Introduction to Probability: Lecture Notes

1 Discrete probability spaces

1.1 Infrastructure

A probabilistic model of an experiment is defined by a *probability space* consisting of a set (*sample space* Ω) of sample points or outcomes (exhaustive collection of elementary outcomes of the experiment) and a probability law P which assigns to each *event* in (subset of) Ω a probability satisfying three axioms: (1) nonnegativity, i.e., all probabilities are nonnegative; (2) additivity, i.e., the probability of the union of disjoint events is the sum of the probabilities of the events taken alone; and (3) normalization, i.e., $P(\Omega) = 1$: the sum of the probabilities of all the outcomes is 1. In cases of interest here, the probability law is defined more simply by probabilities assigned to each of the outcomes.

The *law of total probability* refers to a partition[1] $\{A_i\}$ of Ω into n subsets and states that, for any event B,

$$P(B) = \sum_{1\le i\le n} P(B\cap A_i)$$

Conditional probabilities are defined on subsets of Ω: $P(A|B)$ is a probability on the (sub-) sample space $B \subset \Omega$ and is the sum of the probabilities of the sample points in $A\cap B$ normalized by the probability $P(B)$ in the original probability space, assuming that $P(B) > 0$, i.e.,

$$P(A|B) = \sum_{\omega\in A\cap B} \frac{P(\{\omega\})}{P(B)}.$$

We call $P(A|B)$ the conditional probability of event A given event B (or given that event B has “occurred” or “holds”). The *multiplication rule* generalizes $P(A\cap B) = P(A|B)P(B) = P(B|A)P(A)$ and is given by

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

Events A and B are *independent* if $P(A\cap B) = P(A)P(B)$; more generally, events $A_1, A_2, \ldots, A_n$ are independent if the probability of the intersection of any subset of the events is the product of the probabilities of the individual events.

Bayes' rule says that, for any partition $\{A_1,\ldots,A_n\}$ of Ω,

$$P(B|A) = \frac{P(A|B)P(B)}{\sum_{1\le i\le n} P(A|A_i)P(A_i)}$$

[1] Recall that a partition of a set S is a collection of mutually disjoint (mutually exclusive) subsets whose union is S.

Typically, as in the text, event B is one of the events in the partition $\{A_i\}$. Also, $n = 2$ is common, so that the rule, with $B = E, A_1 = E, A_2 = E^c$, has the form

$$P(E|A) = \frac{P(A|E)P(E)}{P(A|E)P(E) + P(A|E^c)P(E^c)}$$

Boole's inequality generalizes $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$ and reads

$$P(A_1 \cup \ldots \cup A_n) \leq \sum_{1 \leq i \leq n} P(A_i)$$

with equality (according to the additivity axiom) when the A_i's are mutually disjoint.

1.2 Counting

We need to count permutations, combinations, and partitions, and to distinguish sampling with and without replacement.

Permutations. The number of permutations of n elements taken k at a time is

$$(n)_k = n(n-1)(n-2)\ldots(n-k+1)$$

Thus, the number of different orderings of n elements is $n!$.

In population sampling terms, if you choose k people sequentially *without* replacement out of a population of n, then there are $(n)_k$ ordered samples. If the sampling of k people from the population of n is done *with* replacement, then people can be chosen more than once and the number of possible samples is n^k. Thus, the probability that a random sample[2] with replacement actually chooses k distinct people is $(n)_k/n^k$.

Example: *The birthday paradox* is created by the observation that the above probability is less than 1/2 for $n = 365$ possible birthdays, and a random sample containing only $k = 23$ birthdays. If you choose k people at random, you are picking k out of 365 possible birthdays (nearly) at random, and so if $k = 23$, the chances are better than 1/2 that at least two of the k have a common birthday(!).

Combinations. The number $(n)_k$ counts $k!$ permutations for each distinct subset of k elements. Thus the number of k element subsets of an n-element set, i.e., the number of combinations of n objects taken k at a time, is

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

[2] This will always refer to a sample taken *uniformly* at random with all samples equally likely.

which is the familiar binomial coefficient. This is also called the number of combinations of n elements taken k at a time. In population sampling terms, if you choose k people sequentially without replacement out of a population of n, then if the ordering of the sample is ignored, there are $\binom{n}{k}$ distinct samples.

Partitions. The number of ordered partitions of an n element set into subsets of sizes $n_1, n_2, \ldots, n_r$, $(\sum_{1\le i\le r} = n)$ is given by the *multinomial coefficient*

$$\begin{aligned} \binom{n}{n_1\ n_2 \cdots n_r} &:= \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r} \\ &= \frac{n!}{n_1!\cdots n_r!} \end{aligned}$$

Now consider a set of *indistinguishable* elements (an urn of identical balls) and ask how many ways can the elements be partitioned into an ordered collection of r subsets with sizes summing to n? That is, how many choices for $(n_1, \ldots, n_r)$ are there such that $n_1+\cdots+n_r = n$? Consider lining up in $n+r-1$ positions, in some order, the n elements along with $r-1$ separators. The subsets are defined as the sequences of elements between successive separators; adjacent separators are allowed and define empty subsets. An illustration is given by the obvious notation $*||****|*|***$, where $n = 9, r = 5$ and the sequence of r subset sizes is 1,0,4,1,3. There are $(n+r-1)_{r-1}$ ways of choosing the separator locations, but each ordered collection of subsets is represented $r!$ times in this count, so one obtains the answer $\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$. A similar argument (left as an exercise) shows that if empty sets are disallowed, the answer changes to $\binom{n-1}{r-1}$.

1.3 Discrete Random Variables

Given some underlying probability space, *random variables* (rv's) are functions from the sample space Ω to the set $\mathbb{R}$ of real numbers. For a discrete rv X, which is our initial interest, the range of X is discrete and typically a subset of the integers, but the domain may be continuous or discrete. We usually succeed in using capital letters towards the end of the alphabet to denote rv's. Many times the sample points $\omega \in \Omega$ are integers and the mapping X is the identity map: $X(\omega) = \omega$, for all $\omega \in \Omega$.

Example 1: Consider the toss of the conventional 6-sided die, where $\Omega = \{1, \ldots, 6\}$ and $P(A) = \frac{|A|}{6}$, $A \in 2^\Omega$, is the probability law. The definition $X(\omega) = \omega$, $1 \le \omega \le 6$, is more an introduction of new terminology than an introduction of new probabilistic concepts. ■

More generally, the subsets $\{\omega \in \Omega | X(\omega) = x\}$ form events typically defined by some property.

Example 2: Consider the toss of 2 dice with the sample points being all pairs (a, b) with a and b integers in $\{1, 2, 3, 4, 5, 6\}$. Define

$$X((a, b)) = a + b$$

so the property is the sum of the component values; the sets $\{\omega \in \Omega | X(\omega) = x\}$ are those points all having the sum x of component values.
Exercise: Find the probabilities of all values in the range of X.
■

The mapping from the range R_X of X into the set of probabilities $P(\{\omega \in \Omega | X(\omega) = x\})$ is called the *probability mass function* (or pmf) of X and is abbreviated $p_X(x)$ with the underlying probability space understood. By our definitions, it is clear that $\sum_{x \in R_X} p_X(x) = 1$ (verify this). We also adopt the simpler notation, for a given set $S \subseteq \mathbb{R}$,

$$P(X \in S) = P(\{\omega \in \Omega | X(\omega) \in S\})$$

again with the underlying probability space understood. Set membership will be denoted in customary ways, e.g., when $X \in S$ if and only if $1 \leq X \leq n$, then we will usually write $P(1 \leq X \leq n)$. In Example 2 above, we write $P(X = 2) = 1/36,\ P(X = 3) = 1/18,\ P(X = 4) = 1/12$, etc.

