Chapter 1

Basic Probability Models

Further details concerning the first chapter of the appendix can be found in most Introductory texts in probability and mathematical statistics. The material in the second and third chapters can be supplemented with Steele(2001) for further details and many of the proofs.

1.1 Basic Definitions

Probabilities are defined on sets or events, usually denoted with capital letters early in the alphabet such as A, B, C. These sets are subset of a *Sample Space or Probability Space* Ω, which one can think of as a space or set containing all possible outcomes of an experiment. We will say that an event $A \subset \Omega$ occurs if one of the outcomes in A (rather than one of the outcomes in Ω but outside of A) occurs. Not only should we be able to describe the probability of individual events, we should also be able to define probabilities of various combinations of them including

1. Union of sets or events $A \cup B$ = A or B (occurs whenever A occurs or B occurs or both A and B occur.)

2. Intersection of sets $A \cap B$ = A and B (occurs whenever A and B occur).

3. Complement : A^c = not A (occurs when the outcome is not in A).

4. Set differences : $A \setminus B = A \cap B^c$ (occurs when A occurs but B does not)

5. Empty set : $\phi = \Omega^c$ (an impossible event-it never occurs since it contains no outcomes)

Recall *De Morgan's rules* of set theory: $(\cup_i A_i)^c = \cap_i A_i^c$ and $(\cap_i A_i)^c = \cup_i A_i^c$

Events are subsets of Ω. We will call $\mathcal{F}$ the class of all events (including ϕ and Ω).
Definition
A *probability measure* is a set function $P : \mathcal{F} \rightarrow [0, 1]$such that

6. $P(\Omega) = 1$
7. If A_k is a disjoint sequence of events so $A_k \cap A_j = \phi$, for $k \neq j$, then

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

These are the basic axioms of a probability model. From these it is not difficult to prove the following properties:

1. $P(\phi) = 0$.
2. If $A_k, k = 1, \ldots N$ is a finite or countable sequence of disjoint events so $A_k \cap A_j = \phi, k \neq j$, then

$$P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i)$$

3. $P(A^c) = 1 - P(A)$.
4. Suppose $A \subset B$. Then $P(A) \leq P(B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. The inclusion-exclusion principle:

$$P(\cup_k A_k) = \sum_k P(A_k) - \sum\sum_{i<j} P(A_i \cap A_j) + \sum\sum\sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots.$$

7. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_i P(A_i)$.
8. Suppose $A_1 \subset A_2 \subset \ldots$. Then $P(\cup_{i=1}^{\infty} A_i) = lim_{i \to \infty} P(A_i)$.

Counting Techniques

Permutations.

The number of ways of permuting or arranging n distinct objects in a row is $n! = n(n-1)\ldots 1$ and $0! = 1$. Define $n^{(r)} = n(n-1)\ldots(n-r+1)$ (called "n to r factors") for arbitrary n, and r a non-negative integer. This is the number of permutations of n objects taken r at a time. Define $n^{(0)} = 1$ and notice that values like $(\frac{1}{2})^{(3)}$ are well-defined (indeed $(\frac{1}{2})^{(3)} = (\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2}) = \frac{3}{8}$).

For example, the number of distinct ways of rearranging the 15 letters

$$AAAAABBBBCCCDDE$$

would be $15!$ if all 15 letters could be distinguished. Since they cannot, this calculation counts the two possible orderings of the $D's$, e.g. $D_1 D_2$ or $D_2 D_2$ separately, and each of the 3! reorderings of the $C's$ are counted separately, etc. Therefore, dividing by the number of times each letter has been overcounted, the number of distinct rearrangements is

$$\frac{15!}{5!4!3!2!} = \begin{pmatrix} & & 15 & \\ 5 & 4 & 3 & 2 \end{pmatrix}$$

Combinations

Suppose the order of selection is not considered to be important. We wish, for example, to distinguish only different *sets* selected, without regard to the order in which they were selected. Then the number of distinct sets of r objects that can be constructed from n distinct objects is

$$\binom{n}{r} = \frac{n^{(r)}}{r!}$$

Note this is well defined for r a non-negative integer for any real number n.

Independent Events.

Two events A, B are said to be *independent* if

$$P(A \cap B) = P(A)P(B) \tag{1.1}$$

Compare this definition with that of *mutually exclusive or disjoint* events A, B. Events A, B are mutually exclusive if $A \cap B = \phi$.

Independent experiments are often built from *Cartesian Products* of sample spaces. For example if $\Omega_1$and Ω_2 are two sample spaces, and $A_1 \subset \Omega_1, A_2 \subset \Omega_2$, then an experiment consisting of *both of the above* would have sample space the Cartesian product

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

and probabilities of events such as $A_1 \times A_2$ are easily defined, in this case as $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$. Verify in this case that an event entirely determined by the first experiment such as $A = A_1 \times \Omega_2$ is independent of one determined by the second $B = \Omega_1 \times A_2$.

Definition.

A finite or countably infinite set of events $A_1, A_2, \ldots..$ are said to be mutually independent if

$$P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}) \tag{1.2}$$

for any $k \geq 2$ and $i_1 < i_2 < \ldots i_k$.

Independent events have the properties that:

1. A, B independent implies A, B^c independent.

2. Any A_{i_j} can be replaced by $A_{i_j}^c$ in equation (1.2).

Why not simply require that every pair of events is independent? This is, as it turns out, too weak an assumption for many of the results we need in probability and statistics, and does not describe what we intuitively mean by independence either. For example suppose two fair coins are tossed. Let A = first coin is heads, B= second coin is heads, C= we obtain exactly one heads. Then A is independent of B and A is independent of C but A, B, C are **not mutually independent**. Thus *pairwise*

independence does not imply independence. Does it make intuitive sense to say that A, B, C are independent above? If you know whether A and B occur, then you automatically know whether or not the event C occurs so there is a strong dependence among these three events.

"Lim Sup" of events

For a sequence of events $A_n, n = 1, 2, ...$ we define another event $[A_n \ i.o.] = \lim$ sup $A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$. Note that this is the set of all points x which lie in infinitely many of the events $A_1, A_2, ...$ The notation i.o. stands for "infinitely often" because the $\lim$ sup A_n is the set of all points ω which are in infinitely many of the $A_n, n = 1, 2, ...$ There is a similar notion, $\liminf A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$ and it is not difficult to show that the latter set is smaller

$$\liminf A_n \subset \limsup A_n.$$

A point ω is in $\liminf A_n$ if and only if it is in all of the sets A_n except possibly a finite number. For this reason we sometimes denote $\liminf A_n$ as $[A_n$ a.b.f.o.] where $a.b.f.o.$ stands for "all but finitely often".

Borel Cantelli Lemmas

Clearly if events are individually too small, then there little or no probability that their lim sup will occur, i.e. that they will occur infinitely often. This is the essential message of the first of the Borel-Cantelli Lemmas below:
Lemma 1: For an arbitrary sequence of events A_n, if $\sum_n P(A_n) < \infty$ then $P[A_n i.o.] = 0$.
Lemma 2: For a sequence of *independent events* A_n, $\sum_n P(A_n) = \infty$ implies $P[A_n i.o] = 1$.

Conditional Probability.

Suppose we are interested in the probability of the event A but we are given some relevant information, namely that another related event B occurred. How do we revise the probabilities assigned to points of Ω in view of this information? If the information does not effect the relative probability of points in B then the new probabilities of points outside of B should be set to 0 and those within B simply rescaled to add to 1. This is essentially the definition of conditional probability given B. Given that B occurred, reassign probability 0 to those points outside of B and rescale those within so that the total probability is one.

Definition: Conditional Probability:

For $B \in \mathcal{F}$ with $P(B) > 0$, define a new probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.3}$$

This is also a probability measure on the same space $(\Omega, \mathcal{F})$, and satisfies the same properties. Note that $P(B|B) = 1, P(B^c|B) = 0$.

Theorem A1: Bayes Rule

If $P(\cup_n B_n) = 1$ for a *disjoint* finite or countable sequence of events B_n all with positive probability, then

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)} \tag{1.4}$$

Theorem A2: Multiplication rule.

If $A_1 \ldots A_n$are arbitrary events,

$$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 A_1) \ldots P(A_n|A_1 A_2 \ldots A_{n-1}) \tag{1.5}$$

Random Variables

Properties of $\mathcal{F}$.

The class of events $\mathcal{F}$ (called a $\sigma-$algebra or $\sigma-$field) should be such that the operations normally conducted on events, for example countable unions or intersections, or complements, keeps us within that class. In particular it is such that

(a) $\varphi \in \mathcal{F}$

(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(c) If $A_n \in \mathcal{F}$ for all $n = 1, 2, ...$, then $\cup_{n=1}^{\infty} \in \mathcal{F}$.

It follows from these properties that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is also closed under countable intersections, or countable intersections of unions, etc.

Definition

Let X be a function from a probability space Ω into the real numbers. We say that the function is *measurable* (in which case we call it a random variable) if for $x \in \Re$, the set $\{\omega; X(\omega) \leq x\} \in \mathcal{F}$. Since events in $\mathcal{F}$ are those to which we can attach a probability, this permits us to obtain probabilities for the event that the random variable X is less than or equal to any number x.

Definition: Indicator random variables

For an arbitrary set $A \in \mathcal{F}$ define $I_A(\omega) = 1$ if $\omega \in A$and 0 otherwise. This is called an *indicator random variable*. (sometimes a *characteristic function* in measure theory, but not here).

Definition: Simple Random variables.

Consider events $A_i \in \mathcal{F}$ such that $\cup_i A_i = \Omega$. Define $X(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega)$ where $c_i \in \Re$. Then X is measurable and is consequently a random variable. We normally assume that the sets A_i are disjoint. Because this is a random variable which can take only finitely many different values, then it is called *simple*. Any random variable taking only finitely many possible values can be written in this form.

We will often denote the event $\{\omega \in \Omega; X(\omega) \leq x\}$ more compactly by $[X \leq x]$. In general functions of one or more random variables gives us another random variable (provided that function is measurable). For example, if X_1, X_2 are random variables, so is

1. $X_1 + X_2$
2. $X_1 X_2$
3. $min(X_1, X_2)$.

The *cumulative distribution function* (cumulative distribution function) of a *Random variable*X is defined to be the function $F(x) = P[X \leq x]$, for $x \in \Re$.

Properties of the cumulative distribution function.

