



Random Variables and Applications

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

November 11, 2024

TSABSHIN

Random Variables and Applications

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University



CUEE

November 11, 2024

Outline

- 1 Background and notations (not taught)
- 2 Random variables
- 3 Inequalities
- 4 Transform methods
- 5 Function of random variables
- 6 Random vectors
- 7 Simulation

How to read this handout

- 1 readers are assumed to have a background on univariate random variables and statistics in undergrad level (sophomore year)
- 2 the note is used with lecture in EE501 (you cannot master this topic just by reading this note) – class lectures include
 - graphical concepts, math derivation of details/steps in between
 - computer codes to illustrate examples
- 3 pay attention to the symbol ; you should be able to prove such  result
- 4 each chapter has a list of references; find more formal details/proofs from in-text citations
- 5 almost all results in this note can be Googled; readers are encouraged to ‘stimulate neurons’ in your brain by proving results without seeking help from the Internet first
- 6 typos and mistakes can be reported to jitkomut@gmail.com

Background and notations (not taught)

Outlines

- random experiments
- the axioms of probability
- conditional probability
- independence of events
- sequential experiments

Random experiments

an experiment in which the outcome varies in an unpredictable fashion when the experiment is repeated under the same conditions

examples:

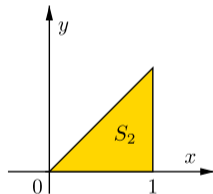
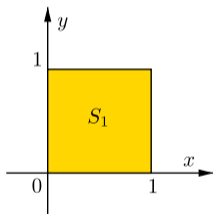
- select a ball from an urn containing balls numbered 1 to n
- toss a coin and note the outcome
- roll a dice and note the outcome
- measure the time between page requests in a Web server
- pick a number at random between 0 and 1

Sample space

the set of all possible outcomes, denoted by S

- obtained by listing all the elements, e.g., $S = \{H, T\}$, or
- giving a property that specifies the elements, e.g., $S = \{x \mid 0 \leq x \leq 3\}$

same experimental procedure may have different sample spaces



- experiment 1: pick two numbers at random between zero and one
- experiment 2: pick a number X at random between 0 and 1, then pick a number Y at random between 0 and X

Examples of sample spaces

three possibilities for the number of outcomes in sample spaces

finite, countably infinite, uncountably infinite

examples:

$$S_1 = \{1, 2, 3, \dots, 10\}$$

$$S_2 = \{HH, HT, TT, TH\}$$

$$S_3 = \{x \in \mathbb{Z} \mid 0 \leq x \leq 10\}$$

$$S_4 = \{1, 2, 3, \dots\}$$

$$S_5 = \{(x, y) \in \mathbb{R} \times \mathbb{R} \mid 0 \leq y \leq x \leq 1\}$$

$$S_6 = \text{Set of functions } X(t) \text{ for which } X(t) = 0 \text{ for } t \geq t_0$$

discrete sample space: if S is countable (S_1, S_2, S_3, S_4)

continuous sample space: if S is not countable (S_5, S_6)

Events

a subset of a sample space when the outcome satisfies certain conditions

examples: A_k denotes an event corresponding to the experiment E_k

E_1 : select a ball from an urn containing balls numbered 1 to 10

A_1 : an even-numbered ball (from 1 to 10) is selected

$$S_1 = \{1, 2, 3, \dots, 10\}, \quad A_1 = \{2, 4, 6, 8, 10\}$$

E_2 : toss a coin twice and note the sequence of heads and tails

A_2 : the two tosses give the same outcome

$$S_2 = \{HH, HT, TT, TH\}, \quad A_2 = \{HH, TT\}$$

E_3 : count # of voice packets containing only silence from 10 speakers

A_3 : no active packets are produced

$$S_3 = \{x \in \mathbb{Z} \mid 0 \leq x \leq 10\}, \quad A_3 = \{0\}$$

two events of special interest:

- **certain event**, S , which consists of all outcomes and hence always occurs
- **impossible event** or **null event**, \emptyset , which contains no outcomes and never occurs

Review of set theory

- $A = B$ if and only if $A \subset B$ and $B \subset A$
- $A \cup B$ (union): set of outcomes that are in A or in B
- $A \cap B$ (intersection): set of outcomes that are in A and in B
- A and B are *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$
- A^c (complement): set of all elements not in A
- $A \cup B = B \cup A$ and $A \cap B = B \cap A$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- DeMorgan's Rules

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

Axioms of Probability

probabilities are numbers assigned to events indicating how likely it is that the events will occur

a **probability law** is a rule that assigns a number $P(A)$ to each event A

$P(A)$ is called the *the probability of A* and satisfies the following axioms

axiom 1 $P(A) \geq 0$

axiom 2 $P(S) = 1$

axiom 3 If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$

If A_1, A_2, \dots is a sequence of events such that $A_i \cap A_j = \emptyset$ for $i \neq j$ then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

Probability Facts

- $P(A^c) = 1 - P(A)$
- $P(A) \leq 1$
- $P(\emptyset) = 0$
- If A_1, A_2, \dots, A_n are pairwise mutually exclusive then

$$P\left(\bigcup_{k=1}^n A_k\right) = \sum_{k=1}^n P(A_k)$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $A \subset B$ then $P(A) \leq P(B)$

Conditional Probability

the probability of event A given that event B has occurred

the conditional probability, $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{for } P(B) > 0$$

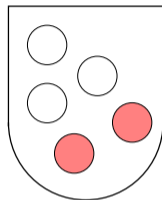
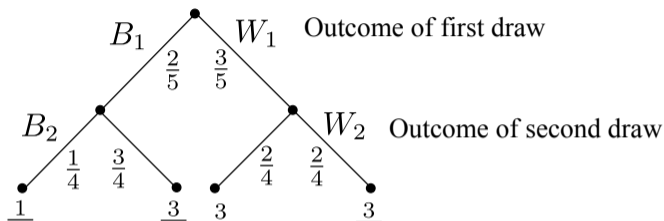
if B is known to have occurred, then A can occur only if $A \cap B$ occurs

simply renormalizes the probability of events that occur jointly with B

useful in finding probabilities in sequential experiments

Example: Tree diagram of picking balls

selecting two balls at random without replacement



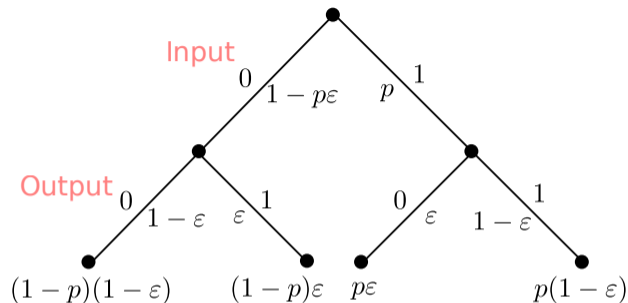
B_1, B_2 are the events of getting a black ball in the first and second draw

$$P(B_2|B_1) = \frac{1}{4}, \quad P(W_2|B_1) = \frac{3}{4}, \quad P(B_2|W_1) = \frac{2}{4}, \quad P(W_2|W_1) = \frac{2}{4}$$

the probability of a path is the *product* of the probabilities in the transition

$$P(B_1 \cap B_2) = P(B_2|B_1)P(B_1) = \frac{1}{4} \frac{2}{5} = \frac{1}{10}$$

Example: Tree diagram of Binary Communication



A_i : event the input was i ,

B_i : event the receiver was i

$$P(A_0 \cap B_0) = (1-p)(1-\varepsilon)$$

$$P(A_0 \cap B_1) = (1-p)\varepsilon$$

$$P(A_1 \cap B_0) = p\varepsilon$$

$$P(A_1 \cap B_1) = p(1-\varepsilon)$$

Theorem on Total Probability

let B_1, B_2, \dots, B_n be mutually exclusive events such that

$$S = B_1 \cup B_2 \cup \dots \cup B_n$$

(their union equals the sample space)

event A can be partitioned as

$$A = A \cap S = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$$

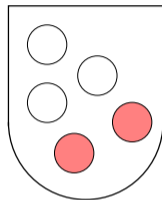
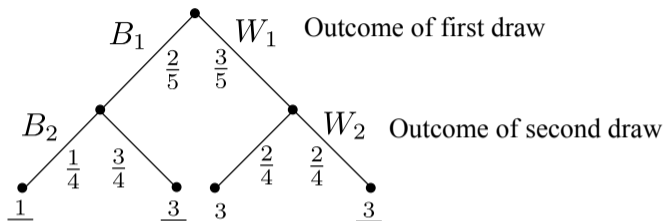
since $A \cap B_k$ are disjoint, the probability of A is

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

or equivalently,

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

Example: revisit the tree diagram of picking two balls



find the probability of the event that the second ball is white

$$\begin{aligned} P(W_2) &= P(W_2|B_1)P(B_1) + P(W_2|W_1)P(W_1) \\ &= \frac{3}{4} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{3}{5} = \frac{3}{5} \end{aligned}$$

Bayes' Rule

the conditional probability of event A given B is related to the inverse conditional probability of event B given A by

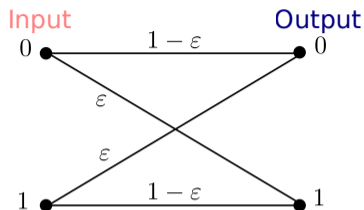
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ is called a **priori** probability
- $P(A|B)$ is called a **posteriori** probability

let A_1, A_2, \dots, A_n be a partition of S

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^n P(B|A_k)P(A_k)}$$

Example: Binary Channel



A_i event the input was i

B_i event the receiver output was i

input is equally likely to be 0 or 1

$$P(B_1) = P(B_1|A_0)P(A_0) + P(B_1|A_1)P(A_1) = \varepsilon(1/2) + (1 - \varepsilon)(1/2) = 1/2$$

applying Bayes' rule, we obtain

$$P(A_0|B_1) = \frac{P(B_1|A_0)P(A_0)}{P(B_1)} = \frac{\varepsilon/2}{1/2} = \varepsilon$$

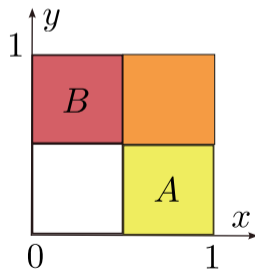
if $\varepsilon < 1/2$, input 1 is more likely than 0 when 1 is observed

Independence of events

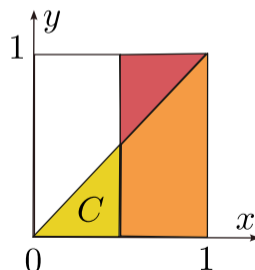
events A and B are *independent* if

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

- knowledge of event B does not alter the probability of event A
- this implies $P(A|B) = P(A)$

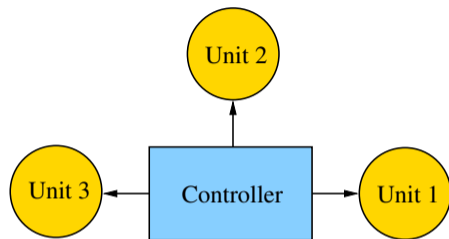


A and B are independent



A and C are not independent

Example: System reliability



- system is 'up' if the controller and at least *two* units are functioning
- controller fails with probability p
- peripheral unit fails with probability a
- all components fail independently

A : event the controller is functioning, B_i : event unit i is functioning

F : event two or more peripheral units are functioning

find the probability that the system is up

the event F can be partition as

$$\begin{aligned}F &= (B_1 \cap B_2 \cap B_3^c) \cup (B_1 \cap B_2^c \cap B_3) \cup (B_1^c \cap B_2 \cap B_3) \cup (B_1 \cap B_2 \cap B_3) \\P(F) &= P(B_1)P(B_2)P(B_3^c) + P(B_1)P(B_2^c)P(B_3) \\&\quad + P(B_1^c)P(B_2)P(B_3) + P(B_1)P(B_2)P(B_3) \\&= 3(1-a)^2a + (1-a)^3\end{aligned}$$

therefore,

$$\begin{aligned}P(\text{system is up}) &= P(A \cap F) = P(A)P(F) \\&= (1-p)P(F) = (1-p)\{3(1-a)^2a + (1-a)^3\}\end{aligned}$$

Sequential independent experiments

- consider a random experiment consisting of n **independent** experiments
- let A_1, A_2, \dots, A_n be events of the experiments
- we can compute the probability of events of the sequential experiment

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n)$$

- example: Bernoulli trial
 - perform an experiment and note if the event A occurs
 - the outcome is “success” or “failure”
 - the probability of success is p and failure is $1 - p$

Binomial probability

- perform n Bernoulli trials and observe the number of successes
- let X be the number of successes in n trials
- the probability of X is given by the **Binomial probability law**

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

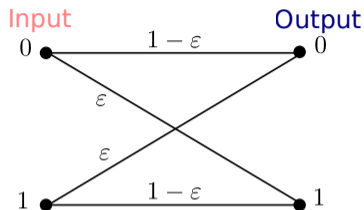
for $k = 0, 1, \dots, n$

- the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is the number of ways of picking k out of n for the successes

Example: Error Correction Coding



- transmit each bit three times
- decoder takes a majority vote of the received bits

compute the probability that the receiver makes an incorrect decision

- view each transmission as a Bernoulli trial
- let X be the number of wrong bits from the receiver

$$P(X \geq 2) = \binom{3}{2} \varepsilon^2 (1 - \varepsilon) + \binom{3}{3} \varepsilon^3$$