Functions of rv's give rv's. If $Y = f(X)$ then $P(X = x) = P(Y = f(x))$ and $P(Y = y) = P(X \in \{x | f(x) = y\})$.

Example 3: Let X have a uniform pmf on the integers $\{1, \ldots, 2n\}$ for some positive integer n, and let Y be the function $Y = \lfloor X/2 \rfloor$, the largest integer no larger than $X/2$. Clearly, $0 \leq Y \leq n$, and except for $x = 1$ and $x = n$, $\{x | f(x) = y\}$ has two integers, $2y$ and $2y + 1$. We get

$$P(Y = y) = \begin{cases} \frac{1}{2n}, & y = 0 \\ \frac{1}{n}, & 1 \leq y \leq n - 1 \\ \frac{1}{2n}, & y = n \end{cases}$$

■

Multiple rv's defined on the same probability space have *joint* pmf's and are handled in analogy with the probabilities $P(A \cap B)$ of Chapter 1. The notation for the joint probability that $X = x$ and $Y = y$ is

$$p_{X,Y}(x, y) := P(X = x, Y = y) = P(\Omega_{\{X=x\}} \cap \Omega_{\{Y=y\}})$$

where we define, for any such rv Z

$$\Omega_{\{Z=z\}} := \{\omega \in \Omega | Z(\omega) = z\}$$

In analogy with the Theorem of Total Probability of Chapter 1, we see that

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

where, in this context of multiple rv's, $p_X(x)$ is called a *marginal* pmf (distribution).

1.3.1 Conditional probabilities and independence

The related concepts like conditional probabilities and independence extend in the obvious ways, as does the notion of a function of an rv. Let S be a subset of the range of an rv X in a given probability space. Then $P(X = x|S)$ is to be interpreted as $P(B|A)$ (as described in Chapter 1) with $B = \Omega_{\{X=x\}}$ and $A = \{\omega \in \Omega | X(\omega) \in S\}$. Thus,

$$P(X = x|S) = \begin{cases} \frac{p_X(x)}{\sum_{y \in S} p_X(y)} & x \in S \\ 0 & x \notin S \end{cases}$$

Note that the above function is a bona fide pmf; in particular the normalization axiom holds.

For two rv's X and Y defined on the same probability space, the conditional probability

$$p_{X|Y}(x|y) := P(X = x|Y = y) = P(X = x, Y = y)/P(Y = y)$$

is the above conditional probability with $S = \Omega_{\{Y=y\}}$. The rv's X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y in the ranges of X and Y, respectively. The concept extends to multiple rv's as it did to multiple events in Chapter 1.

Example 4: The roll of a pair of identical n-sided dice called die 1 and die 2 is modeled by two rv's $X_i,\ i = 1, 2$, giving the respective results for die 1 and die 2. The dice behave independently and we have, for each possible result x_1, x_2,

$$\begin{aligned} p_{X_1,X_2}(x_1, x_2) &= p_{X_1|X_2}(x_1|x_2)p_{X_2}(x_2) \\ &= p_{X_2|X_1}(x_2|x_1)p_{X_1}(x_1) \\ &= p_{X_1}(x_1)p_{X_2}(x_2) \\ &= 1/n^2 \end{aligned}$$

Now define a function $Y = \min(X_1, X_2)$ and compute as an exercise

$$P(Y = k) = \frac{2(n-k)+1}{n^2}, \quad 1 \le k \le n$$

■

Note that we have bid a fond farewell to the explicit mention of probability spaces; we now speak primarily of rv's and their pmf's.

1.3.2 Expectation

Random variables typically represent measurements, counts, etc. so it is of interest to know their averages. To motivate our definitions, suppose we want the average daily peak wind speed recorded in New York over the past n days. Most people would understand that to mean the sum of the n peaks divided by n. Now if n is large and speeds are rounded to integers denoting miles per hour (mph), which is standard, we will have many repetitions of the same speed. If n_i is the number of times the peak is i mph in the past n days, then we can organize the calculation of the average α_n as

$$\begin{aligned} \alpha_n &= \frac{1}{n}\sum_{i\geq 0} n_i \times i \\ &= \sum_{i\geq 0} \frac{n_i}{n} \times i \end{aligned}$$

where $\sum_{i\geq 0} n_i/n = 1$. Now for fairly large n we can interpret n_i/n, the fraction of the days with a peak wind speed of i mph, as an estimate of the probability that a randomly picked one of the last n days has a peak wind speed of i mph. Formally, a model of this situation defines X as an rv giving the daily peak wind speed, lets $p_X(i) \approx n_i/n$ be its pmf, and defines the *expected value* of X to be

$$\mathbb{E}X := \sum_{i\geq 0} i p_X(i)$$

which is to be compared with α_n above. $\mathbb{E}X$ is also called the mean, expectation, or simply average of X. The expected value of some function $g(X)$ of the peak wind speed is defined as you would expect,

$$\mathbb{E}g(X) = \sum_{i\geq 0} g(i) p_X(i)$$

In particular, $\mathbb{E}X^k$ is called the k-th *moment* of X, so the mean is also known as the first moment.

The *centered* version of the rv X is $\hat{X} = X - \mathbb{E}X$ and is so named because its mean is obviously 0. The second moment $var(X) := \mathbb{E}\hat{X}^2$ of the centered version is called the *variance* of X. The square root of the variance is called the *standard deviation* of X and clearly has the same units as X. The standard deviation, denoted by σ_X, is a classical measure of how much mass the pmf concentrates away from the mean, and how spread out the pmf is.

Expressing the variance in terms of the moments gives

$$\begin{aligned} Var_X &= \sum_x (x - \mathbb{E}X)^2 p_X(x) \\ &= \mathbb{E}X^2 - 2\mathbb{E}X \cdot \mathbb{E}X + [\mathbb{E}X]^2 \\ &= \mathbb{E}X^2 - [\mathbb{E}X]^2 \end{aligned}$$

The mean and standard deviation of a linear function $Y = aX + b$ of the rv X:

$$\begin{aligned}\mathbb{E}Y &= a\mathbb{E}X + b \\ \sigma_Y &= a\sigma_X\end{aligned}$$

The standard deviation is not influenced by the location parameter b (as it just measures dispersion around the mean), but it scales, as does the mean, by the factor a.

The conditional expected value of X given the event A, i.e., the mean of the distribution $P(X|A)$, is denoted by

$$\mathbb{E}[X|A] = \sum_x xP(X = x|A)$$

Similarly, for X and Y defined on the same probability space, we have the conditional expectation

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

from which, in analogy with the theorem of total probability,

$$\begin{aligned}\mathbb{E}X &= \sum_y \sum_x xp_{X,Y}(x, y) \\ &= \sum_y \sum_x xp_{X|Y}(x|y)p_Y(y)\end{aligned}$$

which can be put in the form of the *theorem of total expectation*

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

It follows directly from the definition of independence of rv's that, for any given functions g, h,

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

Moments need not exist. For example, the zeta pmf is

$$p_X(k) = \frac{1}{\zeta(s)} \cdot \frac{1}{k^s},\ k \geq 0$$

with $s > 1$ a parameter, and with $\zeta(s)$, the Riemann zeta function, being the *normalization constant*

$$\zeta(s) = \sum_{k\geq 0} \frac{1}{k^s}$$

so called because it ensures the normalization axiom of the probability law. (Restricted to a finite set, the zeta pmf becomes Zipf's law.) Such pmf's are said to have *heavy tails*, where tails are defined formally by the numbers $P(X > n) = \sum_{k>n} p_X(k)$. A tail is heavy in the sense that it approaches the axis only rather slowly and so has a lot of the probability mass.[3] If one chooses $s = 2$, one finds that the mean $\sum_{k\geq 1} 1/k$ does not converge; the partial sums $\sum_{1\leq k\leq n} 1/k$, which are known as the harmonic numbers, grow as $\ln n$, for large n.

