Probability

Xi Tan (tan19@purdue.edu)

January 22, 2017

Contents

Preface

This booklet is divided into 7 Chapters. The first chapter introduces the definitions of basic concepts, such as event, sample space, and probability space. Followed in the next chapter, we will discuss the relationship between two or more events when they interplay with each other. The third chapter formally brings in random variables and vectors, as a basis to develop their quantitative measure and characteristic functions later in chapter four. Chapter five includes some well-known limit theorems, which is useful for asymptotic analysis. The last two chapters will discuss several selected topics in probability theory, and provide a summary of common distributions.

Part I

Elementary Theory of Probability

Chapter 1

Combinatorial Analysis

1.1 Axioms

There are two important rules in combinatorics: the rule of sum, and the rule of product.

The rule of sum says, if we have a ways to finish a task using one method and alternatively, b ways to finish the same task using another method, then there are ab ways of finish this task. More generally,

$$|S_1 \cup S_2 \cup \ldots \cup S_n| = |S_1| + |S_2| + \ldots + |S_n| \tag{1.1}$$

One extension of the rule of sum is the inclusion-exclusion principle, which does not require sets A_i to be disjoint. This does include the rule of sum, in that if sets A_i are disjoint, the terms from the second to the last are all zero.

$$|\bigcup_{i=1}^{n} A_i| = \sum_{i=1}^{n} |A_i| - \sum_{1 \le i < j \le n} |A_i \cap A_j| + \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| - \cdots$$
$$+ (-1)^{n-1} |A_1 \cap \cdots \cap A_n| \tag{1.2}$$

The rule of product says, if finishing one task requires two steps, and there are a ways to choose in the first step and b ways to choose in the second step, then there are ab ways to finish this task. More generally,

$$|S_1 \times S_2 \times \cdots \times S_n| = |S_1| \cdot |S_2| \cdots |S_n| \tag{1.3}$$

1.2 Binomial Coefficient and Its Applications

Binomial Coefficient

We list here some of the useful binomial identities, all numbers are nature number (not including 0).

$$\binom{n}{k} = \binom{n}{n-k} \tag{1.4}$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \tag{1.5}$$

$$\binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1} \tag{1.6}$$

From the famous Pascal's rule,

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1} \tag{1.7}$$

There is another form which is equivalent to equation 1.7,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{1.8}$$

Here is an example that uses the *logarithmic differentiation*, $f' = f\big[\ln(f)\big]'$.

$$\frac{d}{dt}\binom{t}{k} = \binom{t}{k}\sum_{i=0}^{k-1}\frac{1}{t-i} \tag{1.9}$$

A list of series that involves binomial coefficients,

$$\sum_{k=0}^{n}\binom{n}{k} = 2^n \tag{1.10}$$

$$\sum_{k=0}^{n}k\binom{n}{k} = n2^{n-1} \tag{1.11}$$

$$\sum_{k=0}^{n}k^2\binom{n}{k} = (n+n^2)2^{n-2} \tag{1.12}$$

These can all be obtained by examining the function value or derivatives of the function $(1+x)^\alpha$, where α could be any real number, and $|x| < 1$.

There are some identities that could be proved using combinatorial analysis, such as *double counting*. Here is an example,

$$\sum_{k=1}^{n}\binom{n}{k}\binom{k}{q} = 2^{n-q}\binom{n}{q} \tag{1.13}$$

The left side of equation 1.13 counts the number of ways of selecting k elements first, and then choosing q elements from the resulting subset. These q elements could be identical for different k. The right hand side of the equation says this is equivalent to first choosing q elements directly from the set, and merging them into one of the 2^{n-q} subset of the set containing all but those selected q elements.

Another example is,

$$\sum_{m_1=0}^{n_1}\binom{n_1}{m_1}\binom{n_2}{m-m_1} = \binom{n}{m} \tag{1.14}$$

This simply means choosing $m = m_1 + m_2$ objects from a set of $n = n_1 + n_2$ objects is equivalent to choosing m_1 objects from n_1 objects, and m_2 objects from n_2 objects.

Sometimes, knowing the bounds and asymptotic formulae could be helpful.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{n\cdot e}{k}\right)^k \tag{1.15}$$

$$\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}, \text{ as } n \to \infty \tag{1.16}$$

Bernoulli Distribution

The i.i.d. Case: Binomial Distribution

The Batch Mode Case: Hypergeometric Distribution

1.3 Multinomial Coefficient and Its Applications

Multinomial Coefficient

The notion of *multinomial coefficient* is a generalization of binomial coefficient, which is defined in the multinomial theorem:

$$(x_1 + x_2 + \ldots + x_r)^n = \sum_{(n_1, \ldots, n_r): n_1 + \ldots + n_r = n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}$$

We call $\binom{n}{n_1, n_2, \ldots, n_r}$ the multinomial coefficient.

Problem 1.3.1 A set of n distinct items is to be divided into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$, where $\sum_{i=1}^{r} n_i = n$. How many different divisions are possible?

