let $X_1, \ldots, X_n$ be i.i.d., $X_i \sim \text{expon}(\lambda), \forall i \in \{1, \ldots, n\}$

Nothing to prove when $n = 1$.

Assume $f_{X_1+\cdots+X_{n-1}}(t) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-2}}{(n-2)!}$

Then

$$\begin{aligned}
f_{X_1+\cdots+X_n}(t) &= \int_0^\infty f_{X_n}(t-s) f_{X_1+\cdots+X_{n-1}}(s)ds \\
&= \int_0^t \lambda e^{-\lambda(t-s)} \lambda e^{-\lambda s}\frac{(\lambda s)^{n-2}}{(n-2)!}ds \\
&= \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}
\end{aligned}$$

Probability that an exponential variable is less than another

Let $X_1 \sim \text{expon}(\lambda_1)$ and $X_2 \sim \text{expon}(\lambda_2)$

$$\begin{aligned}
\mathbb{P}\{X_1 < X_2\} &= \int_0^\infty \mathbb{P}\{X_1 < X_2 | X_1 = x\} \lambda_1 e^{-\lambda_1 x} dx \\
&= \int_0^\infty \mathbb{P}\{x < X_2\} \lambda_1 e^{-\lambda_1 x} dx \\
&= \int_0^\infty e^{-\lambda_2 x} \lambda_1 e^{-\lambda_1 x} dx \\
&= \int_0^\infty \lambda_1 e^{-(\lambda_1 + \lambda_2) x} dx \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}
\end{aligned}$$

Smallest of independent exponential random variables

Suppose $X_1, \ldots, X_n$ are indep. exponential random variables, $X_i \sim \text{expon}(\mu_i)$.

$$\begin{aligned}
\mathbb{P}\{min(X_1, \ldots, X_n) > x\} &= \mathbb{P}\{X_1 > x, X_2 > x, \ldots, X_n > x\} \\
&= \prod_{i=1}^n \mathbb{P}\{X_i > x\} \\
&= \prod_{i=1}^n e^{-\mu_i x} \\
&= \exp\left\{ -\left(\sum_{i=1}^n \mu_i \right) x \right\}
\end{aligned}$$

Thus, $min(X_1, \ldots, X_n) \sim \text{expon}\left(\sum_{i=1}^n \mu_i\right)$

Example Greedy Algorithm (see book)

5.3 The Poisson Process

5.3.1 Counting Processes

Definition A stochastic process $\{N(t), t \geq 0\}$ is a **counting process** if $N(t)$ represents the total number of events that occur by time t. Then these must hold:

- $N(t) \geq 0$
- $N(t) \in \mathbb{Z}$
- $(s < t) \Rightarrow (N(s) \leq N(t))$
- For $s < t$, but $N(t) - N(s)$ is the number of events in $(s, t]$

Definition A counting process has **independent increments** if the number of events that occur in disjoint time intervals are independent.

Definition A counting process has **stationary increments** if the distribution of the number f events that occur in any interval of time depends only on the length of the interval.

5.3.2 Definition of Poisson Process

Definition Counting process $\{N(t), t \geq 0\}$ is a **Poisson process** having rate $\lambda > 0$ if

- $N(0) = 0$
- independent increments
- $N(t+s) - N(s) \sim \text{Pois}(\lambda t)$ which implies stationary increments and that $\mathbb{E}[N(t)] = \lambda t$

Definition The function f is said to be $o(h)$ if $\lim_{h \to 0} \frac{f(h)}{h} = 0$

Definition Counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate $\lambda > 0$ if

- $N(0) = 0$
- stationary and independent increments
- $\mathbb{P}\{N(h) = 1\} = \lambda h + o(h)$
- $\mathbb{P}\{N(h) \geq 2\} = o(h)$

Theorem The two definitions of Poisson process are equivalent.

Proof in book

5.3.3 Interarrival and Waiting Time Distributions

Let T_1 be the time of the first event. Let T_n be the time between the $(n-1)$st and the nth event for $n > 1$. We call $\{T_n\}$ the **sequence of interarrival times**

$\mathbb{P}\{T_1 > t\} = \mathbb{P}\{N(t) = 0\} = e^{-\lambda t}$ so $T_1 \sim \text{expon}(\lambda)$

$\mathbb{P}\{T_2 > t\} = \mathbb{E}[\mathbb{P}\{T_2 > t | T_1\}] = \mathbb{E}[\mathbb{P}\{0 \text{ events in } (s, s+t] | T_1\}] = \mathbb{E}[\mathbb{P}\{0 \text{ events in } (s, s+t]\}] = e^{-\lambda t}$

$T_2 \sim \text{expon}(\lambda)$, and T_2 is indep. of T_1

Proposition: $T_n \sim \text{expon}(\lambda), \forall n \in \mathbb{N}$

We call $S_n = \sum_{i=1}^{n} T_i$ the **waiting time** until the nth event. By this proposition, and the result from §5.2.3 and §2.2, $S_n \sim \text{gamma}(n, \lambda)$, that is

$$f_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}$$

Alternate method:

$N(t) \geq n \Leftrightarrow S_n \leq t$

$$F_{S_n}(t) = \mathbb{P}\{S_n \leq t\} = \mathbb{P}\{N(t) > n\} = \sum_{j=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

differentiate:

$$
\begin{aligned}
f_{S_n}(t) &= -\sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} + \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} \\
&= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} + \sum_{j=n+1}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\
&= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}
\end{aligned}
$$

5.3.4 Further Properties of Poisson Processes

Suppose each event in Poisson process $\{N(t), t \geq 0\}$ with rate λ is classified as either type I (with probability p) or type II (with probability $1-p$). Let $N_1(t)$ and $N_2(t)$ denote the number of type I and type II events occurring in $[0, t]$. Then $N(t) = N_1(t) + N_2(t)$.

Proposition $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes having respective rates λp and $\lambda(1-p)$.
Proof:

Verify that $\{N_1(t), t \geq 0\}$ satisfies the [second] definition of Poisson process:

- $N_1(0) = 0$ follows from $N(0) = 0$
- stationarity and independence of increments is inherited because the number of type I events in an interval can be obtained by conditioning on the number of events in that interval, and the distribution of the number of events in that interval depends only on the length of the interval and is independent of what has occurred in any nonoverlapping interval
- $$
\begin{aligned}
\mathbb{P}\{N_1(h) = 1\} &= \mathbb{P}\{N_1(h) = 1 | N(h) = 1\}\mathbb{P}\{N(h) = 1\} + \mathbb{P}\{N_1(h) = 1 | N(h) \geq 2\}\mathbb{P}\{N(h) \geq 2\} \\
&= p(\lambda h + o(h)) + o(h) \\
&= \lambda p h + o(h)
\end{aligned}
$$
- $\mathbb{P}\{N_1(h) \geq 2\} \leq \mathbb{P}\{N(h) \geq 2\} = o(h)$

So, $\{N_1(t), t \geq 0\}$ is a Poisson process with rate λp. Similarly, $\{N_2(t), t \geq 0\}$ is a Poisson process with rate $\lambda(1-p)$

See §3.8 (or example 3.23 in the book) for why they are independent.

See book for excellent examples

5.3.5 Conditional Distribution of the Arrival Times

Suppose $N(t) = 1$. What is T_1?

$$\begin{aligned}
\mathbb{P}\{T_1 < s | N(t) = 1\} &= \frac{\mathbb{P}\{T_1 < s, N(t) = 1\}}{\mathbb{P}\{N(t) = 1\}} \\
&= \frac{\mathbb{P}\{N(s) = 1\}\mathbb{P}\{N(t) - N(s) = 0\}}{\mathbb{P}\{N(t) = 1\}} \\
&= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\
&= \frac{s}{t}
\end{aligned}$$

So $T_1 \sim \mathcal{U}(0, t)$

see book for order statistics

Theorem (Dart Theorem) Given $N(t) = n$, the n arrival times $S_1, \ldots S_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval $(0, t)$. Phrased differently, this theorem states that under the condition that $N(t) = n$, the times $S_1, \ldots, S_n$ at which events occur, considered as unordered random variables, are distributed independently and uniformly in the interval $(0, t)$.

Proposition If $N_i(t)$, $i \in \{1, \ldots, k\}$ represents the number of type i events occurring by time t, then $N_i(t)$, $i \in \{1, \ldots, k\}$ are independent Poisson random variables having means $\mathbb{E}[N_i(t)] = \lambda \int_0^t P_i(s)ds$ where $P_i(t)$ is hte probability that an event occurring t time y will be classified as type i, independently of anything that previously occurred.

5.3.6 Estimating Software Reliability

m is the number of bugs in a package

Bug i will cause errors to occur in accordance with a Poisson process having unknown rate λ_i, $i \in \{1, \ldots, m\}$.

Then, number of errors due to bug i that occurs in any s units of operating time is $\sim$ Pois$(\lambda_i s)$

The Poisson processes caused by bugs i for $i \in \{1, \ldots, m\}$ are independent

Let program run, then after time t, remove all bugs that caused an error by time t. Then the only bugs that remain are those whose errors all occur after t.

What is error rate of the revised package?

Let $\phi_i(t) = \begin{cases} 1, & \text{bug } i \text{ has not caused an error by } t \\ 0, & \text{otherwise} \end{cases}$

The error rate of the revised package is $\Lambda(t) = \sum_i \lambda_i \phi_i(t)$ (add up all the rates of Poisson processes associated with bugs that haven't been caught yet)

$\mathbb{E}[\Lambda(t)] = \sum_i \lambda_i \mathbb{E}[\phi_i(t)] = \sum_I \lambda_i e^{-\lambda_i t}$

Each discovered bug is responsible for certain number of errors. Denote by $M_j(t)$ the number of bugs that caused exactly j errors by time t. (So, $M_1(t)$ is the number of bugs that caused exactly one error, $M_2(t)$ the number of bugs that caused exactly two errors, etc.) Then $\sum_j jM_j(t)$ is the total number of errors found before time t.

Let $I_i(t) = \begin{cases} 1, & \text{bug } i \text{ causes exactly 1 error by time } t \\ 0, & \text{otherwise} \end{cases}$

Then $M_1(t) = \sum_i I_i(t)$

$\mathbb{E}[M_1(t)] = \sum_i \mathbb{E}[I_i(t)] = \sum_i \lambda_i t e^{-\lambda_i t}$

So, then we have $\mathbb{E}\left[\Lambda(t) - \frac{M_1(t)}{t}\right] = 0$

$$\begin{aligned}
\mathrm{Var}(\Lambda(t)) &= \sum_i \lambda_i^2 \mathrm{Var}(\phi_i(t)) = \sum_i \lambda_i^2 e^{-\lambda_i t}(1 - e^{-\lambda_i t}) \\
\mathrm{Var}(M_1(t)) &= \sum_i \mathrm{Var}(I_i(t)) = \sum_i \lambda_i t e^{-\lambda_i t}(1 - \lambda_i t e^{-\lambda_i t})
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\Lambda(t), M_1(t)) &= \mathrm{Cov}\left(\sum_i \lambda_i \phi_i(t), \sum_j I_j(t)\right) \\
&= \sum_i \sum_j \mathrm{Cov}(\lambda_i \phi_i(t), I_j(t)) \\
&= \sum_i \lambda_i \mathrm{Cov}(\phi_i(t), I_i(t)) \\
&= -\sum_i \lambda_i e^{-\lambda_i t} \lambda_i t e^{-\lambda_i t}
\end{aligned}$$

Where the last two equalities follow because $(i \neq j) \Rightarrow (\phi_i(t)$ and $I_j(t)$ are independent).

Then

$$\mathrm{Var}\left(\Lambda(t) - \frac{M_1(t)}{t}\right) = \mathbb{E}\left[\left(\Lambda(t) - \frac{M_1(t)}{t}\right)^2\right] = \sum_i \lambda_i^2 e^{-\lambda_i t} + \frac{1}{t}\sum_i \lambda_i e^{-\lambda_i t} = \frac{\mathbb{E}[M_1(t) + 2M_2(t)]}{t^2}$$

where we use $\mathbb{E}[M_2(t)] = \frac{1}{2}\sum_i (\lambda - it)^2 e^{-\lambda_i t}$

5.4 Generalizations of the Poisson Process

5.4.1 Nonhomogeneous Poisson Process

5.4.2 Compound Poisson Process

Definition A stochastic process $\{X(t), t \geq 0\}$ is said to be a **compound Poisson process** if it can be represented as

$$X(t) = \sum_{i=1}^{N(t)} Y_i, \quad t \geq 0$$

where $\{N(t)\}$ is a Poisson process, and $\{Y_i, i \geq 1\}$ is a family of i.i.d. random variables, independent of $\{N(t)\}$.

5.4.3 Conditional or Mixed Poisson Processes

Definition Let $\{N(t)\}$ be a counting process whose probabilities are defined as follows: there is a poisitive random variable L such that, conditional on $L = \lambda$, the counting prcess is a Poisson process with rate λ. This counting process is called a **conditional** or **mixed Poisson process**. Such a process has staionary increments, but generally does not have independent increments.

5.5 Extra stuff

Stronger version of memorylessness

We showed earlier that if $X \sim \text{expon}(\lambda)$, then

$$\mathbb{P}\{X > t + s | X > t\} = \mathbb{P}\{X > s\}$$

We want to show that for nonnegative random variables Y and Z, independent of each other and of X,

$$\mathbb{P}\{X > Y + Z | X > Y\} = \mathbb{P}\{X > Z\}$$

$$\begin{aligned}
\mathbb{P}\{X > Y + Z | X > Y\} &= \int_0^\infty \int_0^\infty \mathbb{P}\{X > Y + Z | X > Y, Y = y, Z + z\}\mathbb{P}\{Y \in dy\}\mathbb{P}\{Z \in dz\} \\
&= \int_0^\infty \int_0^\infty \mathbb{P}\{X > y + z | X > y\}\mathbb{P}\{Y \in dy\}\mathbb{P}\{Z \in dz\} \\
&= \int_0^\infty \int_0^\infty \mathbb{P}\{X > z\}\mathbb{P}\{Y \in dy\}\mathbb{P}\{Z \in dz\} \\
&= \mathbb{P}\{X > Z\}
\end{aligned}$$

Alternate way to view random (geometric) sum of i.i.d. exponential random variables

Let $X_i \sim \text{expon}(\nu)$, $\forall i \in \{1, 2, \ldots\}$

Let $\mathbb{P}\{K = n\} = (1-p)^{n-1}p$

Let $S = \sum_{j=1}^K X_j$

instead of conditioning on K, think of the X_i as the inter arrival times in Poisson process with rate ν, where each arrival is either a success with probability p or a failure with probability $1-p$.