Multinomial probability

- generalize the binomial probability law to the occurrence of more than one event
- let B_1, B_2, \dots, B_m be possible events with

$$P(B_k) = p_k, \quad \text{and} \quad p_1 + p_2 + \dots + p_m = 1$$

- suppose n independent repetitions of the experiment are performed
- let X_j be the number of times each B_j occurs
- the probability of the vector (X_1, X_2, \dots, X_m) is given by

$$P(X_1 = k_1, X_2 = k_2, \dots, X_m = k_m) = \frac{n!}{k_1! k_2! \dots k_m!} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$$

where $k_1 + k_2 + \dots + k_m = n$

Geometric probability

- repeat independent Bernoulli trials until the the first success occurs
- let X be the number of trials until the occurrence of the first success
- the probability of this event is called the *geometric probability law*

$$P(X = k) = (1 - p)^{k-1}p, \quad \text{for } k = 1, 2, \dots$$

- the geometric probabilities sum to 1:

$$\sum_{k=1}^{\infty} P(X = k) = p \sum_{k=1}^{\infty} q^{k-1} = \frac{p}{1 - q} = 1$$

where $q = 1 - p$

- the probability that more than n trials are required before a success

$$P(X > n) = (1 - p)^n$$

Example: Error control by retransmission

- A sends a message to B over a radio link
 - B can detect if the messages have errors
 - the probability of transmission error is q
 - find the probability that a message needs to be transmitted more than two times
- each transmission is a Bernoulli trial with probability of success $p = 1 - q$

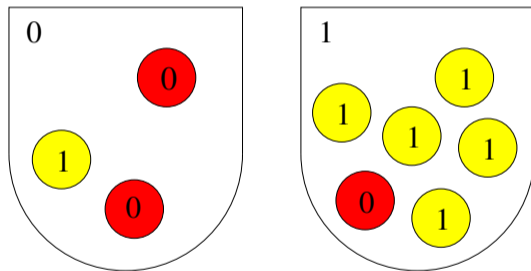
the probability that more than 2 transmissions are required is

$$P(X > 2) = q^2$$

Sequential dependent experiments

sequence of subexperiments in which the outcome of a given subexperiment determine which subexperiment is performed next

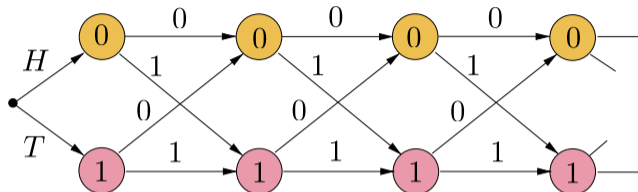
example: select the urn for the first draw by flipping a fair coin



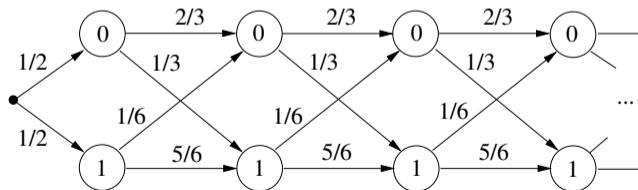
draw a ball, note the number on the ball and replace it back in its urn
the urn used in the next experiment depends on # of the ball selected

Trellis Diagram

Sequence of outcomes



Probability of a sequence of outcomes



is the product of probabilities along the path

Markov chains

let A_1, A_2, \dots, A_n be a sequence of events from n sequential experiments

the probability of a sequence of events is given by

$$P(A_1 A_2 \cdots A_n) = P(A_n | A_1 A_2 \cdots A_{n-1}) P(A_1 A_2 \cdots A_{n-1})$$

if the outcome of A_{n-1} only determines the n^{th} experiment and A_n then

$$P(A_n | A_1 A_2 \cdots A_{n-1}) = P(A_n | A_{n-1})$$

and the sequential experiments are called **Markov Chains**

thus,

$$P(A_1 A_2 \cdots A_n) = P(A_n | A_{n-1}) P(A_{n-1} | A_{n-2}) \cdots P(A_2 | A_1) P(A_1)$$

Example: find $P(0011)$ in the urn example

the probability of the sequence 0011 is given by

$$P(0011) = P(1|1)P(1|0)P(0|0)P(0)$$

where the transition probabilities are

$$P(1|1) = \frac{5}{6}, \quad P(1|0) = \frac{1}{3}, \quad P(0|0) = \frac{2}{3}$$

and the initial probability is given by

$$P(0) = \frac{1}{2}$$

hence,

$$P(0011) = \frac{5}{6} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{5}{54}$$

Discrete-time Markov chain

a Markov chain is a random sequence that has n possible states:

$$x(t) \in \{1, 2, \dots, n\}$$

with the property that

$$\mathbf{prob}(x(t + 1) = i \mid x(t) = j) = p_{ij}$$

where $P = [p_{ij}] \in \mathbf{R}^{n \times n}$

- p_{ij} is the **transition probability** from state j to state i
- P is called the **transition matrix** of the Markov chain
- the state $x(t)$ still cannot be determined with *certainty*

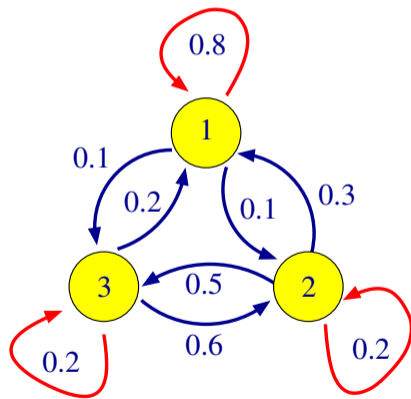
Example:

a customer may rent a car from any of three locations and return to any of the three locations

Rented from location

1	2	3	
0.8	0.3	0.2	1
0.1	0.2	0.6	2
0.1	0.5	0.2	3

Returned to location



Properties of transition matrix

let P be the transition matrix of a Markov chain

- all entries of P are real *nonnegative* numbers
- the entries in any column are summed to 1 or $\mathbf{1}^T P = \mathbf{1}^T$:

$$p_{1j} + p_{2j} + \cdots + p_{nj} = 1$$

(a property of a **stochastic matrix**)

- 1 is an eigenvalue of P
- if q is an eigenvector of P corresponding to eigenvalue 1, then

$$P^k q = q, \quad \text{for any } k = 0, 1, 2, \dots$$

Probability vector

we can represent probability distribution of $x(t)$ as n -vector

$$p(t) = \begin{bmatrix} \mathbf{prob}(x(t) = 1) \\ \vdots \\ \mathbf{prob}(x(t) = n) \end{bmatrix}$$

- $p(t)$ is called a **state probability vector** at time t
- $\sum_{i=1}^n p_i(t) = 1$ or $\mathbf{1}^T p(t) = 1$
- the state probability propagates like a linear system:

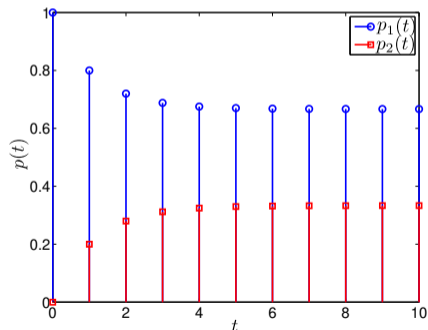
$$p(t+1) = Pp(t)$$

- the state PMF at time t is obtained by multiplying the initial PMF by P^t

$$p(t) = P^t p(0), \quad \text{for } t = 0, 1, \dots$$

Example: Markov model for packet speech

- two states of packet speech: contain 'silent activity' or 'speech activity'
- the transition matrix is $P = \begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix}$
- the initial state probability is $p(0) = (1, 0)$
- the packet in the first state is 'silent' with certainty



- eigenvalues of P are 1 and 0.4
- calculate P^t by using 'diagonalization' or 'Cayley-Hamilton theorem'

$$P^t = \begin{bmatrix} (5/3)(0.4 + 0.2 \cdot 0.4^t) & (2/3)(1 - 0.4^t) \\ (1/3)(1 - 0.4^t) & (5/3)(0.2 + 0.4^{t+1}) \end{bmatrix}$$

- $P^t \rightarrow \begin{bmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{bmatrix}$ as $t \rightarrow \infty$ (all columns are the same in limit!)
- $\lim_{t \rightarrow \infty} p(t) = \begin{bmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} p_1(0) \\ 1 - p_1(0) \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$

$p(t)$ does not depend on the *initial state probability* as $t \rightarrow \infty$

what if $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$?

- we can see that

$$P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \dots$$

- P^t does not converge but oscillates between two values

under what condition $p(t)$ converges to a constant vector as $t \rightarrow \infty$?

definition: a transition matrix is **regular** if some integer power of it has all *positive* entries

Fact: if P is regular and let w be *any* probability vector, then

$$\lim_{t \rightarrow \infty} P^t w = q$$

where q is a **fixed** probability vector, independent of t

Steady state probabilities

we are interested in the **steady state probability vector**

$$q = \lim_{t \rightarrow \infty} p(t) \quad (\text{if converges})$$

- the steady-state vector q of a regular transition matrix P satisfies

$$\lim_{t \rightarrow \infty} p(t+1) = P \lim_{t \rightarrow \infty} p(t) \quad \implies \quad Pq = q$$

(in other words, q is an eigenvector of P corresponding to eigenvalue 1)

- if we start with $p(0) = q$ then

$$p(t) = P^t p(0) = 1^t q = q, \quad \text{for all } t$$

q is also called the **stationary state PMF** of the Markov chain

Example: weather model ('rainy' or 'sunny')

probabilities of weather conditions given the weather on the preceding day:

$$P = \begin{bmatrix} 0.4 & 0.2 \\ 0.6 & 0.8 \end{bmatrix}$$

(probability that it will rain tomorrow given today is sunny, is 0.2)

given today is sunny with probability 1, calculate the probability of a rainy day in long term

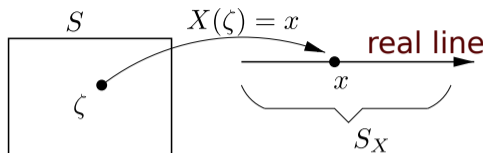
References

Chapter 2 in

A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition, Pearson Prentice Hall, 2009

Random variables

Definition



a random variable X is a *function* mapping an outcome to a real number

- the sample space, S , is the *domain* of the random variable
- S_X is the range of the random variable

example: toss a coin three times and note the sequence of heads and tails

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

Let X be the number of heads in the three tosses

$$S_X = \{0, 1, 2, 3\}$$

Types of Random Variables

Discrete RVs take values from a countable set

example: let X be the number of times a message needs to be transmitted until it arrives correctly

$$S_X = \{1, 2, 3, \dots\}$$

Continuous RVs take an infinite number of possible values

example: let X be the time it takes before receiving the next phone calls

Mixed RVs have some part taking values over an interval like typical continuous variables, and part of it concentrated on particular values like discrete variables

Probability measures

Cumulative distribution function (CDF)

$$F(a) = P(X \leq a)$$

Probability mass function (PMF) for discrete RVs

$$p(k) = P(X = k)$$

Probability density function (PDF) for continuous RVs

$$f(x) = \frac{dF(x)}{dx}$$

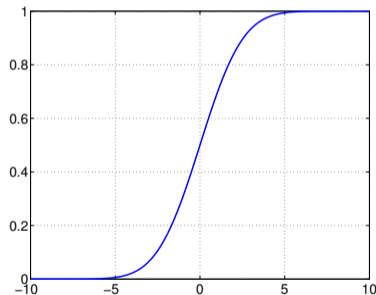
Cumulative Distribution Function (CDF)

Properties

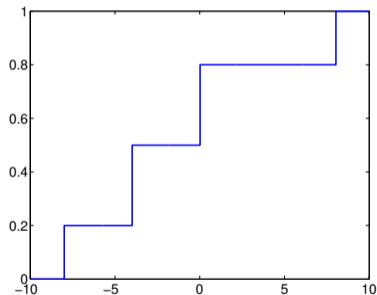
$$0 \leq F(a) \leq 1$$

$$F(a) \rightarrow 1, \quad \text{as } a \rightarrow \infty$$

$$F(a) \rightarrow 0, \quad \text{as } a \rightarrow -\infty$$



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



$$F(a) = \sum_{k \leq a} p(k)$$

Probability density function

probability density function (PDF)

- $f(x) \geq 0$
- $P(a \leq X \leq b) = \int_a^b f(x)dx$
- $F(x) = \int_{-\infty}^x f(u)du$

probability mass function (PMF)

- $p(k) \geq 0$ for all k
- $\sum_{k \in S} p(k) = 1$

Expected values

let $g(X)$ be a function of random variable X

$$\mathbf{E}[g(X)] = \begin{cases} \sum_{x \in S} g(x)p(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & X \text{ is continuous} \end{cases}$$

Mean

$$\mu = \mathbf{E}[X] = \begin{cases} \sum_{x \in S} xp(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} xf(x)dx & X \text{ is continuous} \end{cases}$$

Variance

$$\sigma^2 = \mathbf{var}[X] = \mathbf{E}[(X - \mu)^2]$$

n^{th} moment

$$\mathbf{E}[X^n]$$

Facts

Let $Y = g(X) = aX + b$, a, b are constants

- $\mathbf{E}[Y] = a\mathbf{E}[X] + b$
- $\mathbf{var}[Y] = a^2 \mathbf{var}[X]$
- $\mathbf{var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