1. A cumulative distribution function $F(x)$ is non-decreasing. i.e. $F(x) \geq F(y)$ whenever $x \geq y$.
2. $F(x) \to 0$, as $x \to -\infty$.
3. $F(x) \to 1$, $x \to \infty$.
4. $F(x)$ is right continuous. i.e. $F(x) = \lim_{h \to 0^+} F(x + h)$ (i.e. the limit as h decreases to 0).

There are two primary types of distributions considered here, discrete distributions and continuous ones. Discrete distributions are those whose cumulative distribution function at any point x can be expressed as a finite or countable sum of values. For example

$$F(x) = \sum_{\{i; x_i \leq x\}} p_i$$

for some probabilities p_i which sum to one. In this case the cumulative distribution is piecewise constant, with jumps at the values x_i that the random variable can assume. The values of those jumps are the individual probabilities p_i. For example $P[X = x]$ is equal to the size of the jump in the graph of the cumulative distribution function at the point x. We refer to the function $f(x) = P[X = x]$ as the *probability function* of the distribution when the distribution is discrete.

1.2 Some Special Discrete Distributions

The Discrete Uniform Distribution

Many of the distributions considered so far are such that each point is equally likely. For example, suppose the random variable X takes each of the points $a, a+1, \ldots b$ with the same probability $\frac{1}{b-a+1}$. Then the cumulative distribution function is

$$F(x) = \frac{x-a+1}{b-a+1}, x = a, a+1, \ldots b$$

and the probability function is $f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, \ldots b$ and 0 otherwise.

The Hypergeometric Distribution

Suppose we have a collection (the *population*) of N objects which can be classified into two groups S or F where there are r of type "S" and $N-R$ of type "F". Suppose we take a random sample of n items without replacement from this population. Then the probability that we obtain exactly x items of type "S"' is

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, x = 0, 1, \ldots$$

The Binomial Distribution

The setup is identical to that in the last paragraph only now we sample *with replacement*. Thus, for each member of the sample, the probability of an S is $p = r/N$. Then the probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots n$$

With any distribution, the sum of *all* the probabilities should be 1. Check that this is the case for the binomial, i.e. that

$$\sum_{x=0}^{n} f(x) = 1.$$

The Hypergeometric distribution is often approximated by the binomial distribution in the case N large. For the binomial distribution, the two *parameters* n, p are fixed, and usually known. For fixed sample size n we count $X =$ "the number of S's in n *trials* of a simple experiment" (e.g. tossing a coin).

The Negative Binomial distribution

The binomial distribution was generated by assuming that we repeated trials a fixed number n of times and then counted the total number of successes X in those n trials. Suppose we decide in advance that we wish a fixed number (k) of successes instead,

and sample repeatedly until we obtain exactly this number. Then the number of trials X is random.

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, x = k, k+1, \ldots$$

A special case of most interest is the case $k = 1$called the *Geometric* distribution. Then

$$f(x) = p(1-p)^{x-1}, x = 1, 2, \ldots$$

The Poisson Distribution.

Suppose that a disease strikes members of a large population (of n individuals) independently, but in each case it strikes with very small probability p. If we count X the number of cases of the disease in the population, then X has the binomial (n, p) distribution. For very large n and small p this distribution can be again approximated as follows:

Theorem A3

Suppose $f_n(x)$ is the probability function of a binomial distribution with $p = \lambda/n$ for some fixed λ. Then as $n \to \infty$,

$$f_n(x) \to f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for each $x = 0, 1, 2, \ldots$.

The function $f(x)$ above is the probability function of a *Poisson Distribution* named after a French mathematician. This distribution has a single parameter λ, which makes it easier to use than the binomial, since the binomial requires knowledge or estimation of two parameters. For example the size n of the population of individuals who are susceptible to the disease might be unknown but the "average" number of cases in a population of a certain size might be obtainable from medical data.

1.3 Expected Value

An indicator random variable I_A takes two values, the value 1 with probability $P(A)$ and the value 0 otherwise. Its expected value, or average over many (independent) trials would therefore be $0(1 - P(A)) + 1P(A) = P(A)$. This is the simplest case of an integral or expectation.

Recall that a simple random variable is one which has only finitely many distinct values c_i on the sets A_i where these sets form a partition of the sample space (i.e. they are disjoint and their union is Ω).

Expectation of simple random Variables.

For a simple random variable $X = \sum_i c_i I_{A_i}$, define $E(X) = \sum_i c_i P(A_i)$. The form is standard:

$$E(X) = \sum(\text{values of } X) \times (\text{Probability of values})$$

Thus, for example, if a random variable X has probability function $f(x) = P[X = x]$, then $E(X) = \sum_x xf(x)$.

Example The expected value of X, a random variable having the Binomial(n, p) distribution is $E(X) = np$.

Expectation of non-negative measurable random variables.

Definition: Suppose X is a non-negative random variable so that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then we define

$$E(X) = sup\{E(Y); Y \text{ simple and } Y \leq X\}.$$

Expected value: discrete case.

If a random variable X has probability function $f(x) = P[X = x]$, then the definition of expected value in the case of *finitely many* possible values of x is essentially $E(X) = \sum_x xf(x)$. This formula continues to hold even when X may take a countably infinite number of values provided that the series $\sum_x xf(x)$ is absolutely convergent.

Notation.

Note that by $\int_A XdP$ we mean $E(XI_A)$ where I_A is the indicator of the event A.

Properties of Expectation.

Assume X, Y are non-negative random variables. Then ;

1. If $X = \sum_i c_i I_{A_i}$ is simple, then $E(X) = \sum_i c_i P(A_i)$.
2. If $X(\omega) \leq Y(\omega)$ for all ω, then $E(X) \leq E(Y)$.
3. If X_n is increasing to X, then $E(X_n)$ increases to $E(X)$ (this is usually called the *Monotone Convergence Theorem*).
4. For non-negative numbers α, β, $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

General Definition of Expected Value.

For an arbitrary random variable X, define two new random variables $X^+ = max(X, 0)$, and $X^- = max(0, -X)$. Note that $X = X^+ - X^-$. Then we define $E(X) = E(X^+) - E(X^-)$. This is well defined even if one of $E(X^+)$ or $E(X^-)$ are equal to ∞ as long as both or not infinite since the form $\infty - \infty$ is meaningless. If both $E(X^+) < \infty$ and $E(X^-) < \infty$ then we say X is *integrable*.

Example:

Define a random variable X such that $P[X = x] = \frac{1}{x(x+1)}, x = 1, 2, \ldots.$.

Is this random variable integrable? If we write out the expected value

$$\sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} \frac{1}{x+1}$$

and this is a divergent sequence so in this case the random variable is not integrable.

General Properties of Expectation.

In the general case, expectation satisfies 1-4 above plus the additional properties:

5. If $P(A) = 0$, then $\int_A X(\omega)dP = 0$
6. If $P[X = c] = 1$ for some constant c, then $E(X) = c$.
7. If $P[X \geq 0] = 1$ then $E(X) \geq 0$.

Other interpretations of Expected Value

For a discrete distribution, the distribution is often represented graphically with a bar graph or histogram. If the values of the random variable are $x_1 < x_2 < x_3 < \ldots$then rectangles are constructed around each value, x_i, with *area* equal to the probability $P[X = x_i]$. In the usual case that the x_i are equally spaced, the rectangle around x_i has as base $(\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2})$. In this case, the expected value $E(X)$ is the x-coordinate of the center of gravity of the probability histogram.

We may also think of expected value as a long run average over many independent repetitions of the experiment. Thus, $f(x) = P[X = x]$ is approximately the long run proportion of occasions on which we observed the value $X = x$ so the *long run average* of many independent replications of X is $\sum_x xf(x) = E(X)$.

Lemma (Fatou's lemma: limits of integrals)

If X_n is a sequence of non-negative random variables

$$E[\liminf X_n] \leq \liminf EX_n$$

It is possible for $X_n(\omega) \to X(\omega)$ for all ω and yet $E(X_n)$ does not converge to $E(X)$. For example let $\Omega = (0, 1)$ and the probability measure be Lebesgue measure on the interval. Define $X(\omega) = n$ if $0 < \omega < 1/n$ and otherwise $X(\omega) = 0$. Then $X_n(\omega) \to 0$ for all ω but $E(X_n) = 1$ does not converge to the expected value of the limit. This example shows that some additional condition is required beyond (almost sure) convergence of the random variables in order to conclude that the expected values converge. One such condition is given in the following important result.

Theorem A4 (Lebesgue dominated convergence Theorem)

If $X_n(\omega) \to X(\omega)$ for each ω, and there exists integrable Y with $|X_n(\omega)| \leq Y(\omega)$ for all n, ω, then X is integrable and $E(X_n) \to E(X)$.

Lebesgue-Stieltjes Integral

A basic requirement of any sigma-algebra of subsets of the real line before it is much use is that it contain the intervals, since we often wish to compute probabilities of intervals like $[a < X < b]$.

Definition (Borel Sigma Algebra)

The smallest sigma algebra which contains all of the open intervals is called the Borel sigma algebra. The sets in this sigma algebra are referred to as Borel sets.

Fortunately it is easy to show that this sigma algebra also contains all of the closed intervals, in fact all countable unions of intervals of any kind, open, closed or half open. We call a function $g(x)$ on the real numbers (i.e. $\Re \to \Re$) Borel measurable if for any Borel subset $B \subset \Re$, the set $\{x; g(x) \in B\}$ is also a Borel set.

We now consider integration of functions on the real line or Euclidean space. Suppose $g(x)$ is a Borel measurable function $\Re \to \Re$. Suppose $F(x)$ is a Borel measurable function satisfying

1. $F(x)$ is non-decreasing. i.e. $F(x) \geq F(y)$ whenever $x \geq y$.
2. $F(x)$ is right continuous. i.e. $F(x) = \lim F(x + h)$ as h decreases to 0.

Notice that we can use F to define a measure μ on the real line; for example the measure of the interval $(a, b]$ we can take to be $\mu((a, b]) = F(b) - F(a)$. The measure is extended from these intervals to all Borel sets in the usual way, by first defining the measure on the algebra of finite unions of intervals, and then extending this measure to the Borel sigma algebra generated by this algebra. We will define $\int g(x)dF(x)$ or $\int g(x)d\mu$ exactly as we did expected values but with the probability measure P replaced by μ and $X(\omega)$ replaced by $g(x)$. In particular, for a simple function $g(x) = \sum_i c_i I_{A_i}(x)$, we define $\int g(x)dF = \sum_i c_i \mu(A_i)$.