Then S is the arrival time of the first success. $S \sim \text{expon}(p\nu)$

Superposition of two Poisson processes

Let L be a Poisson process with rate λ

Let M be a Poisson process with rate μ, independent of L

Then superposition process N (defined as $N_t = L_t + M_t$) is Poisson process with rate $\lambda + \mu$

$\mathbb{P}\{\text{a given arrival in the superposition process is from } L\} = \frac{\lambda}{\lambda+\mu}$

$\mathbb{P}\{\text{a given arrival in the superposition process is from } M\} = \frac{\mu}{\lambda+\mu}$

Alternate way to compute $\mathbb{P}\{X < Y\}$ when $X \sim$ expon(λ) and $Y \sim$ expon(μ)

Think of X as the time of the first arrival in a Poisson process with rate λ.

Think of Y as the time of the first arrival in a Poisson process with rate μ.

$\mathbb{P}\{X < Y\} = \mathbb{P}\{\text{first arrival in superposition process is from the first process}\} = \frac{\lambda}{\lambda+\mu}$

Decomposition of two Poisson processes

Each arrival in Poisson process N with rate λ has probability p_k of being "type" k, with $\sum_{k=1}^{n} p_k = 1$.

We can decompose into independent Poisson processes of only arrivals of type k, call it $N^{(k)}$ with rate $p_k\lambda$.

Non-stationary Poisson processes

is a counting process with independent increments, but **not** stationary increments

λ_t is the rate per unit time at t

$$\mathbb{P}\{N_A = k\} = e^{-\mu(A)} \frac{(\mu(A))^k}{k!}$$

where $\mu(A) = \int_A \lambda_u du$

Intuition: plot λ_t over t, then for an infinitesimally small time interval, the function is approximately constant, and the mean of the Poisson distribution of arrivals in this interval is the area under the curve.

6 Continuous-Time Markov Chains

6.1 Introduction

Poisson process (where $N(t)$ is the state of the process) is a continuous -time Markov chain with states $\{0, 1, 2, \ldots\}$ that always proceeds from state n to $n+1$ where $n \geq 0$.

A process is a **pure birth process** if the state of the system is always increased by one in any transition.

A process is a **birth and death model** if the state of the system is $n+1$ or $n-1$ after a transition from state n.

6.2 Continuous-Time Markov Chains

Definition A continuous-time stochastic process $\{X(t), t \geq 0\}$ whose state space is $\{0, 1, 2, \ldots\}$ is a **continuous-time Markov Chain** if for all $s, t \geq 0$¡ and nonnegative integers $i, j, x(u), 0 \leq u < s$,

$$\mathbb{P}\{X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s\} = \mathbb{P}\{X(t+s) = j | X(s) = i\}$$

or in other words, it has the Markovian property that the conditional distribution of the future $X(t+s)$ given the present $X(s)$ and the past $X(u), 0 \leq u < s$ depends only on the present and is independent of the past.

If in addition,

$$\mathbb{P}\{X(t+s) = j | X(s) = i\}$$

is independent of s, then the continuous-time Markov chain is said to have stationary or homogeneous transition probabilities. All Markov chains considered here will be assumed to have stationary transition probabilities.

If we let T_i denote the amount of time that the process stays in state i before making transition into a different state, then

$$\mathbb{P}\{T_i > s + t | T_i > s\} = \mathbb{P}\{T_i > t\}$$

Therefore T_i is memoryless and must be exponentially distributed.

Definition An alternate way to define a continuous-time Markov chain: a stochastic process that has the properties that as each time it enters state i, the amount of timeit spendsin the state before transitioning into a different state is exponentially distributed with mean $1/v_i$ and when the process leaves state i, it enters state j with home probability P_{ij} which satisfies, for all i, that $P_{ij} = 0$ and $\sum_j P_{ij} = 1$.

In other words, a continuous-time Markov chain is a stochastic process that moves from state to state in accordance with a discrete-time Markov chain, and the amount of time spent in each state before proceeding to the next state is exponentially distributed. Additionally, the amount of time the process spends in state i and in the next state visited must be independent (Markov property).

6.3 Birth and Death Processes

6.4 The Transition Probability Function $P_{ij}(t)$

6.5 Limiting Probabilities

6.6 Time Reversibility

6.7 Uniformization

6.8 Computing the Transition Probabilities

7

8

9 Reliability Theory

9.1 Introduction

Reliability theory: probability that a system will function

9.2 Structure Functions

system of n components, each component is either functioning or failed, indicator variable x_i for the ith component:

$$x_i = \begin{cases} 1, & i\text{th component functioning} \\ 0, & i\text{th component failed} \end{cases}$$

state vector:

$$\mathbf{x} = (x_1, \ldots, x_n)$$

structure function of the system:

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{system is functioning when state vector is } \mathbf{x} \\ 0, & \text{system failed when state vector is } \mathbf{x} \end{cases}$$

series structure functions iff all components function:

$$\phi(\mathbf{x}) = \min(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i$$

parallel structure functions iff at least one component is functioning:

$$\phi(\mathbf{x}) = \max(x_1, \ldots, x_n) = 1 - \prod_{i=1}^{n}(1 - x_i)$$

k**-out-of-**n **structure** functions off at least k of the n components are functioning:

$$\phi(\mathbf{x}) = \begin{cases} 1, & \sum_{i=1}^{n} x_i \geq k \\ 0, & \sum_{i=1}^{n} x_i < k \end{cases}$$

Ex: four-component structure; 1 and 2 both function, at least one of 3 and 4 function

$$\phi(\mathbf{x}) = x_1 x_2 \max(x_3, x_4) = x_1 x_2 (1 - (1 - x_1)(1 - x_2))$$

$\phi(\mathbf{x})$ is an increasing function of $\mathbf{x}$; replacing a failed component by a functioning one will never lead to a deterioration of the system

i.e., $x_i \leq y_i, \ \forall i \in \{1, \ldots, n\} \Rightarrow \phi(\mathbf{x}) \leq \phi(\mathbf{y})$

a system is thus **monotone**

9.2.1 Minimal Path and Minimal Cut Sets

$\mathbf{x}$ is a **path vector** if $\phi(\mathbf{x}) = 1$

$\mathbf{x}$ is a **minimal path vector** if $\phi(\mathbf{y}) = 0$ for all $\mathbf{y} < \mathbf{x}$

note: $\mathbf{y} < \mathbf{x}$ if $y_i \leq x_i, \ \forall i \in \{1, \ldots, n\}$ with $y_i < x_i$ for some i

If $\mathbf{x}$ is a minimal path vector, then $A = \{i : x_i = 1\}$ is a **minimal path set**; minimal set of components whose functioning ensures the system's functioning

In a k-out-of-n system, there are $\binom{n}{k}$ minimal path sets (all sets consisting of exactly k components)

let $A_1, \ldots, A_s$ denote minimal path sets. Define $\alpha_j(\mathbf{x})$ as the indicator function of the jth minimal path set:

$$\alpha_j(\mathbf{x}) = \begin{cases} 1, & \text{all components of } A_j \text{ are functioning} \\ 0, & \text{otherwise} \end{cases} = \prod_{i \in A_j} x_i$$

A system functions iff all components of at least one minimal path set are functioning:

$$\phi(\mathbf{x}) = \begin{cases} 1, & \alpha_j(\mathbf{x}) = 1 \text{ for some } j \\ 0, & \alpha_j(\mathbf{x}) = 0 \text{ for all } j \end{cases} = \max{}_j \alpha_j(\mathbf{x}) = \max{}_j \prod_{i \in A_j} x_i$$

$\mathbf{x}$ is a **cut vector** if $\phi(\mathbf{x}) = 0$

$\mathbf{x}$ is a **minimal cut vector** if $phi(\mathbf{y}) = 1$ for all $\mathbf{y} > \mathbf{x}$

If $\mathbf{x}$ is a minimal cut vector, then $C = \{i : x_i = 0\}$ is a **minimal cut set**; a minimal set of components whose failure ensures the failure of the system let $C_1, \dots, C_k$ denote minimal cut sets. Define $\beta_j(\mathbf{x})$ as the indicator function of the jth minimal cut set:

$$\beta(\mathbf{x}) = \begin{cases} 1, & \text{at least one component of } C_j \text{ is functioning} \\ 0, & \text{all components in } C_j \text{ are not functioning} \end{cases} = \max_{i \in C_j} x_i$$

A system fails iff all components of at least one minimal cut set are not functioning:

$$\phi(\mathbf{x}) = \prod_{j=1}^{k} \beta_j(\mathbf{x}) = \prod_{j=1}^{k} \max_{i \in C_j} x_i$$

9.3 Reliability of Systems of Independent Components

X_i, the state of the ith component is a random variable such that

$$\mathbb{P}\{X_i = 1\} = p_i = 1 - \mathbb{P}\{X_i = 0\}$$

where p_i is the **reliability** of the ith component.

Note that

$$\phi(\mathbf{X}) = X_i \phi(1_i, \mathbf{X}) + (1 - X_i)\phi(0_i, \mathbf{X})$$

(see below for explanation of notation).

Define the **reliability** of the system r by

$$r := \mathbb{P}\{\phi(\mathbf{X}) = 1\} = \mathbb{E}[\phi(\mathbf{X})]$$

where $\mathbf{X} = (X_1, \dots, X_n)$.

Define **reliability function** $r(P) = r$ where $P = (p_1, \dots, p_n)$

reliability of a series system:

$$r(P) = \mathbb{P}\{\phi(\mathbf{X}) = 1\} = \mathbb{P}\{X_i = 1, \ \forall i =\in \{1, \dots, n\}\} = \prod_{i=1}^{n} p_i$$

reliability of a parallel system:

$$r(P) = \mathbb{P}\{\phi(\mathbf{X}) = 1\} = \mathbb{P}\{X_i = 1 \text{ for some } i \in \{1, \dots, n\}\} = 1 - \prod_{i=1}^{n}(1 - p_1)$$

reliability of a k-out-of-n system with equal probabilities:

$$r(p, \dots, p) = \mathbb{P}\{\phi(\mathbf{X}) = 1\} = \mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq k\right\} = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Theorem If $r(P)$ is the reliability function of a system of independent components, then $r(P)$ is an increasing function of P.

Proof:

$$\begin{aligned} r(P) &= \mathbb{E}[\phi(\mathbf{X})] = p_i \mathbb{E}[\phi(\mathbf{X}) | X_i = 1] + (1 - p_i)\mathbb{E}[\phi(\mathbf{X}) | X_i = 0] \\ &= p_i \mathbb{E}[\phi(1_i, \mathbf{X})] + (1 - p_i)\mathbb{E}[\phi(0_i, \mathbf{X})] \end{aligned}$$

where $(1_i, \mathbf{X}) = (X_1, \ldots, X_{i-1}, 1, X_{i+1}, \ldots, X_n)$ and $(0_i, \mathbf{X}) = (X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)$ so,

$$r(P) = p_i \mathbb{E}[\phi(1_i, \mathbf{X}) - \phi(0_i, \mathbf{X})] + \mathbb{E}[\phi(0_i, \mathbf{X})]$$

but since ϕ is an increasing function, $\mathbb{E}[\phi(1_i, \mathbf{X}) - \phi(0_i, \mathbf{X})] \geq 0$ so $r(P)$ increases in p_i for all i.

Notation: $\mathbf{xy}$ denotes inner product; $\max(\mathbf{x}, \mathbf{y}) = (\max(x_1, y_1), \ldots, \max(x_n, y_n))$

Warning $\phi(\mathbf{X})$ is a function of $X_1, \ldots, X_n$. To find $\mathbb{E}[\phi(\mathbf{X})]$, you cannot just replace X_i with p_i, because, for example, $p_1^2 \neq \mathbb{E}[X_1^2] = \mathbb{E}[X_1] = p_1$. You can only replace the X_i with p_i if it is not being multiplied by itself. See bridge problem in homework for details.

Suppose a system of n different components is to be built from a stockpile containing exactly 2 of each type of component. Which is better?

- Build 2 separate systems, then

$$\mathbb{P}\{\text{at least one system functions}\} = 1 - \mathbb{P}\{\text{neither system functions}\} = 1 - ((1 - r(P))(1 - r(P')))$$

- Build one system, then

$$\mathbb{P}\{\phi(\mathbf{X}) = 1\} = r[\mathbf{1} - (\mathbf{1} - P)(\mathbf{1} - P')]$$

since $1 - (1 - p_i)(1 - p_i')$ is the probability that the ith component functions

Theorem For any reliability function r and vectors P, P',

$$r[\mathbf{1} - (\mathbf{1} - P)(\mathbf{1} - P)] \geq 1 - [1 - r(P)][1 - r(P')]$$

or equivalently

$$\mathbb{E}[\phi(\max(\mathbf{X}, \mathbf{X}'))] \geq \mathbb{E}[\max(\phi(\mathbf{X}), \phi(\mathbf{X}'))]$$

Proof:

Let $X_1, \ldots, X_n, X_1', \ldots, X_n'$ be mutually independent 0-1 random variables with

$$p_i = \mathbb{P}\{X_i = 1\};\ p_i' = \mathbb{P}\{X_i' = 1\}$$

$$\mathbb{P}\{\max(X_i, X_i') = 1\} = 1 - (1 - p_i)(1 - p_i')$$

$$\Rightarrow r[\mathbf{1} - (\mathbf{1} - P)(\mathbf{1} - P')] = E(\phi[\max(\mathbf{X}, \mathbf{X}')])$$

Because ϕ is monotonically increasing, $\phi[\max(\mathbf{X}, \mathbf{X}')]$ greater than or equal to both $\phi(\mathbf{X})$ and $\phi(\mathbf{X}')$, so

$$\phi[\max(\mathbf{X}, \mathbf{X}')] \geq \max[\phi(\mathbf{X}), \phi(\mathbf{X}')]$$

So,

$$r[\mathbf{1} - (\mathbf{1} - P)(\mathbf{1} - P')] \geq \mathbb{E}[\max[\phi(\mathbf{X}), \phi(\mathbf{X}')]] = \mathbb{P}\{\max[\phi(\mathbf{X}), \phi(\mathbf{X}')] = 1\} = 1 - [1 - r(P)][1 - r(P')]$$

9.4 Bounds on the Reliability Function

9.5 System Life as a Function of Component Lives

For a distribution function G, define $\overline{G}(a) \equiv 1 - G(a)$ as the probability that the random variable is greater than a.