Note that there are $\binom{n}{n_1}$ possible choices for the first group; for each choice of the first group there are $\binom{n}{n - n_1}$ possible choices for the second group; and so on. Hence it follows that there are

$$\binom{n}{n_1} \binom{n - n_1}{n_2} \cdots \binom{n - n_1 - n_2 - \ldots - n_{r-1}}{n_r}$$
$$= \frac{n!}{(n - n_1)! n_1!} \frac{(n - n_1)!}{(n - n_1 - n_2)! n_2!} \cdots \frac{(n - n_1 - n_2 - \ldots - n_{r-1}!}{(0)! n_r!}$$
$$= \frac{n!}{n_1! n_2! \ldots n_r!}$$

possible divisions.

Alternatively, we can first permute these n items, where there are $n!$ such orderings. The first n_1 elements are assigned to group 1, the next n_2 elements are assigned to group 2, and so on. However, for example, keeping all but n_i group fixed, this method would generate $n_i!$ equivalent divisions (note the order within a group does not matter). Therefore, we need to cancel out the equivalent-group effect by dividing $n_1! n_2! \ldots n_r!$. Finally, the multinomial coefficient is,

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!} \tag{1.17}$$

1.4 Categorical Distribution

1.5 Multinomial Distribution

1.6 Multiset Coefficient and Its Applications

Multiset Coefficient

The notion of multiset (or bag) is a generalization of the notion of set in which members are allowed to appear more than once.

The number of times an element belongs to the multiset is the *multiplicity* of that member. The total number of elements in a multiset, including repeated memberships, is the *cardinality* of the multiset. For example, in the multiset {a, a, b, b, b, c} the multiplicities of the members a, b, and c are respectively 2, 3, and 1, and the cardinality of the multiset is 6.

The number of multisets of cardinality k, with elements taken from a finite set of cardinality n, is called the *multiset* coefficient or *multiset number*, and is denoted as $\left(\!\binom{n}{k}\!\right)$. It is equivalent to asking, with

replacement, the number of all possible combinations of making k draws from a urn with n distinguishable balls labeled $1 \ldots n$.

Problem 1.6.1 With replacement, how many possible combinations to make k draws from a urn with n distinguishable balls labeled $1 \ldots n$?

NOT the sum
ll the multi-
al coefficients,
h can be seen
omputing $(1 +$
$1)^k$, that is,
$\binom{k}{N_1,N_2,\ldots,N_n}$ $=$

cts being ad-
t ensures the
emptiness.

If we translate this prob directly to the combinatorial language, we may end up with counting the total number of the multinomial coefficients (actually, it is also correct). To solve this prob, let's first see a similar example.

Problem 1.6.2 Suppose k balls that are indistinguishable from each other are to be distributed into n distinguishable (non-empty) urns, how many different outcomes are possible?

We note that this prob is equivalent to selecting $n-1$ of the $k-1$ spaces between (fixed) adjacent objects as our dividing points, e.g., $OOO|OOO|OO$. We count the "bars" as urns (in total n) and big "O" as balls (in total k). Therefore, there are $\binom{k-1}{n-1}$ such outcomes.

Now, we are ready to go back to our original prob. prob 1.6.1 could be asked this way: if however, we allow empty urns, how many outcomes are possible?

of probability 1
emove one ball
each urn.

One possible way of borrowing the non-empty case solution to solve the empty one is by starting with $k+n$ balls (instead of k), and place them into n urns, however at last remove one ball from each urn (so in total n). Because some of the urns may only contain one ball, this would give us the number of orderings $\binom{n+r-1}{r-1}$. Another way to look at it is to allow all $k+n-1$ positions (not gaps) available to both symbols, and we count all balls between two bars the same type. Hence, the number of all possible combinations is $\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$. Note, this scheme is allowing emptiness, because two "bars" can be adjacent to each other.

Another beautiful explanation is to construct an equivalent mapping. Note, the rule of drawing a series of k numbers $a_1, a_2, \ldots, a_k$ from the set $\{1, 2, \ldots, n\}$ with repetition is

$$1 \le a_1 \le a_2 \le \ldots \le a_k \le n$$

Now, a new series of k numbers $b_1, b_2, \ldots, b_k$ can be constructed as follows

$$
\begin{array}{cccc}
1 \le a_1 < & a_2+1 < & \ldots < & a_k+k-1 \le n+k-1 \\
\downarrow & \downarrow & & \downarrow \\
b_1 & b_2 & \ldots & b_k
\end{array}
$$

Note that $b_1 < b_2 < \ldots < b_k$. This is a model without replacement, and is a one-to-one mapping of the original prob. Under the model of drawing k times from $n+k-1$ balls without replacement, the number of all possible combinations is $\binom{n+k-1}{k}$, which is the same as what we obtained earlier.

A last thing worth noting is that, as we mentioned earlier, the number of all multinomial coefficients is also $\binom{k+n-1}{n-1}$.

1.7 Selected Topics

Double Factorial

Stirling Numbers

1.8 The Bertrand's Ballot prob

The Bertrand's ballot prob was first introduced by Joseph Bertrand in 1887, in the form of: "In an election where candidate A receives p votes and candidate B receives q votes with $p > q$, what is the probability that A will be strictly ahead of B throughout the count?"