Example of Random Variables

Discrete RVs

- Bernoulli
- Binomial
- Multinomial
- Geometric
- Negative binomial
- Poisson
- Uniform

Continuous RVs

- Uniform
- Exponential
- Gaussian (Normal)
- Gamma
- Beta
- Rayleigh
- Cauchy
- Laplacian

Bernoulli random variables

let A be an event of interest

a Bernoulli random variable X is defined as

$$X = 1 \text{ if } A \text{ occurs} \quad \text{and} \quad X = 0 \text{ otherwise}$$

it can also be given by the *indicator function* for A

$$X(\zeta) = \begin{cases} 0, & \text{if } \zeta \text{ not in } A \\ 1, & \text{if } \zeta \text{ in } A \end{cases}$$

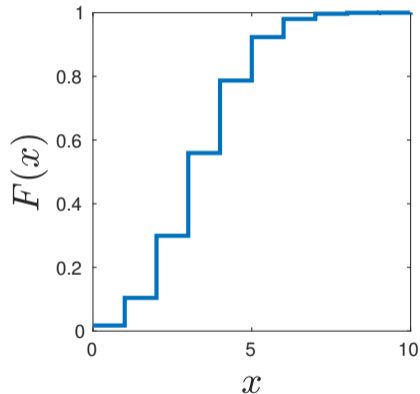
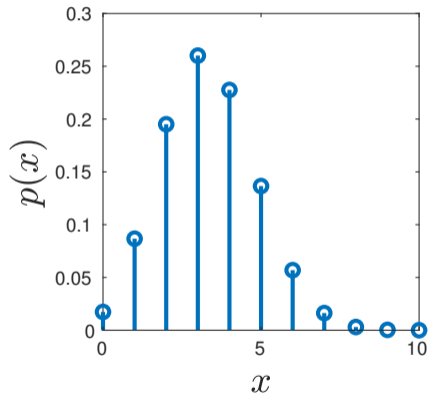
PMF: $p(1) = p, \quad p(0) = 1 - p, \quad 0 \leq p \leq 1$

Mean: $\mathbf{E}[X] = p$

Variance: $\mathbf{var}[X] = p(1 - p)$

Example

Bernoulli PMF: $p = 1/3$



Binomial random variables

- X is the number of successes in a sequence of n independent trials
- each experiment yields success with probability p
- when $n = 1$, X is a Bernoulli random variable
- $S_X = \{0, 1, 2, \dots, n\}$
- ex. Transmission errors in a binary channel: X is the number of errors in n independent transmissions

PMF
$$p(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n$$

Mean

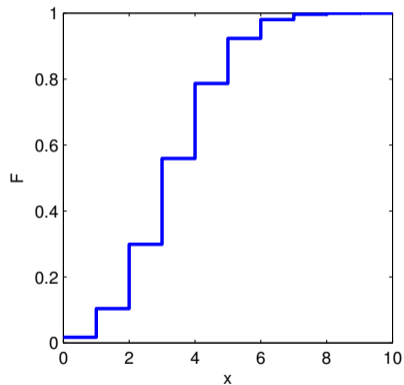
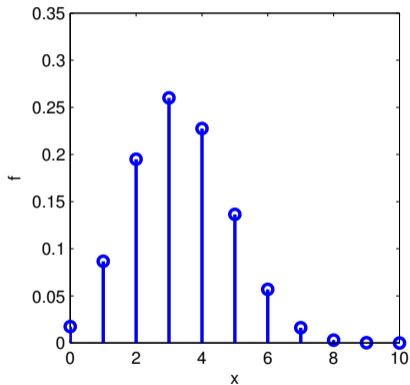
$$\mathbf{E}[X] = np$$

Variance

$$\mathbf{var}[X] = np(1 - p)$$

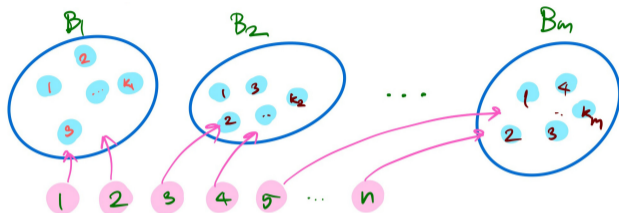
Example of Binomial PMF

$$p = 1/3, n = 10$$



Multinomial coefficient

suppose we partition a set of n objects into m subsets B_1, B_2, \dots, B_m



- B_i is assigned k_i elements and $k_1 + k_2 + \dots + k_m = n$
- denote N_i the number of possible assignments to the subset B_i

$$N_1 = \binom{n}{k_1}, N_2 = \binom{n - k_1}{k_2}, \dots, N_{m-1} = \binom{n - k_1 - k_2 - \dots - k_{m-2}}{k_{m-1}}$$

- the number of possible partitions is $N_1 N_2 \dots N_{m-1} = \frac{n!}{k_1! k_2! \dots k_m!}$ and is called the **multinomial coefficient**

Multinomial random variables

- a generalization of binomial random variables to consider a trial having more than two possible outcomes
- in each trial, there are m possible events, denoted by B_1, B_2, \dots, B_m with

$$P(B_k) = p_k, \quad \text{and} \quad p_1 + p_2 + \dots + p_m = 1$$

- suppose n independent repetitions of the experiment are performed
- let X_j be the number of times each B_j occurs

$$P(X_1 = k_1, X_2 = k_2, \dots, X_m = k_m) = \frac{n!}{k_1! k_2! \dots k_m!} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$$

where $k_1 + k_2 + \dots + k_m = n$

- the multinomial coefficient is the number of possible orderings that $X_1 = k, \dots, X_m = k_m$

PMF: the joint probability of vector $X = (X_1, X_2, \dots, X_m)$

$$P(X_1 = k_1, X_2 = k_2, \dots, X_m = k_m) = \frac{n!}{k_1! k_2! \dots k_m!} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$$

where $k_i \in \{0, 1, \dots, n\}$ and $k_1 + k_2 + \dots + k_m = n$

Mean

$$\mathbf{E}[X_i] = np_i$$

Variance

$$\mathbf{var}[X_i] = np_i(1 - p_i), \quad \mathbf{cov}(X_i, X_j) = -np_i p_j, \quad i \neq j$$

some applications:

- the data of N samples can be categorized into K classes, e.g., N subjects with blood types of A, B, AB, and O
- multinomial logistic regression in K -class classification

Geometric random variables

- repeat independent Bernoulli trials, each has probability of success p
- X is the number of experiments required until the first success occurs
- $S_X = \{1, 2, 3, \dots\}$
- ex. Message transmissions: X is the number of times a message needs to be transmitted until it arrives correctly

PMF

$$p(k) = P(X = k) = (1 - p)^{k-1}p$$

Mean

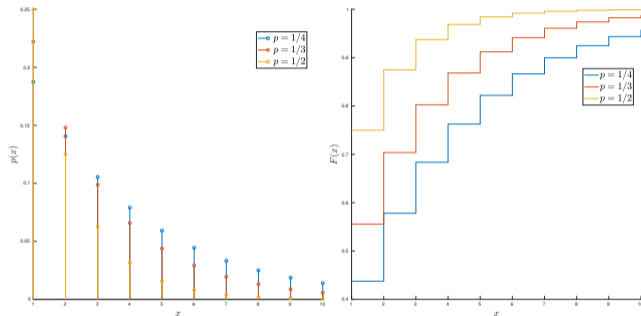
$$\mathbf{E}[X] = \frac{1}{p}$$

Variance

$$\mathbf{var}[X] = \frac{1 - p}{p^2}$$

Example of Geometric PMF

$$p = 1/4, 1/3, 1/2$$



■ parameters:

Negative binomial (Pascal) random variables

- repeat independent Bernoulli trials until observing the r^{th} success
- X is the number of trials required until the r^{th} success occurs
- X can be viewed as the sum of r geometrically RVs
- $S_X = \{r, r + 1, r + 2, \dots\}$

PMF

$$p(k) = P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \dots$$

Mean

$$\mathbf{E}[X] = \frac{r}{p}$$

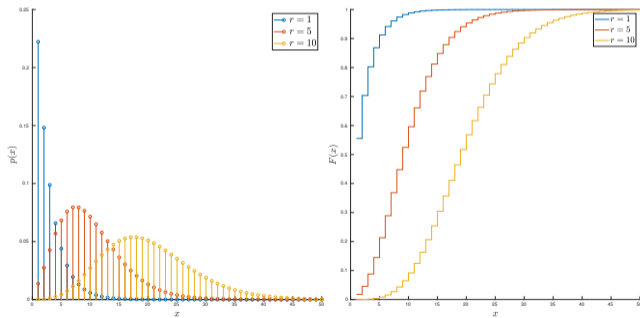
Variance

$$\mathbf{var}[X] = \frac{r(1-p)}{p^2}$$

some text defines k as the number of failures until the r th success

$$P(X = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad k = 0, 1, 2, \dots$$

example of negative binomial PMF: $r = 1, 5, 10$ and $p = 1/3$



Poisson random variables

- X is a number of events occurring in a certain period of time
- events occur with a known average rate
- the expected number of occurrences in the interval is λ
- $S_X = \{0, 1, 2, \dots\}$
- examples:
 - number of emissions of a radioactive mass during a time interval
 - number of queries arriving in t seconds at a call center
 - number of packet arrivals in t seconds at a multiplexer

PMF

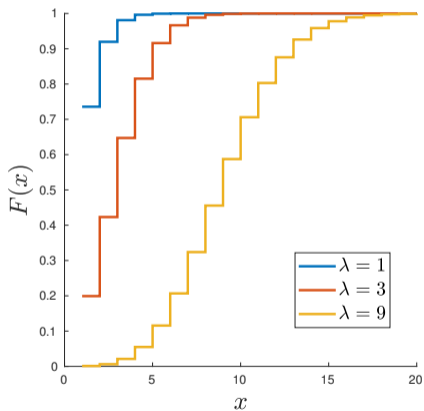
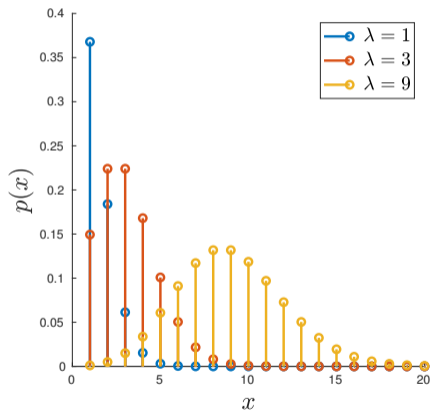
$$p(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

Mean $\mathbf{E}[X] = \lambda$

Variance $\mathbf{var}[X] = \lambda$

Example of Poisson PMF

$$\lambda = 1, 3, 9$$



Derivation of Poisson distribution

- approximate a binomial RV when n is large and p is small
- define $\lambda = np$, in 1898 Bortkiewicz showed that

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

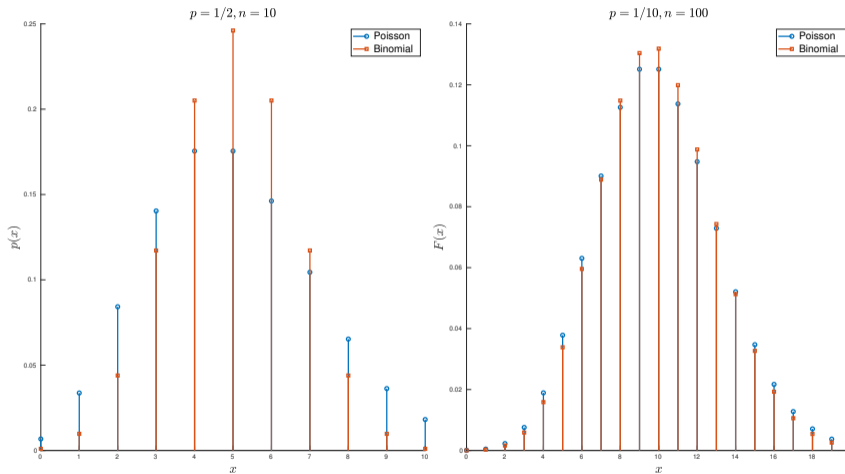
Proof.

$$p(0) = (1-p)^n = (1-\lambda/n)^n \approx e^{-\lambda}, \quad n \rightarrow \infty$$
$$\frac{p(k+1)}{p(k)} = \frac{(n-k)p}{(k+1)(1-p)} = \frac{(1-k/n)\lambda}{(k+1)(1-\lambda/n)}$$

take the limit $n \rightarrow \infty$

$$p(k+1) = \frac{\lambda}{k+1} p(k) = \left(\frac{\lambda}{k+1}\right) \left(\frac{\lambda}{k}\right) \cdots \left(\frac{\lambda}{1}\right) p(0) = \frac{\lambda^{k+1}}{(k+1)!} e^{-\lambda}$$

Comparison of Poisson and Binomial PMFs



Exponential random variables

- arise when describing the time between occurrence of events
- examples:
 - the time between customer demands for call connections
 - the time used for a bank teller to serve a customer
- λ is the rate at which events occur
- a continuous counterpart of the geometric random variable