The ith component functions for a random length of time (distribution F_i), then fails.

Let F denote the distribution of system lifetime, then

$$\overline{F}(t) = \mathbb{P}\{\text{system life } > t\} = \mathbb{P}\{\text{system is functioning at time } t\} = r(P_1(t), \dots, P_n(t))$$

where

$$P_i(t) = \mathbb{P}\{\text{component } i \text{ is functioning at } t\} = \mathbb{P}\{\text{lifetime of } i > t\} = \overline{F}_i(t)$$

So,

$$\overline{F}(t) = r(\overline{F}_1(t), \dots, \overline{F}_n(t))$$

system life in a series system: $r(P) = \prod_{i=1}^n p_i$, so

$$\overline{F}(t) = \prod_{i=1}^n \overline{F}_i(t)$$

system life in a parallel system: $r(P) = 1 - \prod_{i=1}^n (1 - p_i)$, so

$$\overline{F}(t) = 1 - \prod_{i=1}^n F_i(t)$$

failure rate function of G:

$$\lambda(t) = \frac{g(t)}{\overline{G}(t)} = \frac{\frac{d}{dt}G(t)}{\overline{G}(t)}$$

where $g(t) = \frac{d}{dt}G(t)$

in §5.2.2, if G is the distribution of the lifetime of an item, then $\lambda(t)$ represents the probability intensity that a t-year-old item will fail

G is an **increasing failure rate (IFR)** distribution if $\lambda(t)$ is an increasing function of t

G is an **decreasing failure rate (DFR)** distribution if $\lambda(t)$ is an decreasing function of t

A random variable has **Weibull distribution** if its distribution is, for some $\lambda > 0$, $\alpha > 0$,

$$G(t) = 1 - e^{-(\lambda t)^\alpha}$$

for $t \geq 0$. The failure rate function is

$$\lambda(t) = \frac{e^{-(\lambda t)^\alpha} \alpha(\lambda t)^{\alpha-1}\lambda}{e^{-(\lambda t)^\alpha}} = \alpha\lambda(\lambda t)^{\alpha-1}$$

Weibull distribution is IFR when $\alpha \geq 1$, and DFR when $0 < \alpha \leq 1$; when $\alpha = 1$, $G(t) = 1 - e^{-\lambda t}$, the exponential distribution (IFR and DFR).

see book for Gamma distribution

Suppose lifetime distribution of each component in a monotone system is IFR. Does this imply that the system lifetime is also IFR? Suppose each component has same lifetime distribution G. That is, $F_i(t) = G(t)$, $\forall i \in \{1, \dots, n\}$. Compute failure rate function of F.

If we define $r(\overline{G}(t)) \equiv r(\overline{G}(t), \ldots, \overline{G}(t))$, then

$$\lambda_F(t) = \frac{\frac{d}{dt}F(t)}{\overline{F}(t)} = \frac{\frac{d}{dt}(1 - r(\overline{G}(t)))}{r(\overline{G}(t))} = \frac{r'(\overline{G}(t))}{r(\overline{G}(t))}G'(t) = \frac{\overline{G}(t)r'(\overline{G}(t))}{r(\overline{G}(t))}\frac{G'(t)}{\overline{G}(t)} = \lambda_G(t)\frac{pr'(p)}{r(p)}\bigg|_{p=\overline{G}(t)}$$

Since $\overline{G}(t)$ is a decreasing function of t, if each component of a coherent system has the same IFR lifetime distribution, then the distribution of system lifetime will be IFR if $pr'(p)/r(p)$ is a decreasing function of p.

See book for example on IFR k-out-of-n system and a non-IFR parallel system

See p 606 for discussion on **mixtures**

If a distribution $F(t)$ has density $f(t) = F'(t)$, then

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

$$\int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1 - F(s)}ds = -\log \overline{F}(t)$$

So,

$$\overline{F}(t) = e^{-\Lambda(t)}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ the [**cumulative**] **hazard function** of distribution F.

A distribution F has an **increasing failure rate on the average (IFRA)** if

$$\frac{\Lambda(t)}{t} = \frac{\int_0^t \lambda(s)ds}{t}$$

increases in t for $t \geq 0$. The average failure rate up to time t increases as t increases.

F is IFR $\Rightarrow$ F is IFRA, but not necessarily the converse:

F is IFRA

$\Leftrightarrow \Lambda(s)/s \leq \Lambda(t)/t$ whenever $0 \leq s \leq t$

$\Leftrightarrow \frac{\Lambda(\alpha t)}{\alpha t} \leq \frac{\Lambda(t)}{t}$ for $0 \leq \alpha \leq 1, \ \forall t \geq 0$

$\Leftrightarrow -\log \overline{F}(\alpha t) \leq -\alpha \log \overline{F}(t)$

$\Leftrightarrow \log \overline{F}(\alpha t) \geq \log \overline{F}^{\alpha}(t)$

$\Leftrightarrow \overline{F}(\alpha t) \geq \overline{F}^{\alpha}(t),$ for $0 \leq \alpha \leq 1, \ \forall t \geq 0$ because log is a monotone function

For a vector $P = (p_1, \ldots, p_n)$, define $P^\alpha = (p_1^\alpha, \ldots, p_n^\alpha)$

Proposition Any reliability function $r(P)$ satisfies

$$r(P^\alpha) \geq [r(P)]^\alpha$$

for $0 \leq \alpha \leq 1$.

Proof

> If $n = 1$, then $r(p) \equiv 0$ or $r(p) \equiv 1$ or $r(p) \equiv p$. In all three cases, the inequality is satisfied.
>
> Assume the proposition holds for $n - 1$ components. Consider a system of n components with structure function ϕ. Condition on whether or not the nth component is functioning:
>
> $$r(P^\alpha) = p_n^\alpha r(1_n, P^\alpha) + (1 - p_n^\alpha)r(0_n, P^\alpha)$$

Consider a system of components 1 through $n-1$ having a structure function $\phi_1(\mathbf{x}) = \phi(1_n, \mathbf{x})$ (watch subscripts), so the reliability function is $r_1(P) = r(1_n, P)$, so from the inductive assumption,

$$r(1_n, P^\alpha) \geq [r(1_n, P)]^\alpha$$

Similarly, consider a system of components 1 through $n-1$ having a structure function $\phi_0(\mathbf{x}) = \phi(0_n, \mathbf{x})$, then

$$r(0_n, P^\alpha) \geq [r(0_n, P)]^\alpha$$

So,

$$r(P^\alpha) \geq p_n^\alpha [r(1_n, P)]^\alpha + (1 - p_n^\alpha)[r(1_n, P)]^\alpha \geq [p_n r(1_n, P) + (1 - p_n) r(0_n, P)]^\alpha = [r(P)]^\alpha$$

(see lemma).

Lemma If $0 \leq \alpha \leq 1$, $0 \leq \lambda \leq 1$, then

$$h(y) = \lambda^\alpha x^\alpha + (1 - \lambda^\alpha) y^\alpha - (\lambda x + (1 - \lambda) y)^\alpha \geq 0$$

for $0 \leq y \leq x$.

Theorem For a monotone system of independent components, if each component has an IFRA lifetime distribution, then the distribution of system lifetime is itself IFRA.

Proof

distribution of system lifetime F is

$$\overline{F}(\alpha t) = r(\overline{F}_1(\alpha t), \ldots, \overline{F}_n(\alpha t))$$

Since r is monotonic, since each of the component distributions $\overline{F}_i$ is IFRA, then using the fact that $\overline{F}(\alpha t) \geq \overline{F}^\alpha(t)$ as shown earlier, along with the proposition just proved,

$$\overline{F}(\alpha t) \geq r(\overline{F}_1^\alpha(t), \ldots, \overline{F}_n^\alpha(t)) \geq [r(\overline{F}_1(t), \ldots, \overline{F}_n(t))]^\alpha = \overline{F}^\alpha(t)$$

9.6 Expected System Lifetime

$$\mathbb{P}\{\text{system life} > t\} = r(\overline{f}(t))$$

where $\overline{f}(t) = (\overline{F}_1(t), \ldots, \overline{F}_n(t))$

See §2.10 for why

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > x\} dx$$

Thus,

$$\mathbb{E}[\text{system life}] = \int_0^\infty r(\overline{f}(t)) dt$$

Consider a k-out-of-n system of i.i.d. exponential components. If θ is the mean lifetime of each component, then

$$\overline{F}_i(t) = \int_0^t \frac{1}{\theta} e^{-x/\theta} dx = e^{-t/\theta}$$

Since for a k-out-of-n system,

$$r(p, p, \ldots, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1 - p)^{n-i}$$

we have

$$\mathbb{E}[\text{system life}] = \int_0^\infty \sum_{i=k}^n \binom{n}{i} (e^{-t/\theta})^i (1-e^{-t/\theta})^{n-i} dt$$

Let $y = e^{-t/\theta} \Rightarrow dy = -\frac{y}{\theta} dt$, then

$$\mathbb{E}[\text{system life}] = \theta \sum_{i=k}^n \binom{n}{i} \int_0^1 y^{i-1}(1-y)^{n-i} dy$$

$$= \theta \sum_{i=k}^n \frac{n!}{(n-i)!i!} \frac{(i-1)!(n-i)!}{n!} = \theta \sum_{i=k}^n \frac{1}{i}$$

(see lemma below).

Lemma

$$\int_0^1 y^n (1-y)^m dy = \frac{m!n!}{(m+n+1)!}$$

Proof

Let $C(n,m) = \int_0^1 y^n(1-y)^m dy$

Integration by parts:

$$C(n,m) = \left. \frac{1}{n+1} y^{n+1}(1-y)^m \right|_{y=0}^1 - \int_0^1 \frac{1}{n+1} y^{n+1} m(1-y)^{m-1}(-1) dy$$

$$= \frac{m}{n+1} \int_0^1 y^{n+1}(1-y)^{m-1} dy$$

$$C(n,m) = \frac{m}{n+1} C(n+1, m-1)$$

Since $C(n,0) = \frac{1}{n+1}$, use induction to prove the result.

Another approach

Lifetime of a k-out-of-n system can be written as $T_1 + \cdots + T_{n-k+1}$ where T_i represents time between the $(i-1)$st and ith failure. $T_1 + \cdots + T_{n-k+1}$ represents the time when the $(n-k+1)$st component fails which is the moment that the number of functioning components is less than k. When all n components are functioning, the rate at which failures occur is n/θ; i.e., T_1 is exponentially distributed with mean θ/n. Therefore, T_i represents the time until the next failure when there are $n-(i-1)$ functioning components; T_i is exponentially distributed with mean $\theta/(n-i+1)$. So,

$$\mathbb{E}[T_1 + \cdots + T_{n-k+1}] = \theta \left[\frac{1}{n} + \cdots + \frac{1}{k} \right]$$

9.6.1 An Upper Bound on the Expected Life of a Parallel System

9.7 Systems with Repair

9.8 Extra Stuff

(Covers §9.5)

Let L represent lifetime. $\mathbb{P}\{L > t\}$ is such that $\mathbb{P}\{L > 0\} = 1$, and $\lim_{t\to\infty} \mathbb{P}\{L > t\} = 0$.

We can express it as

$$\mathbb{P}\{L > t\} = e^{-H(t)}$$

where the cumulative hazard function $H(t) = -\log \mathbb{P}\{L > t\}$ is increasing(?) in t, has $H(0) = 0$, and $\lim_{t\to\infty} H(t) = \infty$.

Thus, if $X \sim$ exponential distribution with parameter 1, $\mathbb{P}\{X > u\} = e^{-u}$, so

$$\mathbb{P}\{X > H(t)\} = e^{-H(t)} = \mathbb{P}\{L > t\}$$

Let $H(t) = \int_0^t h(s)ds$ and $h(t) = \frac{d}{dt}H(t)$.

Exponential distribution:

$$\mathbb{P}\{L > t\} = e^{-\lambda t},\ t \geq 0$$

$$H(t) = \lambda t,\ t \geq 0$$

$$h(t) = \lambda$$

Interpretation: for a lifetime that follows an exponential distribution, since the hazard function is constant, its probability of dying at any given moment is independent of how long it has lived so far.

Weibull Distribution:

$$\mathbb{P}\{L > t\} = e^{-(\lambda t)^\alpha},\ t \geq 0$$

$$H(t) = (\lambda t)^\alpha,\ t \geq 0$$

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1}$$

Let $F(t) = \mathbb{P}\{L \leq t\}$ and let $f(t) = \frac{d}{dt}F(t)$ and let $\overline{F}(t) = 1 - F(t)$

Then

$$h(t) = \frac{d}{dt}(-\log \mathbb{P}\{L > t\}) = \frac{-\frac{d}{dt}(\mathbb{P}\{L > t\})}{\mathbb{P}\{L > t\}} = \frac{-\frac{d}{dt}(1 - F(t))}{\overline{F}(t)} = \frac{f(t)}{\overline{F}(t)}$$

Probability of death based on current condition:

$$\begin{aligned}
\mathbb{P}\{L \leq t + \Delta t | L > t\} &= 1 - \mathbb{P}\{L > t + \Delta t | L > t\} \\
&= 1 - \frac{\mathbb{P}\{L > t + \Delta t\}}{\mathbb{P}\{L > t\}} \\
&= 1 - \exp(-H(t + \Delta t) + H(t)) \\
&= 1 - \exp\left(-\int_t^{t+\Delta t} h(s)ds\right) \\
&\approx 1 - \exp(-h(t)\Delta t) \\
&\approx h(t)\Delta t
\end{aligned}$$

for small Δt (use Taylor series for last step)

$$\lim_{u\to\infty} \frac{1}{u}\{L \leq t + u | L > t\} = h(t)$$

Note: if something has an increasing failure rate (IFR), the hazard function $h(t)$ must be differentiable and $\frac{\partial}{\partial t}\lambda(t) > 0$ for all t.

Example: two ways to compute lifetime of parallel structure

Three components in parallel with lifetimes that follow exponential distributions of parameters λ_1, λ_2, λ_3.