For another amusing example called the St. Petersburg paradox, suppose you pay \$1000 to play the following game. You start out being given \$1 as your initial fortune, and then have your fortune doubled for each successive head you toss; when you toss your first tails, the game is over. Your expected winnings are $\alpha - \$1000$ where

$$\begin{aligned} \alpha &= 1 \cdot \frac{1}{2} + 2 \cdot (\frac{1}{2})2 + \ldots + 2^{i-1} \cdot (\frac{1}{2})^i + \ldots \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots \\ &= \infty \end{aligned}$$

so this sounds like a good game for you. But you win back your fee of \$1000 only if you get 10 heads in a row, at least, and the odds *against* this happening are worse than 1000 to 1! The game doesn't sound so good now. The probability of big wins is positive but very small; on the other hand, it's not so small as to create a finite expected value. Moral: As a measure of a pmf, the mean may have limited usefulness in practice. To a certain extent this usefulness will be seen by looking at the variance, which in this example, also entails a divergent sum.

An example of this last comment, one similar to experiences you have all had, requires the balance of two courses of action, each with the same expectation, but much different variability. There are two routes between New York City and Newark, the first always requiring 30 minutes. The second requires only 15 minutes 2/3 of the time, but 1/3 of the time it requires a whole hour. Which route do you prefer? They both have a mean of 30 minutes; the first one has a standard deviation of 0 minutes and is totally reliable, while the second has a standard deviation of over 20 minutes to go along with the fact that it takes 1/2 the average 2/3 of the time.

Let X be defined on the positive integers $\{1, 2, \ldots, n\}$. Then the mean can also be expressed in terms of the tails as follows:

$$\mathbb{E}X = \sum_{0\leq k\leq n-1} P(X > k)$$

(If one expands the summands, the proof is simply a reorganization of sums, which is left as an exercise.)

1.3.3 Important pmf's with applications

A *Bernoulli* rv B takes on the value 1 or 0 with probability p or $1-p$, respectively. B can be looked upon as a mapping of coin tosses with p the bias for heads, as

[3] For rv's taking on both positive and negative values, there will be two tails: the one given above and $P(X < -n)$

we discussed earlier. We have easily

$$\begin{aligned}\mathbb{E}B &= p \\ Var(B) &= p(1-p)\end{aligned}$$

Sums of independent and identical Bernoulli rv's have great importance. In a bit string $B_1, \ldots, B_n$ of length n the sum S_n of the B_i's gives the number of one bits. The probability that the sum is k is the probability that k of the bit positions contain a 1 and $n-k$ have a 0, and since there are $\binom{n}{k}$ different ways of choosing the positions of the 1 bits, we get

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k},\ 0 \le k \le n$$

which is the *binomial* pmf with parameters n, p. Clearly, this pmf also gives the probabilities of k heads in n tosses of a coin. More generally, sequences like $B_1, \ldots, B_n$ are often called sequences of *Bernoulli trials* with probability of "success" p and probability of "failure" $1-p$.

Before giving the mean and variance of a binomial rv, we make the following two observations:

- The mean of a sum of rv's (independent or not) is the sum of their means.
- The variance of a sum of *independent* rv's is the sum of their variances.

The first observation is trivial to prove. To prove the second for two rv's X and Y, routine algebra leads to

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

and so, since $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$ by independence, we get the desired result. Inductively, our second observation above follows for any number of independent rv's.

Using observations 1 and 2 on a sum of Bernoulli rv's, we see that, for a binomial rv,

$$\begin{aligned}\mathbb{E}S_n &= np \\ Var(S_n) &= np(1-p)\end{aligned}$$

Exercise: Prove these directly using the pmf for a binomial rv. ■

If the number of throws of a standard 6-sided die is unconstrained, what is the pmf for the number, Y, of (independent) throws that have been made when the first ace appears? Since the throws are independent, this is given by the *geometric* pmf $p_Y(k) = (1-5/6)^{k-1}1/6$. More generally, we speak of Bernoulli trials with success probability p, and the pmf

$$p_Y(k) = (1-p)^{k-1}p,\ k = 1, 2, \ldots$$

defined on the positive integers. If we had asked for the pmf of the number Y^* of failures before getting the first success, then we would have written

$$p_{Y^*}(k) = (1-p)^k p, \; k = 0, 1, 2, \ldots$$

which is also the geometric pmf, but defined on the nonnegative integers. (You should be able to verify in your head that $\sum_{k\geq 0} p_{Y^*}(k) = \sum_{k\geq 1} p_Y(k) = 1$.) Using our observations on linear functions of rv's, we have, after routine calculations,

$$\begin{aligned} \mathbb{E}Y^* &= \mathbb{E}Y - 1 = \frac{1}{p} - 1 = \frac{1-p}{p} \\ Var(Y) &= var(Y^*) = \frac{1-p}{p^2} \end{aligned}$$

The *Pascal* pmf, commonly known as the *negative binomial* pmf, has connections with the binomial and geometric pmf's, and gives the probability that the k-th success of a Bernoulli process occurs at the t-th trial. This event occurs if and only if the t-th trial yields a success (with probability p) and the preceding $t-1$ had $k-1$ successes (with probability $\binom{t-1}{k-1}p^{k-1}(1-p)^{t-k}$), so by the independence of the t-th trial from all those preceding it, the trial X_k giving the k-th success has the pmf

$$p_{X_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \; t \geq k \geq 1$$

The numbers of trials between adjacent successes is geometric so we can find the moments of X_k from those for a sum of k independent geometric rv's on the positive integers with parameter p. Then

$$\begin{aligned} \mathbb{E}X_k &= k\frac{1}{p} \\ Var(X_k) &= k\frac{1-p}{p^2} \end{aligned}$$

To compute moments directly, it is most convenient to derive a recurrence. We have

$$\begin{aligned} \mathbb{E}X^r &= \sum_{t=k}^{\infty} t^r \binom{t-1}{k-1} p^k (1-p)^{t-k} \\ &= \frac{k}{p} \sum_{t=k}^{\infty} t^{r-1} \binom{t}{k} p^{k+1} (1-p)^{t-k} \\ &= \frac{k}{p} \sum_{u=k+1}^{\infty} (u-1)^{r-1} \binom{u-1}{k} p^{k+1} (1-p)^{u-k+1} \\ &= \frac{k}{p} \mathbb{E}[(Y-1)^{r-1}] \end{aligned}$$

where Y is negative binomial with parameters $k+1$ and p. Set $r=1$ to get the mean above and then use the recurrence with $k=2$ to obtain the second moment, whence the variance.