Definition (Integration of Borel measurable functions)

Suppose $g(x)$ is a non-negative Borel measurable function so that $g(x) \geq 0$ for all $x \in \Re$. Then we define

$$\int g(x)d\mu = \sup\{\int h(x)d\mu; h \text{ is simple and } h \leq g\}.$$

For a general function $f(x)$ we write $f(x) = f^+(x) - f^-(x)$ where both f^+ and f^- are non-negative functions. We then define $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$ provided that this makes sense (i.e. is not of the form $\infty - \infty$). Finally we say that f is integrable if both f^+ and f^- have *finite integrals*, or equivalently, if $\int |f(x)|d\mu < \infty$.

Definition (absolutely continuous)

A measure μ on $\Re$ is *absolutely continuous* with respect to Lebesgue measure λ (denoted $\mu << \lambda$) if there is an integrable function $f(x)$ such that $\mu(B) = \int_B f(x)d\lambda$ for all Borel sets B. The function f is called the *density* of the measure μ with respect to λ.

Intuitively, two measures μ, λ on the same measurable space $(\Omega, \mathcal{F})$ (not necessarily the real line) satisfy $\mu << \lambda$ if the support of the measure λ includes the support of the measure μ. For a discrete space, the measure μ simply reweights those points with non-zero probabilities under λ. For example if λ represents the discrete uniform distribution on the set $\Omega = \{1, 2, 3, ..., N\}$ (so that $\lambda(B)$ is $N^{-1}\times$the number of integers in $B \cap \{1, 2, 3, ..., N\}$) and $f(x) = x$, then if $\mu(B) = \int_B f(x)d\lambda$,we have $\mu(B) = \sum_{x \in B \cap \{1,2,3,...,N\}} x$. Note that the measure μ assigns weights $\frac{1}{N}, \frac{2}{N}, ...1$ to the points $\{1, 2, 3, ..., N\}$ respectively.

The so-called *continuous distributions* such as the normal, gamma, exponential, beta, chi-squared, student's t, etc. should be called *absolutely continuous with respect to Lebesgue measure* rather than just continuous.

Theorem A5 (The Radon-Nykodym Theorem)

For arbitrary measures μ and λ defined on the same measure space, the two conditions below are equivalent:

1. μ is absolutely continuous with respect to λ so that there exists a function $f(x)$ such that
$$\mu(B) = \int_B f(x)d\lambda$$
2. For all B, $\lambda(B) = 0$ implies $\mu(B) = 0$.

The first condition above asserts the existence of a "density function" as it is usually called in statistics but it is the second condition above that is usually referred to as absolute continuity. The function $f(x)$ is called the *Radon Nikodym* derivative of μ with respect to λ. We sometimes write $f = \frac{d\mu}{d\lambda}$ but f is not in general unique. Indeed there are many $f(x)$ corresponding to a single μ, i.e. many functions f satisfying $\mu(B) = \int_B f(x)d\lambda$ for all Borel B. However, for any two such functions f_1, f_2, $\lambda\{x; f_1(x) \neq f_2(x)\} = 0$. This means that f_1 and f_2 are *equal almost everywhere* (λ).

The so-called discrete distributions in statistics such as the binomial distribution, the negative binomial, the geometric, the hypergeometric, the Poisson or indeed any distribution concentrated on the integers is absolutely continuous with respect to the counting measure $\lambda(A) =$number of integers in A.

If the measure induced by a cumulative distribution function $F(x)$ is absolutely continuous with respect to Lebesgue measure, then $F(x)$ is a continuous function. However it is possible that $F(x)$ be a continuous function without the corresponding measure being absolutely continuous with respect to Lebesgue measure.

Definition (equivalent measures)

Two measures μ and λ defined on the same measure space are said to be *equivalent* if both $\mu << \lambda$ and $\lambda << \mu$. Alternatively they are equivalent if if $\mu(A) = 0$ if and only if $\lambda(A) = 0$ for all A.Intuitively this means that the two measures share exactly the same support or that the measures are either both positive on a given event or they are both zero an that event.

In general there are three different types of probability distributions, when expressed in terms of the cumulative distribution function.

1. Discrete: For countable $x_n,\ p_n$, $F(x) = \sum_{\{n;x_n \leq x\}} p_n$. The corresponding measure has countably many atoms.

2. Continuous singular. $F(x)$ is a continuous function but for some Borel set B having Lebesgue measure zero, $\lambda(B) = 0$, we have $P(X \in B) = \int_B dF(x) = 1$.

3. Absolutely continuous (with respect to Lebesgue measure). $F(x) = \int_{-\infty}^{x} f(x)d\lambda$ for some function f called the *probability density function*.

There is a general result called the Lebesgue decomposition which asserts that any cumulative distribution function can be expressed as a mixture of those of the above three types, i.e. a (sigma-finite) measure μ on the real line can be written

$$\mu = \mu_d + \mu_{ac} + \mu_s,$$

the sum of a discrete measure μ_d, a measure μ_{ac} absolutely continuous with respect to Lebesgue measure and a measure μ_s that is continuous singular. For a variety of reasons of dubious validity, statisticians concentrate on absolutely continuous and discrete distributions, excluding, as a general rule, those that are singular.

1.4 Discrete Bivariate and Multivariate Distributions

Definitions.

For discrete random variables X, Y defined on the same probability space, the function $f(x, y) = P[X = x, Y = y]$ giving the probability of all combinations of values of the random variables X, Y is called the *joint probability function* of X and Y. (Read the comma "," as the word "and", the intersection of two events). The function $F(x, y) = P[X \leq x, Y \leq y]$ is called the *joint cumulative distribution function*. The joint probability function allows us to compute the probability functions of both X and Y. For example

$$P[X = x] = \sum_{\text{all } y} f(x, y).$$

We call this the *marginal* probability function of X, denoted by $f_X(x) = P[X = x] = \sum_{\text{all } y} f(x, y)$. Similarly, $f_Y(y)$ is obtained by adding the joint probability function over all values of x. Finally we are often interested in the conditional probabilities of the form

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

This is called the *conditional probability function* of X given Y.

Expected Values

For a single (discrete) random variable we determined the expected value of a function of X, say $h(X)$ by

$$E[h(X)] = \sum_{\text{all x}} (\text{value of } h) \times (\text{Probability of value}) = \sum_x h(x) f(x)$$

For two or more random variables we should use a similar approach. However, when we add over all cases, this requires adding over all values of x and y. Thus, if h is a function of both X and Y,

$$E[h(X, Y)] = \sum_{\text{all } x \text{ and } y} h(x, y) f(x, y).$$

Independent Random Variables

Two discrete random variables X, Y are said to be *independent* if the events $[X = x]$ and $[Y = y]$ are independent for all x, y, i.e. if

$$P[X = x, Y = y] = P[X = x]P[Y = y] \text{ for all } x, y$$

or equivalently if

$$f(x, y) = f_X(x) f_Y(y) \text{ for all } x, y.$$

This definition extends in a natural way to more than two random variables. For example we say random variables $X_1, X_2, \ldots X_n$ are (mutually) independent if, for every choice of values $x_1, x_2, \ldots x_n$, the events $[X_1 = x_1], [X_2 = x_2], \ldots [X_n = x_n]$ are independent events. This holds if the joint probability function of all n random variables factors into the product of the n marginal probability functions.

Theorem A6

If X, Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

Definition: Variance

The variance of a random variable measures its variability about its own expected value. Thus if one random variable has larger variance than another, it *tends* to be farther from its own expectation. If we denote the expected value of X by $E(X) = \mu$, then

$$Var(X) = E[(X - \mu)^2].$$

Adding a constant to a random variable does not change its variance, but multiplying it by a constant does; it multiplies the original variance by the constant squared.

Example

Suppose the random variable X has the binomial(n, p) distribution. Then

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np\{\sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j}\} \\
&= np\{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}\} \\
&= np
\end{aligned}$$

and so $E(X) = np$. A similar calculation allows us to obtain $E(X(X-1)] = n(n-1)p^2$ from which we can obtain $var(X) = np(1-p)$.

Definition: Covariance

Define the covariance between 2 random variables X, Y as

$$cov(X, Y) = E[(X - EX)(Y - EY)]$$

Covariance measures the linear association between two random variables. Note that the covariance between two *independent random variables* is 0. If the covariance is large and positive, there is a tendency for large values of X to be associated with large values of Y. On the other hand, if large values of X are associated with small values of Y, the covariance will tend to be negative. There is an alternate form for covariance, generally easier for hand calculation but more subject to computer overflow problems: $cov(X, Y) = E(XY) - (EX)(EY)$.

Theorem A7

For any two random variables X, Y

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$$

One special case is of fundamental importance: the case when X, Y are independent random variables and $var(X+Y) = var(X) + var(Y)$ since $cov(X,Y) = 0$.

Properties of Variance and Covariance

For any random variables X_i and constants a_i

1. $Var(X_1) = cov(X_1, X_1)$.
2. $var(a_1X_1 + a_2) = a_1^2 var(X_1)$.
3. $cov(X_1, X_2) = cov(X_2, X_1)$.
4. $cov(X_1, X_2 + X_3) = cov(X_1, X_2) + cov(X_1, X_3)$.
5. $cov(a_1X_1, a_2X_2) = a_1a_2cov(X_1, X_2)$.
6. Similarly $var(\sum_{i=1}^n a_iX_i) = \sum a_i^2 var(X_i) + 2\sum\sum_{\{(i,j);i<j\}} a_ia_jcov(X_i, X_j)$.

Correlation Coefficient

The covariance has an arbitrary scale factor because of property 5 above. This means that if we change the units in which something is measured, (for example a change from imperial to metric units of weight), the covariance will change. It is desirable to measure covariance in units free of the effect of scale. To this end, define the *standard deviation* of X by $SD(X) = \sqrt{var(X)}$. Then the *correlation coefficient* between X and Y is

$$\rho = \frac{cov(X,Y)}{SD(X)SD(Y)}$$

For any pair of random variables X, Y, we have $-1 \leq \rho \leq 1$ with $\rho = \pm 1$ if and only if the points (X, Y) always lie on a line so $Y = aX + b$ (almost surely) for some constants a, b. The fact that $\rho \leq 1$ follows from the following argument, and the argument for $-1 \leq \rho$ is similar. Consider for any t,

$$\begin{aligned} \text{var}(X - tY) &= \text{cov}(X - tY, X - tY) \\ &= \text{var}(X) - 2t\text{cov}(X,Y) + t^2\text{var}(Y) \end{aligned}$$

Since variance is always ≥ 0, this quadratic equation in t cannot have two real roots so the discriminant must be non-positive,

$$[2\text{cov}(X,Y)]^2 - 4\text{var}(X)\text{var}(Y) \leq 0$$

i.e.

$$|\text{cov}(X,Y)| \leq \sqrt{\text{var}(X)\text{var}(Y)}$$

The Multinomial Distribution

Suppose an experiment is repeated n times (called "trials") where n is fixed in advance. On each "trial" of the experiment, we obtain an outcome in one of k different categories $A_1, A_2, \ldots A_k$ with the probability of outcome A_i given by p_i. Here $\sum_{i=1}^{k} p_i = 1$. At the end of the n trials of the experiment consider the count of $X_i =$ "number of outcomes in category i", for $i = 1, 2, \ldots k$. Then the random variables $(X_1, X_2, \ldots X_k)$ have a joint *multinomial* distribution given by the joint probability function

$$P[X_1 = x_1, X_2 = x_2, \ldots X_k = x_k] = \binom{n}{x_1 x_2 \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$
$$= \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

whenever $\sum_i x_i = n$. Otherwise $P[X_1 = x_1, X_2 = x_2, \ldots X_k = x_k]$ is 0. Note that the marginal distribution of each X_i is binomial (n, p_i) and so $E(X_i) = np_i$.