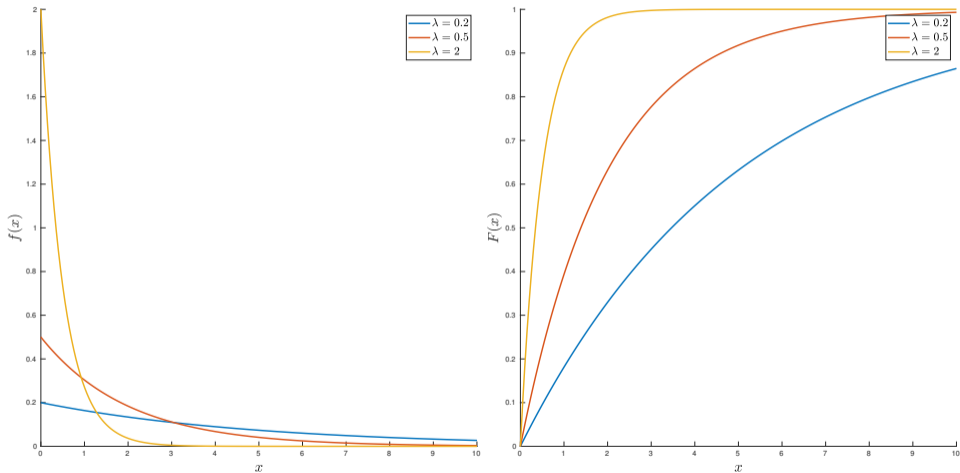
PDF

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

Mean $\mathbf{E}[X] = \frac{1}{\lambda}$

Variance $\mathbf{var}[X] = \frac{1}{\lambda^2}$

Example of Exponential PDF



Memoryless property

the property states that

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

- $P(X > t + h | X > t)$ is the probability of having to wait additionally at least h seconds given that one has already been waiting t seconds
- $P(X > h)$ is the probability of waiting at least h seconds when one first begins to wait
- thus, the probability of waiting at least an additional h seconds is the same regardless of how long one has already been waiting

Proof of memoryless property

$$\begin{aligned}P(X > t + h | X > t) &= \frac{P\{(X > t + h) \cap (X > t)\}}{P(X > t)}, \quad \text{for } h > 0 \\ &= \frac{P(X > t + h)}{P(X > t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h}\end{aligned}$$

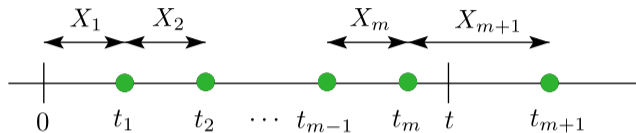
this is not the case for other non-negative continuous RVs

in fact, the conditional probability

$$P(X > t + h | X > t) = \frac{1 - P(X \leq t + h)}{1 - P(X \leq t)} = \frac{1 - F(t + h)}{1 - F(t)}$$

depends on t in general

m -Erlang random variables



- the k th event occurs at time t_k
- the times X_1, X_2, \dots, X_m between events are exponential RVs
- $N(t)$ denotes the number of events in t seconds, which is a Poisson RV
- $S_m = X_1 + X_2 + \dots + X_m$ is the elapsed time until the m th occurs

we can show that S_m is an m -Erlang random variable

Derivation of Erlang pdf

$S_m \leq t$ iff m or more events occur in t seconds

$$F(t) = P(S_m \leq t) = P(N(t) \geq m) = 1 - \sum_{k=0}^{m-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

to get the density function of S_m , we take the derivative of $F(t)$:

$$\begin{aligned} f(t) &= \frac{dF(t)}{dt} = \sum_{k=0}^{m-1} \frac{e^{-\lambda t}}{k!} \left(\lambda(\lambda t)^k - k\lambda(\lambda t)^{k-1} \right) \\ &= \frac{\lambda(\lambda t)^{m-1} e^{-\lambda t}}{(m-1)!} \Rightarrow \text{Erlang distribution with parameters } m, \lambda \end{aligned}$$

- the sum of m exponential RVs with rate λ is an m -Erlang RV
- if m becomes large, the m -Erlang RV should approach the normal RV
- from the pdf, m -erlang is a special case of gamma variable with parameter $\alpha = m$

Uniform random variables

Discrete Uniform RVs

- X has n possible values, x_1, \dots, x_n that are equally probable
- **PMF**

$$p(x) = \begin{cases} \frac{1}{n}, & \text{if } x \in \{x_1, \dots, x_n\} \\ 0, & \text{otherwise} \end{cases}$$

Continuous Uniform RVs

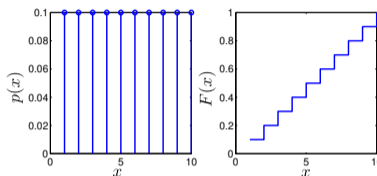
- X takes any values on an interval $[a, b]$ that are equally probable
- **PDF**

$$f(x) = \begin{cases} \frac{1}{(b-a)}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

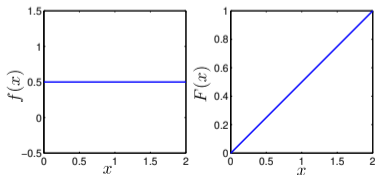
- **Mean:** $\mathbf{E}[X] = (a + b)/2$
- **Variance:** $\mathbf{var}[X] = (b - a)^2/12$

Example of discrete uniform PMF

$$X = 0, 1, 2, \dots, 10$$



Example of Continuous Uniform PMF: $X \in [0, 2]$



Gaussian (Normal) random variables

- arise as the outcome of the *central limit theorem*
- the sum of a *large* number of RVs is distributed approximately normally
- many results involving Gaussian RVs can be derived in analytical form
- let X be a Gaussian RV with parameters mean μ and variance σ^2

Notation

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

PDF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x - \mu)^2}{2\sigma^2}, \quad -\infty < x < \infty$$

Mean $\mathbf{E}[X] = \mu$

Variance $\mathbf{var}[X] = \sigma^2$

Standard Gaussian

let $Z \sim \mathcal{N}(0, 1)$ be the normalized Gaussian variable

CDF of Z is

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt \triangleq \Phi(z)$$

then CDF of $X \sim \mathcal{N}(\mu, \sigma^2)$ can be obtained by

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

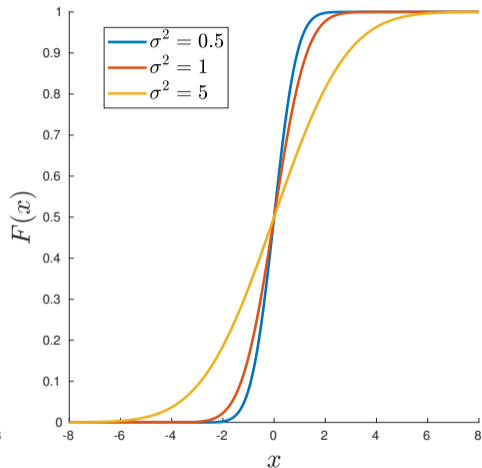
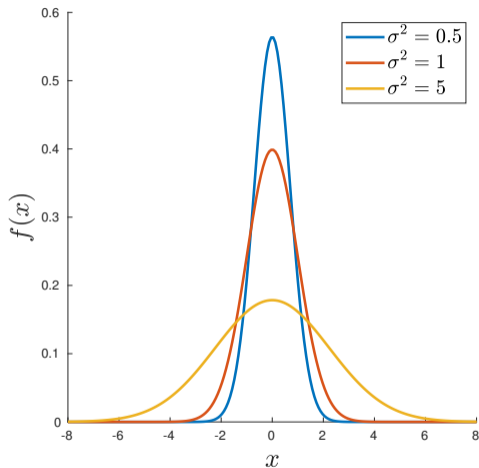
in MATLAB, the error function is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

hence, $\Phi(z)$ can be computed via the `erf` command as

$$\Phi(z) = \frac{1}{2} \left[1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right) \right]$$

Example of Gaussian PDF



- parameters: $\mu = 0, \sigma^2 = 0.5, 1, 5$

Gamma random variables

- appears in many applications:
 - the time required to service customers in queuing system
 - the lifetime of devices in reliability studies
 - the defect clustering behavior in VLSI chips
- let X be a Gamma variable with parameters α, λ

PDF

$$f(x) = \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0; \quad \alpha, \lambda > 0$$

where $\Gamma(z)$ is the gamma function, defined by

$$\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-x}dx, \quad z > 0$$

Mean $E[X] = \frac{\alpha}{\lambda}$

Variance $\text{var}[X] = \frac{\alpha}{\lambda^2}$

Properties of the gamma function

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

$$\Gamma(z + 1) = z\Gamma(z) \quad \text{for } z > 0$$

$$\Gamma(m + 1) = m!, \quad \text{for } m \text{ a nonnegative integer}$$

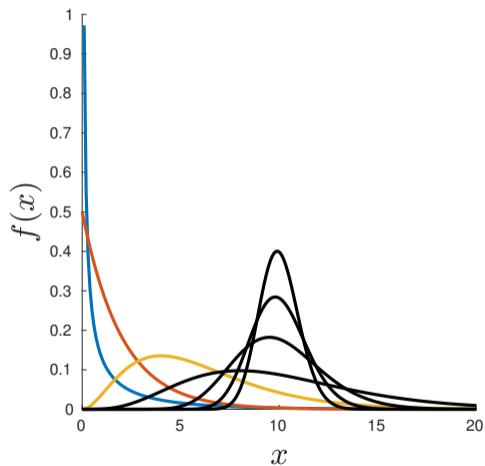
the value of $\Gamma(1/2)$ is obtained by a change of variable $u = \sqrt{x}$ to Gaussian

Special cases

a Gamma RV becomes

- exponential RV when $\alpha = 1$
- m -Erlang RV when $\alpha = m$, a positive integer
- chi-square RV with k DOF when $\alpha = k/2, \lambda = 1/2$

Example of Gamma PDF



- blue: $\alpha = 0.2, \lambda = 0.2$ (long tail)
- green: $\alpha = 1, \lambda = 0.5$ (exponential)
- red: $\alpha = 3, \lambda = 1/2$ (Chi square with 6 DOF)
- black: $\alpha = 5, 20, 50, 100$ and $\alpha/\lambda = 10$ (α -Erlang with mean 10)

Beta random variables

- used to model the randomness of percentages, proportions or ratios
- ranges of beta variables are in $[0, 1]$
- let X be a beta variable with parameters $\alpha, \beta > 0$

PDF

$$f(x) = \frac{x^{\alpha-1}(1-x)^{(\beta-1)}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1, \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (\text{beta function})$$

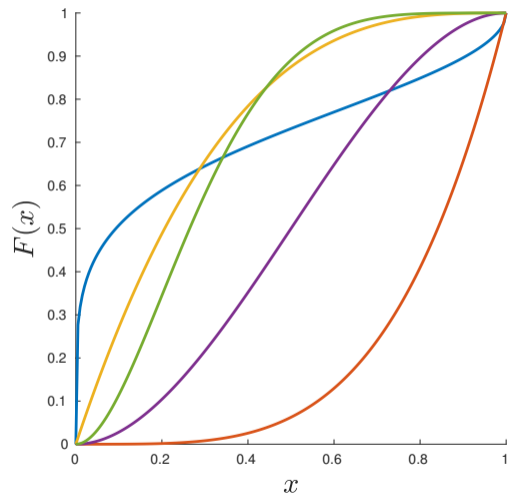
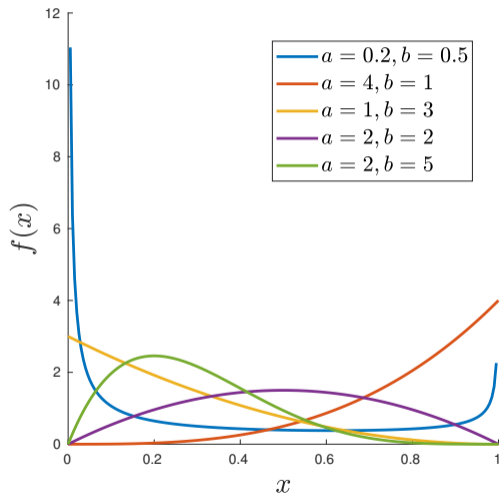
Mean: $\mathbf{E}[X] = \frac{\alpha}{\alpha + \beta}$

Variance: $\mathbf{var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

special case: $\beta = 1$ and $\alpha = \mathbf{Z}^+$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + 1)} = \frac{1}{\alpha}, \quad f(x) = \alpha x^{\alpha-1}, \quad 0 \leq x \leq 1$$

Example of beta distributions



Chi-squared random variables

- arise as a sum of k i.i.d. Gaussian variables
- ex. sample variance of i.i.d. Gaussian samples $\{X_1, \dots, X_N\}$ with variance σ^2 ; it is well-known that $(N - 1)s^2/\sigma^2$ is \mathcal{X}_{N-1}^2
- appear in asymptotic properties of estimators
- $X \sim \mathcal{X}_k^2$: chi-square variable with degree of freedom k

PDF

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x \geq 0, \quad k \in \mathbf{Z}^+$$