The Poisson distribution is an excellent model of many natural and engineering processes (e.g., the number of calls arriving at a telephone exchange in a given time interval, the number of α particles irradiated by a radioactive substance in a unit of volume over a given time interval, the rainfall (number of drops) on a unit area over some time interval, flaws in crystal structures, ... A little later in the course we will see what properties characterize such applications. For a given parameter β, an rv X has a *Poisson* distribution if

$$p_X(k) = \frac{\beta^k}{k!}e^{-\beta}, \ k = 0, 1, 2, \dots \tag{1}$$

The mean is simply the parameter β, as is the variance:

$$\begin{aligned}\mathbb{E}X &= \sum_{k\geq 0} k\frac{\beta^k}{k!}e^{-\beta} \\ &= \beta\sum_{k\geq 1}\frac{\beta^{k-1}}{(k-1)!}e^{-\beta} \\ &= \beta\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}X^2 &= \beta^2\sum_{k\geq 2}k(k-1)\frac{\beta^{k-2}}{k!}e^{-\beta}+\beta \\ &= \beta^2\sum_{k\geq 2}k(k-1)\frac{\beta^{k-2}}{(k-2)!}e^{-\beta}+\beta \\ &= \beta^2+\beta\end{aligned}$$

and so

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \beta$$

The binomial distribution with parameters n, p is well approximated by the Poisson pmf for large n and small p. The following limit puts this statement on a firm foundation: Let $p = \lambda/n$ for some fixed constant $\lambda > 0$. Then

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

To prove this, write the binomial probability as

$$p_X(k) = \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}$$

and verify that, for any *fixed* k, the first factor tends to 1 and the third factor tends to $e^{-\lambda}$ as $n \to \infty$

2 Continuous Random Variables

2.1 Infrastructure

Suppose we want a discrete probability model for choosing a number uniformly at random from some continuous interval – say $[0,1]$ for simplicity. We could adopt a discretization parameter Δ, assume for convenience that $\Delta = 1/N$, let the range of the rv be all multiples of Δ up to 1 (i.e., $\Delta, 2\Delta, 3\Delta, \ldots, 1$), and assign the pmf $P(\hat{X} = i\Delta) = 1/N$. This would give us a model $\hat{X}$ as nearly exact as we wish, by choosing Δ small enough, or equivalently, N large enough.

But mathematically it is usually convenient to work with the continuous limit itself, although as a computational matter, limited-precision confines us to the discrete world. But this transition means that we must move from the class of discrete probability mass functions to a new class of continuous functions called *probability density functions* (pdf's). To introduce pdf's, first define, for the above discretization,

$$F_{\hat{X}}(k\Delta) := \sum_{i \le k} P(\hat{X} \le k\Delta),$$

so that

$$P(\hat{X} = k\Delta) = F_{\hat{X}}(k\Delta) - F_{\hat{X}}((k-1)\Delta)$$

A probability, like $F_{\hat{X}(x)}$, that an rv takes on a value at most some given value x is called the *cumulative distribution function* of the rv (it is the probability mass accumulated by the distribution up to and including the given value x; we will review its properties a little later.

The pdf corresponds to $\frac{F_{\hat{X}}(k\Delta) - F_{\hat{X}}((k-1)\Delta)}{\Delta}$ and gives the "rate" at which probability mass is assigned to the intervals $((k-1)\Delta, k\Delta]$. If, say, $\hat{X}$ measures length, then the probability density at a point x gives the probability mass per unit length at point x. Informally, as $k \to \infty, \Delta \to 0$, with $k\Delta = x$ held fixed, we have $\hat{X} \to X$ where X is a continuous rv with the limiting pdf

$$\frac{F_{\hat{X}}(x) - F_{\hat{X}}((x-\Delta)}{\Delta} \to \frac{dF_X(x)}{dx} = f_X(x),$$

where $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$ is the cdf of the continuous rv X. Implicitly, the limit is assumed to exist, and the cumulative distribution is assumed to be differentiable. There are important applications in which F_X is not differentiable everywhere; these will be discussed later.

We note that, as a density, $f_X(x)$ is nonnegative, but unbounded, in contrast to probabilities, which are bounded by 1; its fundamental role lies in giving the probability of continuous events A, which will usually consist of one or more intervals.

$$P(X \in A) = \int_A f_X(x)dx$$

The notion of the probability of X taking on a specific value is no longer applicable; the limit of $P(\hat{X} = k\Delta)$ is the probability mass at a point, which is 0, as one expects in a continuous, uncountable sample space. This explains in part why probability laws have to be defined in terms of events (sets) in general. It also explains why we don't have to worry about whether intervals in A have boundary points (are open or closed).

In the usual calculus context for densities when Δ is being taken infinitesimally small, one often sees an informal use of the differential dx rather than Δ; in these terms, the quantity $f_X(x)dx$ is an estimate of the probability that X falls in an interval $[x, x + dx]$. *Thus, the connection between the discrete and continuous theories lies in the analogy between pmf's and differential probabilities $f_X(x)dx$ with sums in the first case becoming integrals in the second.* It is then clear that the normalization condition survives as

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

and a brief argument shows that expectations of functions $g(X)$ survive as

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

2.1.1 Important probability density functions

As can be seen from the opening example of the discrete uniform law, $F(x) - F(x - \Delta) = \Delta$. so the corresponding pdf on the continuous interval [0,1] has the value 1, $0 \le x \le 1$, and 0 otherwise. The cumulative distribution is $F_X(x) = x, \ 0 \le x \le 1$, and it takes the value 0 for $x < 0$ and the value 1 for $x > 1$. The mean and variance are easily computed,

$$\mathbb{E}X = \int_0^1 xf_X(x)dx = \int_0^1 xdx = 1/2$$

$$Var(X) = \int_0^1 x^2dx - (1/2)^2 = 1/12$$

Exercise: Work out the corresponding results when the interval (often referred to as the *support* of the distribution) is changed to $[a, b]$

The continuous analog of the geometric distribution provides a more interesting limit. Let Z have the geometric law

$$p_Z(k) = p(1-p)^{k-1}, \ k \ge 1,$$

and suppose the lifetime $\hat{T}$ of some device is modeled as $Z\Delta$ for some basic time unit Δ such as hours or days. We would like a more realistic model that allows for lifetimes, T, of any duration, i.e., any positive real, with the same expected value. Let $\mathbb{E}Z\Delta = \Delta/p$ be a constant $1/\lambda$, so $p = \lambda\Delta$. The probability density is then estimated by

$$\frac{F_{\hat{T}}(k\Delta) - F_{\hat{T}}((k-1)\Delta)}{\Delta} = \lambda(1 - \lambda\Delta)^{t/\Delta}$$

for k large and Δ small, $t = k\Delta$. Now take the limit $k \to \infty$, $\Delta \to 0$ with $t = k\Delta$ held constant to obtain $\hat{T} \to T$ and the density function for the *exponential distribution*[4]

$$f_T(t) = \lambda e^{-\lambda t},\ t \geq 0$$

It is easy to see that this function integrates to 1 over $[0, \infty)$, and has the cdf $F_X(t) = \int_0^t \lambda e^{-\lambda y} dy = 1 - e^{-\lambda t},\ t \geq 0$. For the mean, an integration by parts gives

$$\begin{aligned}
\mathbb{E}X &= \int_0^\infty \lambda t e^{-\lambda t} dt = -\int_0^\infty t \cdot d(e^{-\lambda t}) \\
&= -te^{-\lambda t}|_0^\infty + \int_0^\infty e^{-\lambda t} dt \\
&= 1/\lambda
\end{aligned}$$

and, using the integral for the mean,

$$\begin{aligned}
Var(X) &= \int_0^\infty \lambda t^2 e^{-\lambda t} dt - 1/\lambda^2 \\
&= -\int_0^\infty t^2 \cdot d(e^{-\lambda t}) dt - 1/\lambda^2 \\
&= -t^2 e^{-\lambda t}|_0^\infty + 2\int_0^\infty t e^{\lambda t} dt - 1/\lambda^2 \\
&= 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2
\end{aligned}$$

Exercise: Extend these results to the Laplace, or two-sided exponential, distribution

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda |x|},\ -\infty < x < \infty.$$

Another interesting, but more involved limit starts with the Poisson distribution given in (1) and yields the *normal* or *Gaussian* law. We omit the details[5] and give the limiting density function:

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

where, as computations will show, μ and σ^2 are the mean and variance of the distribution. Later, we will see that the normal distribution also plays a vital role in limit theorems for sums of rv's.