$\mathbb{P}\{L \le t\} = \mathbb{P}\{L_1 \le t\}\mathbb{P}\{L_2 \le t\}\mathbb{P}\{L_3 \le t\} = (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t})(1 - e^{-\lambda_3 t})$

$$\mathbb{E}[L] = \int_0^\infty \mathbb{P}\{L > t\}dt$$
$$= 1 - \int_0^\infty (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t})(1 - e^{-\lambda_3 t})dt$$
$$= 1 - \int_0^\infty 1 - e^{-\lambda_1 t} - e^{-\lambda_2 t} - e^{-\lambda_3 t} + e^{-(\lambda_1+\lambda_2)t} + e^{-(\lambda_1+\lambda_3)t} + e^{-(\lambda_2+\lambda_3)t} - e^{-(\lambda_1+\lambda_2+\lambda_3)t} dt$$
$$= \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} - \frac{1}{\lambda_1+\lambda_2} - \frac{1}{\lambda_1+\lambda_3} - \frac{1}{\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2+\lambda_3}$$

(Use the same trick for expected value)

Suppose all have the same distribution. Then what?

Method 1:

Use the above formula to get $3/\lambda - 3/(2\lambda) + 1/(3\lambda) = 11/(6\lambda)$

Method 2:

Three time intervals:

L_1: duration for which 3 components are working; $\sim$ expon(3λ)

L_2: duration for which 2 components are working: $\sim$ expon(2λ)

L_3: duration for which 1 component is working: $\sim$ expon(λ)

$\mathbb{E}[L] = \mathbb{E}[L_1 + L_2 + L_3] = 1/(3\lambda) + 1/(2\lambda) + 1/(\lambda) = 11/(6\lambda)$

10 Brownian Motion and Stationary Processes

10.1 Gaussian processes

10.1.1 Gaussian Distribution

Definition We write $X \sim \text{Gsn}(\mu, \sigma^2)$ if $\mathbb{P}\{X \in dx\} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx$

Definition The **standard** Guassian distribution is $\text{Gsn}(0, 1)$, that is, $\mathbb{P}\{X \in dx\} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

De Moivre showed that the binomial distribution with $n \to \infty$ approximates the Guassian (see de Moivre-Laplace theorem)

Bachelier: $\text{Gsn}(0, \frac{1}{2\pi})$: $\mathbb{P}\{X \in dx\} = e^{-\pi x^2} dx$

Moment generating function for $X \sim \text{Gsn}(0, 1)$:

$$
\begin{aligned}
\mathbb{E}[e^{rX}] &= \int_{-\infty}^{\infty} e^{rx}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \\
&= \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2rx)}\frac{1}{\sqrt{2\pi}}dx \\
&= \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2rx+r^2)}e^{r^2/2}\frac{1}{\sqrt{2\pi}}dx \\
&= e^{r^2/2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-(x-r)^2/2}dx \\
&= e^{r^2/2}
\end{aligned}
$$

Differentiating the moment generating function $\phi(r) = e^{r^2/2}$ and setting $r = 0$ helps us find expected values:

$$\left[\frac{d^n}{dr^n}e^{r^2/2}\right]_{r=0} = \mathbb{E}[X^n]$$

From this we have $\mathbb{E}[Z^2] = 1$, $\mathbb{E}[Z^4] = 3$, $[Z^6] = 15$, as well as $\mathbb{E}[Z^1] = \mathbb{E}[Z^3] = \mathbb{E}[Z^5] = \cdots = 0$ which can be seen from the symmetry of the pdf of Z.

$$\mathbb{E}[Z^n] = \begin{cases} 0 & n \text{ is odd} \\ \frac{n!}{2^{n/2}\left(\frac{n}{2}\right)!} & n \text{ is even} \end{cases}$$

Transformation to standard Gaussian:

Let $X \sim \text{Gsn}(\mu, \sigma^2)$

$$
\begin{aligned}
\mathbb{P}\left\{\frac{X-\mu}{\sigma} \leq y\right\} &= \mathbb{P}\{X \leq \mu + \sigma y\} \\
\mathbb{P}\left\{\frac{X-\mu}{\sigma} \leq y\right\} &= \int_{-\infty}^{\mu+\sigma y} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}dx \\
\mathbb{P}\left\{\frac{X-\mu}{\sigma} \in dy\right\} &= \frac{d}{dy}\int_{-\infty}^{\mu+\sigma y} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}dx\,dy \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{((\mu+\sigma y)-\mu)^2/2\sigma^2}\right)\sigma\,dy \\
&= \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,dy
\end{aligned}
$$

Where we use $\frac{d}{dy}(g(h(y))) = g'(h(y))h'(y)$ with $g(z) = \int_0^z \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}dx$ and $h(y) = \mu + \sigma y$.

So

$$X \sim \text{Gsn}(\mu, \sigma^2) \Leftrightarrow X = \mu + \sigma Z$$

where $Z \sim \text{Gsn}(0, 1)$

The generating function for X is $\mathbb{E}[e^{rX}] = \mathbb{E}[e^{r(\mu+\sigma Z)}] = e^{r\mu}e^{(r\sigma)^2/2} = e^{r\mu+(r^2\sigma^2/2)}$, and since the moment generating function is determined uniquely, any distribution with the previous generating function must have distribution $\text{Gsn}(\mu, \sigma^2)$:

$$X \sim \text{Gsn}(\mu, \sigma^2) \Leftrightarrow \mathbb{E}[e^{rX}] = e^{r\mu+(r^2\sigma^2/2)}$$

Linear combinations:

Let $X \sim \text{Gsn}(\mu, \sigma^2)$ be independent of $Y \sim \text{Gsn}(\nu, \tau^2)$. Then

$$\alpha X + \beta Y \sim \text{Gsn}(\alpha\mu + \beta\nu, \alpha^2\sigma^2 + \beta^2\tau^2)$$

This can be shown by using the moment generating function and the independence of X and Y

$$\mathbb{E}[e^{r\alpha X}] = e^{\alpha\mu r + (r^2\alpha^2\sigma^2/2)}; \;\; \mathbb{E}[e^{r\beta Y}] = e^{\beta\nu r + (r^2\beta^2\tau^2/2)}$$
$$\mathbb{E}[e^{r(\alpha X + \beta Y)}] = \mathbb{E}[e^{r\alpha X}] \cdot \mathbb{E}[e^{r\beta Y}] = e^{(\alpha\mu+\beta\nu)+r^2(\alpha^2\sigma^2+\beta^2\tau^2)/2} \Leftrightarrow \alpha X + \beta Y \sim \text{Gsn}(\alpha\mu + \beta\nu, \alpha^2\sigma^2 + \beta^2\tau^2)$$

Gaussian-gamma connection:

$$Z \sim \text{Gsn}(0,1) \Rightarrow Z^2 \sim \text{gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$$

Proof:

$$\mathbb{P}\{Z^2 \le t\} = \mathbb{P}\{-\sqrt{t} \le Z \le \sqrt{t}\} = \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$
$$\frac{d}{dt}\mathbb{P}\{Z^2 \le t\} = \left(\frac{1}{\sqrt{2\pi}} e^{-t/2}\right) t^{-1/2} = \frac{\lambda e^{-\lambda t}(\lambda t)^{\alpha-1}}{\Gamma(\alpha)}$$

where $\lambda = \alpha = \frac{1}{2}$, which also shows that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Also see §2.10.

In §3.8, we showed that if $X \sim \text{gamma}(\alpha, \lambda)$ and $Y \sim \text{gamma}(\beta, \lambda)$ are independent, then $X + Y \sim \text{gamma}(\alpha + \beta, \lambda)$. Then, if $Z_1, \ldots, Z_n$ i.i.d. with distribution $\text{Gsn}(0,1)$, then

$$Z_1^2 + \cdots + Z_n^2 \sim \text{gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$$

This is the **chi-square distribution** with n degrees of freedom.

Gaussian and exponentials:

If X and Y are i.i.d. with distribution $\text{Gsn}(0,1)$, then $X^2 + Y^2 \sim \text{gamma}\left(1, \frac{1}{2}\right) \equiv \text{expon}\left(\frac{1}{2}\right)$

Let X and Y be x- and y-coordinates of a random point in $\mathbb{R}^2$; (X, Y) is the landing point of a dart aimed at the origin.

Let $R = \sqrt{X^2 + Y^2}$ be the distance from the origin

Let A be the angle made by the vector $\langle X, Y \rangle$, measured CCW from positive x-axis.

Then

- $R^2 \sim \text{expon}\left(\frac{1}{2}\right)$
- $A \sim \mathcal{U}(0, 2\pi]$
- R and A are independent

This is used in Monte-Carlo studies for generating Gaussian variables: given $R^2 \sim \text{expon}\left(\frac{1}{2}\right)$ and $A \sim \mathcal{U}(0, 2\pi]$, then $R\cos A$ and $R\sin A$ are independent $\text{Gsn}(0,1)$ variables.

We can derive the pdf of the arcsine distribution ($\text{beta}\left(\frac{1}{2}, \frac{1}{2}\right)$) with this interpretation. Since X^2 and Y^2 have distribution $\text{gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, and using the fact that if $A \in [0, \arcsin\sqrt{u}] \cup [\pi - \arcsin\sqrt{u}, \pi] \cup [\pi, \pi + \arcsin\sqrt{u}] \cup [2\pi - \arcsin\sqrt{u}, 2\pi)$, then $(\sin A)^2 \leq u$ to find:

$$\mathbb{P}\left\{\frac{X^2}{X^2+Y^2} \leq u\right\} = \mathbb{P}\{(\sin A)^2 \leq u\} = 4\frac{\arcsin\sqrt{u}}{2\pi} = \frac{2}{\pi}\arcsin\sqrt{u}$$

$$\mathbb{P}\left\{\frac{X^2}{X^2+Y^2} \in du\right\} = \frac{1}{\pi\sqrt{u(1-u)}}$$

Cauchy distribution

If X and Y are independent $\text{Gsn}(0,1)$ variables, the distribution of $T = X/Y$ is called the **standard Cauchy distribution**, which has pdf

$$p(z) = \frac{1}{(1+z^2)\pi}$$

for $z \in \mathbb{R}$.

Note that T and $1/T$ have the same distribution.

$1/T = Y/X = \tan A$ where $A \sim \mathcal{U}(0, 2\pi]$

One way to derive the standard Cauchy density using this:

$$\begin{aligned}
\mathbb{E}[T] = \mathbb{E}[f(\tan A)] &= \int_0^{2\pi} f(\tan a)\frac{1}{2\pi}da \\
&= \int_{(0,\frac{\pi}{2})\cup(\frac{3\pi}{2}2\pi)} f(\tan a)\frac{1}{2\pi}da + \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} f(\tan a)\frac{1}{2\pi}da && \text{resolve uniqueness of arctan} \\
&= \int_{-\infty}^{\infty} f(x)\frac{1}{2\pi}\frac{1}{1+x^2}dx + \int_{-\infty}^{\infty} f(x)\frac{1}{2\pi}\frac{1}{1+x^2}dx && \tan a = x;\ a = \arctan x;\ da = \frac{1}{1+x^2}dx \\
&= \int_{-\infty}^{\infty} f(x)\frac{1}{\pi(1+x^2)}dx
\end{aligned}$$

So we get the pdf of the Cauchy distribution

T has no expected value:

$$\int_{-\infty}^{0} z\ p(z)\ dz = -\infty$$
$$\int_{0}^{\infty} z\ p(z)\ dz = \infty$$

and the sum of these two integrals is undefined.

Exercise Show that if X, Y are independent $\text{Gsn}(0,1)$ variables,

that $X^2+Y^2 \sim \text{expon}\left(\frac{1}{2}\right)$ and $Y/X \sim$ Cauchy, and are independent.

Hint: $\mathbb{E}[f(X^2+Y^2, Y/X)]$

Relation to binomial distribution: If $S_n \sim \text{binom}(n,p)$, then

$$\frac{S_n - np}{\sqrt{np(1-p)}} \sim Z$$

as $n \to \infty$ and $p \to$ a moderate value.

10.1.2 Gaussian vectors

Definition A random vector $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is said to be **Gaussian** (in $\mathbb{R}^n$) if

$$\alpha \cdot \mathbf{X} = \alpha_1 X_1 + \cdots + \alpha_n X_n$$

is Gaussian for any vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

Recall that if $X_1, \ldots X_n$ are independent Gaussian variables, then any linear combination $\alpha_1 X_1 + \ldots + \alpha_n X_n$ is also Gaussian. This implies that $\mathbf{X} = (X_1, \ldots, X_n)$ is Gaussian. The converse is not necessarily true though.

Theorem A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is Gaussian if and only if it has the form

$$X_i = \mu_i + \sum_{j=1}^{n} a_{ij} Z_j$$

$\forall i \in \{1, \ldots, n\}$ where $Z_1, \ldots, Z_n$ are independent $\mathrm{Gsn}(0,1)$ variables, $\mu_1, \ldots, \mu_n \in \mathbb{R}$, and $a_{ij} \in \mathbb{R}, \forall i, j \in \{1, \ldots, n\}$.

Example

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1.2 \\ -0.8 \\ -0.5 \end{bmatrix} + \begin{bmatrix} 0.3 & 0 & 0.5 \\ 0.8 & 0.1 & 2.0 \\ 0 & -0.8 & -6.4 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix}$$

Each X_i is in the form expressed in the previous theorem. The X_i depend on each other through their dependence on the Z_i.

Note that $\mathbb{E}[X_1] = 1.2$ because $\mathbb{E}[Z_i] = 0$ for all i. We can see that $\mathbb{E}[X_i] = \mu_i$.

Note that

$$\begin{aligned} \mathrm{Cov}(X_1, X_3) &= \mathbb{E}[(X_1 - 1.2)(X_3 + 0.5)] \\ &= \mathbb{E}[(0.3Z_1 + 0.5Z_3)(-0.8Z_2 - 6.4Z_3)] \\ &= \mathbb{E}[(0.5)(-6.4)(Z_3)^2] \end{aligned}$$

because $\mathbb{E}[Z_i Z_j] = 0$ if $i \neq j$ due to independence.