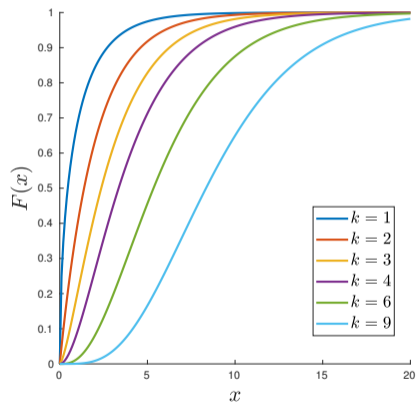
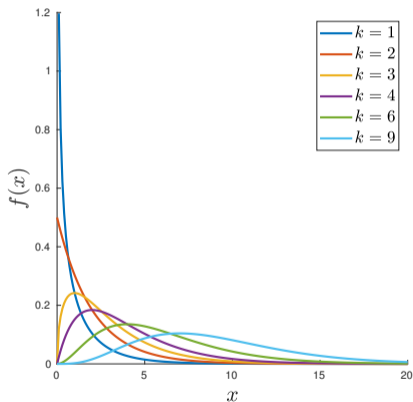
Mean

$$\mathbf{E}[X] = k$$

Variance

$$\mathbf{var}[X] = 2k$$

Example of chi-squared PDF



Rayleigh random variables

- arise when observing the magnitude of a vector
- ex. The absolute values of random complex numbers whose real and imaginary are i.i.d. Gaussian

PDF

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, \quad x \geq 0, \quad \alpha > 0$$

Mean

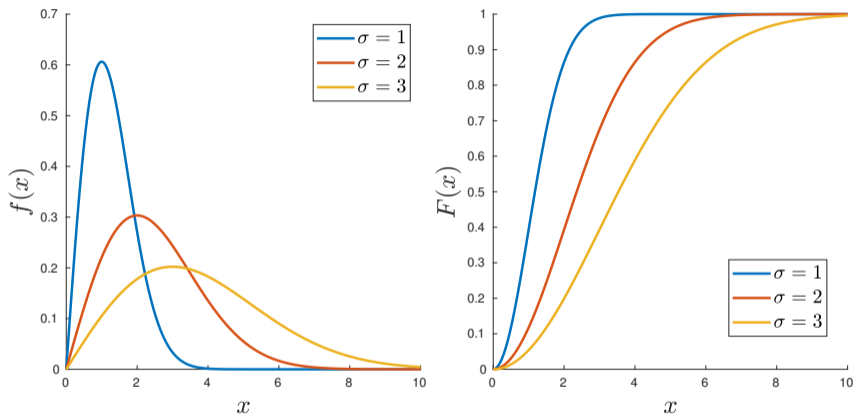
$$\mathbf{E}[X] = \sigma \sqrt{\pi/2}$$

Variance

$$\mathbf{var}[X] = \frac{4 - \pi}{2} \sigma^2$$

if X is Rayleigh, then X^2 is χ_2^2

Example of Rayleigh PDF



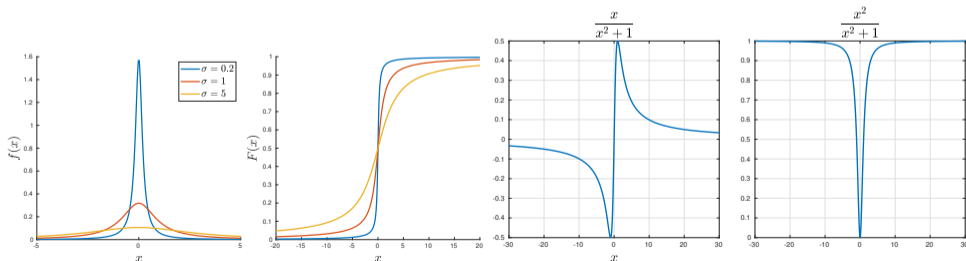
- parameters: $\sigma = 1, 2, 3$

Cauchy random variables

PDF

$$f(x) = \frac{\sigma}{\pi(x^2 + \sigma^2)}, \quad -\infty < x < \infty$$

- Cauchy distribution does not have *any moments*
- no mean, variance or higher moments defined
- $Z = X/Y$ is the standard Cauchy if X and Y are independent Gaussian



Moments of Cauchy variables

if we try to compute $\mathbf{E}[X]$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} \frac{x}{x^2 + 1} dx = \int_{-\infty}^0 \frac{x}{x^2 + 1} dx + \int_0^{\infty} \frac{x}{x^2 + 1} dx$$

the two integrals are not canceled out because each is infinite

$$\int_0^{\infty} \frac{x}{x^2 + 1} dx = (1/2) \int_0^{\infty} \frac{1}{x^2 + 1} d(x^2 + 1) = (1/2)[\log(x^2 + 1)]_0^{\infty} = \infty$$

for the second moment

$$\mathbf{E}[X^2] = 2 \int_0^{\infty} \frac{x^2}{x^2 + 1} dx = 2 \int_0^{\infty} 1 - \frac{1}{x^2 + 1} dx = \infty$$

the higher moments also diverge because the lower moments do

Laplacian random variables

PDF

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x-\mu|}, \quad -\infty < x < \infty$$

Mean

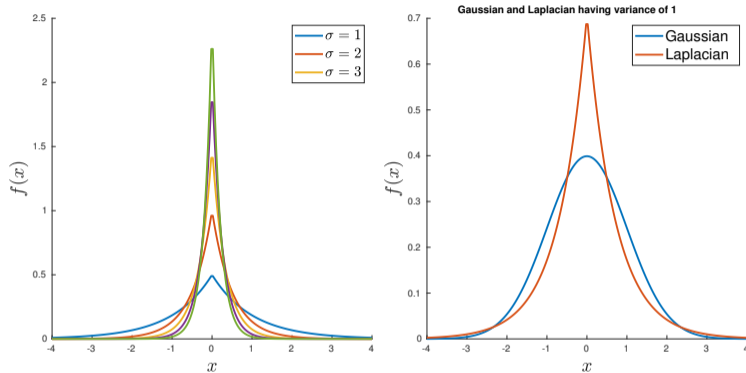
$$\mathbf{E}[X] = \mu$$

Variance

$$\mathbf{var}[X] = \frac{2}{\alpha^2}$$

- arise as the difference between two i.i.d exponential RVs
- unlike Gaussian, the Laplace density is expressed in terms of the *absolute* difference from the mean

Example of Laplacian PDF



- parameters: $\mu = 0, \alpha = 1, 2, 3, 4, 5$
- Laplacian pdf is more concentrated at the mean than pdf of Gaussian (with the same variance)

Related MATLAB commands

- `cdf` returns the values of a specified cumulative distribution function
- `pdf` returns the values of a specified probability density function
- `randn` generates random numbers from the standard Gaussian distribution
- `rand` generates random numbers from the standard uniform distribution
- `random` generates random numbers drawn from a specified distribution
- `histogram` plots a histogram of data samples

References

Chapter 3,4 in

A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition, Pearson Prentice Hall, 2009

Inequalities

Topics

- Markov inequality
- Chebyshev inequality
- Chernoff bound
- Jensen inequality

Markov inequality

let X be a *nonnegative* RV with mean $\mathbf{E}[X]$

$$P(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \quad a > 0$$

Chebyshev inequality

let X be an RV with mean μ and variance σ^2

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Example: Markov inequality

manufacturing of low grade resistors

- assume the average resistance is 100 ohms (measured by a statistical analysis)
- some of resistors have different values of resistance

if all resistors over 200 ohms will be discarded, what is the maximum fraction of resistors to meet such a criterion ?

using Markov inequality with $\mu = 100$ and $a = 200$

$$P(X \geq 200) \leq \frac{100}{200} = 0.5$$

the percentage of discarded resistors cannot exceed 50% of the total

Example: Chebyshev inequality

if the variance of the resistance is known to equal 100, find the probability that the resistance values are between 50 and 150

$$\begin{aligned}P(50 \leq X \leq 150) &= P(|X - 100| \leq 50) \\ &= 1 - P(|X - 100| \geq 50)\end{aligned}$$

by Chebyshev inequality

$$P(|X - 100| \geq 50) \leq \frac{\sigma^2}{(50)^2} = 1/25$$

hence,

$$P(50 \leq X \leq 150) \geq 1 - \frac{1}{25} = \frac{24}{25}$$

Chernoff bound

the Chernoff bound is given by

$$P(X \geq a) \leq \inf_{t \geq 0} \mathbf{E}[e^{t(X-a)}]$$

which can be expressed as

$$\log P(X \geq a) \leq \inf_{t \geq 0} \{-ta + \log \mathbf{E}e^{tX}\}$$

- $\mathbf{E}[e^{tX}]$ is the *moment generating function*
- $\log \mathbf{E}e^{tX}$ is called the *cumulant generating function*
- Chernoff bound is useful when $\mathbf{E}e^{tX}$ has an analytical expression

Example: Chernoff bound of Gaussian

X is Gaussian with zero mean and unit variance
the cumulant generating function is

$$\log \mathbf{E}[e^{tX}] = t^2/2$$

hence,

$$\log P(X \geq a) \leq \inf_{t \geq 0} \{-ta + t^2/2\} = -a^2/2$$

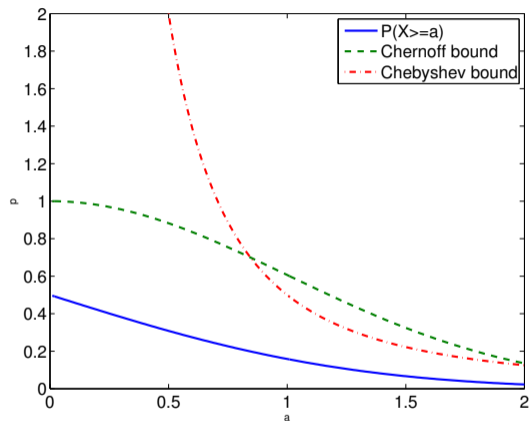
and the Chernoff bound gives

$$P(X \geq a) \leq e^{-a^2/2}$$

which is tighter than the Chebyshev inequality:

$$P(|X| \geq a) \leq 1/a^2 \quad \implies \quad P(X \geq a) \leq 1/2a^2$$

Example: Chernoff bound



when a is small, Chebyshev bound is useless while the Chernoff bound is tighter

Jensen inequality

the idea is related to the convexity of a function f

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \quad 0 \leq \theta \leq 1, \quad \forall x, y \in \mathbf{dom} f$$

Jensen's inequality: let f be a convex function and X be an RV

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$$

finite form: let f be convex and $x_1, \dots, x_n \in \mathbf{dom} f$ and $a_1, \dots, a_n > 0$

$$f\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i f(x_i)}{\sum_i a_i}$$

References

- 1 Chapter 3,4 in
A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition, Pearson Prentice Hall, 2009

Transform methods

Topics

- moment generating function (MGF)
- characteristic function (CF)

Moment generating functions

for a random variable X , the moment generating function (MGF) of X is

$$\Phi(t) = \mathbf{E}[e^{tX}]$$

Continuous

$$\Phi(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

Discrete

$$\Phi(t) = \sum_k e^{tx_k} p(x_k)$$

- except for a sign change, $\Phi(t)$ is the 2-sided Laplace transform of pdf
- the set of t for which the integral is finite forms the domain of $\Phi(t)$

Moment theorem

computing any moments of X is easily obtained by

$$\mathbf{E}[X^n] = \left. \frac{d^n \Phi(t)}{dt^n} \right|_{t=0}$$

because

$$\begin{aligned}\mathbf{E}[e^{tX}] &= \mathbf{E} \left[1 + tX + \frac{(tX)^2}{2!} + \cdots + \frac{(tX)^n}{n!} + \cdots \right] \\ &= 1 + t\mathbf{E}[X] + \frac{t^2}{2!}\mathbf{E}[X^2] + \cdots + \frac{t^n}{n!}\mathbf{E}[X^n] + \cdots\end{aligned}$$

note that $\Phi(0) = 1$

linear transformation: if $Y = aX + b$, then

$$\Phi_y(t) = e^{tb}\Phi_x(at)$$

MGF of Gaussian variables

the MGF of $X \sim \mathcal{N}(0, 1)$ is given by

$$\Phi(t) = e^{t^2/2}$$

it can be derived by completing square in the exponent:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{tx} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx$$

the MGF of $X \sim \mathcal{N}(\mu, \sigma^2)$ (affine transformation of $\mathcal{N}(0, 1)$) is

$$\Phi(t) = e^{(\mu t + \sigma^2 t^2/2)}$$

from the moment theorem, we obtain

$$\Phi'(0) = \mu, \quad \Phi''(0) = \mu^2 + \sigma^2$$

Characteristic functions

the characteristic function (CF) of a random variable X is defined by

Continuous

$$\Phi(\omega) = \mathbf{E}[e^{i\omega X}] = \int_{-\infty}^{\infty} f(x)e^{i\omega x} dx$$

Discrete

$$\Phi(\omega) = \mathbf{E}[e^{i\omega X}] = \sum_k e^{i\omega x_k} p(x_k)$$

- $\Phi(\omega)$ is simply the (inverse) Fourier transform of the PDF or PMF of X
- every pdf and its characteristic function form a unique Fourier pair:

$$\Phi(\omega) \iff f(x)$$

- it looks as if we can obtain $\Phi(\omega)$ by substituting $t = i\omega$ from MGF to CF but the existence of two transformations could be different

Properties of characteristic functions

- CF always exists because of absolute convergence (not true for MGF)

$$|\Phi(\omega)| \leq \int_{-\infty}^{\infty} |e^{i\omega x}| |f(x)| dx = \int_{-\infty}^{\infty} f(x) dx = 1$$