[4] To see this, the natural logarithm and a series expansion give

$$\frac{\lambda t}{\Delta} \ln(1 - \lambda\Delta) = -\lambda t[1 + \lambda\Delta + (\lambda\Delta)^2 + \cdots] \to -\lambda t$$

[5] The limit can be established directly using Stirling's formula, but it is actually a consequence of the central limit theorem that we will cover later in the course

To show that the normal density satisfies the normalization axiom, we first change the variable in $\int_{-\infty}^{\infty} f_X(x)dx$ to $y = \frac{x-\mu}{\sigma}$, and then observe that it is enough to show that

$$\begin{aligned}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}dy\right)^2 &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2}du\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-v^2/2}dv \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(u^2+v^2)/2}dudv = 1\end{aligned}$$

For this, we change to polar coordinates, where you will recall that the differential element of area, $rdrd\theta$, rotates a differential radial element $(r, r+dr)$ through a differential angle $(\theta, \theta+d\theta)$, and where $u = r\cos\theta,\ v = r\sin\theta$ and hence $r^2 = u^2+v^2$. Then the desired result follows from

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(u^2+v^2)/2}dudv &= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} re^{-r^2/2}drd\theta \\ &= -\int_0^{\infty} d(e^{-r^2/2}) = 1\end{aligned}$$

To verify that $\mathbb{E}X = \mu$, it is enough to show that, for the rv $Y = \frac{X-\mu}{\sigma}$,

$$\mathbb{E}Y = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ye^{-y^2/2}dy = -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} d(e^{-y^2/2}) = 0$$

Similarly, one finds that

$$\begin{aligned}\mathbb{E}Y^2 = Var(Y) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2e^{-y^2/2}dy = -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} yd(e^{-y^2/2}) \\ &= -\left.\frac{ye^{-y^2/2}}{\sqrt{2\pi}}\right|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}dy = 1\end{aligned}$$

from which it follows that $Var(X) = \sigma^2$. The zero-mean, unit-variance normal rv Y is called a *standard normal.*

2.1.2 Cumulative distribution functions and tails

Recall that we introduced the tails of distributions for nonnegative rv's; these are just 1 minus the cdf:

$$P(X > x) = 1 - F_X(x)$$

For the continuous case, we have an analogous formula for the mean in terms of the tail of a nonnegative rv:

$$\mathbb{E}X = \int_0^{\infty}[1 - F_X(x)]dx$$

which can be verified by an integration by parts.
Exercise: Extend this result to cover cases where the rv also has a tail $P(X < -x)$. One obtains

$$\mathbb{E}X = \int_0^\infty P(X > x)dx - \int_0^\infty P(X < -x)dx$$

A quick review of the elementary properties of the cdf $F_X(x)$:

1. Densities and pmf's are nonnegative so F_X must be nondecreasing. The cdf is flat (remains constant) where the pmf is 0, in the discrete case, and correspondingly only over intervals where the density has value 0 (i.e., where the value of an rv can not possibly lie) in the continuous case.[6]

2. The normalization condition requires that $F_X(x) \to 1$ as $x \to \infty$ and $F_X(x) \to 0$ as $x \to -\infty$ For distributions with finite support (finite sample space in the discrete case, and finite collection of finite intervals in the continuous case), there will be a maximum point above which the cdf will be 1 and a minimum point below which it will be 0.

3. In the discrete case, a pmf can be computed from its cdf by $p_X(k) = F_X(k) - F_X(k-1)$ and in the continuous case with differentiable cdf's the pdf can be computed from $f_X(x) = \frac{d}{dx}F_X(x)$. Note that the density can also be computed from the tail as $-\frac{d}{dx}P(X > x) = -\frac{d}{dx}[1 - F_X(x)] = f_X(x)$.

The normal distribution has no closed form cdf; the standard normal cdf is commonly denoted by the symbol

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

with the derivative (density) sometimes denoted by $\varphi(\cdot)$. Because of the number and importance of its applications, tables of the standard normal cdf are ubiquitous; if $P(X \leq x)$ for a Gaussian rv X with mean μ and variance σ^2 is desired, then one simply consults a table for the standard normal with the argument $\frac{x-\mu}{\sigma}$ to find

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

Working with cdf's can be much easier than working with pmf's or pdf's for certain problems. For example, the maximum of rv's $X_1, \ldots, X_n$ has a value at most x if and only if each of the X_i's has a value at most x, so if the rv's are independent, the cdf for the maximum is just the product of the cdf's for the individual rv's. If desired, the pmf or pdf for the maximum can then be found

[6] For general rv's to be covered later, cdf's can make discontinuous jumps in cases where discrete probability masses are combined with the continuous case. This generality is disallowed for the moment.

from property 3 above.

Exercises:

1. How would you, in analogy with the technique above, go about finding the minimum of independent rv's?
2. Show that Gaussian distributions are preserved under linear transformations, i.e., if X is Gaussian, then so is $Y = aX + b$.

Example: (*Generating random samples.*) In simulating systems in a random environment, one must be able to generate samples of an rv from a given, general distribution. Mathematical software and programming languages typically provide routines that generate numbers very nearly uniformly distributed on some interval, which we may take as $[0, 1]$ for simplicity. The problem we address here is converting such a sample to one from any given, monotone increasing cumulative distribution function.

If the cdf $F_X(x)$ is strictly increasing then it has an inverse $F_X^{-1}(x)$ that is unique for each x in the range of X. Define the rv $\hat{X} := F_X^{-1}(U)$, where U has the uniform distribution on $[0, 1]$. If we can show that $\hat{X}$ has the distribution $F_X(x)$, then we will have exhibited a method for converting a sample of U to a sample of X, as desired. But since $P(U \leq u) = u,\ 0 \leq u \leq 1$,

$$P(\hat{X} \leq x) = P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$$

and we are done. ■

Analysis in terms of distribution functions is also natural in developing a formula for the densities of monotone functions. Suppose $g(\cdot)$ is monotonically increasing, and hence has an inverse $g^{-1}(\cdot)$. Then, if we let $Y = g(X)$,

$$F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

whereupon differentiation gives

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

If $g(\cdot)$ is monotone decreasing, it again has an inverse, but now $g(X) \leq y$ only if $X > g^{-1}(y)$, so

$$F_Y(y) = P(g(X) \leq y) = P(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

and hence, by differentiation,

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

But the derivative of $g^{-1}(\cdot)$ is negative so the two cases can be assembled as

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

for any monotone strictly increasing or strictly decreasing function $Y = g(X)$.