Covariance of a linear transformation.

Suppose $X = (X_1, \ldots, X_n)'$ is a vector whose components are possibly dependent random variables. We define the expected value of this random vector by

$$\mu = E(X) = \begin{pmatrix} EX_1 \\ . \\ . \\ . \\ . \\ EX_n \end{pmatrix}$$

and the covariance matrix by a matrix

$$V = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & . & . & cov(X_1, X_n) \\ cov(X_2, X_1 & var(X_2) & . & . & cov(X_2, X_n \\ . & & & & . \\ . & & & & . \\ cov(X_n, X_1) & . & . & . & var(X_n) \end{pmatrix}.$$

Then if A is a $q \times n$ matrix of constants, the random vector $Y = AX$ has mean $A\mu$ and covariance matrix AVA'. In particular if $q = 1$, the variance of AX is AVA' .

1.5 Continuous Distributions

Definitions

Suppose a random variable X can take any real number in an interval. Of course the number that we record is often rounded to some appropriate number of decimal places, so we don't actually observe X but $Y = X$ rounded to the nearest $\Delta/2$ units. So, for

example, the probability that we record the number $Y = y$ is the probability that X falls in the interval $y - \Delta/2 < X \leq y + \Delta/2$. If $F(x)$ is the cumulative distribution function of X this probability is $P[Y = y] = F(y + \Delta/2) - F(y - \Delta/2)$. Suppose now that Δ is very small and that the cumulative distribution function is piecewise continuously differentiable with a derivative given in an interval by

$$f(x) = F'(x).$$

Then $F(y+\Delta/2)-F(y-\Delta/2) \approx f(y)\Delta$ and so Y is a discrete random variable with probability function given (approximately) by $P[Y = y] \approx \Delta f(y)$. The derivative of the cumulative distribution function of X is the *probability density function* of the random variable X. Notice that an interval of small length Δ around the point y has approximate probability given by *length of interval*$\times f(y)$. Thus the probability of a (small) interval is approximately proportional to the probability density function in that interval, and this is the motivation behind the term *probability density*.

Example.

Suppose X is a random number chosen in the interval $[0, 1]$. We wish that any interval of length $\Delta \subset [0, 1]$ will have the same probability Δ regardless of where it is located. Then the cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

The probability density function is given by the derivative of the cumulative distribution function $f(x) = 1$ for $0 < x < 1$ and otherwise $f(x) = 0$. Notice that $F(y) = \int_{-\infty}^{y} f(x)dx$ for all y and the probability density function can be used to determine probabilities as follows;

$$P[a < X < b] = P[a \leq X \leq b] = \int_a^b f(x)dx.$$

In particular, notice that $F(b) = \int_{-\infty}^{b} f(x)dx$ for all b.

Example.

Let $F(x)$ be the binomial $(n, 1/2)$ cumulative distribution function. Notice that the derivative $F'(x)$ exists and is continuous in fact is zero) except at finitely many points $x = 0, 1, 2, 3, 4$. Is it true that $F(b) = \int_{-\infty}^{b} F'(x)dx$? In this case the right side is zero since $F'(x) = 0$ except at finitely many points but the left side is not. Equality is only guaranteed under further conditions.

Definition (cumulative distribution function)

Suppose the cumulative distribution function of a random variable $F(x)$ is such that its derivative $f(x) = F'(x)$ exists except at finitely many points. Suppose also that

$$F(b) = \int_{-\infty}^{b} f(x)dx \tag{1.6}$$

for all $b \in \Re$. Then the distribution is *absolutely) continuous* and the function $f(x)$ is called the *probability density function*.

Example.

Is it really necessary to impose the additional requirement (1.6) or this just a consequence of the fundamental theorem of calculus? Consider the case $F(x) = 0, x < 0,$ and $F(x) = 1$,for $x \geq 0$. This cumulative distribution function is piecewise differentiable (the only point where the derivative fails to exist is the point $x = 0$). But is the function the integral of its derivative? Since the derivative is zero except at one point where it is not defined, any sensible notion of integral results in $\int_{-\infty}^{b} F'(x)dx = 0$ for any b.

For a continuous distribution, probabilities are determined by integrating the probability density function. Thus

$$P[a < X < b] = \int_{a}^{b} f(x)dx \tag{1.7}$$

A probability density function is not unique. For example we may change $f(x)$at finitely many points and it will still satisfy (1.7) above and all probabilities, determined by integrating the function, remain unchanged. Whenever possible we will choose a continuous version of a probability density function, but at a finite number of discontinuity points, it does not matter how we define the function.

Properties of a Probability Density Function

1. $f(x) \geq 0$ for all $x \in \Re$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

The Continuous Uniform Distribution.

Consider a random variable X that takes values with a continuous uniform distribution on the interval $[a, b]$. Then the cumulative distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

and so the probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ and elsewhere the probability density function is 0. Again, notice that the definition of f at the points

a and b does not matter, since altering the definition at two points will not alter the integral of the function.

Suppose we were to approximate a continuous random variable X having probability density function $f(x)$ by a discrete random variable Y obtained by rounding X to the nearest Δ units. Then the probability function of the discrete random variable is Y is

$$P[Y = y] = P[y - \Delta/2 \leq X \leq y + \Delta/2] \approx \Delta f(y)$$

and its expected value is

$$E(Y) = \sum_y yP[y - \Delta/2 < X \leq y + \Delta/2] \approx \sum_y y\Delta f(y).$$

Note that as the interval length Δ approaches 0, this sum approaches the integral

$$\int xf(x)dx$$

This argues for the following definition of expected value for continuous random variables, if it is to be compatible with the expected value of its discretized or rounded relative Y. For *continuous random variables*

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

and for any function on the real numbers $h(x)$,

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

Using this definition, we find that for the uniform density $f(x) = \frac{1}{b-a}$ for $a < x < b$, the expected value is the midpoint between the two ends of the interval $\frac{a+b}{2}$.

The Exponential Distribution.

Consider a random variable X having probability density function

$$f(x) = \frac{1}{\mu}e^{-x/\mu}, x > 0$$

The cumulative distribution function is given by

$$F(x) = 1 - e^{-x/\mu}$$

and the moments are

$$E(X) = \mu, var(X) = \mu^2$$

Such a random variable is called the *exponential distribution* and it is commonly used to model lifetimes of simple components such as fuses, transistors.

The Normal distribution

Normal Approximation to the Poisson distribution

Consider a random variable X which has the Poisson distribution with parameter μ. Recall that $E(X) = \mu$ and $var(X) = \mu$ so $SD(X) = \mu^{1/2}$. We wish to approximate the distribution of this random variable for large values of μ. In order to prevent the distribution from disappearing off to $+\infty$, consider the standardized random variable

$$Z = \frac{X - \mu}{\mu^{1/2}}.$$

Then $P[Z = z] = P[X = \mu + z\mu^{1/2}] = \frac{\mu^x}{x!}e^{-\mu}$ where $x = \mu + z\mu^{1/2}$ is an integer. Using Stirling's approximation $x! \sim \sqrt{2\pi x}x^x e^{-x}$and taking the limit of this as $\mu \to \infty$, we obtain

$$\frac{\mu^x}{x!}e^{-\mu} \sim \frac{1}{\sqrt{2\pi\mu}}e^{-z^2/2}$$

where the symbol $\sim$ is taken to mean that the ratio of the left to the right hand side approaches 1. The function on the right hand side is a constant multiple of one of the basic functions in statistics $e^{-x^2/2}$ which, upon normalization so that it integrates to one, is the standard normal probability density function.

The standard normal distribution

Consider a continuous random variable with probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, -\infty < x < \infty$$

Such a distribution we call the *standard normal distribution* or the $N(0, 1)$ distribution. The cumulative distribution function

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

is not obtainable in simple closed form, and requires either numerical approximation or a table of values. The probability density function $f(x)$ is symmetric about 0 and appears roughly as follows:

The integral of the standard normal probability density function is 1, but to show this requires conversion to polar coordinates. If we square the integral of the normal probability density function, we obtain

$$\begin{aligned}\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx\right) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dxdy \\ &= \frac{1}{2\pi}\int_{0}^{\infty}\int_{0}^{2\pi} e^{-\frac{1}{2}r^2}rd\theta dr \text{ where } x = r\cos\theta \text{ and } y = r\sin\theta \\ &= 1\end{aligned}$$