- CF is maximum at origin because $f(x) \geq 0$:

$$|\Phi(\omega)| \leq \Phi(0) = 1$$

- CF is self-adjoint: $\Phi(-\omega) = \Phi^*(\omega)$ (where $*$ is complex conjugate)
- CF is non-negative definite: for any real numbers w_1, w_2, \dots, w_n and complex numbers z_1, z_2, \dots, z_n

$$\sum_{j=1}^n \sum_{k=1}^n \Phi(w_j - w_k) z_j z_k^* \geq 0$$

Example: CF

Linear transformation: if $Y = aX + b$, then

$$\Phi_y(\omega) = e^{ib\omega} \Phi_x(a\omega)$$

Gaussian variables: let $X \sim \mathcal{N}(\mu, \sigma^2)$

the characteristic function of X is

$$\Phi(\omega) = e^{i\mu\omega} \cdot e^{-\sigma^2\omega^2/2}$$

(more details of applying CF to show the central limit theorem)

Binomial variables: parameters are n, p and $q = 1 - p$

$$\Phi(\omega) = (pe^{i\omega} + q)^n$$

Poisson variables: with parameter λ

$$\Phi(\omega) = e^{\lambda(e^{i\omega} - 1)}$$

Generating function

let X be a **nonnegative integer-valued** random variable

the generating function of X is defined as the z -transform of its PMF:

$$G(z) = \mathbf{E}[z^X] = \sum_{k=0}^{\infty} p(k)z^k$$

the characteristic function of X is given by $\Phi(\omega) = G(e^{i\omega})$

$G(z)$ is called the generating function due to the fact that

$$p(k) = \left. \frac{1}{k!} \frac{d^k}{dz^k} G(z) \right|_{z=0}$$

by using a similar derivation to that used in the moment theorem

Laplace Transform

let X be a *nonnegative continuous* random variable

the Laplace transform of the pdf of X is defined as

$$\mathcal{L}(s) = \mathbf{E}[e^{-sX}] = \int_0^{\infty} f(x)e^{-sx} dx$$

the moment theorem also holds for $\mathcal{L}(s)$:

$$\mathbf{E}[X^n] = (-1)^n \left. \frac{d^n}{ds^n} \mathcal{L}(s) \right|_{s=0}$$

Benefits of the transform methods

- moments of RVs are obtained by differentiating the transform
- transform of convolution integral is simply the product of transforms
i.e., the Laplace transform pair:

$$h(t) * u(t) \iff H(s)U(s)$$

- distribution of transformed variable can be derived through its characteristic function

Function of random variables

Topics

- linear and quadratic transformations
- general transformations

Functions of random variables

let X be an RV and $g(x)$ be a real-valued function defined on the real line

- $Y = g(X)$, Y is also an RV
- CDF of Y will depend on $g(x)$ and CDF of X

Example: define $g(x)$ as

$$g(x) = (x)^+ = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

- an input voltage X passes thru a halfwave rectifier
- A/D converter: a uniform quantizer maps input to the closet point
- Y is # of active speakers in excess of M , i.e., $Y = (X - M)^+$

CDF of $Y = g(X)$

probability of equivalent events:

$$P(Y \text{ in } C) = P(g(X) \text{ in } C) = P(X \text{ in } B)$$

where B is the equivalent event of values of X such that $g(X)$ is in C

Example: Voice Transmission System

- X is # of active speakers in a group of N speakers
- let p be the probability that a speaker is active
- a voice transmission system can transmit up to M signals at a time
- let Y be the number of signal discarded, so $Y = (X - M)^+$

Y take values from the set $S_Y = \{0, 1, \dots, N - M\}$

we can compute PMF of Y as

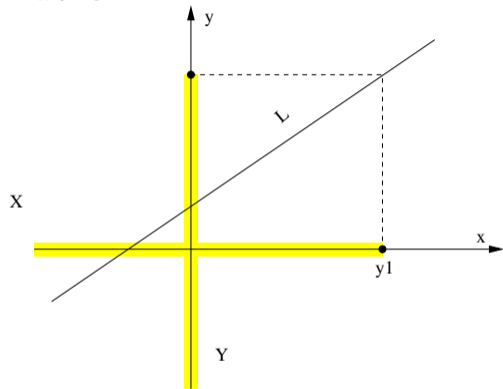
$$P(Y = 0) = P(X \text{ in } \{0, 1, \dots, M\}) = \sum_{k=0}^M p_X(k)$$

$$P(Y = k) = P(X = M + k) = p_X(M + k), \quad 0 < k \leq N - M,$$

Affine functions

define $Y = aX + b$, $a > 0$. Find CDF and PDF of Y

If $a > 0$



$$\begin{aligned}P(Y \leq y) &= P(aX + b \leq y) \\ &= P(X \leq (y - b)/a)\end{aligned}$$

thus,

$$F_Y(y) = F_X\left(\frac{y - b}{a}\right)$$

pdf of Y is obtained by differentiating the CDF wrt. to y

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y - b}{a}\right)$$

Affine function of a Gaussian

let $X \sim \mathcal{N}(m, \sigma^2)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x - m)^2}{2\sigma^2}$$

let $Y = aX + b$, with $a > 0$

from page 122,

$$f_Y(y) = \frac{1}{a} f_X \left(\frac{y - b}{a} \right) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp - \frac{(y - b - am)^2}{2(a\sigma)^2}$$

- Y has also a Gaussian distribution with mean $b + am$ and variance $(a\sigma)^2$
- thus, a linear function of a Gaussian is also a Gaussian

Example: Quadratic functions

define $Y = X^2$. find CDF and PDF of Y

for a positive y , we have

$$\{Y \leq y\} \iff \{-\sqrt{y} \leq X \leq \sqrt{y}\}$$

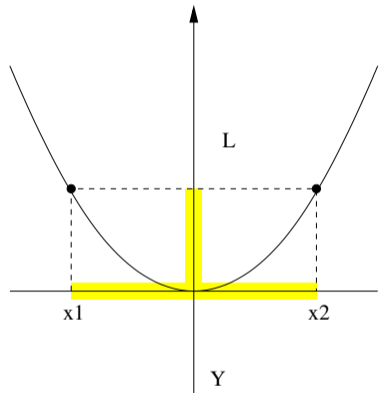
thus,

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}), & y > 0 \end{cases}$$

differentiating wrt. to y gives

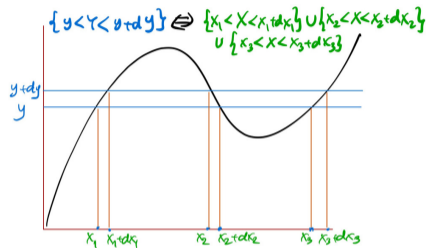
$$f_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}}$$

for $X \sim \mathcal{N}(0, 1)$, Y is a chi-square random variable with one dof



General functions of random variables

suppose $Y = g(X)$ is a transformation (could be many-to-one)



suppose $y = g(x)$ has n roots:

$$y = g(x_1) = g(x_2) = \cdots = g(x_n)$$

two equivalent events:

$$\{y < Y < y + dy\} \Leftrightarrow \bigcup_{k=1}^n \{x_k < X < x_k + dx_k\}$$

the probabilities of two equivalent events are approximately

$$f_Y(y)|dy| = f_X(x_1)|dx_1| + f_X(x_2)|dx_2| + \cdots + f_X(x_n)|dx_n|$$

$$f_Y(y) = \frac{f_X(x_1)}{|g'(x_1)|} + \cdots + \frac{f_X(x_n)}{|g'(x_n)|}$$

where $g'(x)$ is the derivative (Jacobian) of $g(x)$

Examples: affine and quadratic

affine: $Y = aX + b$, $g'(x) = a$

the equation $y = ax + b$ has a single solution $x = (y - b)/a$ for every y , so

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right)$$

quadratic: $Y = aX^2$, $a > 0$, $g'(x) = 2ax$

if $y \leq 0$, then the equation $y = ax^2$ has no real solutions, so $f_Y(y) = 0$

if $y > 0$, then it has two solutions

$$x_1 = \sqrt{y/a}, \quad x_2 = -\sqrt{y/a}$$

and therefore

$$f_Y(y) = \frac{1}{2\sqrt{ay}} \left(f_X(\sqrt{y/a}) + f_X(-\sqrt{y/a}) \right)$$

Log of uniform variables

verify that if X has a standard uniform distribution $\mathcal{U}(0, 1)$, then

$$Y = -\log(X)/\lambda$$

has an exponential distribution with parameter λ

for $Y = y$, we can solve $X = x = e^{-\lambda y} \Rightarrow$ unique root

- the Jacobian is $g'(x) = -\frac{1}{\lambda x} = -e^{\lambda y}/\lambda$
- when $y < 0$, $x = e^{-\lambda y} \notin [0, 1]$; hence, $f_Y(y) = 0$
- when $y \geq 0$ (or $e^{-\lambda y} \in [0, 1]$), we will have

$$f_Y(y) = \frac{f_X(e^{-\lambda y})}{|-1/\lambda x|} = \lambda e^{-\lambda y}$$

Amplitude samples of a sinusoidal waveform

let $Y = \cos X$ where $X \sim \mathcal{U}(0, 2\pi]$, find the pdf of Y

for $|y| > 1$ there is no solution of $x \Rightarrow f_Y(y) = 0$

for $|y| < 1$ the equation $y = \cos x$ has two solutions:

$$x_1 = \cos^{-1}(y), \quad x_2 = 2\pi - x_1$$

the Jacobians are

$$g'(x_1) = -\sin(x_1) = -\sin(\cos^{-1}(y)) = -\sqrt{1-y^2}, \quad g'(x_2) = \sqrt{1-y^2}$$

since $f_X(x) = 1/2\pi$ in the interval $(0, 2\pi]$, so

$$f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}, \quad \text{for } -1 < y < 1$$

note that although $f_Y(\pm 1) = \infty$ the probability that $y = \pm 1$ is 0

Random vectors

Random vectors

we denote X a random vector

X is a function that maps each outcome ζ to a vector of real numbers

an n -dimensional random variable has n components:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

also called a *multivariate* or *multiple* random variable

Probabilities

Joint CDF

$$F(X) \triangleq F_X(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

Joint PMF

$$p(X) \triangleq p_X(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

Joint PDF

$$f(X) \triangleq f_X(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F(X)$$

Marginal PMF

$$p_{X_j}(x_j) = P(X_j = x_j) = \sum_{x_1} \dots \sum_{x_{j-1}} \sum_{x_{j+1}} \dots \sum_{x_n} p_X(x_1, x_2, \dots, x_n)$$

Marginal PDF

$$f_{X_j}(x_j) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(x_1, x_2, \dots, x_n) dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_n$$

Conditional PDF: the PDF of X_n given X_1, \dots, X_{n-1} is

$$f(x_n | x_1, \dots, x_{n-1}) = \frac{f_X(x_1, \dots, x_n)}{f_{X_1, \dots, X_{n-1}}(x_1, \dots, x_{n-1})}$$

Characteristic Function

the characteristic function of an n -dimensional RV is defined by

$$\begin{aligned}\Phi(\omega) = \Phi(\omega_1, \dots, \omega_n) &= \mathbf{E}[e^{i(\omega_1 X_1 + \dots + \omega_n X_n)}] \\ &= \int_X e^{i\omega^T X} f(X) dX\end{aligned}$$

where

$$\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$\Phi(\omega)$ is the n -dimensional Fourier transform of $f(X)$

Independence

the random variables X_1, \dots, X_n are **independent** if

the joint pdf (or pmf) is equal to the product of their marginal's

Discrete

$$p_X(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

Continuous

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

we can specify an RV by the characteristic function in place of the pdf,

X_1, \dots, X_n are *independent* if

$$\Phi(\omega) = \Phi_1(\omega_1) \cdots \Phi_n(\omega_n)$$

Expected Values

the expected value of a function

$$g(X) = g(X_1, \dots, X_n)$$

of a vector random variable X is defined by

$$\mathbf{E}[g(X)] = \int_x g(x) f(x) dx \quad \text{Continuous}$$

$$\mathbf{E}[g(X)] = \sum_x g(x) p(x) \quad \text{Discrete}$$

Mean vector

$$\mu = \mathbf{E}[X] = \mathbf{E} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{E}[X_1] \\ \mathbf{E}[X_2] \\ \vdots \\ \mathbf{E}[X_n] \end{bmatrix}$$

Correlation and Covariance matrices

Correlation matrix has the second moments of X as its entries:

$$R \triangleq \mathbf{E}[XX^T] = \begin{bmatrix} \mathbf{E}[X_1X_1] & \mathbf{E}[X_1X_2] & \cdots & \mathbf{E}[X_1X_n] \\ \mathbf{E}[X_2X_1] & \mathbf{E}[X_2X_2] & \cdots & \mathbf{E}[X_2X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}[X_nX_1] & \mathbf{E}[X_nX_2] & \cdots & \mathbf{E}[X_nX_n] \end{bmatrix}$$

with

$$R_{ij} = \mathbf{E}[X_iX_j]$$

Covariance matrix has the second-order central moments as its entries:

$$C \triangleq \mathbf{E}[(X - \mu)(X - \mu)^T]$$

with

$$C_{ij} = \mathbf{cov}(X_i, X_j) = \mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

Properties of correlation and covariance matrices

let X be a (real) n -dimensional random vector with mean μ

Facts:

- R and C are $n \times n$ symmetric matrices
- R and C are positive semidefinite
- If X_1, \dots, X_n are independent, then C is diagonal
- the diagonals of C are given by the variances of X_k
- if X has zero mean, then $R = C$
- $C = R - \mu\mu^T$