2.1.3 Conditioning

Conditional pdf's follow the earlier development with no conceptual change; indeed, it is but an exercise to formulate the notion in the continuous case with differentiable cdf's. If X is a continuous rv with density $f_X(x)$, then the conditional density of X given the event $X \in A$, with $P(A) > 0$, is

$$\begin{aligned} f_{X|A}(x) &= \frac{f_X(x)}{P(X \in A)}, \ x \in A \\ &= 0, \text{ otherwise} \end{aligned}$$

which clearly integrates to 1 over the set A and hence is a legitimate probability density defined on A with expectations of functions $g(\cdot)$ written as

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

In the original probability space, the probability of the event $X \in B$ conditioned on the event $X \in A$ is simply $\int_B f_{X|A}(x)dx$. The region of integration can in fact be restricted to $A \cap B$, since the density is 0 outside A.

In the continuous case, the law of total probability is

$$f_X(x) = \sum_{i=1}^{n} f_{X|A_i}(x) P(A_i)$$

where the A_i form a partition of the sample space with $P(A_i) > 0$ for all i, $1 \le i \le n$. The corresponding law of total expectation generalized to functions $g(\cdot)$ is

$$\mathbb{E}g(X) = \sum_{i=1}^{n} \mathbb{E}[X|A_i] P(A_i)$$

Examples.

1. Let X be uniform on $[0,1]$ and consider the conditional density of X given the event A that it falls in some subinterval, say $[0,a]$, $0 < a < 1$. Then by the formula given for conditional probabilities, $f_{X|A}(x) = 1/a$, $0 \le x \le a$ and is 0 otherwise. We note that this new density is also uniform, but on a smaller interval. This recursive property of the uniform distribution can be critical in applications and extends easily to general conditioning intervals or collections of intervals.

2. Let X have the exponential distribution with parameter μ and consider the conditioning event A to be $X > a$. The tail $1 - F_X(a) = e^{-\mu a}$ gives the probability of A, and so

$$f_{X|A}(x) = \frac{\mu e^{-\mu x}}{e^{-\mu a}} = \mu e^{-\mu(x-a)}, \ x > a$$

and hence, conditioned on the event $X > a$, the rv $Y = X - a$ has the density

$$f_Y(y) = \mu e^{-\mu y},\ y \geq 0$$

which is the same distribution as X. Now if X models a waiting time or lifetime, so that Y is the time remaining in X, given that X is greater than a, we discover the remarkable fact that the time remaining is not a function of a and has the same distribution as X. This may seem unexpected, but in fact it is exactly what we do expect when we consider that the exponential distribution is the continuous analog of the geometric distribution, and when we recall the connection the geometric distribution has with sequences of Bernoulli trials. To put this in more practical terms, suppose some device has an exponential lifetime with parameter μ. If the device is observed at some time and found to be still functioning, then its remaining lifetime has exactly the same distribution as the one it started out with. It is obvious why this property of age independence is called the *memoryless* property; in the continuous world, it is limited to the exponential distribution, but no attempt to prove this uniqueness result is made here. This property is the seed for a stupendous literature on the modeling and analysis of systems. We will apply it to the Poisson process later in the course.

■

2.1.4 Joint distributions

Joint distributions for multiple continuous rv's are handled just as in the discrete case, with once again the major differences being the replacement of pmf's by pdf's and sums by integrals. By a natural extension of earlier arguments, $f_{X,Y}(x,y)dxdy$ estimates the joint probability $P(x \leq X \leq x + dx, y \leq Y \leq y + dy)$. Joint, conditional, and marginal probabilities are related by

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

and

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy$$

with $f_{X|Y}(x|y)$ defined only for those y such that $f_Y(y) > 0$.

Independence of rv's X and Y implies, and is implied by, the same relation for densities as for pmf's, i.e., $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and hence

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

Two dimensional densities can be found, as for the one-dimensional case, by differentiation of cdf's

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

where

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(_{X,Y}(s,t),$$

and the continuous form of the Bayes rule has the obvious form in terms of densities

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|t)f_X(t)dt}$$

which gives us the desired conditional densities given Y in terms of the conditional densities given X, which, in the context of the inference problem, we already know.

Examples.

1. We consider the recursive property of the uniform distribution in the setting of two independent uniform random draws X and Y from the interval [0,1]. Their joint density is $f_{X,Y}(x,y) = 1$ over the unit square. Let $Y > X$, so we confine ourselves to the triangle above the diagonal $x = y$ in the unit square. The conditional density is then

   $$f_{X,Y|Y>X}(x,y|Y>X) = \frac{f_{X,Y}(x,y)}{P(Y>X)} = \frac{1}{1/2} = 2$$

   The maximum of X, Y has the cumulative distribution $F_X(x)F_Y(x) = x^2$, as noted before[7] and hence the density $2x$. The conditional density of the position of the minimum X given the maximum $Y = y$ is therefore $2/2y = 1/y,\ 0 \leq x \leq y$, the uniform distribution, once again.

   Pursue this setting further and consider the notion of random interval that one finds most often in the applied probability literature. The interval between the two points chosen independently and uniformly at random from the interval $[0,1]$ defines a *random interval*, i.e., a random subinterval of [0,1]. To find the distribution governing its length L, we make use of the density function for the maximum $Z = \max(X,Y)$, $f_Z(z) = 2z,\ 0 \leq z \leq 1$. Write

   $$P(L > x) = \int_0^1 P(L > x|Z = z)f_Z(z)dz$$

   Let M denote the minimum of the randomly chosen points and observe that, given the maximum $Z = z$, the conditional pdf of M is simply the uniform density $1/z$ on $[0,z]$ by the earlier argument. Then

   $$\begin{aligned} P(L > x|Z = z) &= P(M \leq z - x|Z = z) = \int_0^{z-x} \frac{dx}{z} = \frac{z-x}{z},\ z > x, \\ &= 0,\ z \leq x \end{aligned}$$

[7] Recall that $\max(X,Y) \leq x$ if and only if both $X \leq x$, and $Y \leq x$, so $P(\max(X,Y) \leq x) = P(X \leq x)P(Y \leq x) = x^2$

Substituting, one finds

$$P(L > x) = \int_x^1 \frac{z-x}{z} \cdot 2zdz = 1 - 2x + x^2$$

whereupon we arrive at the triangular density

$$f_L(x) = \frac{d}{dx}[1 - P(L > x)] = 2 - 2x,$$

With the mean $\mathbb{E}L = 1/3$.

A much simpler route to the same result (but one that does not illustrate joint and conditional probabilities as desired here) identifies the region of the unit square where $L > x$, observes that the total area is $P(L > x) = (1-x)^2$, and differentiates to obtain

$$f_L(x) = -\frac{d}{dx}P(L > x) = 2(1-x)$$

Exercise. Find the probability density of the area of a *random rectangle* defined by two independent random intervals, one giving the vertical sides and the other giving the horizontal sides.