J. Bertrand himself gave the solution $\frac{p-q}{p+q}$ by using mathematical induction in the original paper. First of all, let's consider the initial case. At the first count, the vote can be either for candidate A or B. If

it *could* be for candidate A, the simplest scenario is $p = 1, q = 0$, which is of probability 1 for candidate A to win the vote. This agrees with the formula. If it *could* be for candidate B, the simplest scenario is $p = 2, q = 1$, now there are $\binom{2+1}{1} = 3$ counting orders, i.e., AAB, ABA, or BAA. Note "AAA" is the only favorable order out of three possible orders. This again agrees with the formula. Assume the theorem is true when $p = a - 1$ and $q = b$ (last vote would be for candidate A), and when $p = a$ and $q = b - 1$ (last vote would be for candidate B), which is the two possible scenarios at the second to last count. Now, considering the case with $p = a$ and $q = b$, the last vote is either for candidate A with probability $a/(a + b)$, or for candidate B with probability $b/(a + b)$. So the probability of candidate A to always lead the count is:

$$\frac{a}{a+b}\frac{(a-1)-b}{(a-1)+b} + \frac{b}{a+b}\frac{a-(b-1)}{a+(b-1)} = \frac{a-b}{a+b}$$

This proves that the theorem is true for all $p > q \geq 0$.

Désiré André in the same year gave an elegant proof of this prob, and the method is now called "the principle of reflection", or "the André's reflection method", or "the Bertrand's ballot theorem". There are three facts need to noted. One is that, the sequences start with A or B with probability $p/(p + q)$ and $q/(p + q)$, respectively. The second fact is that, sequences start with B will for sure to tie at some point because A will finally win, and they are all unfavorable, because A already "loses" at the first count. The third is that, sequences that start with A can be classified into two cases, one is the case that A leads the counting process from beginning to the end which is the favorable case, the other is the case when the sequence will tie at some point which is the unfavorable case, importantly, the number of sequences in this latter case is the same as that of the sequences starting with B, because there is a bijection mapping. The probability is $1 - 2 \times \frac{q}{p+q} = \frac{p-q}{p+q}$.

The number of the unfavorable cases is $2 \times \binom{p+q-1}{q-1}$, of which half starts with "A" and another half starts with "B", as explained above. The number of favorable cases is $\binom{p+q-1}{p-1} - \binom{p+q-1}{q-1}$. Actually, $\frac{\binom{p+q-1}{p-1} - \binom{p+q-1}{q-1}}{\binom{p+q}{p}} = \frac{p-q}{p+q}$.

Consider now the prob to find the probability that the second candidate is never ahead (i.e. ties are allowed); the solution is $\frac{p+1-q}{p+1}$. This is simply seen by awaring the following equivalent description:

- same as the basic version, ties are NOT allowed; but,

- there are $p + 1$ votes for candidate A and q votes for candidate B;

- the first vote is for candidate A;

The probability can then be computed as:

$$P(\text{A winning } with \text{ ties}) = P(\text{A winning } without \text{ ties} \mid \text{the first vote is A})$$
$$= \frac{P(\text{A winning without ties } and \text{ the first vote is A})}{P(\text{the first vote is A})}$$
$$= \frac{(p+1-q)/(p+1+q)}{(p+1)/(p+1+q)} = \frac{p+1-q}{p+1}$$

Another way to look at this prob is to model it as the following: represent a voting sequence as a lattice path on the Cartesian plane and,

- Start the path at $(0, 0)$;

- Each time a vote for the first candidate is received move right 1 unit;

- Each time a vote for the second candidate is received move up 1 unit.

Each such path corresponds to a unique sequence of votes and will end at (p, q). A sequence is "good" exactly when the corresponding path never goes above the diagonal line $y = x$; equivalently, a sequence is "bad" exactly when the corresponding path touches the line $y = x + 1$. For each "bad" path P, define a new path P' by reflecting the part of P up to the first point it touches the line across it. P' is a path from (-1, 1) to

$\frac{(a-1)-b}{(a-1)+b}$ and $\frac{a-(}{a+(}$ are conditional abilities, which conditioned on last count.

One possible ping is to denot sequence as where "L" and " the left and right of the first tie tion (must be a and then conve each character to its altern e.g., AABBA would be BBABABAA.

(p, q). The same operation applied again restores the original P. This produces a one-to-one correspondence between the "bad" paths and the paths from (-1, 1) to (p, q). The number of these paths is $\binom{p+q}{q-1}$. So the probability asked is $\frac{\binom{p+q}{q}-\binom{p+q}{q-1}}{\binom{p+q}{p}} = \frac{p+1-q}{p+1}$.
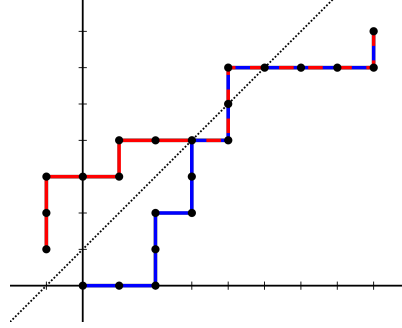

Figure 1.1: Bertrand's Ballot prob allowing Ties

An interesting application of this is the famous Catalan number formula, which can be introduced under the random walk story. A random walk on the integers is to take n steps of unit length, beginning at the origin and ending at the point m, that never become negative. Assuming n and m have the same parity and $n \geq m \geq 0$, this number is, according to the Bertrand's Ballot prob allowing ties,

$$\binom{n}{\frac{n+m}{2}} - \binom{n}{\frac{n+m}{2}+1} = \frac{m+1}{\frac{n+m}{2}+1}\binom{n}{\frac{n+m}{2}}$$

Here, $p + q = n$ and $p - q = m$, compared to our used settings. When $m = 0$ and n is even, this gives the Catalan number $\frac{1}{\frac{n}{2}+1}\binom{n}{\frac{n}{2}}$.