Cross Correlation and Cross Covariance

let X, Y be vector random variables with means μ_X, μ_Y respectively

Cross Correlation

$$\mathbf{cor}(X, Y) = \mathbf{E}[XY^T]$$

if $\mathbf{cor}(X, Y) = 0$ then X and Y are said to be **orthogonal**

Cross Covariance

$$\begin{aligned}\mathbf{cov}(X, Y) &= \mathbf{E}[(X - \mu_X)(Y - \mu_Y)^T] \\ &= \mathbf{cor}(X, Y) - \mu_X \mu_Y^T\end{aligned}$$

if $\mathbf{cov}(X, Y) = 0$ then X and Y are said to be **uncorrelated**

Affine transformation

let Y be an affine transformation of X :

$$Y = AX + b$$

where A and b are deterministic matrices

- $\mu_Y = A\mu_X + b$

$$\mu_Y = \mathbf{E}[AX + b] = A\mathbf{E}[X] + \mathbf{E}[b] = A\mu_X + b$$

- $C_Y = AC_X A^T$

$$\begin{aligned} C_Y &= \mathbf{E}[(Y - \mu_Y)(Y - \mu_Y)^T] = \mathbf{E}[(A(X - \mu_X))(A(X - \mu_X))^T] \\ &= A\mathbf{E}[(X - \mu_X)(X - \mu_X)^T]A^T = AC_X A^T \end{aligned}$$

Gaussian random vector

X_1, \dots, X_n are said to be **jointly Gaussian** if their joint pdf is given by

$$f(X) \triangleq f_X(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp - \frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)$$

μ is the mean ($n \times 1$) and $\Sigma \succ 0$ is the covariance matrix ($n \times n$):

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \cdots & \Sigma_{nn} \end{bmatrix}$$

and

$$\mu_k = \mathbf{E}[X_k], \quad \Sigma_{ij} = \mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

Example

the joint density function of X (not normalized) is given by

$$f(x_1, x_2, x_3) = \exp - \frac{x_1^2 + 3x_2^2 + 2(x_3 - 1)^2 + 2x_1(x_3 - 1)}{2}$$

- f is an exponential of *negative quadratic* in x so X must be a Gaussian

$$f(x_1, x_2, x_3) = \exp - \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \\ x_3 - 1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 - 1 \end{bmatrix}$$

- the mean vector is $(0, 0, 1)$ and the covariance matrix is

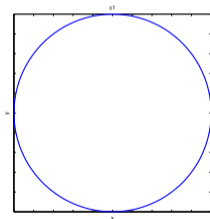
$$C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 1/3 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

- the variance of x_1 is highest while x_2 is smallest
- x_1 and x_2 are uncorrelated, so are x_2 and x_3

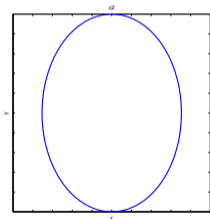
Gaussian density contour

examples of Gaussian density contour (the exponent of exponential)

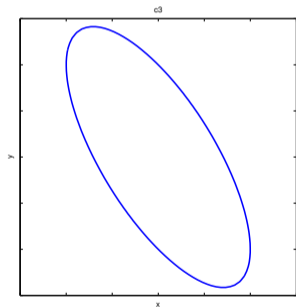
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

Properties of Gaussian variables

many results on Gaussian RVs can be obtained analytically:

- marginal's of X is also Gaussian
- conditional pdf of X_k given the other variables is a Gaussian distribution
- uncorrelated Gaussian random variables are *independent*
- any affine transformation of a Gaussian is also a Gaussian

these are well-known facts

and more can be found in the areas of estimation, statistical learning, etc.

Characteristic function of Gaussian

$$\Phi(\omega) = \Phi(\omega_1, \omega_2, \dots, \omega_n) = e^{i\mu^T \omega} e^{-\frac{\omega^T \Sigma \omega}{2}}$$

Proof. By definition and arranging the quadratic term in the power of \exp

$$\begin{aligned}\Phi(\omega) &= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \int_X e^{iX^T \omega} e^{-\frac{(X-\mu)^T \Sigma^{-1} (X-\mu)}{2}} dx \\ &= \frac{e^{i\mu^T \omega} e^{-\frac{\omega^T \Sigma \omega}{2}}}{(2\pi)^{n/2} |\Sigma|^{1/2}} \int_X e^{-\frac{(X-\mu-i\Sigma\omega)^T \Sigma^{-1} (X-\mu-i\Sigma\omega)}{2}} dx \\ &= \exp(i\mu^T \omega) \exp\left(-\frac{1}{2} \omega^T \Sigma \omega\right)\end{aligned}$$

(the integral equals 1 since it is a form of Gaussian distribution)
for one-dimensional Gaussian with zero mean and variance $\Sigma = \sigma^2$,

$$\Phi(\omega) = e^{-\frac{\sigma^2 \omega^2}{2}}$$

Affine Transformation of a Gaussian is Gaussian

let X be an n -dimensional Gaussian, $X \sim \mathcal{N}(\mu, \Sigma)$ and define

$$Y = AX + b$$

where A is $m \times n$ and b is $m \times 1$ (so Y is $m \times 1$)

$$\begin{aligned}\Phi_Y(\omega) &= \mathbf{E}[e^{i\omega^T Y}] = \mathbf{E}[e^{i\omega^T (AX+b)}] \\ &= \mathbf{E}[e^{i\omega^T AX} \cdot e^{i\omega^T b}] = e^{i\omega^T b} \Phi_X(A^T \omega) \\ &= e^{i\omega^T b} \cdot e^{i\mu^T A^T \omega} \cdot e^{-\omega^T A \Sigma A^T \omega / 2} \\ &= e^{i\omega^T (A\mu + b)} \cdot e^{-\omega^T A \Sigma A^T \omega / 2}\end{aligned}$$

we read off that Y is Gaussian with mean $A\mu + b$ and covariance $A\Sigma A^T$

Marginal of Gaussian is Gaussian

the k^{th} component of X is obtained by

$$X_k = [0 \ \cdots \ 1 \ 0] X \triangleq \mathbf{e}_k^T X$$

(\mathbf{e}_k is a standard unit column vector; all entries are zero except the k^{th} position)

hence, X_k is simply a linear transformation (in fact, a projection) of X

X_k is then a Gaussian with mean

$$\mathbf{e}_k^T \mu = \mu_k$$

and covariance

$$\mathbf{e}_k^T \Sigma \mathbf{e}_k = \Sigma_{kk}$$

Uncorrelated Gaussians are independent

suppose (X, Y) is a jointly Gaussian vector with

$$\text{mean } \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \quad \text{and covariance } \begin{bmatrix} C_X & 0 \\ 0 & C_Y \end{bmatrix}$$

in other words, X and Y are *uncorrelated* Gaussians:

$$\text{cov}(X, Y) = \mathbf{E}[XY^T] - \mathbf{E}[X]\mathbf{E}[Y]^T = 0$$

the joint density can be written as

$$\begin{aligned} f_{XY}(x, y) &= \frac{1}{(2\pi)^n |C_X|^{1/2} |C_Y|^{1/2}} \exp -\frac{1}{2} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^T \begin{bmatrix} C_X^{-1} & 0 \\ 0 & C_Y^{-1} \end{bmatrix} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} \\ &= \frac{1}{(2\pi)^{n/2} |C_X|^{1/2}} e^{-\frac{1}{2}(x-\mu_x)^T C_X^{-1}(x-\mu_x)} \cdot \frac{1}{(2\pi)^{n/2} |C_Y|^{1/2}} e^{-\frac{1}{2}(y-\mu_y)^T C_Y^{-1}(y-\mu_y)} \end{aligned}$$

proving the independence

Conditional of Gaussian is Gaussian

let Z be an n -dimensional Gaussian which can be decomposed as

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix} \right)$$

the conditional pdf of X given Y is also Gaussian with conditional mean

$$\mu_{X|Y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (Y - \mu_y)$$

and conditional covariance

$$\Sigma_{X|Y} = \Sigma_x - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{xy}^T$$

Proof:

from the **matrix inversion lemma**, Σ^{-1} can be written as

$$\Sigma^{-1} = \begin{bmatrix} S^{-1} & -S^{-1}\Sigma_{xy}\Sigma_{yy}^{-1} \\ -\Sigma_{yy}^{-1}\Sigma_{xy}^T S^{-1} & \Sigma_{yy}^{-1} + \Sigma_{yy}^{-1}\Sigma_{xy}^T S^{-1}\Sigma_{xy}\Sigma_{yy}^{-1} \end{bmatrix}$$

where S is called the **Schur complement of Σ_{xx} in Σ** and

$$\begin{aligned} S &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T \\ \det \Sigma &= \det S \cdot \det \Sigma_{yy} \end{aligned}$$

we can show that $\Sigma \succ 0$ if and only if $S \succ 0$ and $\Sigma_{yy} \succ 0$

from $f_{X|Y}(x|y) = f_X(x, y)/f_Y(y)$, we calculate the exponent terms

$$\begin{aligned} & \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} - (y - \mu_y)^T \Sigma_{yy}^{-1} (y - \mu_y) \\ &= (x - \mu_x)^T S^{-1} (x - \mu_x) - (x - \mu_x)^T S^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y) \\ & \quad - (y - \mu_y)^T \Sigma_{yy}^{-1} \Sigma_{xy}^T S^{-1} (x - \mu_x) \\ & \quad + (y - \mu_y)^T (\Sigma_{yy}^{-1} \Sigma_{xy}^T S^{-1} \Sigma_{xy} \Sigma_{yy}^{-1}) (y - \mu_y) \\ &= [x - \mu_x - \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y)]^T S^{-1} [x - \mu_x - \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y)] \\ &\triangleq (x - \mu_{X|Y})^T \Sigma_{X|Y}^{-1} (x - \mu_{X|Y}) \end{aligned}$$

$f_{X|Y}(x|y)$ is an exponential of quadratic function in x

so it has a form of Gaussian

Standard Gaussian vectors

for an n -dimensional Gaussian vector $X \sim \mathcal{N}(\mu, C)$ with $C \succ 0$

let A be an $n \times n$ invertible matrix such that

$$AA^T = C$$

(A is called a **factor** of C)

then the random vector

$$Z = A^{-1}(X - \mu)$$

is a standard Gaussian vector, *i.e.*,

$$Z \sim \mathcal{N}(0, I)$$

(obtain A via eigenvalue decomposition or Cholesky factorization)

Quadratic Form Theorems

let $X = (X_1, \dots, X_n)$ be a standard n -dimensional Gaussian vector:

$$X \sim \mathcal{N}(0, I)$$

then the following results hold

- $X^T X \sim \chi^2(n)$
- let A be a symmetric and *idempotent* matrix and $m = \text{tr}(A)$ then

$$X^T A X \sim \chi^2(m)$$

Proof

an eigenvalue decomposition of A : $A = UDU^T$ where

$$\lambda(A) = 0, 1 \quad U^T U = U U^T = I$$

it follows that

$$X^T A X = X^T U D U^T X = Y^T D Y = \sum_{i=1}^n d_{ii} Y_i^2$$

- since U is orthogonal, Y is also a standard Gaussian vector
- since A is idempotent, d_{ii} is either 0 or 1 and $\text{tr}(D) = m$

therefore $X^T A X$ is the m -sum of standard normal RVs

Simulation

Outlines

- statistic, sampling distribution
- why is simulation useful?
 - proof of concept
 - when analysis is difficult to obtain
- pseudo-random number generation

Statistic

definition: suppose X_1, X_2, \dots, X_N are the observed random variables, then the random variable

$$T = g(X_1, X_2, \dots, X_N)$$

is called a **statistic**

examples:

- 1 sample mean, sample median, sample mode, sample variance
- 2 sample moments: kurtosis, skewness
- 3 order statistic: sample maximum and minimum
- 4 test statistic: t -statistic, chi-squared statistic, F statistic
- 5 sample quantiles

a statistic can provide an inference for the random variables X_k 's

Sampling distribution

setting: suppose X_1, X_2, \dots, X_N are random samples from a distribution involving a parameter θ

definition: let T be a function of X_1, X_2, \dots, X_N and possibly θ

$$T = g(X_1, X_2, \dots, X_N, \theta) \quad (\text{a statistic})$$

the distribution of T (given θ) is called the **sampling distribution** of T

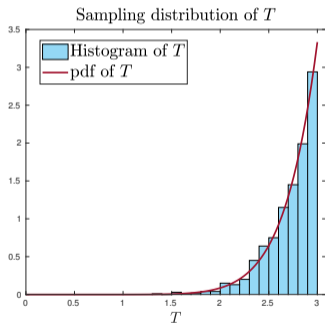
typically, a sampling distribution depends on


- original distribution of X_k 's
- the number of observations N
- type of statistic (here, function g)


Example: Sampling distribution of order statistic

X_k 's are independent uniform on $[0, \theta]$, $\theta = 3$

$T = \max(X_1, X_2, \dots, X_N)$ (T can provide as an estimate of θ)

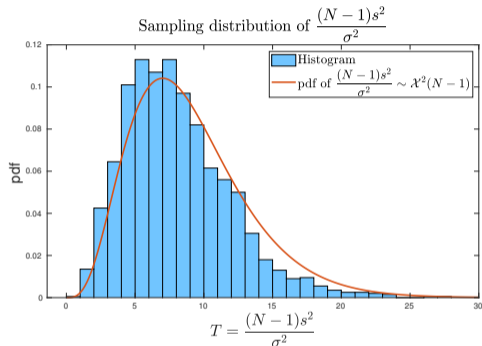
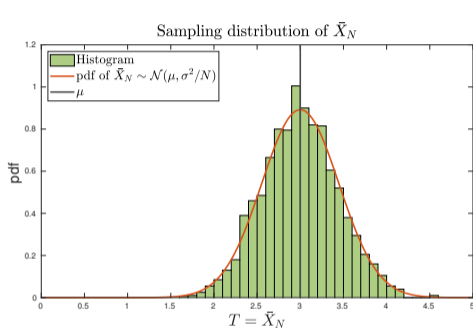


- $f_T(t) = \frac{Nt^{N-1}}{\theta^N}$ for $0 \leq t \leq \theta$  pls verify
- use $N = 10$ and simulate 1000 realizations of T
- the sampling distribution is NOT the same as X_k 's (which is uniform)

check point : f_T depends on i) max function ii) N and iii) distribution of X_k 's

Example: Sampling distribution of sample mean and variance

X_1, \dots, X_N are normal with mean $\mu = 3$ and $\sigma^2 = 2$ and $N = 10$



- $T = \bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i \sim \mathcal{N}(\mu, \sigma^2/N)$
- $T = s^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)^2$ and $\frac{(N-1)s^2}{\sigma^2} \sim \chi^2(N-1)$

Why is simulation useful?