2. The problem is to find the probability of the event T that 3 independent uniform random draws from $[0,1]$ can be made to form a triangle. A geometric approach identifies the region in 3 space where every point defines a triangle and then uses volume arguments to determine $P(T)$. Another approach is to condition on the maximum of the three draws and find the probability that the remaining two sum to greater than the maximum. Suppose the maximum has the value z. The key observation is that the remaining two values must be distributed as independent samples from a uniform distribution on $[0,z]$. Normalize these, i.e., divide them by z, making them samples of independent rv's, say X, Y, uniformly distributed on $[0,1]$. Then

$$\begin{aligned} P(T) &= P(X+Y>1) \\ &= \int_0^1 P(Y > 1-x|X=x)f_X(x)dx \\ &= \int_0^1 [1-(1-x)]dx = 1/2 \end{aligned}$$

As can be seen the value of the maximum, as a conditioning event, operates only as a scale factor and does not enter into the analysis once the problem in terms of the other two rv's has been rescaled by the maximum.

We have already seen that, to find densities, it is often most convenient to compute cdf's (or tails) and then get densities by differentiation. This was illustrated earlier for functions of two rv's by our calculation of the density for the length of a random interval. The regions of integration in higher dimensions are often tricky to identify; where it changes shape is often a key, early observation.

EXAMPLE Compute the density of the difference $Z = X - Y$, where X, Y are independent exponentially distributed rv's with mean $1/\lambda$. This is example 3.28, page 189 of B&T, but we provide here an alternative solution. In calculating the cdf or the tail of Z, the shape of the region of integration changes at $z = 0$ (see Figure 3.25, B&T).

Define $z_* := \max(0, z)$, and note from the figure that, after routine calculations,

$$\begin{aligned} P(X - Y > z) &= \int_{z_*}^{\infty} \int_0^{x-z} f_{X,Y}(x, y) dx dy \\ &= \int_{z_*}^{\infty} \lambda e^{-\lambda x} dx \int_0^{x-z} \lambda e^{-\lambda y} dy \\ &= e^{-\lambda z_*}[1 - \frac{1}{2} e^{\lambda(z - z_*)}] \end{aligned}$$

Substitute for z_* to obtain

$$P(Z > z) = \begin{cases} \frac{1}{2} e^{-\lambda z} & z_* = z \\ 1 - \frac{1}{2} e^{\lambda z} & z_* = 0 \end{cases}$$

whereupon differentiation of $-P(Z > z)$ gives the Laplace density

$$f_Z(z) = \frac{\lambda}{2} e^{-\lambda |z|}, \quad -\infty < z < \infty$$

■

General rv's have both discrete and continuous components; representations are less elegant, but rarely do difficulties arise in applications. Queueing theory underlies the following illustration.

EXAMPLE: Consider checking in at an airline counter, checking out at a market, arriving at a car wash, messages arriving at a communications buffer, just to mention a few of the huge number of such settings. These applications are characterized by customers either beginning service immediately on arrival with some positive probability p, or waiting in a queue for a time modeled as a continuous rv, say W with pdf $f_W(w)$, until a server becomes free. Finding p and the density f_W as a function of the parameters of arrival and service processes is the province of queueing theory; for the moment assume they are given. Describing W just in terms of densities is unsatisfactory, as there is a positive probability (mass) concentrated at the value 0, which must be represented by an infinite spike.[8] Matters improve if we focus on cdf's, since point (i.e., discrete) probability masses now show up as discontinuities, or jumps, with sizes equal to the point probability masses. In the present case, the cdf is $F_W(w) = p +$

[8] The impulse functions of spectral analysis are useful here, but we can do without them).

$(1-p)\int_0^w f_W(w)$ with a discontinuity at the origin. Formulation issues aside[9], moments are easily calculated in situations such as this. Here, we have, by the law of total expectation,

$$\begin{aligned}\mathbb{E}W^k &= p\mathbb{E}[W^k|W=0]+(1-p)\mathbb{E}[W^k|W>0] \\ &= (1-p)\int_0^\infty w^k f_W(w)dw\end{aligned}$$

Mixed distributions are also common in applications. Suppose there is a discrete rv Z on the integers $1,\dots,n$, with the respective probabilities $p_1,\dots,p_n$, and suppose the rv X has the conditional density $f_{X_k}(x)$ of rv X_k given that $Z=k$. Then the pdf for X is

$$f_X(x)=p_1 f_{X_1}(x)+\cdots+p_n f_{X_n}(x)$$

and it is trivial to see that it satisfies the properties of density functions if the $f_{X_i}(\cdot)$ do.
EXAMPLE. A general scenario consists of a service facility with multiple servers having varying rates of service. If there are n servers, the overall expected service time, using the notation above, can be written

$$\begin{aligned}\mathbb{E}S_n &= \sum_{k=1}^n \mathbb{E}[X_k|Z=k]P(Z=k) \\ &= \sum_{k=1}^n p_k\int_0^\infty x f_{X_k}(x)dx\end{aligned}$$

2.1.5 Moment generating functions

A standard *transform* of a sequence $a_0,a_1,\dots$ is

$$M(s)=\sum_{k\geq 0}a_k e^{ks}$$

also called a *z-transform* if e^s is replaced by z. If the a_k's form a probability distribution for an rv X, then the above expression is simply

$$M_X(s)=\mathbb{E}e^{sX}$$

In this context, which will be the one to apply hereafter, it is often called a *moment generating function* (mgf) for the simple reason that the n-th derivative of $M(s)$ with respect to s evaluated at $s=0$ is the n-th moment.[10] This is easy

[9] One needs a more general version of integration, e.g., the Laplace-Stieltjes integral, which can accommodate distributions with both discrete and continuous components.

[10] The Laplace transform of a continuous distribution, viz., $\mathbb{E}e^{-sX}$, is also commonly found in the literature; moments are found as here but with an alternation in sign. Use of the z-transform (with e^s replaced by z) for discrete rv's complicates the moment calculations somewhat; derivatives evaluated at the origin give the so-called *factorial moments.*

to see, assuming that expectation and differentiation are interchangeable, since then

$$\frac{d^n}{ds^n} M_X(s) = \mathbb{E}[X^n e^{sX}]$$

Note that $M_X(0) = 1$.
EXAMPLES.

1. A uniform law on $\{1, 2, \ldots, n\}$ has the mgf

$$M_X(s) = \frac{1}{n}e^s + \frac{1}{n}e^{2s} + \frac{1}{n}e^{3s} + \ldots + \frac{1}{n}e^{ns} = \frac{e^s}{n}\frac{1-e^{ns}}{1-e^s}$$

2. A geometric law on the positive integers with parameter p has the mgf

$$M_X(s) = \sum_{k\geq 1} p(1-p)^{k-1}e^{sk} = \frac{pe^s}{1-(1-p)e^s}$$

The mgf for the corresponding geometric law on the *nonnegative* integers is also easily calculated and given by

$$M_X(s) = \sum_{k\geq 0} p(1-p)^{k}e^{sk} = \frac{p}{1-(1-p)e^s}$$

3. The mgf of a Poisson law with parameter λ is

$$\begin{aligned} M_X(s) &= \sum_{k\geq 0} \frac{\lambda^k}{k!}e^{-\lambda}e^{ks} \\ &= e^{-\lambda(1-e^s)} \sum_{k\geq 0} \frac{(\lambda e^s)^k}{k!}e^{-\lambda e^s} \\ &= e^{-\lambda(1-e^s)} \end{aligned}$$

The basic definitions also apply to the mgf's of continuous rv's

$$M_X(s) := \int_{-\infty}^{\infty} e^{sx} f_X(x)dx$$

with moments computed as before.
EXAMPLES

1. The exponential distribution with parameter μ has the mgf

$$M_X(s) = \int_0^{\infty} \mu e^{-\mu x} e^{sx} dx = \frac{\mu}{\mu - s}$$

2. The mgf of the standard normal rv is

$$\begin{aligned}M_X(s) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}e^{sx}dx \\ &= \frac{1}{\sqrt{2\pi}}e^{s^2/2}\int_{-\infty}^{\infty} e^{-x^2/2+sx-s^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}e^{s^2/2}\int_{-\infty}^{\infty} e^{-(x-s)^2/2}dx \\ &= e^{s^2/2}\end{aligned}$$

Exercises.