Let's tweak this prob a little bit more. Let's say candidate A starts at a "bonus" votes, not 0. That is to say, the system goes from $(0, a)$ to $(p + q, p + a - q)$. If we do not allow ties, all paths hit the x-axis will be unfavorable, and the number of these paths equals the number of paths from $(0, a)$ to its "mirror" point $(p + q, -p - a + q)$. So, there are in total $\binom{p+q}{p+a}$ unfavorable paths, and the probability is $1 - \frac{\binom{p+q}{p+a}}{\binom{p+q}{p}}$. Note when $a = 0$, there is already a tie, and the question really should be asked as: "What if the first count is A, and then the process never has a tie". So the probability should compute as:

$$1 - \frac{p}{p+q}\frac{\binom{p-1+q}{p-1}}{\binom{p-1+q}{p-1}} = \frac{p-q}{p+q}$$

It should have no prob with $a > 0$.

If we allow ties, the "mirror" point should be reflected against $y = -1$, so it becomes $(p+q, -2-p-a+q)$. Now, suppose we go up u steps and go down d steps. Solving the following equations:

$$u + d = p + q$$
$$u + a - d = -2 - p - a + q$$

gives us $u = q - a - 1$ and $d = p + a + 1$. So there are $\binom{p+q}{p+a+1}$ paths unfavorable, and the probability is then $1 - \frac{\binom{p+q}{p+a+1}}{\binom{p+q}{p}}$. When $a = 0$, this becomes $\frac{p+1-q}{p+1}$, which agrees with what we obtained earlier.

In summary, if one starts at $(0, a)$ and ends at $(p+q, p+a-q)$, the number of unfavorable paths is $\binom{p+q}{p+a}$ if not allowing ties, $\binom{p+q}{p+a+1}$ if allowing ties.

1.9   Catalan Number

In chapter 1.8, we first met Catalan number from the generalized Bertrand's Ballot prob. We write here again the definition of Catalan number, with an intuitive interpretation.

The Catalan number is defined as,

$$C_n = \frac{1}{n+1} \binom{2n}{n} \tag{1.18}$$

The underlying story reads: Given two urns, one with n red balls and the other with n black balls, we want to draw one ball at a time (either red or black), such that at no time the number of pre-specified color is less than its alternative.

Since the Catalan number is associated with two equal-sized sets, it is oftentimes co-occurrent with the words "pair", "full binary", and etc.

Chapter 2

Conditional Probability

Definition 2.0.1 If $P(F) > 0$, then

$$P(E|F) = \frac{P(E, F)}{P(F)} \tag{2.1}$$

$P(E|F)$ is called the conditional probability of E given F. Conditional probability agrees with Definition 3.1.3, and should be treated in the same way.

Definition 2.0.2 The multiplication rule

$$P(E_1, E_2, \ldots, E_n) = P(E_1)P(E_2|E_1)\ldots P(E_n|E_1, \ldots, E_{n-1}) \tag{2.2}$$

Definition 2.0.3 Bayes' Formula

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{2.3}$$

where $P(A_i)$ is sometimes called the prior distribution, $P(B|A_i)$ the likelihood function, and $P(A_i|B)$ the posterior distribution. The partition function $P(B)$ can be computed using the law of total probability

$$P(B) = \sum_{j=1}^{n} P(B|A_j)P(A_j) \tag{2.4}$$

Definition 2.0.4 Two events E and F are said to be *independent* if

$$P(E, F) = P(E)P(F) \tag{2.5}$$

A set of events are independent if every finite subset of these events is independent.

2.1 Conditional Probability

2.2 Conditional Expectation

2.3 Conditional Independence

Chapter 3

Probaility Space

3.1 Sample Space, Events, and Probability

Definition 3.1.1 A sample space S is a set of all possible outcomes of an experiment. An outcome is also called a sample point.

Note, when an experiment consists of several repetitions, each one of them is called a *trial*. As an example, if one decides to toss a coin 42 times, we can call each toss a trial of the experiment composed of 42 ones.

Definition 3.1.2 An event E is a subset of the sample space S. If the outcome of the experiment is contained in E, then we say that E occurred.

Definition 3.1.3 In short, a probability space is a measure space such that the measure of the whole space is equal to one. The expanded definition is following: a probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of:

- the sample space Ω — an arbitrary non-empty set,

- the $\sigma - algebra$ $\mathcal{F} \in 2^{\Omega}$ (also called $\sigma - filed$) — a set of subsets of Ω, called events, such that:

  - $\mathcal{F}$ contains the empty set: $\emptyset \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then also $(\Omega \setminus A) \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then also $(\bigcup A_i) \in \mathcal{F}$,

- the probability measure $P : \mathcal{F} \to [0, 1]$ — a function on $\mathcal{F}$ such that:

- P is countably additive: if $\{A_i\} \in \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup A_i) = \sum P(A_i)$, where $\bigcup$ denotes the disjoint union,

- the measure of entire sample space is equal to one: $P(\Omega) = 1$.