- proof of concept
- when analysis becomes too difficult
- when T is a statistic of samples X_1, \dots, X_N and we need to discuss about statistical properties of T (mean, median, variance, etc)
 - 1 samples X_1, \dots, X_N are generated and compute T
 - 2 repeat step 1 B times, we have $T^{(1)}, T^{(2)}, \dots, T^{(B)}$
 - 3 calculate a summary statistic Z of $T^{(1)}, T^{(2)}, \dots, T^{(B)}$

the distribution of Z will be called **simulation distribution** or **Monte Carlo distribution**

example: on page 159, we used $B = 2000$, simulation mean of \bar{X}_N is 2.9968 and simulation variance of \bar{X}_N is 0.2042

Sampling distribution of quantile

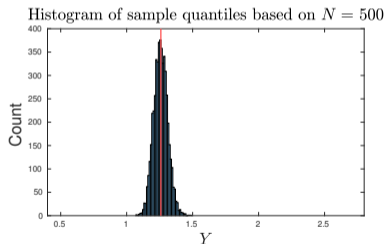
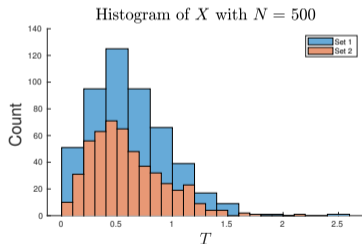
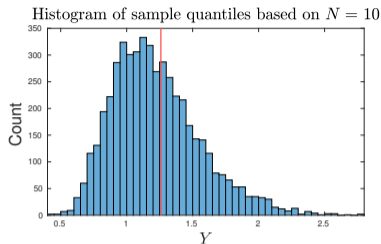
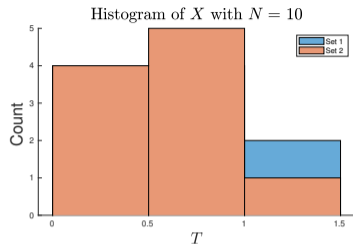
let X_1, \dots, X_N be gamma variables with $\alpha = 3$ and $\beta = 5$

- Y is the sample τ -quantile with $\tau = 0.95$ based on N samples X_k 's
- the exact τ -quantile is $q_\tau = F_X^{-1}(\tau) = 1.2592$
- repeat and generate $B = 5000$ replications of $Y^{(1)}, \dots, Y^{(B)}$
- plot the sampling distribution of Y and compute statistic such as the simulation variance
- fact: the sample τ -quantile from N samples has asymptotic distribution

$$\sqrt{N}(Y - q_\tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{[f_X(q_\tau)]^2}\right)$$

- simulation variance of Y should decrease as N is large

Simulation: sample quantile



- simulation variances of Y are 0.1091 ($N = 10$) and 0.0027 ($N = 500$)
- when N is large, the sampling distribution of Y is close to normal

Median of a complicated distribution

let $X|\lambda$ be an exponential with parameter λ but λ itself is random with pdf f_λ

what if we want to find the median of X ? \Rightarrow maybe we should derive $f_X(x)$ first ?

example: $f_\lambda(\lambda) = 3\lambda^2$ for $0 \leq \lambda \leq 1$

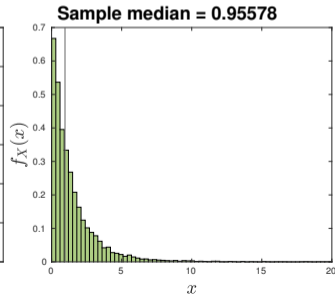
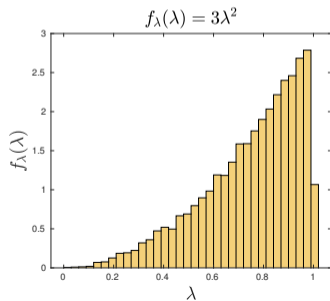
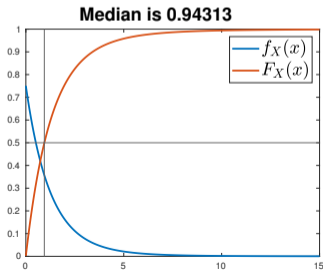
$$\begin{aligned}f_X(x) &= \int_0^1 f_{X|\lambda}(x, \lambda) d\lambda = \int_0^1 f_{X|\lambda}(x|\lambda) f_\lambda(\lambda) d\lambda = \int_0^1 \lambda e^{-\lambda x} 3\lambda^2 d\lambda \\&= \int_0^1 3\lambda^3 e^{-\lambda x} d\lambda = -\frac{e^{-\lambda x}}{x^4} [6 + 6\lambda x + 3\lambda^2 x^2 + \lambda^3 x^3]_0^1 \\&= \frac{3}{x^4} [6 - e^{-x}(6 + 6x + 3x^2 + x^3)]\end{aligned}$$

it is complicated to integrate f_X (and then calculate $F_X^{-1}(1/2)$)

Simulation: Median of complicated distribution

alternative to finding analytical expression of the marginal f_X , we can

- 1 generate B samples of $\lambda \lambda^{(1)}, \dots, \lambda^{(B)}$ according to f_λ
- 2 **simulate** $X^{(i)}$ having exponential distribution with parameter $\lambda^{(i)}$, for $i = 1, 2, \dots, B$
- 3 $X^{(1)}, \dots, X^{(B)}$ would be a random sample of f_X
- 4 the median of samples $X^{(1)}, \dots, X^{(B)}$ should be close to the median of distribution f_X



Mean of $|X - Y|$

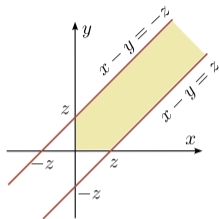
setting:

- servers A and B serve customers with the same rate of λ customers per minute
- when each one finish serving m customers, they take a break and agree to meet
- they would finish the tasks at different time, so one has to wait for another
- question: on average, how long will one of the servers have to wait for the other?

modeling:

- service times of both servers are independent exponential RVs with parameter λ
- let X and Y be the times that server A and B takes to finish serving m customers
- X and Y are m -Erlang RV with rate λ (special case of Gamma distribution)
- the time one server has to wait for the other is $|X - Y|$
- we want to find $\mathbf{E}[|X - Y|]$

analysis I: define $Z = |X - Y|$, derive pdf of Z , and find the integral



$$\begin{aligned} F_Z(z) &= P(|X - Y| \leq z) = P(-z \leq X - Y \leq z) \\ &= \int \int f_{XY}(x, y) dy dx \quad (\text{over shaded region}) \\ &= \int_0^z \int_0^{x+z} f_{XY}(x, y) dy dx + \int_z^\infty \int_{x-z}^{x+z} f_{XY}(x, y) dy dx \end{aligned}$$

even $f_{XY}(x, y)$ can be simplified to product of two Erlang pdfs, we still have to

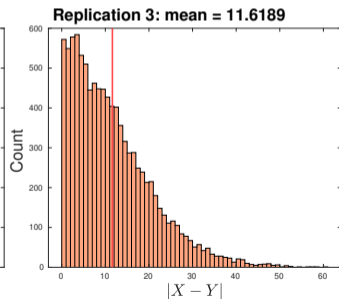
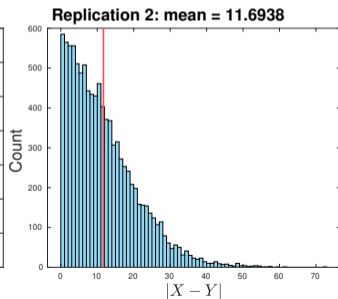
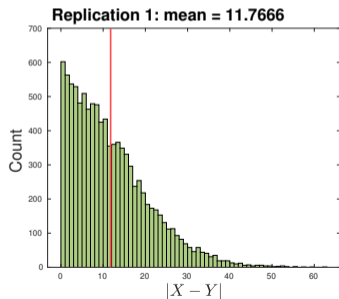
- derive pdf of Z (taking derivative of $F_Z(z)$)
- derive $\mathbf{E}[Z] = \int z f_Z(z) dz$

analysis II: derive MGF of $|X - Y|$

$$\Phi(t) = \mathbf{E}[e^{t|X-Y|}] = \int_0^\infty \int_0^\infty e^{t|x-y|} f_X(x) f_Y(y) dy dx$$

Simulation: mean of $|X - Y|$

- generate 10,000 samples of X and Y as independent 10-Erlang with $\lambda = 0.3$
- compute $Z = |X - Y|$
- plot the histogram of Z and sample mean of Z
- this is called 1 replication; repeat for 3 replications to see how the approximate of $\mathbf{E}[Z]$ could vary



Generate samples from distribution

there are many algorithms

- inverse transformation method
- acceptance-rejection method

in this handout, we generate samples distributed from a given analytical expression of pdf/cdf of continuous RVs

- there are other efficient methods for particular distributions
- also other methods for generate samples from pmf of discrete RVs

Inverse of the probability integral transform

Theorem: for a continuous RV X with cdf F_X , the transformed variable $U = F_X(X)$ is uniform on $[0, 1]$

↷ also, if U is a uniform $[0, 1]$ the transformed variable $F^{-1}(U)$ has distribution function F

proof: let $U \sim \mathcal{U}[0, 1]$ if we can find a strictly monotone transformation

$$T : [0, 1] \rightarrow \mathbf{R} \quad \text{such that} \quad T(U) = X$$

we will have the following

- $F_X(x) = P(X \leq x) = P(T(U) \leq x) = P(U \leq T^{-1}(x)) = T^{-1}(x)$
(the last step uses the fact that U is uniform)
- we have $F_X(X) = T^{-1}(X) = U$

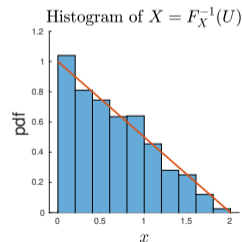
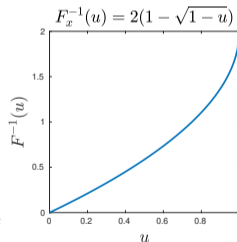
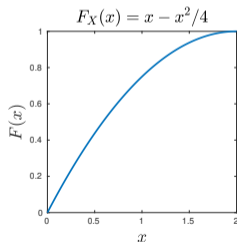
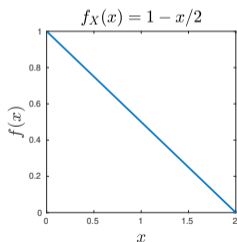
this result is used to generate RV with distribution F provided that F^{-1} is known

Inverse transform sampling

setting: generate X with cdf $F_X(x)$ (that is continuous and $F_X^{-1}(x)$ is known)

inversion method:

- 1 generate U , a uniform variable in $[0, 1]$
- 2 return $X = F_X^{-1}(U)$



Examples of F_X^{-1} formula

list of distributions where the inverse cdf can be provided explicitly

name	range	$f_X(x)$	$F_X(x)$	$X = F_X^{-1}(U)$	simplified
expo	$x \geq 0$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$-\frac{1}{\lambda} \log(1 - U)$	$-\frac{1}{\lambda} \log(U)$
Rayleigh	$x \geq 0$	$\frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$	$1 - e^{-\frac{x^2}{2\sigma^2}}$	$\sigma \sqrt{-\log(1 - U)}$	$\sigma \sqrt{-\log(U)}$
triangular	$0 \leq x \leq a$	$\frac{2}{a} \left(1 - \frac{x}{a}\right)$	$\frac{2}{a} \left(x - \frac{x^2}{2a}\right)$	$a(1 - \sqrt{1 - U})$	$a(1 - \sqrt{U})$
beta	$0 \leq x \leq 1$	$\alpha x^{\alpha-1}$	x^α	$U^{1/\alpha}$	
Cauchy	R	$\frac{\sigma}{\pi(x^2 + \sigma^2)}$	$\frac{1}{2} + \frac{1}{\pi} \tan^{-1}(x/\sigma)$	$\sigma \tan(\pi(U - 1/2))$	$\sigma \tan(\pi U)$

the simplified version is obtained by noting that $1 - U$ is distributed as U

when F_X^{-1} is not available analytically, numerical solution to $F_X(X) = U$ is applied

References

- 1 Luc Devroye, *Non-uniform random variate generation*, Springer, 1986
- 2 Chapter 12 in D. Schervish, *Probability and Statistics*, fourth edition, Pearson, 2012