Let $\mathbf{r} \in \mathbb{R}^n$.

$$\mathbb{E}[\mathbf{r} \cdot \mathbf{X}] = \mathbb{E}[r_1 X_1 + \cdots + r_n X_n] = r_1 \mu_1 + \cdots r_n \mu_n = \mathbf{r} \cdot \vec{\mu}$$

$$\begin{aligned} \mathrm{Var}(\mathbf{r} \cdot \mathbf{X}) &= \mathbb{E}[(\mathbf{r} \cdot \mathbf{X} - \mathbf{r} \cdot \vec{\mu})^2] \\ &= \mathbb{E}[(r_1(X_1 - \mu_1) + \cdots + r_n(X_n - \mu_n))^2] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} r_i r_j \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} r_i \sigma_{ij} r_j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{r} \cdot (\Sigma \mathbf{r}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{r}^\intercal \Sigma \mathbf{r} \end{aligned}$$

where $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, and $\Sigma = [\sigma_{ij}]_{1 \le i,j \le n}$ is the covariance matrix.

$$\mathbb{E}[e^{\mathbf{r} \cdot \mathbf{X}}] = \exp\left\{\mathbf{r} \cdot \vec{\mu} + \frac{1}{2}\mathbf{r}^\intercal \Sigma \mathbf{r}\right\}$$

Theorem If $\mathbf{X} = (X_1, \ldots, X_n)$ is Gaussian, X_i and X_j are independent if and only if $\mathrm{Cov}(X_i, X_j) = 0$. Note that $\mathbf{X}$ must be Gaussian; otherwise, it is possible that X_i and X_j are Gaussian with covariance 0, but still be dependent on each other. More generally, if $\mathbf{X}$ is Gaussian, then $X_1, \ldots X_n$ are independent if and only if the covariance matrix Σ is diagonal, that is, $\sigma_{ij} = 0$ for all $i \neq j$.

10.1.3 Gaussian processes

Definition Let $X = \{X_t\}_{t \ge 0}$ be a stochastic process with state space $\mathbb{R}$ (that is, $X_t \in \mathbb{R}$). Then X is a **Gaussian process** if $\mathbf{X} = (X_{t_1}, \ldots, X_{t_n})$ is Gaussian for all choices of $n \ge 2$ and $t_1, \ldots, t_n$.

If X is a Gaussian process, the distribution of $\mathbf{X}$ is specified for all n and $t_1, \ldots, t_n$ once we specify the **mean function** $m(t) = \mathbb{E}[X_t]$ and the **covariance function** $v(s,t) = \mathrm{Cov}(X_s, X_t)$. m is arbitrary, but v must be symmetric (i.e., $v(s,t) = v(t,s)$) and positive-definite (i.e., $\mathbf{r}^\intercal \Sigma \mathbf{r} > 0$, see explanation of $\mathrm{Var}(\mathbf{r} \cdot \mathbf{X})$.

10.1.4 Gauss-Markov Chains

Let $0 < r < 1$ and $c > 0$ be fixed. Let $Z_1, Z_2, \ldots$ be i.i.d. $\mathrm{Gsn}(0,1)$, let $Y_0 \sim \mathrm{Gsn}(a_0, b_0^2)$ be independent of the Z_n. Define process $Y = \{Y_n\}_{n \in \{0,1,\ldots\}}$ recursively by

$$Y_{n+1} = rY_n + cZ_{n+1}$$

Or in words, "the current value is equal to a proportion of the previous value plus an additive randomness."

Note that $cZ_n \sim \mathrm{Gsn}(0, c^2)$.

$$Y_n = r^n Y_0 + cr^{n-1}Z_1 + cr^{n-2}Z_2 + \cdots + crZ_{n-1} + cZ_n$$

The vector $(Y_0, Y_1, \ldots, Y_n)$ is Gaussian for any n because each Y_i is a linear combination of independent Gaussian random variables. Therefore, $Y = \{Y_n\}_{n \in \{0,1,\ldots\}}$ is a Gaussian process.

$$a_n = \mathbb{E}[Y_n] = r^n a_0$$

$$\begin{aligned} v_{n,n} &= \mathrm{Var}(Y_n) \\ &= r^{2n}b_0^2 + c^2 r^{2n-2} + \cdots + c^2 \\ &= r^{2n}b_0^2 + c^2\frac{1 - r^{2n}}{1 - r^2} \end{aligned}$$

$\lim_{n\to\infty} a_n = 0$

$\lim_{n\to\infty} v_{n,n} = \frac{c^2}{1-r^2}$

Note that the limit is free of Y_0.

$\lim_{n\to\infty} Y_n \sim \mathrm{Gsn}(0, \frac{c^2}{1-r^2})$

10.2 Brownian Motion

10.2.1 Brownian motion and Wiener process

Defining Brownian motion axiomatically as a special stochastic process

Definition A stochastic process $X = \{X_t\}_{t\geq 0}$ is a **Brownian motion** if

- $t \mapsto X_t$ is continuous (not a counting process, which is discrete and strictly nondecreasing)
- independent increments
- stationary increments

Theorem $X_{s+t} - X_s \sim \text{Gsn}(\mu t, \sigma^2 t)$ where $\mu \in \mathbb{R}$, $\sigma^2 \in [0, \infty)$, both free of s and t.

Proof outline:

> By stationarity, $X_{s+t} - X_s$ has the same distribution as $X_t - X_0$, and we concentrate on the latter here. Take $0 = t_0 < t_1 < \cdots < t_n = t$ such that $t_i - t_{i=1} = \frac{t}{n}$. Then
>
> $$X_t - X_0 = (X_{t_1} - X_{t_0}) + (X_{t_2} - X_{t_1}) + \cdots + (X_{t_n} - X_{t_{n-1}})$$
>
> expresses $X_t - X_0$ as the sum of n i.i.d. random variables. Moreover, since X is continuous, the terms on the irght side are small when n is large. By the central limit theorem, the distribution of the right side is approximately Gaussian. Since n can be as large as desired, $X_t - X_0$ must have a Gaussian distribution, say, with mean $a(t)$ and variance $b(t)$. To determine their forms, write
>
> $$X_{s+t} - X_0 = (X_s - X_0) + (X_{s+t} - X_s)$$
>
> and note that the two terms on the right side are independent with means $a(s)$ and $a(t)$, and variances $b(s)$ and $b(t)$. Thus,
>
> $$a(s+t) = a(s) + a(t), \quad b(s+t) = b(s) + b(t).$$
>
> It follows that $a(t) = \mu t$ and $b(t) = \sigma^2 t$ for some constant μ and some positive constant σ^2.

Definition A process $W = \{W_t\}_{t\geq 0}$ is a **Wiener process** if it is a Brownian motion with $W_0 = 0, \mathbb{E}[W_t] = 0, \text{Var}(W_t) = t$

Then $W_t \sim \text{Gsn}(0, t)$ and $W_t - W_s \sim \text{Gsn}(0, t - s)$

Relationship between Brown and Wiener

Let X be a Brownian motion with $X_{s+t} - X_t$ having mean μt and variance $\sigma^2 t$. Then from $X_t - X_0 \sim \text{Gsn}(\mu t, \sigma^2 t)$ and $W_t \sim \text{Gsn}(0, t)$

$$\frac{(X_t - X_0) - \mu t}{\sigma} = W_t$$
$$X_t = X_0 + \mu t + \sigma W_t$$

μ is drift coefficient, σ is volatility

Interpret this as a line $X_0 + \mu t$ with deviations vertically about the line.

Further,

$$\mathbb{E}[e^{rW_t}] = \mathbb{E}[e^{r\sqrt{t}}Z] = e^{r^2 t/2}$$

10.2.2 Poisson approximation

Large particle surrounded by small molecules

$L_t \equiv$ number of molecules that hit particle from left in $[0, t]$

$M_t \equiv$ number of molecules that hit particle from right in $[0, t]$

each hit moves large particle ε amount

$X_t^\varepsilon =$ displacement over $[0, t] = \varepsilon L_t - \varepsilon M_t$

$\{L_t\}$ and $\{M_t\}$ are independent Poisson processes both with rate λ

$\mathbb{E}[L_t] = \mathbb{E}[M_t] = \mathrm{Var}(L_t) = \mathrm{Var}(M_t) = \lambda t$

$\mathbb{E}[X_t^\varepsilon] = \mathbb{E}[\varepsilon L_t - \varepsilon M_t] = \varepsilon\lambda t - \varepsilon\lambda t = 0$

$\mathrm{Var}(X_t^\varepsilon) = \varepsilon^2\lambda t + \varepsilon^2\lambda t = 2\varepsilon^2\lambda t$

Let $\mathrm{Var}(X_t^\varepsilon) \equiv \sigma^2 t$ (Motivation: we want the variance to be linear with t), then

$$\lambda = \frac{\sigma^2}{2\varepsilon^2}$$

Moment generating function:

$$\begin{aligned}
\mathbb{E}\left[e^{rX_t^\varepsilon}\right] &= \mathbb{E}\left[e^{r(\varepsilon L_t - \varepsilon M_t)}\right] \\
&= \mathbb{E}\left[e^{r\varepsilon L_t}\right]\mathbb{E}\left[e^{-r\varepsilon M_t}\right] \\
&\text{using the fact that } \mathbb{E}\left[e^{r\varepsilon L_t}\right] = \sum_{k=0}^{n} e^{r\varepsilon k} e^{-\lambda t}\frac{(\lambda t)^k}{k!} = e^{-\lambda t} e^{\lambda t e^{r\varepsilon}} = \exp\{-\lambda t(1 - e^{r\varepsilon})\}, \\
&\text{and similarly, that } \mathbb{E}\left[e^{-r\varepsilon M_t}\right] = \exp\{-\lambda t(1 - e^{r\epsilon})\}, \\
&= \exp\{-\lambda t(1 - e^{-r\varepsilon})\}\exp\{-\lambda t(1 - e^{r\epsilon})\} \\
&= \exp\{\lambda t(e^{r\varepsilon} + e^{-r\varepsilon} - 2)\} \\
&= \exp\left\{\frac{\sigma^2 t}{2\varepsilon^2}(e^{r\varepsilon} + e^{-r\varepsilon} - 2)\right\} \\
&= \exp\left\{\frac{\sigma^2 t}{2} \cdot \frac{e^{r\varepsilon} + e^{-r\varepsilon} - 2}{\varepsilon^2}\right\} \\
&\text{using l'Hopital's rule twice,} \\
\lim_{\varepsilon\to 0}\mathbb{E}\left[e^{rX_t^\varepsilon}\right] &= \exp\left\{\frac{1}{2}\sigma^2 t r^2\right\} \\
&= \mathbb{E}\left[e^{rY}\right] \text{ where } Y \sim \mathrm{Gsn}(0, \sigma^2 t)
\end{aligned}$$

In the last step, we recognized that if $Y \sim \mathrm{Gsn}(0, \sigma^2 t)$, then $\frac{Y}{\sqrt{\sigma^2 t}} = Z \sim \mathrm{Gsn}(0, 1)$ and that , as we showed previously, $\mathbb{E}\left[e^{rZ}\right] = e^{r^2/2}$, so we have $\mathbb{E}\left[e^{rY}\right] = \mathbb{E}\left[e^{\left(r\sqrt{\sigma^2 t}\right)Z}\right] = e^{r^2\sigma^2 t/2}$

So we have

$$\lim_{\varepsilon\to 0}\mathbb{P}\left\{X_t^\varepsilon \le x\right\} = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2 t}}\exp\left\{-\frac{y^2}{2\sigma^2 t}\right\}dy$$

If we let $X_t = \lim_{\varepsilon\to 0} X_t^\varepsilon$, we have the following properties for the process $\{X_t\}_{t\ge 0}$:

- $t \mapsto X_t$ is continuous (not a counting process, which is discrete and strictly nondecreasing)
- independent increments
- stationary increments

We call this process a **Brownian motion process**.

Consider the total distance traveled (not displacement) in $[0, t]$ (call it T_t). Since $T_t^\epsilon = \varepsilon L_t + \varepsilon M_t$, we have

$$\mathbb{E}[T_t^\varepsilon] = \mathbb{E}[\varepsilon L_t + \varepsilon M_t] = 2\varepsilon\lambda t = 2\varepsilon\left(\frac{\sigma^2}{2\epsilon^2}\right)t = \frac{\sigma^2 t}{\varepsilon}$$

$$\lim_{\varepsilon\to 0}\mathbb{E}[T_t^\varepsilon] = \lim_{\varepsilon\to 0}\mathbb{E}[\varepsilon L_t + \varepsilon M_t] = \lim_{\varepsilon\to 0}\frac{\sigma^2 t}{\varepsilon} = \infty$$

So on any finite time interval, the total distance traveled is infinite.

Thus we can say that $t \mapsto X_t$ is not differentiable anywhere and is highly oscillatory...

10.2.3 Brownian motion as Gaussian

Let $X = \{X_t\}$ be a Brownian motion with $X_0 = 0$. Fix $n \geq 1$ and $0 \leq t_1 < t_2 < \ldots < t_n$. Then X_{t_1}, $X_{t_2} - X_{t_1}, \ldots,$ $X_{t_n} - X_{t_{n-1}}$ are independent and Gaussian. Since the vector $(X_{t_1}, \ldots, X_{t_n})$ is obtained from a linear transformation of those increments, $(X_{t_1}, \ldots, X_{t_n})$ is n-dimensional Gaussian.

Theorem Let X be a Brownian motion with $X_0 = 0$, drift μ, volatility σ. Then X is a Gaussian process with mean function $m(t) = \mu t$ and covariance function $v(s,t) = \sigma^2(s \wedge t)$. Conversely, if X is a Gaussian process with these mean and covariance functions, *and* if X is continuous, then X is a Brownian motion with $X_0 = 0$, drift μ, and volatility σ.

Corollary A process $W = \{W_t\}_{t\geq 0}$ is a Wiener process if and only if it is continuous and Gaussian with $\mathbb{E}[W_t] = 0$ and $\mathrm{Cov}(W_s, W_t) = s \wedge t$.

$$\begin{aligned}
\mathrm{Cov}(W_s, W_t) &= \mathbb{E}[(W_s - \mathbb{E}[W_s])(W_t - \mathbb{E}[W_t])] \\
&= \mathbb{E}[W_s W_t] - \mathbb{E}[W_s]\mathbb{E}[W_t] \\
&= \mathbb{E}[W_s W_t] \\
&= \mathbb{E}[W_s(W_s + (W_t - W_s))] && \text{assume WLOG that } s < t \\
&= \mathbb{E}[W_s^2] + \mathbb{E}[W_s(W_t - W_s)] \\
&= \mathbb{E}[W_s^2] + \mathbb{E}[W_s]\mathbb{E}[(W_t - W_s)] && \text{indep. increments} \\
&= \mathbb{E}[W_s^2] \\
&= \mathbb{E}[(W_s - 0)^2] \\
&= \mathbb{E}[(W_s - \mathbb{E}[W_s])^2] \\
&= \mathrm{Var}(W_s) \\
&= s
\end{aligned}$$

Had we chosen $t < s$, then we would have obtained t instead. So $\mathrm{Cov}(W_s, W_t) = s \wedge t$.