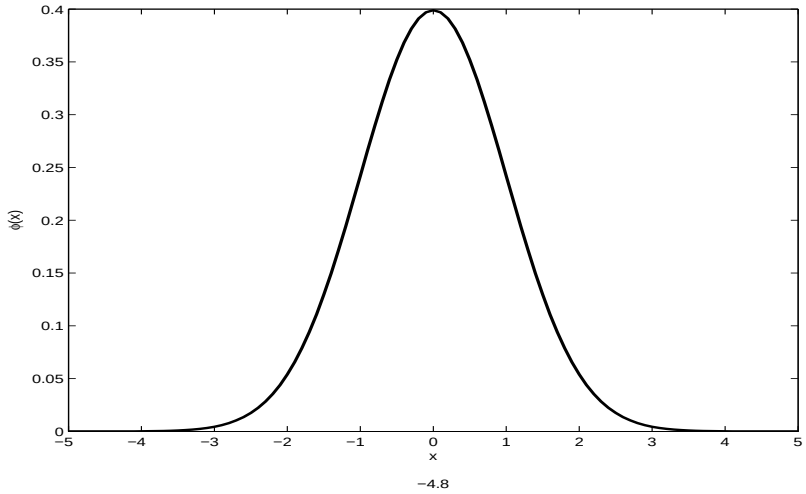

Figure 1.1: The Standard Normal Probability Density Function

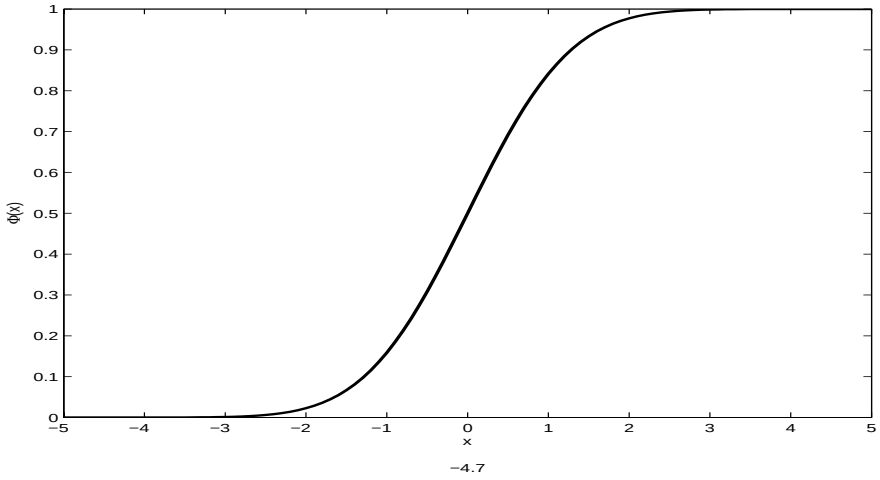

Figure 1.2: The Standard Normal Cumulative Distribution Function

The normal cumulative distribution function is as given in Figure 1.2.

We usually provide the values of the normal cumulative distribution function either through a function such as normcdf in Matlab or through a table of values such as Table A1 (a much more compact version than that found at the back of most statistics books).

x	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.	.500	.540	.579	.618	.655	.692	.726	.758	.788	.816
1.	.841	.864	.885	.903	.919	.933	.945	.955	.964	.971
2.	.977	.982	.986	.989	.992	.994	.995	.997	.997	.998
3.	.9987	.9990	.9993	.9995	.9997	.9998	.9998	.9999	.9999	.99995

Table A1:Values of the Standard Normal Cumulative Distribution Function Φ(x)

For example we can obtain

$\Phi(1.1) = 0.864,$ $\Phi(0.6) = 0.726,$

$\Phi(-0.5) = 1 - \Phi(0.5) \quad = 1 - 0.692$

Note, for example that $\Phi(-x) = 1 - \Phi(x)$ for all x and if Z has a standard normal distribution, we can find probabilities of intervals such as

$$P[-1 < Z < 1] \approx .68 \text{ and } P[-2 < Z < 2] \approx .95.$$

The General Normal Distribution.

If we introduce a shift in the location in the graph of the normal density as well as a change in scale, then the resulting random variable is of the form

$$X = \mu + \sigma Z, Z \sim N(0, 1)$$

for some constants $-\infty < \mu < \infty, \sigma > 0$. In this case, since

$$P(X \leq x) = P(Z \leq \frac{x - \mu}{\sigma}) = \Phi(\frac{x - \mu}{\sigma})$$

it is easy to show by differentiating this with respect to x that the probability density function of X is

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

If a random variable X has the above normal distribution with location μ and scale σ we will denote this by $X \sim N(\mu, \sigma^2)$.

Moments

Show that the function $f(x; \mu, \sigma)$ integrates to 1 and is therefore a probability density function. It is not too hard to find the expected value and variance of a random variable

having the probability density function $f(x;\mu,\sigma)$ by integration:

$$E(X) = \int_{-\infty}^{\infty} xf(x;\mu,\sigma)dx = \mu$$
$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\mu,\sigma)dx = \sigma^2$$

and this gives meaning to the parameters μ and σ^2, the former being the mean or expected value of the distribution and the latter the variance.

Linear Combinations of Normal Random Variables

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent random variables. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. More generally if we sum independent random variables each having a normal distribution, the sum itself also has a normal distribution. The expected value of the sum is the sum of the expected values of the individual random variables and the variance of the sum is the sum of the variances.

Problem.

Suppose $X_i \sim N(\mu, \sigma^2)$ are independent random variables. What is the distribution of the sample mean

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

Assume $\sigma = 1$ and find the probability $P[|\bar{X}_n - \mu| > 0.1]$ for various values of n. What happens to this probability as $n \to \infty$?

The Central Limit Theorem

The major reason that the normal distribution is the single most commonly used distribution is the fact that it tends to approximate the distribution of sums of random variables. For example, if we throw n dice and S_n is the sum of the outcomes, what is the distribution of S_n?. The tables below provide the number of ways in which a given value can be obtained. The corresponding probability is obtained by dividing by 6^n. For example on the throw of $n = 1$ dice the probable outcomes are 1,2,...,6 with probabilities all $1/6$ as indicated in the histogram below:

If we sum the values on two fair dice, the possible outcomes are the values 2,3,...,12 as shown in the following table and the probabilities are the values below:

Values	2	3	4	5	6	7	8	9	10	11	12
Probabilities$\times 36$	1	2	3	4	5	6	5	4	3	2	1

The probability histogram of these values is also shown:

Finally for the sum of the values on three independent dice, the values range from 3 to 18 and have probabilities which, when multiplied by 6^3 result in the values

1 3 6 10 15 21 25 27 27 25 21 15 10 6 3 1

to which we can fit three separate quadratic functions one in the middle region and one

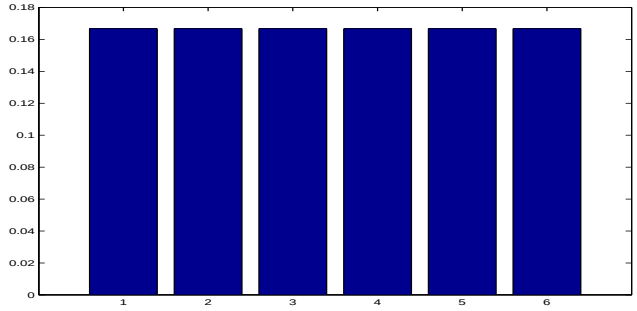

Figure 1.3: Th

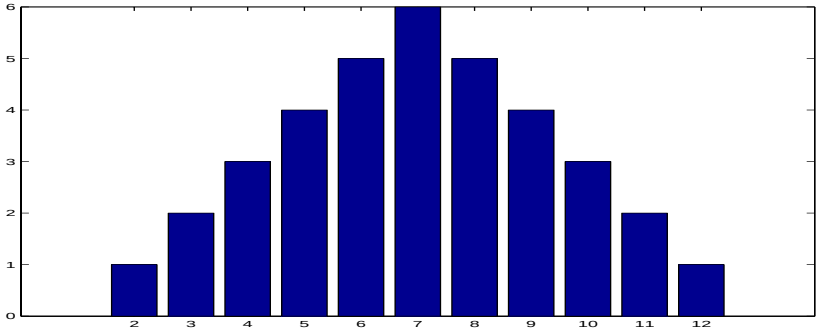

Figure 1.4: The

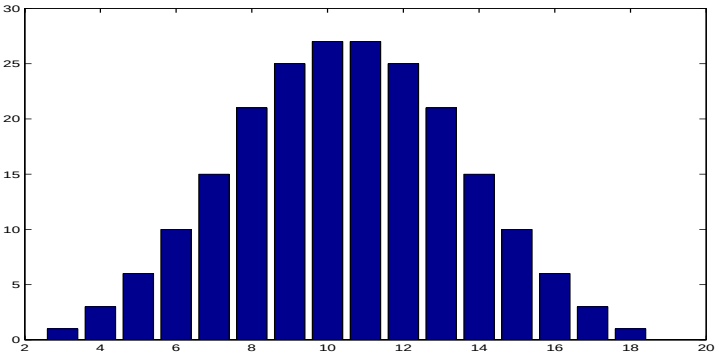

Figure 1.5: The distribution of the sum of three discrete Uniform {1,2,3,4,5,6} Random Variables

in each of the two tails. The histogram of these values shown in Figure 1.5 already resembles a normal probability density function.

In general, these distributions show a simple pattern. For $n = 1$, the probability function is a constant (polynomial degree 0). For $n = 2$, two linear functions spliced together. For $n = 3$ a spline consisting of three quadratic pieces (polynomials of degree $n-1$). In general the histogram for S_n, the sum of the values on n independent dice, consists of n piecewise polynomials of degree $n-1$. These histograms rapidly approach the shape of the normal probability density function.

Example

Let $X_i = 0$ or 1 when the i'th toss of a biased coin is Tails or Heads respectively. What is the distribution of $S_n = \sum_{i=1}^{n} X_i$? Consider the standardized random variable obtained by subtracting $E(S_n)$ and dividing by its standard deviation or the square root of $var(S_n)$:

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$$

Suppose we approximate the distribution of $S_n^*$for large values of n.

First let an integer $x \sim np + z\sqrt{np(1-p)}$ for fixed z and $0 < p < 1$. Then as $n \to \infty$, $x/n \to p$. Stirling's approximation implies that

$$\binom{n}{x} \sim \frac{\sqrt{2\pi}n^{n+1/2}e^{-n}}{2\pi x^{x+1/2}(n-x)^{n-x+1/2}} \sim \frac{1}{\sqrt{2\pi np(1-p)}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}.$$

Also using the series expansion $\ln(1+x) = x - \frac{1}{2}x^2 + O\left(x^3\right)$, putting $\sigma = \sqrt{\frac{p(1-p)}{n}}$, and noting that $\sigma \to 0$ as $n \to \infty$,

$$\begin{aligned}
\ln\{\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}\} &= x\ln(\frac{p}{p+z\sigma}) + (n-x)\ln(\frac{1-p}{1-p-z\sigma}) \\
&= -x\ln(1+\frac{z\sigma}{p}) - (n-x)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\ln(1+\frac{z\sigma}{p}) - n(1-p-z\sigma)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\{(\frac{z\sigma}{p}) - \frac{1}{2}(\frac{z\sigma}{p})^2 + O(\frac{z\sigma}{p})^3\} \\
&- n(1-p-z\sigma)\{-(\frac{z\sigma}{1-p}) - \frac{1}{2}(\frac{z\sigma}{1-p})^2 + O(\frac{z\sigma}{1-p})^3\} \\
&= -n\{z\sigma + \frac{z^2\sigma^2}{p} - \frac{1}{2}\frac{z^2\sigma^2}{p} - z\sigma + \frac{z^2\sigma^2}{1-p} - \frac{1}{2}\frac{z^2\sigma^2}{1-p} + O(\sigma^3)\} \\
&= -\frac{1}{2}z^2\sigma^2(\frac{n}{p} + \frac{n}{1-p}) + O(n^{-1/2}) \\
&= -\frac{z^2}{2} + O(n^{-1/2})
\end{aligned}$$

Therefore,

$$P[S_n = x] = P[S_n^* = z] = \binom{n}{x} p^x (1-p)^{n-x}$$
$$\sim \binom{n}{x} (\frac{x}{n})^x (1-\frac{x}{n})^{n-x} \frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x (1-\frac{x}{n})^{n-x}}$$
$$\sim \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

This is the standard normal probability density function multiplied by the distance between consecutive values of S_n^*. In other words, this result says that the area under the probability histogram for S_n^* for the bar around the point z can be approximated by the area under the normal curve between the same two points $(z \pm \frac{1}{2\sqrt{np(1-p)}})$.