3.2 Probability Axioms

Law of Total Probability

Suppose $B_n (n = 1, 2, 3, \cdots$ is a finite or countably infinity partition of the sample space, then for an event A

$$Pr(A) = \sum_n Pr(A, B_n) \tag{3.1}$$

or,

$$Pr(A) = \sum_n Pr(A|B_n) Pr(B_n) \tag{3.2}$$

It can also be stated for conditional probabilities. Suppose X is an event in the same sample space, we have

$$Pr(A|X) = \sum_n Pr(A|B_n, X)Pr(B_n|X) \tag{3.3}$$

One application of Eq 3.2 is when calculating $Pr(A)$ is difficult, we can introduce an "auxiliary variable" B, in the hope that the conditional probability $Pr(A|B_n)$ is easier to compute.

Law of Total Variance

Law of Total Covariance

Law of Total Expectation

Law of Total Cumulance

Probability Inequalities

3.3   Types of Probabilities: Frequentism and Bayesian

Chapter 4

Random Variables

The Laplace distribution can be written as an infinite mixture of Gaussians with variance w distribution according to an exponential distribution. An exponential distribution can be written as a χ^2 distribution with two degrees of freedom.

4.1 Continuous Random Variables

4.2 Discrete Random Variables

4.3 Joint Distributed Random Variables

4.4 Random Vectors/Matrices

4.5 Function of Random Variables

Transformation

Convolutions: Sum of Normally Distributed Random Variables

Product Distribution

Ratio Distribution

Chapter 5

Useful Distributions

5.1 Discrete Distributions

Poisson Distribution

Bernoulli Distribution

Binomial Distribution

Negative Binomial Distribution

Categorical Distribution

Multinomial Distribution

Geometric Distribution

Hyper-Geometric Distribution

Poisson Distribution

5.2 Continuous Distributions

Uniform Distribution

Exponential Distribution

χ^2 **Distribution**

Gaussian Distribution

Univariate Gaussian

Multivariate Gaussian

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{5.1}$$

Using the technique of 'completing the square', we have

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_2)} \tag{5.2}$$

$$= \frac{1/\sqrt{2\pi|\boldsymbol{\Sigma}|} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}}{1/\sqrt{2\pi|\boldsymbol{\Sigma}_{22}|} \exp\left\{-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu})\right\}} \tag{5.3}$$

$$= C \cdot \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{5.4}$$

$$= C \cdot \exp\left\{-\frac{1}{2}[(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Lambda}_{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1)]\right\} \tag{5.5}$$

$$= C \cdot \exp\left\{-\frac{1}{2}\mathbf{x}_1^T \boldsymbol{\Lambda}_{11}\mathbf{x}_1 + \mathbf{x}_1^T[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)]\right\} \tag{5.6}$$

Hence

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Lambda}_{11}^{-1} \tag{5.7}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{5.8}$$

or

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{5.9}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{5.10}$$

Dirichlet Distribution

Gamma Function and Beta Function

The *gamma function* is defined for all complex numbers except the non-positive integers. For complex numbers with a positive real part, it is defined via an improper integral that converges:

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt \tag{5.11}$$

For all positive numbers z,

$$\Gamma(z) = (z-1)\Gamma(z-1) \tag{5.12}$$

and in particular, if n is a positive integer:

$$\Gamma(n) = (n-1)! \tag{5.13}$$

Note, the gamma function shifts the normal definition of factorial by 1.

The *beta function* is defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt \tag{5.14}$$

for $Re(x), Re(y) > 0$.

It can also be written as

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{5.15}$$

Introduction

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space $\mathcal{R}^{K-1}$ given by

$$f(x_1, \ldots, x_{K-1}; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{i=1}^K x_i^{\alpha_i - 1} \tag{5.16}$$

for all $x_1, \ldots, x_K > 0$ and $x_1 + \cdots + x_K = 1$. The density is zero outside this open[1] $(K-1)$-dimensional simplex.

The normalizing constant is the multinomial beta function, which can be expressed in terms of the gamma function:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}, \quad \text{where } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \tag{5.17}$$

---

[1] "Open" here means none of the x_i's can be 1, actually $x_i \in (0,1)$.

t Distribution

Inverse Gaussian Distribution

log-normal Distribution

Laplace Distribution

Beta Distribution

Gamma Distribution

Wishart Distribution

Chapter 6

Quantitative Measure and Characteristic Functions

Chapter 7

Limit Theorems

7.1 Markov and Chebyshev Inequalities

Theorem 7.1.1 <u>Markov's Inequality</u>. If X is a random variable that takes only nonnegative values, then for any value $a > 0$,

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

Theorem 7.1.2 <u>Chebyshev's Inequality</u>. If X is a random variable with finite mean μ and variance σ^2, then, for any value $k > 0$,

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

7.2 Weak Law of Large Numbers

Theorem 7.2.1 <u>The Weak Law of Large Numbers</u>. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables, each having finite mean $E[X_i] = \mu$. Then, for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \text{ as } n \to \infty$$

7.3 Central Limit theorem

Theorem 7.3.1 <u>Central Limit theorem</u>. Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{\frac{X_1 + \cdots + X_n}{n} - \mu}{\sqrt{\sigma^2/n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$.

Chapter 8

Selected Topics of Probability

Part II

Elementary Theory of Stochastic Processes

Chapter 9

Introduction

Definition 9.0.1 A *stochastic process* is a collection of random variables $X(t)$ for $t \in T$, where the *time* parameter T is usually[1] a subset of $\mathcal{R}$. The set $\mathscr{S}$ containing all possible values of $X(t)$ is called the *state space* of the process.