Theorem Let W be a Wiener process. Then the following hold:

- symmetry: Let $\overline{W}_t = -W_t$ for $t \geq 0$, then $\overline{W}$ is a Wiener process
- scale invariance: Let $\widehat{W}_t = \frac{1}{\sqrt{c}} W_{ct}$ for some fixed constant $c > 0$, then $\widehat{W}$ is a Wiener process
- time inversion: $\widetilde{W}_t = tW_{s/t}$ for $t > 0$ (note strict inequality) and $\widetilde{W}_0 = 0$, then $\widetilde{W}$ is a Wiener process

Proof

symmetry: $\overline{W}$ satisfies the definition of Wiener process: $\overline{W}_0 = 0, \mathbb{E}[\overline{W}_t] = 0, \mathrm{Var}(\overline{W}_t) = t$

scale: $\widehat{W}$ also satisfies the definition. $\mathbb{E}[\widehat{W}_t] = \frac{1}{\sqrt{c}} \mathbb{E}[W_{ct}] = 0$ and $\mathrm{Var}(W_{ct}) = \frac{1}{c} \mathrm{Var}(W_{ct}) = t$

time inversion: $\mathbb{E}[\tilde{W}_t] = 0$, and

$$\mathrm{Cov}(\widetilde{W}_s, \widetilde{W}_t) = \mathbb{E}[\widetilde{W}_s \widetilde{W}_t] = st\mathbb{E}[W_{1/s} W_{1/t}] = st\left(\frac{1}{s} \wedge \frac{1}{t}\right) = s \wedge t$$

which is true for Wiener processes. However, we need to check the continuity of $\widetilde{W}$ at $t = 0$.

$$\lim_{t \to 0} \widetilde{W}_t = \lim_{t \to 0} tW_{1/t} = \lim_{u \to \infty} \frac{1}{u} W_u = 0 = \widetilde{W}_0$$

Where we used $u = \frac{1}{t}$.

The intuitive idea behind this is the strong law of large numbers. Pick $\delta > 0$, $u = n\delta$.

$$W_{n\delta} = W_\delta + (W_{2\delta} - W_\delta) + \cdots + (W_{n\delta} - W_{(n-1)\delta}) = Y_1 + \cdots Y_n$$

Then from the law of large numbers,

$$\lim_{n \to \infty} \frac{1}{n\delta} W_{n\delta} = \frac{1}{\delta} \lim_{n \to \infty} \frac{Y_1 + \cdots + Y_n}{n} = \frac{1}{\delta} 0 = 0$$

Example Let W be a Wiener process. Let $0 < s < t$.

$$\begin{aligned}
\mathbb{P}\{W_s \leq x | W_t = z\} &= \mathbb{P}\{\widetilde{W}_s \leq x | \widetilde{W}_t = z\} \\
&= \mathbb{P}\{sW_{1/s} \leq x | tW_{1/t} = z\} \\
&= \mathbb{P}\{W_{1/s} \leq x/s | W_{1/t} = z/t\} \\
&= \mathbb{P}\{W_{1/s} - W_{1/t} = x/s - z/t\}
\end{aligned}$$

Time inversion allowed us to change a condition on the future to a condition on the past.

Brownian bridge

Let $X_t = W_t - tW_1$ for $t \in [0, 1]$. It is called a "bridge" because it is tied down for $t = 0$ and $t = 1$, that is, $X_0 = X_1 = 0$. It is used to model things, such as bonds, whose values change over time, but whose values are known at the end of time ($t = 1$).

$$\mathbb{E}[X_t] = 0$$

$$\mathrm{Var}(X_t) = \mathrm{Var}(W_t - tW_1) = \mathrm{Var}(W_t - t(W_t + (W_1 - W_t)) = (1-t)^2 t - t^2(1-t) = (1-t)t$$

The variance makes sense: it is 0 at $t = 0$ and $t = 1$, and reaches its maximum of $\frac{1}{4}$ at $t = \frac{1}{2}$.

10.2.4 Hitting Times for Wiener Processes

We are interested in the hitting time

$$T_a = \inf\{t \geq 0 : W_t > a\}$$

which is the first time the Wiener process has displacement $a > 0$.

Let us denote $\mathcal{F}_s = \{W_r : r \leq s\}$ as the "past" from time s.

Markov property of W

Fix s and define

$$\widehat{W}_t = W_{s+t} - W_s$$

for $t \geq 0$.

$\widehat{W} = \{\widehat{W}_t\}$ is a wiener process because disjoint increments are independent and because $\widehat{W}_{t_1} - \widehat{W}_{t_0} \sim \text{Gsn}(0, t_1 - t_0)$.

Conceptually, this is saying that if we have a Wiener process, then freeze at time s and reset the origins of time and space to your current time/position, then continue the process, it is still a Wiener process with respect to the new origin.

Strong Markov property of W

Replace deterministic time s with random time T_a.

Distribution of T_a

$\mathbb{P}\{W_t > 0\} = \frac{1}{2}$ because $W_t \sim \text{Gsn}(0, t)$.

$\mathbb{P}\{\widehat{W}_t > 0\} = \frac{1}{2}$

$\mathbb{P}\{W_t > a | T_a < t\} = \frac{1}{2}$

$\mathbb{P}\{W_t > a, T_a < t\} = \frac{1}{2}\mathbb{P}\{T_a < t\}$

$\mathbb{P}\{W_t > a\} = \frac{1}{2}\mathbb{P}\{T_a < t\}$ since $W_0 = 0$ and $W_t > a$, the Wiener process must have already hit a.

We have found the distribution of T_a: $\mathbb{P}\{T_a < t\} = 2\mathbb{P}\{W_t > a\}$

Since $W_t \sim \text{Gsn}(0, t)$, then $W_t = \sqrt{t}Z$ where $Z \sim \text{Gsn}(0, 1)$, so

$$\mathbb{P}\{T_a < t\} = 2\mathbb{P}\{\sqrt{t}Z > a\} = 2\mathbb{P}\left\{Z > \frac{a}{\sqrt{t}}\right\} = 2\int_{a/\sqrt{t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \mathbb{P}\{T_a \leq t\}$$

Remarks

$$\mathbb{P}\{T_a < \infty\} = \lim_{t\to\infty} \mathbb{P}\{T_a < t\} = 2\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 2 \cdot \frac{1}{2} = 1$$

This states that for any $a < \infty$, the particle will hit a with probability 1.

pdf of T_a:

$$\frac{d}{dt}\mathbb{P}\{T_a \leq t\} = \frac{ae^{-a^2/2t}}{\sqrt{2\pi t^3}}$$

for $t \geq 0$

expected value:

$$\mathbb{E}[T_a] = \int_0^\infty t \frac{ae^{-a^2/2t}}{\sqrt{2\pi t^3}} dt = \infty$$

Although particle will hit every level a with probability 1, the expected time it takes to do so is ∞, no matter how small a is.

$$\mathbb{P}\{T_a < t\} = 2\mathbb{P}\{W_t > a\} = 2\mathbb{P}\{Z > a/\sqrt{t}\} = \mathbb{P}\{|Z| > a/\sqrt{t}\} = \mathbb{P}\{Z^2 > a^2/t\} = \mathbb{P}\{a^2/Z^2 < t\}$$

so $\boxed{T_a \text{ has the same distribution as } a^2/Z^2.}$

Maximum process

Define $M_t = \max_{s \le t} W_s$ for $t \ge 0$, the highest level reached during $[0, t]$ by the Wiener particle. $M_0 = 0$, and $t \to M_t$ is continuous and nondecreasing, and $\lim_{n\to\infty} M_t = +\infty$, since the particle will hit any level a. Between any two points, there are infinitely many "flats."

Similarly if we define $m_t = \min_{s \le t} W_s$, the fact that $\lim_{n\to\infty} m_t = -\infty$ shows that the set $\{t \ge 0 : W_t = 0\}$ is infinite, as the particle will cross 0 infinitely many times.

To derive the distribution,

$$\mathbb{P}\{M_t > a\} = \mathbb{P}\{T_a < t\} = 2\mathbb{P}\{W_t > a\} = \mathbb{P}\{|W_t| > a\}$$

for $a \ge 0$. So, $M_t \approx |W_t|$ (we use $\approx$ to denote "has the same distribution as"). Since $|W_t| = \sqrt{t}|Z|$, then

$$\mathbb{P}\{M_t \in dx\} = \mathbb{P}\{|W_t| \in dx\} = 2\frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

for $x \ge 0$.

Expected value and variance:

$$\mathbb{E}[M_t^2] = \mathbb{E}[W_t^2] = \mathrm{Var}(W_t) = t$$

because $\mathbb{E}[W_t] = 0$.

$$\begin{aligned}
\mathbb{E}[M_t] = \mathbb{E}[|W_t|] &= \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx \\
&= 2\int_0^\infty x \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx \\
&= \frac{2t}{\sqrt{2\pi t}} \int_0^\infty \frac{x}{t} e^{-x^2/2t} dx \\
&= \sqrt{\frac{2t}{\pi}} \left[-e^{-x^2/2t}\right]_0^\infty \\
&= \sqrt{\frac{2t}{\pi}}
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}(M_t) &= t - \frac{2t}{\pi} \\
&= \mathbb{E}[M_t^2] - \mathbb{E}[M_t]^2 \\
&= \mathbb{E}[|W_t|^2] - \frac{2t}{\pi} \\
&= \mathbb{E}[W_t^2] - \frac{2t}{\pi} \\
&= t - \frac{2t}{\pi}
\end{aligned}
$$

Hitting 0 and the Arcsine Law of Brownian Motion

Define $R_t = (\min\{u \geq t : W_u = 0\}) - t$ as the time *from* t until the next time W touches 0. By resetting the origin of time/space at time t,

$$\mathbb{P}\{R_t \in du | W_t = x\} = \mathbb{P}\{T_{-x} \in du\} = \mathbb{P}\{x^2/Z^2 \in du\}$$

or in other words, the probability that it takes u time for W to hit 0 given that it has position x at time t is the same as the probability that another Wiener particle reaches position $-x$ at time u. So, $R_t \approx W_t^2/Z^2$ where Z and W_t are independent.

Since $W_t \approx \sqrt{t}Y$ where $Y \sim \mathrm{Gsn}(0,1)$, then

$$R_t \approx tY^2/Z^2$$

Note that $Y/Z \sim$ Cauchy.

Define $D_t = R_t + t = \inf\{u > t : W_u = 0\}$ as the first time that W hits 0 after t.

So

$$D_t \approx t + tY^2/Z^2 = t\frac{Y^2 + Z^2}{Z^2}$$

Note that $\frac{Z^2}{Y^2+Z^2} \sim \mathrm{beta}\left(\frac{1}{2}, \frac{1}{2}\right)$, i.e. $\mathbb{P}\left\{\frac{Z^2}{Y^2+Z^2} \in du\right\} = \frac{1}{\pi\sqrt{u(1-u)}}$

Define $G_t = \sup\{s > t : W_s = 0\}$ as the last time that W hits 0 before t.

Suppose $s < t$. Because $G_t < s \Leftrightarrow D_s > t$,

$$
\begin{aligned}
\mathbb{P}\{G_t < s\} = \mathbb{P}\{D_s > t\} &= \mathbb{P}\left\{s\frac{Y^2+Z^2}{Z^2} > t\right\} \\
&= \mathbb{P}\left\{\frac{Z^2}{Y^2+Z^2} < \frac{s}{t}\right\} \\
&= \int_0^{s/t} \frac{1}{\pi\sqrt{u(1-u)}} du \\
&= \int_0^{\arcsin\sqrt{s/t}} \frac{2}{\pi} dx \qquad\qquad u = \sin^2 x;\ du = 2\sin x \cos x dx \\
&= \frac{2}{\pi} \arcsin\sqrt{s/t}
\end{aligned}
$$

Note that the event $\{G_t < s\}$ can be interpreted as the event that W does not hit 0 in the time interval (s,t). Notice also that

$$G_t \approx t\frac{Z^2}{Y^2+Z^2}$$

10.2.5 Geometric Brownian Motion

Let W be a Wiener process. Fix μ in $\mathbb{R}$ and $\sigma > 0$. Let

$$X_t = X_0 e^{\mu t + \sigma W_t}$$

for $t \geq 0$. Then $X = \{X_t\}_{t \geq 0}$ is a **geometric Brownian motion** with drift μ and volatility σ. Letting $Y = \log X$, we have

$$Y_t = Y_0 + \mu t + \sigma W_t$$

$$\frac{Y_t - (Y_0 + \mu t)}{\sigma} = W_t \sim \mathrm{Gsn}(0, t)$$

where Y is a Brownian motion with drift μ and volatility σ.

Treat X_0 as fixed at some value x_0. We have that $Y_t = \log X_t \sim \mathrm{Gsn}(\log x_0 + \mu t, \sigma^2 t)$. We say that X_t follows the **log-normal distribution**. It follows from $\mathbb{E}[e^{rW_t}] = e^{r^2 t/2}$ that (for fixed t),

$$\mathbb{E}[X_t] = \mathbb{E}[x_0 e^{\mu t + \sigma W_t}] = x_0 e^{\mu t} \mathbb{E}[e^{\sigma W_t}] = x_0 e^{\mu t + \sigma^2 t/2}$$

.

If $X_0 = x_0 > 0$, then $X_t > 0$ for all $t \geq 0$.

Modeling stock prices

Interpret X_t as the price at time t of a share of stock.

$$R_{s,t} = \frac{X_{s+t}}{X_s} = \exp[\mu t + \sigma(W_{s+t} - W_s)]$$

represents the return at $s+t$ of a dollar invested at time s. Since $W_{s+t} - W_s$ is independent of the past until s, the return $R_{s,t}$ is independent of the price history of the stock up to and including time s (that is, it is independent of X_s as well).

Theorem Let $X = \{X_t\}$ be the price process. Suppose that X is continuous and that for $0 \leq t_0 < t_1 < \cdots < t_n$, the returns

$$\frac{X_{t_1}}{X_{t_0}}, \frac{X_{t_2}}{X_{t_1}}, \cdots \frac{X_{t_3}}{X_{t_2}}$$

over disjoint time intervals are independent, and that the distribution of X_t/X_s depends on s and t only through $t - s$.

Then X is necessarily a geometric Brownian motion.

Proof: Let $Y_t = \log X_t$. Then Y is continuous and has stationary and independent increments, so Y is a Brownian motion. Then $Y_t = Y_0 + \mu t + \sigma W_t$ for some wiener process W. Then $X = e^Y$ has the form $X_t = X_0 e^{\mu t + \sigma W_t}$.