Theorem A8

Let $X_i, i = 1, \ldots n$ be independent random variables all with the same distribution, and with mean μ and variance σ^2. Then the cumulative distribution function of

$$S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges to the cumulative distribution function of a standard normal random variable.

Consider, for example, the case where the X_i are independent each with a Bernoulli (p) distribution. Then the sum $\sum_{i=1}^n X_i$ has a binomial distribution with parameters n, p and the above theorem asserts that if we subtract the mean and we divide by the standard deviation of a binomial random variable, then the result is approximately standard normal. In other words, for large values of n a binomial random variable is approximately normal $(np, np(1-p))$. To verify this fact, we plot both the binomial$(100, 0.4)$ histogram as well as the normal probability density function below.

Problem

Use the central limit theorem and the normal approximation to a probability histogram to estimate the probability that the sum of the numbers on 5 dice is 15. Compare your answer with the exact probability.

The Distribution of a Function of a Random Variable.

We have seen that if X has a normal distribution, then a linear function of X, say $aX + b$ also has a normal distribution. The parameters are easily determined since $E(aX + b) = aE(X) + b$ and $var(aX + b) = a^2 var(X)$. Is this true of arbitrary functions and general distributions? For example is X^2 normally distributed? The answer in general is NO. For example, the distribution of X^2 must be concentrated entirely on the positive values of x, whereas the normal distributions are all supported

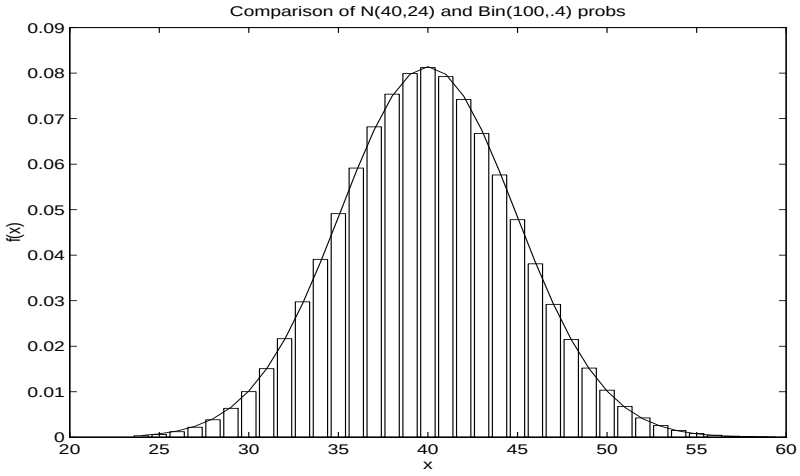

Figure 1.6: Binomial $(100, 0.4)$ Probability histogram together with $N(40, 24)$ probability desity function

on the whole real line (i.e. the probability density function $f(x) > 0, \text{ all } x \in \mathcal{R}$). In general, the safest method for finding the distribution of the function of a random variable in the continuous case is to first find the cumulative distribution of the function and then differentiate to obtain the probability density function. This allows us to verify the result below:

Theorem A9

Suppose a continuous random variable X has probability density function $f_X(x)$. Then the probability density function of $Y = h(X)$ where $h(.)$ is a continuous monotone increasing function with inverse function $h^{-1}(y)$ is

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y)$$

1.6 Moment Generating Functions

Consider a random variable X. We have seen several ways of describing its distribution, using either a cumulative distribution function, a probability density function (continuous case) or probability function or a probability histogram or table (discrete case). We may also use some transform of the probability density or probability function. For example, consider the function defined by

$$M_X(t) = Ee^{tX}$$

defined for all values of t such that this expectations exists and is finite. This function is called the moment generating function of the (distribution of the) random variable

X. It is a powerful tool for determining the distribution of sums of independent random variables and for proving the central limit theorem. In the discrete case we can write $M_X(t) = \sum_x e^{xt} P[X = x]$ and in the continuous case $M_X(t) = \int_{-\infty}^{\infty} e^{xt} f(x) dx$. The logarithm of the moment generating function $\ln(M_X(t))$ is called the cumulant generating function.

Properties of the Moment Generating Function

For these properties we assume that the moment generating function exists at least in some neighbourhood of the value $t = 0$, say for $-\epsilon < t < \epsilon$ for some $\epsilon > 0$. We also assume that $\frac{d}{dt} E[X^n e^{tX}\} = E[\frac{d}{dt} X^n e^{tX}]$ for each value of $n = 0, 1, 2,$ for $-\epsilon < t < \epsilon$. The ability to differentiate under an integral or infinite sum is justified under general conditions involving the rate at which the integral or series converges.

1. $M'(0) = E(X)$
2. $M^{(n)}(0) = E(X^n), n = 1, 2, ...$
3. A moment generating function uniquely determines a distribution. In other words if $M_X(t) = M_Y(t)$ for all $-\epsilon < t < \epsilon$, then X and Y have the same distribution.
4. $M_{aX+b}(t) = e^{bt} M_X(at)$ for constants a, b.
5. If X and Y are independent random variables, $M_{X+Y}(t) = M_X(t) M_Y(t)$.

Examples

Let X have a distribution as given in the table below Then the moment generating function of X is:

Distribution	Moment Generating Function $M_X(t)$
Binomial (n, p)	$(pe^t + 1 - p)^n$
Poisson(λ)	$\exp\{\lambda(e^t - 1)\}$
Exponential, mean μ	$\frac{1}{1-\mu t}$ for $t < 1/\mu$
Normal (μ, σ^2)	$\exp\{\mu t + \sigma^2 t^2/2\}$

Moment generating functions are useful for showing that a sequence of cumulative distribution functions converge because of the following result. The result implies that convergence of the moment generating functions can be used to show convergence of the cumulative distribution functions (i.e. convergence of the distributions).

Theorem A10

Suppose Z_n is a sequence of random variables with moment generating functions $M_n(t)$. Let Z be a random variable Z having moment generating function $M(t)$. If $M_n(t) \to M(t)$ for all t in a neighbourhood of 0, then

$$P[Z_n \leq z] \to P[Z \leq z]$$

as $n \to \infty$ for all values of z at which the function $F_Z(z)$ is continuous.

1.7 Joint Distributions and Convergence

Consider constructing measures on a product Euclidean space. Given Lebesgue measure λ, essentially a measure of length on the real line $\Re$ how do we construct a similar measure, compatible with the notion of area in two-dimensional Euclidean space? We naturally begin with the measure of rectangles or indeed any *product* set of the form $A \times B = \{(x, y); x \in A,\ y \in B\}$ for arbitrary (Borel) sets $A \subset \Re, B \subset \Re$. The measure of a product set can be defined as the product of the measure of the two factor sets,$\mu(A \times B) = \lambda(A)\lambda(B)$. This defines a measure for any product set and by an extension theorem, since the product sets form a Boolean algebra, we can extend this measure to the sigma algebra generated by the product sets.

More formally, suppose we are given two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Define the *product space* to be the space consisting of pairs of objects, one from each of M and N,

$$\Omega = M \times N = \{(x, y);\ x \in M,\ y \in N\}.$$

The Cartesian product of two sets $A \subset M$, and $B \subset N$ is denoted $A \times B = \{(a, b); a \in A, b \in B\}$. This is the analogue of a rectangle, a subset of $M \times N$, and it is easy to define a measure for such sets as follows. Define the *product measure* of product sets of the above form by $\pi(A \times B) = \mu(A)\nu(B)$. The following theorem is a simple consequence of the Caratheodory Extension Theorem.

Theorem A11

The product measure π defined on the product sets of the form $\{A \times B; A \in \mathcal{N},\ \mathcal{B} \in \mathcal{M}\}$ can be extended to a measure on the sigma algebra $\sigma\{A \times B; A \in \mathcal{N}, \mathcal{B} \in \mathcal{M}\}$ of subsets of $M \times N$.

There are two cases of product measure of importance. Consider the sigma algebra on $\Re^2$ generated by the product of the Borel sigma algebras on $\Re$. This is called the Borel sigma algebra in $\Re^2$. We can similarly define the Borel sigma algebra on $\Re^n$.

Similarly, if we are given two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ we can construct a *product measure* Q on the Cartesian product space $\Omega_1 \times \Omega_2$ such that $Q(A \times B) = P_1(A)P_2(B)$ for all $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$.This guarantees the existence of a product probability space in which events of the form $A \times \Omega_2$ are independent of events of the form $\Omega_1 \times B$ for $A \in \mathcal{F}_1, B \in \mathcal{F}_2$.

We say a sequence of random variables $X_1, X_2, \ldots$ is *independent* if the family of sigma-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent, that is for Borel sets $B_n, n = 1, \ldots N$ in $\Re$, the events $[X_n \in B_n], n = 1, ..., N$ form a mutually independent sequence of events so that

$$P[X_1 \in B_1, X_2 \in B_2, ...,X_n \in B_n] = P[X_1 \in B_1]P[X_2 \in B_2]...P[X_n \in B_n]$$

. The sequence is said to be *identically distributed* every random variable X_n has the same cumulative distribution function.

We have already seen the following result, but we repeat it here, if only to get the flavour of the proof.