There are four common types of stochastic processes (with corresponding examples):

	Finite or Countable State Space	**Continuous State Space**
Discrete Time	Markov Chains	Harris Chains
Continuous Time	Poisson Processes and Hawkes Processes	Gaussian Processes and Wiener Processes

Table 9.1: Four Common Types of Stochastic Processes.

Definition 9.0.2 A *renewal process* is an idealized stochastic model for events that occur randomly in time (generically called renewals or arrivals). The basic mathematical assumption is that the times between the successive arrivals are independent and identically distributed (i.i.d.).

Definition 9.0.3 A *delayed renewal process* is just like an ordinary renewal process, except that the first arrival time is allowed to have a different distribution than the other inter-arrival times.

Definition 9.0.4 A *counting process* is a stochastic process $\{N(t), t \geq 0\}$ with values that are non-negative, integer, and increasing:

Definition 9.0.5 A *point process* is a collection of mathematical points randomly located on some underlying mathematical space such as the real line, the Cartesian plane, or more abstract spaces.

Definition 9.0.6 A *diffusion process* is a solution to a stochastic differential equation. For example, Brownian motion.

Definition 9.0.7 A *jump process* is a type of stochastic process that has discrete movements, called jumps, with random arrival times, rather than continuous movement, typically modeled as a simple or compound Poisson process.

---

[1]It can be defined in a general space.

Chapter 10

Markov Chains

10.1 Introduction

Definition 10.1.1 The *Markov property* is defined by the requirement that

$$P(X_{n+1} = x_{n+1}|X_0 = x_0, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n) \tag{10.1}$$

Definition 10.1.2 The conditional probabilities $P(X_{n+1} = y|X_n = x)$ are called the *transition probabilities* of the chain.

Definition 10.1.3 We say a Markov chain has *stationary transition probabilities* if the transition probabilities $P(X_{n+1} = y|X_n = x)$ is independent of the time n.

From now on, when we say that $X_n, n \geq 0$ forms a Markov chain, we mean that these random variables satisfy the Markov property and have stationary transition probabilities.

Definition 10.1.4 A state a of a Markov chain is called an *absorbing state* if $P(a, a) = 1$ or, equivalently, if $P(a, y) = 0$ for all $y \neq a$.

10.2 Discrete Time Markov Chains

Gambler's Ruin

Discrete Time Branching Processes

10.3 Continuous Time Markov Chains

Chapter 11

Poisson Processes

Chapter 12

Renewal Theory

12.1 Renewal theory for positive lattice valued random variables as connected with Markov chains: Blackwell's renewal theorem for positive lattice valued random variables

Chapter 13

Selected Topics of Stochastic Processes

13.1 Martingales which are functions of discrete time Markov chains

13.2 Brownian motion: Path properties, reflection principle, random walk approximation.

13.3 Random Fields

13.4 Discrete and Continuous Time Birth and Death processes

13.5 Discrete and Continuous Time Queuing processes

13.6 Finite State Space Pure Jump Processes

13.7 Infinite Server Queue

Part III

Measure-theoretical Probability

Chapter 14

Introduction

14.1 Why Do We Need Rigorous Probability Theory?

Let's recall the axioms from non-measure probability courses. We have a *sample space* Ω, the elements of which $w \in \Omega$ called *outcomes*, and the subsets $E \subset \Omega$ called *events*. *Probability measure* is nothing but assigning probability $P(E)$ to each event $E \subset \Omega$, under the following constraints:

1. $P(E) \in [0,1]$

2. $P(\Omega) = 1$

3. $P(\uplus_{i=1}^{\infty} E_i) = \sum_{i=1}^{n} P(E_i)$

where in the last constraint, '$\uplus$' means the union of *disjoint* sets.

One of the motivations[1] of developing measure-theoretical probability theory is triggered by the following question:

Q: *How do we know the three listed axioms are consistent, in particular, the last constraint?*

Actually, the *Banach-Tarski paradox* and the *Vitali paradox* are such counter examples (TBA). We have two choices: either we could reject the *axiom of choice*, which is one of the basic assumptions made in the two paradoxes; or we could claim there are some subsets $E \subset \Omega$ that is 'non-measurable'. In this course, we take the second solution.

14.2 Normal Numbers

Definition 14.2.1 Every $X \in [0,1]$ has a binary decimal expansion $X = 0.d_1 d_2 d_3 \cdots$ or $X = \sum_{k=1}^{\infty} d_k 2^{-k}$, $d_k \in \{0,1\}$. However, it is possible that X may have two expansions, for example, $X = 0.111\cdots = 0.1000\cdots$. We choose the expansion ending in all 1's to make it well-defined.

Definition 14.2.2 $X \in (0,1)$ is normal if $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d_k = \frac{1}{2}$.

Proposition 14.2.1 If $X \sim U(0,1)$, then $P(X$ is normal$) = 1$.

Lemma 14.2.1 Suppose that $X_1, X_2, \cdots$ are i.i.d $Ber(1/2)$. Then $X = \sum_{n=1}^{\infty} X_n 2^{-n}$ is uniform on $(0,1)$.

Proof We (only) need to show $P(X \in (a,b]) = b - a$.