10.3 Extra stuff

Remember:

$$X \sim Y \Leftrightarrow \mathbb{E}[f(X)] = \mathbb{E}[f(Y)], \forall f$$

$$X \text{ independent of } Y \Leftrightarrow \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(X)], \forall f, g$$

In particular, note that $|W_t| \sim M_t$, and $|W_t| \neq M_t$.

11 Branching Processes

Start with a single progenitor (root node)

Assume the number of children a node has is independent of everything else

Let p_k be the probability a man has k children, $k \in \{0, 1, \ldots\}$

Assume $p_0 > 0$ (otherwise tree will never be extinct)

Assume $p_0 + p_1 \neq 1$ (otherwise, geometric random variable, $\mathbb{P}\{\text{extinct at step } n\} = p_1^{n-1} p_0$, probability of extinction is 1.)

Let X_n be the number of nodes in the nth generation

Let the expected number of sons a man has be $\mu = \sum_{n=0}^{\infty} n(p_n)$

$\mathbb{E}[X_{n+1} | X_n = k] = k\mu$

$\mathbb{E}[X_{n+1}] = \sum_{k=0}^{\infty} \mathbb{E}[X_{n+1}|X_n = k]\mathbb{P}\{X_n = k\} = \sum_{k=0}^{\infty}(k\mu)\mathbb{P}\{X_n = k\} = \mu\mathbb{E}[X_n]$

$\mathbb{E}[X_0] = 1$; $\mathbb{E}[X_1] = \mu$; $\mathbb{E}[X_2] = \mu^2$; $\ldots \mathbb{E}[X_n] = \mu^n$

$(\mu < 1) \Rightarrow (\lim_{n\to\infty} \mathbb{E}[X_n] = 0) \Rightarrow (\lim_{n\to\infty} X_n = 0)$

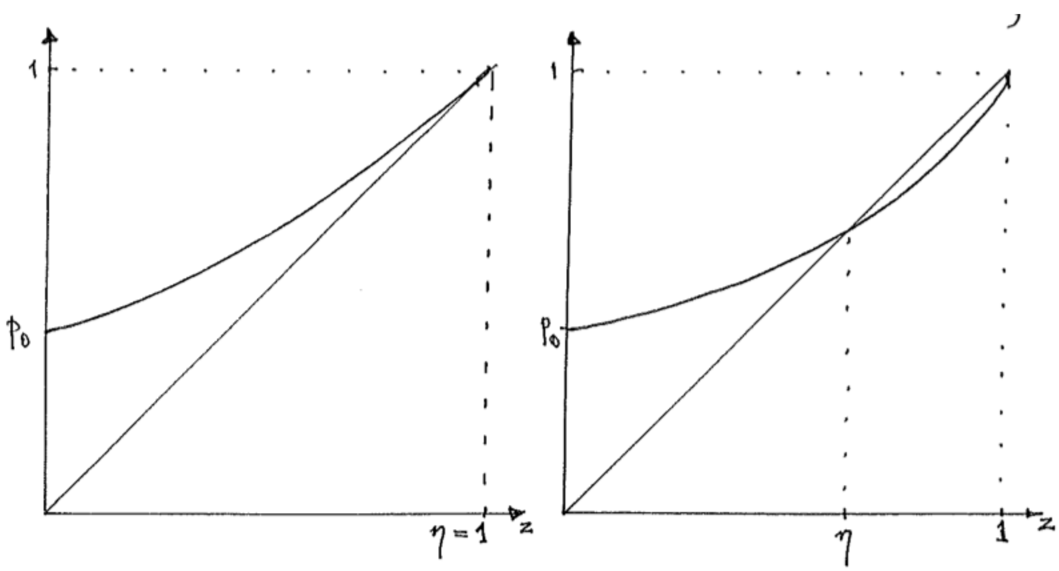
$(\mu > 1) \Rightarrow (\lim_{n\to\infty} \mathbb{E}[X_n] = \infty)$

Definition Let $\eta = \mathbb{P}\{\lim_{n\to\infty} X_n = 0\} = \mathbb{P}\{\text{eventual extinction}\}$

Theorem If we have generating function $g(z) = \sum_{k=0}^{\infty} p_k z^k$ for $z \in [0, 1]$, then η is the smallest such solution to $g(z) = z$.

Theorem: $\mu \leq 1$ if and only if $\eta = 1$; $\mu > 1$ if and only if $0 < \eta < 1$ and η is the unique solution to $z = g(z), z \in (0, 1)$.

Proof:



Suppose that the root node has k children. Then the probability of extinction η is the probability that each of the root's children's trees will eventually be extinct. We can view each of these children (1st generation) as root nodes of their own respective trees, so the probability that each of the children's trees will eventually be extinct is also η. The probability that all k lines are extinct is thus η^k. Then we have $\eta = \sum_{k=0}^{\infty} \mathbb{P}\{\text{extinction}|N = k\}\mathbb{P}\{N = k\} = \sum_{k=0}^{\infty} p_k \eta^k$.

Note the following properties of $g(z) = \sum_{k=0}^{\infty} p_k z^k = (p_0 + p_1 z + p_2 z^2 + \cdots)$, given the assumptions we made earlier:

$g(0) = p_0 > 0$

$g(1) = \sum_{k=0}^{\infty} p_k = 1$

$g(z)$ increases in z

$g'(z) = \sum_{k=0}^{\infty} k p_k z^{k-1} = (p_1 + 2p_2 z + 3p_3 z^2 + \cdots)$ increases in z

$g'(1) = \sum_{k=0}^{\infty} k p_k = \mu$

We are concerned with when $z = g(z)$ (graph both sides, find the intersections). We have two cases, as shown above.

Consider the graph of $z - g(z)$. In the picture on the right, this difference reaches a maximum (between the two intersections). If we define z_0 such that this maximum occurs at $z = z_0$, then we have $\frac{d}{dz}(z - g(z))\big|_{z=z_0} = 0$ which is true iff $\frac{d}{dz}(g(z))\big|_{z=z_0} = 1$. Since $g'(z)$ is increasing in z, we can conclude that in the second picture, $\mu = g'(1) > g'(z_0) = 1$. The first picture is the case that $\mu \le 1$.

Let $\eta_n = \mathbb{P}\{X_n = 0\}$.

Then, $\eta_n = \mathbb{P}\{X_n = 0\} \le \mathbb{P}\{X_{n+1} = 0\} = \eta_{n+1}$, $\forall n$, so $\eta_0 \le \eta_1 \le \cdots$

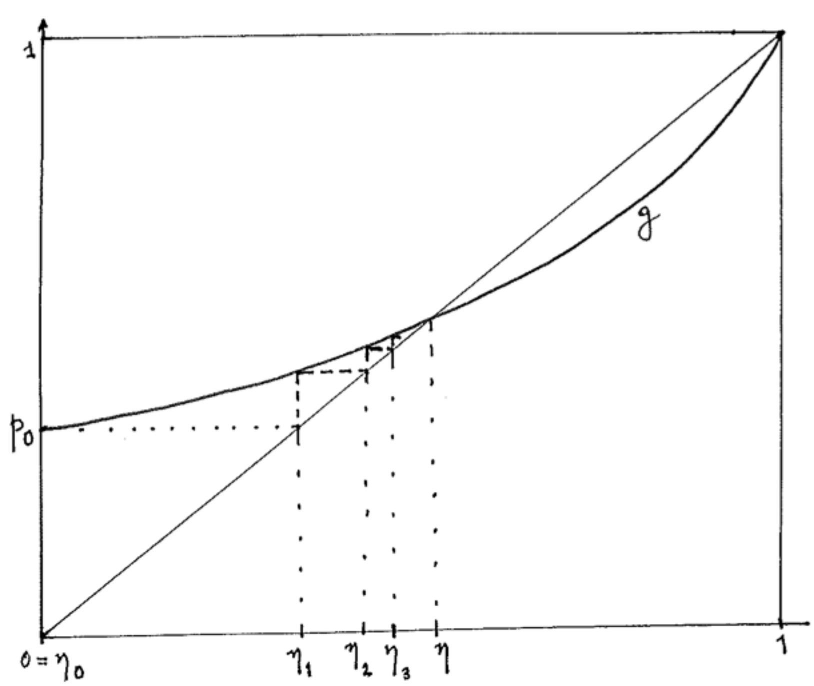
$\eta_0 = \mathbb{P}\{X_0 = 0\} = 0$

$\eta_1 = \mathbb{P}\{X_1 = 0\} = p_0$

We use a similar argument that we used earlier. If the 1st generation has k children, we can view the $n+1$st generation as the nth generations of each of the k children's trees.

Thus, $\eta_{n+1} = \sum_{k=0}^{\infty} p_k (\eta_n)^k = g(\eta_n)$

We can visually represent this recursive process, shown below.



This shows that η is the smallest solution to $z = g(z)$.

Example

Suppose $p_k \sim \text{Pois}(\mu)$ so that $p_k = e^{-\mu}\frac{\mu^k}{k!}$.

Then $g(z) = \sum_{k=0}^{\infty} e^{-\mu}\frac{\mu^k}{k!}z^k = e^{-\mu}\sum_{k=0}^{\infty}\frac{(\mu z)^k}{k!} = e^{-\mu}e^{\mu z} = e^{-\mu(1-z)}$

Solve $z = e^{-\mu(1-z)}$

Or find $\lim_{n\to\infty}\eta_n$:

$\eta_1 = p_0 = e^{-\mu}$

$\eta_{n+1} = e^{-\mu(1-\eta_n)}$

Example Pdf of a binomial distribution

Let there be a series of i.i.d. Bernoulli trials with probability of success p, and let X_n be the indicator variable for the nth trial. Let the number of successes in n trials be $S = \sum_{i=1}^{n} X_i$.

What is the generating function for S?

$$g(z) = \sum_{k=0}^{\infty} z^k \mathbb{P}\{S = k\} = \mathbb{E}[z^S] = \mathbb{E}[z^{X_1}z^{X_2}\cdots z^{X_n}] = \mathbb{E}[z^{X_1}]\mathbb{E}[z^{X_2}]\cdots\mathbb{E}[z^{X_n}]$$

because the X_k are independent.

For all i,

$$\mathbb{E}[z^{X_i}] = z^1\mathbb{P}\{X_i = 1\} + z^0\mathbb{P}\{X_i = 0\} = pz + q$$

So, continuing,

$$g(z) = (pz+q)^n$$
$$\sum_{k=0}^{\infty} z^k\mathbb{P}\{S = k\} = \sum_{k=0}^{n}\binom{n}{k}(pz)^k q^{n-k}$$
$$\mathbb{P}\{S = k\} = \binom{n}{k}p^k q^{n-k}$$

Example Pdf of $X + Y$ if $X \sim \text{Pois}(\mu)$ and $Y \sim \text{Pois}(\nu)$, independent

$$\mathbb{E}[z^X] = \sum_{k=0}^{\infty} z^k\mathbb{P}\{X = k\} = \sum_{k=0}^{\infty} e^{-\mu}\frac{\mu^k}{k!}z^k = e^{-\mu}e^{\mu z} = e^{-\mu(1-z)}$$

Similarly, $\mathbb{E}[z^Y] = e^{-\nu(1-z)}$

Now, combining the two, and using the reverse of the procedure we used above,

$$\mathbb{E}[z^{X+Y}] = \mathbb{E}[z^X]\mathbb{E}[z^Y]$$
$$\sum_{k=0}^{\infty} z^k\mathbb{P}\{X+Y = k\} = e^{-(\mu+\nu)(1-z)}$$
$$\sum_{k=0}^{\infty} z^k\mathbb{P}\{X+Y = k\} = \sum_{k=0}^{\infty} e^{-(\mu+\nu)}\frac{(z(\mu+\nu))^k}{k!}$$
$$\mathbb{P}\{X+Y = k\} = e^{-(\mu+\nu)}\frac{(\mu+\nu)^k}{k!}$$
$$X + Y \sim \text{Pois}(\mu+\nu)$$

12 Gambler's Ruin

12.1 Introduction

You have \$28. At each stage of the game, a coin is flipped; if heads, you get a dollar, if tails, you lose a dollar. The game ends when you have either \$0 or \$100.

Let X_n be your capital after the nth step. The state space is $D = \{0, 1, \ldots, 100\}$. (Note: imagine an opponent that has $\$100 - X_n$.)

$$\mathbb{P}\{X_{n+1} = i+1 | X_n = i\} = \mathbb{P}\{X_{n+1} = i-1 | X_n = i\} = \frac{1}{2}, \quad 1 \leq i \leq 99$$
$$\mathbb{P}\{X_{n+1} = i | X_n = i\} = 1, \quad i \in \{0, 100\}$$

Note that this is a Markov chain, and that we will use the technique of conditioning on the first step throughout this section.

Question 1 Will the game ever end?

Let

$$f(i) := \mathbb{P}\{1 \leq X_n \leq 99, \forall\, n \geq 0 | X_0 = i\}, \quad 0 \leq i \leq 100$$

i.e., the probability that the game continues forever given that $X_0 = i$.

$$f(i) = \begin{cases} 0 & i \in \{0, 100\} \\ \frac{1}{2} f(i+1) + \frac{1}{2} f(i-1) & 1 \leq i \leq 99 \end{cases}$$

Note that $f(i)$ is the average of $f(i-1)$ an $f(i+1)$. This means that the graph of these three points is collinear, and further, that all points $\{f(i) : i \in D\}$ are all collinear. Since $f(0) = f(100) = 0$, that means

$$f(i) = 0, \quad 0 \leq i \leq 100$$

or in other words, the game will definitely end.

Question 2 What is the probability that you get \$100?

Since the game will end, there exists a finite T such that

$$T := \min\{n \geq 0 : X_n \in \{0, 100\}\}$$

Then the probability that you will end the game with \$100, given that $X_0 = i$ is

$$r(i) := \mathbb{P}\{X_T = 100 | X_0 = i\}, \quad 0 \leq i \leq 100$$

$$r(i) = \begin{cases} 0 & i = 0 \\ \frac{1}{2} r(i+1) + \frac{1}{2} r(i-1) & 1 \leq i \leq 99 \\ 1 & i = 100 \end{cases}$$

As before, this equation shows that the graph of $r(i)$ consists of collinear points. Since $f(0) = 0$ and $f(100) = c$, we have

$$r(i) = \frac{i}{100}, \quad 0 \leq i \leq 100$$

Note that because $r(i$ is increasing, the game favors those who are initially rich, so it is not "socially fair." However, it is fair in the sense that the coin flips are fair and that your expected winnings

$$\mathbb{E}[X_T|X_0 = i] = r(i) \cdots 100 = \frac{i}{100} \cdot 100 = i$$

which is your initial capital.