If X, Y are independent integrable random variables on the same probability space, then XY is also integrable and

$$E(XY) = E(X)E(Y).$$

Proof. Suppose first that X and Y are both simple functions, $X = \sum c_i I_{A_i}, Y = \sum d_j I_{B_j}$. Then X and Y are independent if and only if $P(A_i B_j) = P(A_i)P(B_j)$ for all i, j and so

$$\begin{aligned} E(XY) &= E[(\sum c_i I_{A_i})(\sum d_j I_{B_j})] \\ &= \sum \sum c_i d_j E(I_{A_i} I_{B_j}) \\ &= \sum \sum c_i d_j P(A_i) P(B_j) \\ &= E(X)E(Y). \end{aligned}$$

More generally suppose X, Y are non-negative random variables and consider independent simple functions X_n increasing to X and Y_n increasing to Y. Then $X_n Y_n$ is a non-decreasing sequence with limit XY. Therefore, by the monotone convergence theorem,

$$E(X_n Y_n) \to E(XY).$$

On the other hand,

$$E(X_n Y_n) = E(X_n)E(Y_n) \to E(X)E(Y).$$

Therefore $E(XY) = E(X)E(Y)$. The case of general (positive and negative random variables X, Y we leave as a problem. ■

Joint Distributions of more than 2 random variables

Suppose $X_1, ..., X_n$ are random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$ (but not necessarily independent). The joint distribution can be characterized by the *joint cumulative distribution function*, a function on $\Re^n$ defined by

$$F(x_1, ..., x_n) = P[X_1 \leq x_1, ..., X_n \leq x_n] = P([X_1 \leq x_1] \cap .... \cap [X_n \leq x_n]).$$

The joint cumulative distribution function allows us to find $P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n]$. By the inclusion-exclusion principle,

$$\begin{aligned} &P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n] \qquad (1.8) \\ &\quad = F(b_1, b_2, \ldots b_n) - \sum_j F(b_1, ..., a_j, b_{j+1}, ...b_n) \\ &\quad + \sum_{i<j} F(b_1, ..., a_i, b_{i+1} ... a_j, b_{j+1}, ...b_n) - ... \end{aligned}$$

As in the case, we may then build a probability measure on an algebra of subsets of $\Re^n$. This measure is then extended to the Borel sigma-algebra on $\Re^n$.

Theorem A12

The joint cumulative distribution function has the following properties:

(a) $F(x_1, ..., x_n)$ is right-continuous and non-decreasing in each argument x_i when the other arguments $x_j,\ j \neq i$ are fixed.

(b) $F(x_1, ..., x_n) \to 1$ as $min(x_1, ..., x_n) \to \infty$ and $F(x_1, ..., x_n) \to 0$ as $\min(x_1, ..., x_n) \to -\infty$.

(c) The expression on the right hand side of (1.8) is greater than or equal to zero for all $a_1, \ldots a_n,\ b_1, \ldots, b_n$.

The joint probability distribution of the variables $X_1, ..., X_n$ is a measure on $\mathcal{R}^n$. It can be determined from the cumulative distribution function in the usual fashion, first by defining the measure of intervals and then extending this to the sigma algebra generated by these intervals. In order to verify that the random variables are mutually independent, it is sufficient to verify that the joint cumulative distribution function factors;

$$F(x_1, ..., x_n) = F_1(x_1)F_2(x_2)...F_n(x_n) = P[X_1 \leq x_1] \ldots P[X_n \leq x_n]$$

for all $x_1, ..., x_n \in \Re$.

Theorem A13

If the random variables $X_1, ..., X_n$ are mutually independent, then

$$E[\prod_{j=1}^{n} g_j(X_j)] = \prod_{j=1}^{n} E[g_j(X_j)]$$

for any Borel measurable functions $g_1, \ldots, g_n$.

An infinite sequence of random variables $X_1, X_2, \ldots$ is mutually independent if every finite subset is mutually independent.

Definition (Strong (almost sure) Convergence)

Let X and $X_n, n = 1, 2, \ldots$ be random variables all defined on the same probability space $(\Omega, \mathcal{F})$. We say that the sequence X_n converges *almost surely* (or *with probability one*) to X (denoted $X_n \to X$ a.s.) if the event

$$\{\omega; X_n(\omega) \to X(\omega)\} = \cap_{m=1}^{\infty}[|X_n - X| \leq \frac{1}{m} \text{ a.b.f.o.}]$$

has probability one. Here the notation a.b.f.o., standing for "all but finitely often" is the "$\lim$ inf" of the events $[|X_n - X| \leq \frac{1}{m}]$.

In order to show a sequence X_n converges almost surely, we need that X_n are (measurable) random variables for all n, and to show that there is some measurable random variable X for which the set $\{\omega; X_n(\omega) \to X(\omega)\}$ is measurable and hence

an event, and that the probability of this event $P[X_n \to X]$ is 1.Alternatively we can show that for each value of $\epsilon > 0,\ P[|X_n - X| > \epsilon \text{ i.o.}] = 0,$ or in other words that the probability of the set of all points ω such that $X_n(\omega)$ does not converge to $X(\omega)$ is zero. It is sufficient to consider values of ϵ of the form $\epsilon = 1/m$, m=1,2,... above.

The law of large numbers (sometimes called the law of averages) is the best-known result in probability. It says, for example, that the average of independent Bernoulli random variables, or Poisson, or negative binomial, or Gamma random variables, to name a few, converge to their expected value **with probability one**.

Theorem A14 (Strong Law of Large Numbers)

If $X_n, n = 1, 2, \ldots$ is a sequence of independent identically distributed random variables with $E|X_n| < \infty$, (i.e. they are integrable) and $E(X_n) = \mu$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \text{ almost surely as } n \to \infty$$

1.8 Weak Convergence (Convergence in Distribution)

Consider random variables that are constants; $X_n = 1 + \frac{1}{n}$. By any sensible definition of convergence, X_n converges to $X = 1$ as $n \to \infty$. Does the cumulative distribution function of X_n, F_n,say, converge to the cumulative distribution function of X pointwise? In this case it is true that $F_n(x) \to F(x)$ at all values of x except the value $x = 1$ where the function $F(x)$ has a discontinuity. Convergence in distribution (weak convergence, convergence in Law) is defined as pointwise convergence of the c.d.f. at all values of x except those at which $F(x)$ is discontinuous. Of course if the limiting distribution is absolutely continuous (for example the normal distribution as in the Central Limit Theorem), then $F_n(x) \to F(x)$ does hold for all values of x.

Definition (Weak Convergence)

If $F_n(x),\ n = 1, \ldots$ is a sequence of cumulative distribution functions and if F is a cumulative distribution function, we say that F_n converges to F *weakly* or *in distribution* if $F_n(x) \to F(x)$ for all x at which $F(x)$ is continuous. Weak convergence of a sequence of random variables X_n whose c.d.f. converges in the above sense is denoted in a variety of ways such as $X_n \Rightarrow X$ or $X_n \to_D X$ (here "D" stands for "in distribution").

There are simple examples of cumulative distribution functions that converge pointwise but not to a genuine cumulative distribution because some of the mass of the distribution escapes to infinity. For example if F_n is the cumulative distribution function of a point mass at the point n then $F_n(x) \to 0$ for each fixed value of $x < \infty$. An additional condition called tightness is needed to insure that the limiting distribution is a "proper" probability distribution (i.e. has total measure 1). A sequence of probability measures P_n on Euclidean space is *tight* if for all $\epsilon > 0$, there exists a bounded rectangle K such that $P_n(K) > 1 - \epsilon$ for all n. A sequence of cumulative distribution

functions F_n is tight if, for every $\epsilon > 0$, there is a value of $M < \infty$ such that the probabilities of interval $[-M, M]$ are greater than than $1 - \epsilon$,

$$F_n(M) - F_n(-M) \leq 1 - \epsilon \quad \text{for all } n = 1, 2, ...$$

If a sequence F_n converges to some limiting right-continuous function F at continuity points of F and if the sequence is tight, then the limiting function F is a proper cumulative distribution function and the convergence is in distribution. There is a more general definition of weak convergence that is used for stochastic processes or more general spaces of random elements.

Definition (general definition of weak convergence)

A sequence of random elements of a metric space X_n converges weakly to X i.e. $X_n \Rightarrow X$ if and only if $E[f(X_n)] \rightarrow E[f(X)]$ for all bounded continuous functions f.

Definition (convergence in probability)

We say a sequence of random variables $X_n \rightarrow X$ *in probability* if for all $\epsilon > 0$, $P[|X_n - X| > \epsilon] \rightarrow 0$ as $n \rightarrow \infty$.

Convergence in probability is in general a somewhat more demanding concept than weak convergence, but less demanding than almost sure convergence. In other words, convergence almost surely implies convergence in probability and convergence in probability implies weak convergence.

Theorem A15

If $X_n \rightarrow X$ *almost surely* then $X_n \rightarrow X$ *in probability*.

However, convergence in probability does not imply convergence almost surely**,** but it does imply weak convergence.

Theorem A16

If $X_n \rightarrow X$ *in probability*, then $X_n \rightarrow_D X$.

The converse of this theorem holds under one condition, when the convergence in distribution s to a constant.

Theorem A17

If $X_n \rightarrow_D$ *converges in distribution* to some constant c then $X_n \rightarrow c$ in probability.

The next result, Fubini's theorem, allows us to change the order of integration as long as the function being integrated is, in fact, integrable.

Theorem A18 (Fubini's Theorem)

Suppose $g(x,y)$ is integrable with respect to a product measure $\pi = \mu \times \nu$ on $M \times N$. Then

$$\int_{M\times N} g(x,y)d\pi = \int_M [\int_N g(x,y)d\nu]d\mu = \int_N [\int_M g(x,y)d\mu]d\nu.$$

Convolutions

Consider two independent random variables $X,\ Y$, both having a discrete distribution. Suppose we wish to find the probability function of the sum $Z = X + Y$. Then

$$P[Z = z] = \sum_x P[X = x]P[Y = z - x] = \sum_x f_X(x)f_Y(z - x).$$

Similarly, if X, Y are independent absolutely continuous distributions with probability density functions $f_X,\ f_Y$ respectively, then we find the probability density function of the sum $Z = X + Y$ by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx$$

In both the discrete and continuous case, we can rewrite the above in terms of the cumulative distribution function F_Z of Z. In either case,

$$F_Z(z) = E[F_Y(z - X)] = \int_{\Re} F_Y(z - x)F_X(dx)$$

We use the last form as a more general definition of a *convolution* between two cumulative distribution functions F, G. We define the *convolution* of F and G to be $F * G(x) = \int_{-\infty}^{\infty} F(x - y)dG(y)$.