Case 1: $\exists k, s.t.(a,b] = ((k-1)2^{-n}, k2^n]$. Since $(k-1)2^{-n} = 0.d_1 d_2 \cdots d_n 000 \cdots$,

$$P(X \in ((k-1)2^{-n}, k2^{-n}]) = P(X_1 = d_1, X_2 = d_2, \cdots, X_n = d_n, X_i = 1 for i > n) = 2^{-n} = b - a \quad (14.1)$$

---

[1] Add unification of discrete and continuous r.v's.

Case 2: $(a, b] = (l2^{-n}, k2^{-n}], l < k$.

$$P(X \in (l2^{-n}, k2^{-n}]) = (k - l)2^{-n} = b - a \tag{14.2}$$

In general, if $a < b$, let $a_n, b_n \in 2^{-n}\mathbf{Z}$ be s.t. $a_n \le a < a_n + 2^{-n}, b_n - 2^{-n} \le b < b_n$, then

$$P(X \in (a, b]) \le P(X \in (a_n, b_n]) = b_n - a_n \tag{14.3}$$

$b_n - a_n - 2^{-n+1} = P(X \in (a_n + 2^{-n}, b_n - 2^{-n}) \le \epsilon$. Note, $b - a \le b_n - a_n \le b - a + 2^{-n+1}$.
  We need to show that $P(\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{\infty} = \frac{1}{2}) = 1$. This is the SLLN.
  We know that $\{X \in (a, b]\}$ are events. Why is $\{X$ is normal$\}$ an event?

$$\{X \text{ is normal}\} = \left\{ \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{\infty} = \frac{1}{2} \right\} = \bigcap_{l=1}^{\infty} \bigcup_{m=l}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \left| \frac{1}{n} \sum_{k=1}^{\infty} - \frac{1}{2} \right| < \frac{1}{l} \right\} \tag{14.4}$$

This is equivalent to $\forall \epsilon = \frac{1}{l} > 0, \exists m < \infty, s.t., |\frac{1}{n} \sum_{k=1}^{\infty} - \frac{1}{2}| < \epsilon, \forall n \ge m$.

14.3   Formal Definition of Probability Space

Definition 14.3.1 A collection $\mathcal{F}$ of subsets of Ω is a σ-field (or algebra) if

  1. $\mathcal{F}$ is non-empty;

  2. If $A \subset \mathcal{F}$, then $A^c \in \mathcal{F}$ (closed under complement);

  3. If $\{A_i\}$ is a countable sequence of elements of $\mathcal{F}$, then $\cup_i A_i \in \mathcal{F}$ (closed under countable unions).

  Note, 1) $\Omega \in \mathcal{F}, \emptyset \in \mathcal{F}$ since $\Omega = A \cup A^c if A \in \mathcal{F}$. $\emptyset = \Omega^c$. 2) $\mathcal{F}$ is closed under countable intersections.

Definition 14.3.2 A *measure space* is a pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-field of subsets of Ω.

Definition 14.3.3 A non-negative measure μ on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to \bar{\mathcal{R}}_+ = [0, \infty]$ s.t.

  1. $\mu(\emptyset) = 0$;

  2. If A_i is a sequence of disjoint sets in $\mathcal{F}$, then $\mu(\uplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Definition 14.3.4 A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where $(\Omega, \mathcal{F})$ is a measure space, and P is a measure on $(\Omega, \mathcal{F})$ with $P(\Omega) = 1$.

Example 14.3.1 $\mathcal{F} = 2^{\Omega}$ (all subsets of Ω). If $\Omega = \mathbf{Z}$, then usually, $\mathcal{F} = 2^{\Omega}$ is the σ-field we use. If $\Omega = (0, 1]$ or $\mathbf{R}$, then $\mathcal{F}$ is usually too big.

Example 14.3.2 Let $\mathcal{A} \subset 2^{\Omega}$, then $\sigma(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$. If $\mathcal{O} \subset \mathbf{R}$ is the collection of all open subsets of $\mathbf{R}$, then $\sigma(\mathcal{O}) = \mathcal{B}$ is called the *Borel σ-field*.

Chapter 15

Lecture 5 (1/21/2015 Wednesday):

(Today and Friday) in Appendix A.

Definition 15.0.1 A non-empty collection of subsets $\mathcal{A} \subset 2^\Omega$ is called an algebra if

1. $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$

2. $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$

3. $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$

Definition 15.0.2 $\mu : \mathcal{A} \to [0, \infty]$ is a measure on the algebra $\mathcal{A}$ if

1. $\mu(\emptyset) = 0$

2. If $\uplus_{i=1}^\infty A_i \in \mathcal{A}$ then $\mu(\uplus_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i)$

Definition 15.0.3 A measure μ is σ-finite (on an algebra or a σ-field) if $\exists$ a sequence $A_n \nearrow \Omega$ with $\mu(A_n) < \infty$ ($A_n \subset A_{n+1}$ and $\Omega = \cup_{i=1}^\infty A_n$)

Theorem 15.0.1 (Caratheodory Extension) Let μ be a σ-finite measure on an algebra $\mathcal{A}$. Then μ has a unique extension to a measure on $(\Omega, \sigma(\mathcal{A}))$.