Question 3 What is the duration of the game?

Let T be as before. Then the expected value of the duration is

$$\mu_i := \mathbb{E}[T|X_0 = i], \quad 0 \le i \le 100$$

$$\mu_i = \begin{cases} 0 & i \in \{0, 100\} \\ 1 + \frac{1}{2}\mu_{i+1} + \frac{1}{2}\mu_{i-1} & 1 \le i \le 99 \end{cases}$$

For $1 \le i \le 99$, you can write μ_i as $\frac{1}{2}\mu_i + \frac{1}{2}\mu_i$ on the LHS, and move things around to get

$$(\mu_{i+1} - \mu_i) = (\mu_i - \mu_{i-1}) - 2, \quad 1 \le i \le 99$$

Then we have

$$\begin{aligned} \mu_1 = \mu_1 - \mu_0 &= a \\ \mu_2 - \mu_1 &= a - 2 \\ \mu_3 - \mu_2 &= a - 4 \\ &\vdots \\ \mu_i - \mu_{i=1} &= a - 2(i-1) \\ &\vdots \\ \mu_{100} - \mu_{99} &= a - 2 \cdot 99 \end{aligned}$$

Adding the first i equations, we get

$$\mu_i = \mu_i - \mu_0 = a \cdot i - 2\frac{(i-1)i}{2} = a \cdot i - i(i-1)$$

Then we have $0 = \mu_100 = a \cdot 100 - 100 \cdot 99$, so $a = 99$ and

$$\mu_i = 99i - i(i-1) = (100 - i)i$$

Note that μ_i is maximized at $i = 50$.

12.2 Designing a fair game

Suppose we have different coins for different states.

$$\mathbb{P}\{X_{n+1} = i+1|X_n = i\} = p_i, \quad \mathbb{P}\{X_{n+1} = i-1|X_n = i\} = q_i, \quad p_i + q_i = 1, \quad 1 \le i \le 99$$

We can show that the game will end in finite time, as before.

$$f(i) := \mathbb{P}\{1 \leq X_n \leq 99, \forall\, n \geq 0 | X_0 = i\}, \quad 0 \leq i \leq 100$$

i.e., the probability that the game continues forever given that $X_0 = i$.

$$f(i) = \begin{cases} 0 & i \in \{0, 100\} \\ p_i f(i+1) + q_i f(i-1) & 1 \leq i \leq 99 \end{cases}$$

$f(i)$ is still monotonic. Whether $f(i+1) \geq f(i-1)$ or $f(i+1) \leq f(i-1)$, it is clear that $f(i) = p_i f(i+1) + q_i f(i-1)$ lies between them. Then, as before, $f(0) = 0$ for $1 \leq i \leq 100$, and so the game definitely ends.

Letting T and r be as before,

$$r(i) := \mathbb{P}\{X_T = 100 | X_0 = i\}, \quad 0 \leq i \leq 100$$

$$r(i) = \begin{cases} 0 & i = 0 \\ p_i r(i+1) + q_i r(i-1) & 1 \leq i \leq 99 \\ 1 & i = 100 \end{cases}$$

As before, $r(i)$ is monotonic. Suppose we want to exact justice so that you help the oor, but when the rich become poor, they are helped the same amount. For example, you want $r(72) = 1 - r(28)$ because when you have \$72, your opponent has \$28, and should have the same amount of help as you did when you had \$28.

$$r(i) = 1 - r(100 - i), \quad 0 \leq i \leq 100, \quad r(0) = 0$$

Then the graph of $r(i)$, which still hits $(0, 0)$ and $(100, 1)$ and is still monotone, appaears to flatten out when i is near 50.

From earlier,

$$\begin{aligned} r(i) &= p_i r(i+1) + q_i r(i-1) \\ r(i) &= p_i r(i+1) + (1 - p_i) r(i-1) \\ r(i) - r(i-1) &= p_i (r(i+1) - r(i-1)) \\ p_i &= \frac{r(i) - r(i-1)}{r(i+1) - r(i-1)} \end{aligned}$$

With the "justice" enacted, we can see that $p_i = 1 - p_{100-i}$.

However, $r(i)$ can be increasing without implying that p_i is decreasing.

12.3 Unfair game

Let c be the total capital instead of \$100. Let there be one coin for all states, but with probability p and $q = 1 - p =\neq p$ for winning and losing a dollar respectively. Now the duration of the game T is the first time you have \$0 or \$$c$, and as before, is finite. Let

$$r_i := \mathbb{P}\{X_T = c | X_0 = i\}, \quad 0 \leq i \leq c$$

$$r_i = \begin{cases} 0 & i = 0 \\ p r_{i+1} + q r_{i-1} & 1 \leq i \leq c \\ 1 & i = c \end{cases}$$

For $1 \leq i \leq c$, replacing r_i with $pr_i + qr_i$ and moving things around, we have

$$p(r_{i+1} - r_i) = q(r_i - r_{i-1}), \quad 1 \leq i \leq c-1$$

Letting $r = \frac{p}{q}$ for convenience, we have

$$\begin{aligned}
r_1 = r_1 - r_0 &= a \\
r_2 - r_1 &= ar \\
r_3 - r_2 &= ar^2 \\
&\vdots \\
r_i - r_{i-1} &= ar^{i-1} \\
&\vdots \\
r_c - rc - 1 &= ar^{c-1}
\end{aligned}$$

Summing the first i equations, we have

$$r_i = a\frac{1-r^i}{1-r}, \quad 1 \leq i \leq c$$

Since $r_c = 1$, we have $a = \dfrac{1-r}{1-r^c}$, and thus

$$r_i = \frac{1-r^i}{1-r^c} = \frac{1-\left(\frac{p}{q}\right)^i}{1-\left(\frac{p}{q}\right)^c}, \quad 0 \leq i \leq c, \quad p \neq q$$

This also works for $r = 1$:

$$\lim_{r \to 1} \frac{1-r^i}{1-r^c} = \frac{i}{c}$$

as before.

The shape of r_i as i varies depends on whether $r < 1$ or $r > 1$. If $p > q$, the game is favorable, and r_i increases with i as a concave function. If $p < q$, r_i increases with i but is convex

To find

$$\mu_i := \mathbb{E}[T | X_0 = i]$$

we have again

$$\mu_i = \begin{cases} 0 & i \in \{0, c\} \\ 1 + p\mu_{i+1} + q\mu_{i-1} & 1 \leq i \leq c \end{cases}$$

Using the same technique, and letting $r = \frac{q}{p}$ and $s = \frac{1}{p}$,

$$\begin{aligned}
p\mu_i + q\mu_i &= 1 + p\mu_{i+1} + q\mu_{i-1} \\
q(\mu_i - \mu_{i-1}) &= 1 + p(\mu_{i+1} - \mu_i) \\
\mu_{i+1} - \mu_i &= r(\mu_i - \mu_{i-1}) - s, \quad 1 \leq i \leq c-1
\end{aligned}$$

Then

$$
\begin{aligned}
\mu_1 = \mu_1 - \mu_0 &= a \\
\mu_2 - \mu_1 &= ra - s \\
\mu_3 - \mu_2 &= r^2 a - rs - s \\
&\vdots \\
\mu_i - \mu_{i-1} &= r^{i-1}a - r^{i-2}s - r^{i-3}s - \cdots - rs - s \\
&\vdots \\
\mu_c - \mu_{c-1} &= r^{c-1}a - r^{c-2}s - r^{c-3}s - \cdots - rs - s
\end{aligned}
$$

Summing the first i equations, we have

$$
\begin{aligned}
\mu_i &= a(1 + r + r^2 + \cdots + r^{i-1}) - s(1 + (1+r) + (1 + r + r^2) + \cdots + (1 + r + \cdots + r^{i-2})) \\
&= \frac{a}{1-r}(1 - r^i) - \frac{s}{1-r}(1 - r + 1 - r^2 + 1 - r^3 + \cdots + 1 - r^{i-1}) \\
&= \frac{a}{1-r}(1 - r^i) - \frac{s}{1-r}\left(i - \frac{1-r^i}{1-r}\right) \\
&= \frac{a}{1-r}(1 - r^i) - \frac{si}{1-r} + s\frac{1-r^i}{(1-r)^2}, \quad 1 \le i \le c
\end{aligned}
$$

Then $0 = \mu_c$ gives us $a = \dfrac{sc}{1-r^c} - \dfrac{s}{1-r}$, and thus

$$
\mu_i = \frac{1}{p-q}\left(c\frac{1-r^i}{1-r^c} - i\right), \quad 0 \le i \le c
$$

where we replace $s = \frac{1}{p}$ but not $r = \frac{q}{p}$. Recall that for all this, $p \neq q$. Since we found earlier that $r_i = \dfrac{1-r^i}{1-r^c}$, we have

$$
\mu_i = \frac{1}{p-q}(cr_i - i)
$$

What is the intuition behind this?

13 Appendix

13.1 dx notation

$$
\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)dx
$$

$$
\mathbb{P}\{X \in dx\} \equiv f(x)dx
$$

This comes from

$$
\mathbb{P}\{x \le X \le x + \varepsilon\} \approx f(x)\varepsilon
$$

On the left side, dx represents an interval, but on the right side, it represents the length of the interval. In higher-level probability, we use $\lambda(dx)$ to represent the length of the interval, in order to avoid confusion.

Also,

$$\mathbb{P}\{a \le X \le b\} = \int_a^b \mathbb{P}\{X \in dx\}$$

$$\mathbb{P}\{X \in dx, Y \in dy\} \equiv f(x,y)dxdy$$

Suppose $Y = 2X$ and we know the pdf of X: $\mathbb{P}\{X \in dx\} = f(x)dx$. What is the pdf of Y?

$$\mathbb{E}[Y] = \mathbb{E}[2X] = \int_{-\infty}^{\infty} 2xf(x)dx = \int_{-\infty}^{\infty} yf(y/2)\frac{dy}{2}$$

So $\mathbb{P}\{Y \in dy\} = \frac{1}{2}f(y/2)dy$ Don't forget to change the limits of integration!

In general,

$$\mathbb{P}\left\{X \in \frac{dx-a}{b}\right\} = f\left(\frac{x-a}{b}\right)\frac{dx}{b}$$

13.2 Leibniz's Rule

$$\frac{d}{dy}\int_{f(y)}^{g(y)} h(x,y)dx = \frac{dg(y)}{dy}h(g(y),y) - \frac{df(y)}{dy}h(f(y),y) + \int_{f(y)}^{g(y)} \frac{\partial}{\partial y}h(x,y)dx$$

Example that uses Leibniz's Rule

Let $X \sim \text{expon}(1)$, let $0 = a_0 < a_1 < a_2 < \cdots$ s.t. $\forall x \in (0,\infty)$, $\exists i : a_i < x < a_{i+1}$.

Let $Y = y_i$ if $a_i < X \le a_{i+1}$; assume $a_i < y_i \le a_{i+1}$

Choose $\{y_i\}$ to minimize $\mathbb{E}[|X - Y|]$.

Let there be indicator functions

$$I_i = \begin{cases} 1, & a_i < X \le a_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned}\mathbb{E}[|X-Y|] &= \int_0^\infty \mathbb{E}[|x - Y|X = x]\mathbb{P}\{X \in dx\} \\ &= \int_0^\infty \sum_{i=0}^\infty |x - y_i|I_i\mathbb{P}\{X \in dx\} \\ &= \sum_{i=0}^\infty \int_{a_i}^{a_{i+1}} |x - y_i|\mathbb{P}\{X \in dx\}\end{aligned}$$

Minimize $\mathbb{E}[|X - Y|]$ by minimizing each $\int_{a_i}^{a_{i+1}} |x - y_i|\mathbb{P}\{X \in dx\}$

(Remove the absolute value signs)

$$\begin{aligned}\frac{\partial}{\partial y_i}\int_{a_i}^{a_{i+1}} |x - y_i|\mathbb{P}\{X \in dx\} &= \frac{\partial}{\partial y_i}\int_{a_i}^{y_i} (y_i - x)\mathbb{P}\{X \in dx\} + \frac{\partial}{\partial y_i}\int_{y_i}^{a_{i+1}} (x - y_i)\mathbb{P}\{X \in dx\} \\ &= \left(1(y_i - y_i)\mathbb{P}\{X \in dx\} - 0(y_i - y_i)\mathbb{P}\{X \in dx\} + \int_{a_i}^{y_i} 1\mathbb{P}\{X \in dx\}\right) \\ &- \left((1(y_i - y_i)\mathbb{P}\{X \in dx\} - 0(y_i - y_i)\mathbb{P}\{X \in dx\} + \int_{a_{i+1}}^{y_i} 1\mathbb{P}\{X \in dx\}\right) \\ &= \int_{a_i}^{a_{i+1}} \mathbb{P}\{X \in dx\}\end{aligned}$$

etc.

13.3 Average

If $a_n \to L$, then

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m} a_n = L$$

Proof

Given $\varepsilon > 0$, $\exists N_1 > 0$ such that $n \geq N_1$ implies $|a_n - L| < \varepsilon/2$. Then, this implies that for $m \geq N_1$,

$$\begin{aligned}
&\frac{1}{m}((a_{N_1+1} - L) + (a_{N_1+2} - L) + \cdots + (a_m - L)) \\
\leq &\frac{1}{m}(|a_{N_1+1} - L| + \cdots + |a_m - L|) \\
< &\frac{1}{m}(m - N_1)\frac{\varepsilon}{2} \\
\leq &\frac{1}{m} m \frac{\varepsilon}{2} \\
= &\frac{\varepsilon}{2}
\end{aligned}$$

Now that N_1 is fixed, we can choose some N_2 such that

$$N_2 > \frac{2}{\varepsilon}((a_1 - L) + (a_2 - L) + \cdots + a_{N_1} - L))$$

Then, $m \geq N_2$ implies

$$\frac{1}{m}((a_1 - L) + \cdots + (a_{N_1} - L)) < \frac{\varepsilon}{2}$$

Thus, $m \geq N = \max\{N_1, N_2\}$ implies that

$$\frac{1}{m}((a_1 - L) + \cdots + (a_{N_1} - L) + (a_{N_1+1} - L) + \cdots + (a_m - L)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

so,

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m} (a_n - L) = \left(\lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m} \right) - L = 0$$