Properties of convolution

(a) If F, G are cumulative distributions functions, then so is $F * G$

(b) $F * G = G * F$

(c) If either F or G are absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

The convolution of two cumulative distribution functions $F * G$ represents the c.d.f of the sum of two independent random variables, one with c.d.f. F and the other with c.d.f. G.

1.9 Stochastic Processes

A Stochastic process is an indexed family of random variables X_t for t ranging over some index set T such as the integers or an interval of the real line. For example a sequence of independent random variables is a stochastic process, as is a Markov chain. For an example of a continuous time stochastic process, define X_t to be the price of a stock at time t (assuming trading occurs continuously over time).

Markov Chains

Consider a sequence of (discrete) random variables $X_1, X_2, \ldots$ each of which takes integer values $1, 2, \ldots N$ (called *states*). We assume that for a certain matrix P (called the *transition probability matrix*), the conditional probabilities are given by corresponding elements of the matrix; i.e.

$$P[X_{n+1} = j | X_n = i] = P_{ij}, i = 1, \ldots N, j = 1, \ldots N$$

and furthermore that the chain only cares about the last state occupied in determining its future; i.e. that

$$P[X_{n+1} = j | X_n = i, X_{n-1} = i_1, X_{n-2} = i_2 ... X_{n-l} = i_l] = P[X_{n+1} = j | X_n = i] = P_{ij}$$

for all $j, i, i_1, i_2, \ldots i_l$, and $l = 2, 3, ...$ Then the sequence of random variables X_n is called a *Markov Chain*. Markov Chain models are the most common simple models for dependent variables, including weather (precipitation, temperature), movements of security prices etc. They allow the future of the process to depend on the present state of the process, but the past behaviour can influence the future only through the present.

Properties of the Transition Matrix P

Note that $P_{ij} \geq 0$ for all i, j and $\sum_j P_{ij} = 1$ for all i. This last property implies that the $N \times N$ matrix $P - I$ (where I is the identity matrix) has rank at most $N - 1$ because the sum of the N columns of $P - I$ is identically 0.

Example. Rain-No rain

Suppose that the probability that tomorrow is rainy given that today is not is α and the probability that tomorrow is dry given that today is rainy is β. Then if we assume that tomorrow's weather depends on the past only through whether today is wet or dry, the chain given by

$$X_n = \begin{cases} 1 & \text{if} \quad \text{Day } n \text{ is wet} \\ 0 & \text{if} \quad \text{Day } n \text{ is dry} \end{cases}$$

is a Markov chain having transition matrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

The distribution of X_n

Suppose that the chain is started by randomly choosing a state for X_0 with distribution $P[X_0 = i] = q_i, i = 1, 2, \ldots N$. Then the distribution of X_1 is given by

$$\begin{aligned} P(X_1 = j) &= \sum_{i=1}^{N} P(X_1 = j, X_0 = i) \\ &= \sum_{i=1}^{N} P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=1}^{N} P_{ij} q_i \end{aligned}$$

and this is the $j'th$ element of the vector $q'P$ where $\underline{q}$ is the column vector of values q_i. Similarly the distribution of X_n is the vector $\underline{q}'P^n$ where P^n is the product of the matrix P with itself n times. Under very general conditions, it can be shown that these probabilities converge and in many such cases, the limit does not depend on the initial distribution $\underline{q}$.

Definition

A *limiting distribution* of a Markov chain is a vector ($\underline{\pi}$ say) of long run probabilities of the individual states so

$$\pi_i = \lim_{t \to \infty} P[X_t = i].$$

Definition

A *stationary distribution* of a Markov chain is the column vector ($\underline{\pi}$ say) of probabilities of the individual states such that

$\underline{\pi}'P = \underline{\pi}'$.

Theorem A19

Any limiting distribution of a Markov Chain must be a stationary distribution.

Proof.

Note that $\pi' = \lim_{n\to\infty} q'P^n = \lim_{n\to\infty}(q'P^n)P = (\lim_{n\to\infty} q'P^n)P = \pi'P$.

Example: Binary information:

Suppose that $X_1, X_2, \ldots$ is a sequence of binary information (Bernoulli random variables) taking values either 0 or 1. Suppose that the probability that a 0 is followed by a 1 is p and the probability that a 1 is followed by a 0 is given by q where $0 < p, q < 1$. Then the transition matrix for the Markov chain is

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

The limiting distribution for this Markov Chain is

$$\pi = \begin{pmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} \end{pmatrix}$$

so, for example, the long run proportion of zeros in the sequence is $\frac{q}{p+q}$.

When is the limiting distribution of a Markov chain unique and independent of the initial state of the chain?

Definition: irreducible, aperiodic

We say that a Markov chain is *irreducible* if every state can be reached from every other state. In other words for every pair i, j there is some m such that $P_{i,j}^{(m)} > 0$. We say that the chain is *aperiodic* if, for each state i, there is no regular or periodic pattern for the values of k for which $P_{ii}^{(k)} > 0$. For example if $P_{ii}^{(1)} = 0,\ P_{ii}^{(2)} > 0,\ P_{ii}^{(3)} = 0,\ P_{ii}^{(4)} > 0$ and this pattern continues indefinitely then the greatest common divisor of the values k such that $P_{ii}^{(N)} > 0$ is evidently 2. We write this mathematically as $\gcd\{k; P_{ii}^{(k)} > 0\} = 2$ and this chain is not aperiodic, it has period 2. On the other hand if for all states i, $\gcd\{k; P_{ii}^{(k)} > 0\} = 1$, we say the chain is aperiodic. For a *periodic chain* (i.e. one that is not aperiodic, so the period $\gcd\{k; P_{ii}^{(k)} > 0\}$ is greater than one) returns to a state can occur only at multiples of the period $gcd\{N; P_{ii}^{(N)} > 0\}$.

Theorem A20

If a Markov chain is irreducible and aperiodic, then there exists a *unique* limiting distribution $\underline{\pi}$. In this case, as $n \to \infty$, $P^n \to \underline{\pi}'\mathbf{1}$ the matrix whose rows are all identically $\underline{\pi}'$.

Generating Functions

Definition: Generating function

Let $a_0, a_1, a_2, \ldots$ be a finite or infinite sequence of real numbers. Suppose the power series

$$\mathcal{A}(t) = \sum_{i=0}^{\infty} a_i t^i$$

converges for all $-\epsilon < t < \epsilon$ for some value of $\epsilon > 0$. Then we say that the sequence has a *generating function* $\mathcal{A}(t)$.

Note. Every bounded sequence has a generating function since the series $\sum_{i=0}^{\infty} t^i$ converges whenever $|t| < 1$. Thus, discrete probability functions have generating functions. The generating function of a random variable X or its associated probability function $f_X(x) = P[X = x]$ is given by

$$\mathcal{F}_X(t) = \sum_x f_X(x) t^x = E(t^X).$$

Note that if the random variable has finite expected value, then this converges on the interval $t \in [-1, 1]$.

The advantage of generating functions is that they provide a transform of the original distribution to a space where many operations are made much easier. We will give examples of this later. The most important single property is that they are in one-one correspondence with distributions such that the series converges; for each distribution there is a unique generating function and for each generating function there is a unique distribution.

As a consequence of this representation and the following theorem we can use generating functions to determine distributions that would otherwise be difficult to identify.

Theorem A21

Suppose a random variable X has generating function $\mathcal{F}_X(t)$ and Y has generating function $\mathcal{F}_Y(t)$. Suppose that X and Y are independent. Then the generating function of the random variable $W = X + Y$ is $\mathcal{F}_W(t) = \mathcal{F}_X(t)\mathcal{F}_Y(t)$.

Notice that whenever a moment generating function exists, we can recover the generating function from it by replacing e^t by t.

Example

One of six different varieties of coupons is placed in each box of cereal. Find the distribution of the number of cereal boxes you need to buy to obtain all six coupons. It is fairly easy to verify that the (probability) generating function of the number is

$$\frac{5!t^6}{(6-t)(6-2t)(6-3t)(6-4t)(6-5t)} = \frac{5}{324}t^6 + \frac{25}{648}t^7 + \frac{175}{2916}t^8 + \frac{875}{11\,664}t^9 + \frac{11\,585}{139\,968}t^{10} + \frac{875}{10\,368}t^{11} + O\left(t^{12}\right)$$

from which we discover that the probability of only six cereal boxes is $\frac{5}{324}$, the probability of seven is $\frac{25}{648}$, etc.

The Poisson Process

One of the simplest continuous time stochastic processes is the *Poisson Process*. Suppose N_t denotes the total number of arrivals into a system (such as the number of customers arriving at a queue) until time t. Note that the number of arrivals in time interval $(a, b]$ is then $N_b - N_a$. Assume the following properties;

(a) The probability of exactly one arrival in a small interval of length Δt is $\lambda \Delta t + o(\Delta t)$. (Note that the probability does not depend on where the interval is, only on its length).

(b) The probability of two or more arrivals in an interval of length Δt is $o(\Delta t)$ where by definition of the o notation, $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

(c) For disjoint intervals $I_i = (a_i, b_i]$ (so $I_i \cap I_j = \phi, i \neq j$), the number of arrivals in these intervals $Y_i = N_{b_i} - N_{a_i}$ are mutually independent random variables.

Theorem A22

Under the above conditions, (a)-(c), the distribution of the process $N_t, t \in T$ is that of a *Poisson process*. This means that the number of arrivals $N_b - N_a$ in an interval $(a, b]$ has a Poisson distribution with parameter $\lambda(b-a) = \lambda \times$*the length of the interval*, and the number of arrivals in disjoint time intervals are independent random variables. The parameter λ specifies the *rate* of the Poisson process.

We can easily show that if $N(t)$ is a Poisson process and $T_1, T_2, ...$ are the times of the first event, and the time between the first and second events, etc. then $T_1, T_2, ...$ are independent random variables, each with an exponential distribution with expected value $1/\lambda$. Moreover if $T_1, T_2, ... T_n$ are independent random variables each with an exponential (1) distribution, then the sum $\sum_{i=1}^{n} T_i$ has a (gamma) probability density function with probability density function

$$f(x) = \frac{1}{(n-1)!} x^{n-1} e^{-x}, \text{ for } x > 0.$$

This means that the event times for a Poisson process are Gamma distributed.

Poisson Process in space.

In an analogous way we may define a Poisson process in space as a distribution governing the occurrence of random points with the properties indicated above; The number of points in a given set S has a Poisson distribution with parameter $\lambda \times |S|$ where $|S|$ is the area or volume of the set, and if $Y_1, Y_2, ...$are the number of points occurring in disjoint sets $S_1, S_2, \ldots$, they are mutually independent random variables.