 Note: Meausre on algebras also satisfy Themorem 1.1.1

1. $A \subset B$ then $\mu(A) \leq \mu(B)$

2. $\mu(\cup_{i=1}^\infty A_i \leq \sum_i \mu(A_i)$ if $\cup A_i \in \mathcal{A}$

Chapter 16

Probability Space and Measure

16.1   Algebra of Sets

Set Operations

Given two sets $A, B \in \Omega$, there are four basic binary operations on sets:

1. *Union*: $A \cup B = \{x : x \in A \text{ or } x \in B\}$

2. *Intersection*: $A \cap B = \{x : x \in A \text{ and } x \in B\}$

3. *Set difference*: $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$

4. *Set complement*: $A^c = \Omega \setminus A$

Set complement is the "strongest" operation, because if a collection of sets $\mathcal{A}$ is closed under complement, and if it is also closed under any one of the other three operations, $\mathcal{A}$ is closed under the rest two. That is seen from,

$$\text{If closed under } union \begin{cases} A \cap B = (A^c \cup B^c)^c \\ A \setminus B = A \cap B^c = (A^c \cup B)^c \end{cases} \tag{16.1}$$

$$\text{If closed under } intersection \begin{cases} A \cup B = (A^c \cap B^c)^c \\ A \setminus B = A \cap B^c \end{cases} \tag{16.2}$$

$$\text{If closed under } difference \begin{cases} A \cap B = A \setminus (A \setminus B) \\ A \cup B = (A^c \cap B^c)^c = (A^c \setminus (A^c \setminus B^c))^c \end{cases} \tag{16.3}$$

The difference operation is the second "strongest" operation, in that if a collection of sets $\mathcal{A}$ is closed under difference, it is closed under intersection, that is seen from,

$$A \cap B = A \setminus (A \setminus B) \tag{16.4}$$

The third "strongest" operation is the union, which can not be implied from difference or intersection, or their combination.

The "weakest" operation is the intersection, which can be implied from the difference.

To summarize, the "strongest" pair would be the complement plus any one more, which implies everything else; the second "strongest" would be the difference plus the union, which implies the intersection but not the complement (Ω may not be in the collection).

Class of Set Collection

Definition 16.1.1 Given a set Ω, a non-empty collection $\mathcal{P} \subset 2^\Omega$ is called a π-*system* iff:

$$\forall A, B \in \mathcal{P}, \ A \cap B \in \mathcal{P}$$

Notice, this has the weakest requirement.

Definition 16.1.2 Given a set Ω, a non-empty collection $\mathcal{Q} \subset 2^{\Omega}$ is called a *semiring* iff:

$$\forall A, B \in \mathcal{Q} \text{ and } A \supset B$$

$$\exists C_k \subset \mathcal{Q}, s.t., A \setminus B = \bigcup_{k=1}^{n} C_k$$

Definition 16.1.3 Given a set Ω, a non-empty collection $\mathcal{R} \subset 2^{\Omega}$ is called a *ring* iff:

$$\forall A, B \in \mathcal{R}, \ A \cup B \in \mathcal{R} \text{ and } B \setminus A \in \mathcal{R}$$

There are two points need to mention: the empty set is in a ring, since $A \setminus A = \emptyset$; $\mathcal{A}$ is also closed under intersection (the reverse need not be true).

Definition 16.1.4 A *ring* $\mathcal{A}$ is called an *field* iff $\Omega \in \mathcal{A}$.

So a field is closed under all *finite* combination of set operations.

Definition 16.1.5 An *field* is called a σ-*field* if for any sequence A_n of sets in $\mathcal{A}$, $\cup_{n \geq 1} A_n \in \mathcal{A}$.

$\not\exists$

Chapter 17

Integration Theory

Chapter 18

Random Variables

Chapter 19

Law of Large Numbers

19.1 Types of Convergence

Convergence in Distribution

Convergence in Probability

Almost Surely Convergence

L^p **Convergence**

19.2 Weak Law of Large Numbers (WLLN)

19.3 Strong Law of Large Numbers (SLLN)

Chapter 20

Central Limit theorem

Chapter 21

Some Tricks

21.1 Prove by Contraposition

21.2 Construct Finer Partition

Given twofinite partitions $\{A_n\}$ and $\{B_m\}$, a finer partition can be constructed as

$$(\cup_{i=1}^n A_i) \bigcap (\cup_{j=1}^m B_j) = \bigcup (\cap_{i=1}^n \cap_{j=1}^m A_i B_j)$$

21.3 Prove Equality

To prove two numerical quantities are equal $X = Y$, often times we can do this by showing $X \leq Y$ and $X \geq Y$. Similarly, to prove two sets are equal $E = F$, we can show $E \subset F$ and $E \subset F$.

21.4 An Epsilon of Room

If one has to show that $X \leq Y$, try proving that $X \leq Y + \epsilon, \forall \epsilon > 0$. This trick combines well with the "Prove Equality" trick.

In a similar spirit, if one needs to show that a quantity X vanishes, try showing that $|X| \leq \epsilon, \forall \epsilon > 0$.

If one wants to show that a sequence x_n of real numbers converges to zero, try showing that $\limsup_{n \to \infty} |x_n| \leq \epsilon, \forall \epsilon > 0$

One caveat: for finite x, and any $\epsilon > 0$, it is true that $x + \epsilon > x$ and $x - \epsilon < x$, but it is not true when x is equal to $+\infty$ or $-\infty$.

21.5 Interpretations of Probability

Cox's theorem

Principle of Maximum Entropy

Part IV

Measure-theoretical Stochastic Processes

Part V

Appendix