1. Verify that
$$P(X=0) = \lim_{s\to-\infty} M_X(s)$$
and apply this result to a couple of the transforms we have computed so far.

2. Show that, if $X = aY + b$, then $M_X(s) = e^{sb}M_Y(sa)$ and apply this relation to the mgf for the standard normal to find that the mgf for a normal rv with mean μ and variance σ^2 is given by
$$M_X(s) = e^{s\mu}M_Y(s\sigma) = e^{(\sigma^2 s^2/2)+\mu s}$$

3. Let Y be uniformly distributed on $[0,1]$ and suppose that, given $Y = p$, the conditional distribution of X is binomial with parameters n and p. Use generating functions to show that X has a uniform law on the integers $0, 1, \ldots, n$.

Owing to the additional background one needs in classical analysis, our treatment of mgf's will be incomplete in the sense that systematic inversion formulas will not be covered. Suffice it to say here that, so long as an mgf remains finite in a neighborhood of the origin, which is a valid assumption for the distributions covered in these notes, it uniquely specifies a corresponding distribution. On the other hand, our chief interest is in a relatively small collection of distributions in this course, and for any of these we can usually infer the distribution from the shape of the mgf.

Joint mgf's are also useful and are defined as one would expect. The joint mgf for the n rv's $X_1, \ldots, X_2$ is

$$M(s_1, \ldots, s_n) = \mathbb{E}e^{s_1X_1+\cdots+s_nX_n}$$

with the i-th marginal mgf obtainable by setting all but s_i to 0. Again, it can be proved under appropriate conditions that a joint mgf corresponds to a unique joint distribution. The expression for the mgf can be put in the form of an expectation of a product, so if the X_i are independent then this expectation is equal to the product of expectations

$$M(s_1, \ldots, s_n) = M_{X_1}(s_1) \cdots M_{X_n}(s_n)$$

EXERCISE. Let X and Y be independent normal rv's with the same mean μ and variance σ^2. Work in the transform domain to show that $X-Y$ and $X+Y$ are independent.

2.1.6 Sums of random variables

We have been exposed so far to sums of rv's on at least two occasions: The number of successes in a sequence of Bernoulli trials was represented as a sum of Bernoulli rv's, and the waiting time until the k-th success was given by a sum of geometric rv's. In each case, the rv's being summed were i.i.d., i.e., independent and identical in distribution. The sum S of two independent rv's X and Y has a density that can be computed by the convolution of the densities for X and Y. In particular, a simple conditioning argument yields, in the discrete case,

$$p_S(s) = P(X+Y=s) = \sum_x p_X(x) p_Y(s-x)$$

and in the continuous case

$$f_S(s) = \int_{-\infty}^{\infty} f_X(x) f_Y(s-x) dx$$

These operations, both in the discrete and continuous domains, are called *convolutions*. They can be annoyingly tricky to evaluate, but a routine, "graphical" approach can be proposed as a guide: reflect f_Y about the origin and translate it to obtain $f_Y(s-x)$, superimpose and multiply $f_X(x)$ to obtain the function $f_Y(s-x)f_X(x)$, and then integrate to obtain the result. Facility with convolutions comes with practice.

Moment generating functions greatly simplify the analysis of sums, as mgf's of sums become products of individual mgf's. As an example whose importance can not be overestimated, let $X_1, \ldots, X_n$ be independent. Then

$$M_S(s) = \mathbb{E}e^{s(X_1+\cdots+X_n)} = \Pi_{k=1}^n M_{X_k}(s)$$

so if the X_k are identically distributed as well, we get $M_S(s) = M_X^n(s)$ where X has the distribution common to the X_k.
EXAMPLES

1. Let the X_k, $1 \leq k \leq n$, be i.i.d. Bernoulli rv's with parameter p. The mgf of such rv's is simply $1-p+pe^s$ so the mgf of the sum is

   $$M_S(s) = (1-p+pe^s)^n$$

   which we know must be the mgf of the binomial distribution with parameters p and n.

2. Suppose the X_k are independent Poisson rv's with parameters λ_k. Then by our earlier calculation of the mgf for the Poisson law, the mgf of the sum is
$$M_S(s) = \Pi_{k=1}^n e^{\lambda_k(e^s-1)} = e^{\lambda_S(e^s-1)}$$
where $\lambda_S = \lambda_1 + \cdots + \lambda_n$, which shows that *the sum of independent Poisson rv's is Poisson distributed with a parameter equal to the sum of the individual parameters.* This will have strong implications for Poisson processes that we discuss later.

Exercises.

1. Compute the mgf of the binomial distribution directly.
2. Let $S = Z_1^2 + \cdots + Z_n^2$ be the so-called *chi-squared rv with n degrees of freedom*, where the Z_k are independent standard normal rv's. Verify that $M_S(s) = (1-2s)^{-n/2}$.

2.1.7 From conditional to unconditional moments

For two rv's X and Y, the conditional expectation of X given $Y = y$ $\mathbb{E}[X|Y = y]$ is a function of the real number y, which we can make into an identical function of the conditioning rv Y simply by replacing y by Y; but in so doing, the conditional expectation, now denoted by $\mathbb{E}[X|Y]$ has itself become a rv, in particular a function of the rv Y. We can then consider the expectation of this rv. However, this leads us quite simply to

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

in the continuous case. With the analogous result that holds in the discrete case, we have the *Law of Iterated Expectation.*

So far, the novelty is mainly notational, for we have been using the right-hand side of the above equation with no essential need for more compact formulas. On the other hand, getting the unconditional variance $Var(X)$ from the conditional variance is, because of nonlinearity, not so simple. As above, we can again define the rv $Var(X|Y)$ from

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

If we add $(\mathbb{E}[X|Y)^2$ (the square of an rv) to $Var(X|Y)$ we obtain the conditional second moment given Y: $\mathbb{E}[X^2|Y]$. But then, to get $Var(X)$, we subtract the square of the expectation of the conditional mean and take another expectation. Write this out and simplify as follows:

$$\begin{aligned} Var(X) &= \mathbb{E}[\mathbb{E}[X^2|Y]] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\ &= \mathbb{E}[Var(X|Y) + (\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\ &= \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y]) \end{aligned}$$

This last relation is called the Law of Total Variance or simply the total variance formula, and in its expression we see a concrete, useful illustration of the expectation $\mathbb{E}[X|Y]$ as an rv.

An important example is provided by *randomized* sums of i.i.d. rv's, i.e., the case where the number, N, of variables summed is also random. Let

$$Y = X_1 + \cdots + X_N$$

where the X_i's are i.i.d. with mean and variance μ and σ^2, and where N is a rv independent of the X_i's. Clearly, we have $\mathbb{E}[Y|N] = N\mu$ and so, by taking the expected value

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

Further, by the independence of the X_i's, $Var(Y|N) = N\sigma^2$ and so, by the total variance formula,

$$\begin{aligned} Var(Y) &= \mathbb{E}[Var(Y|N)] + Var(\mathbb{E}[Y|N]) \\ &= \mathbb{E}[N\sigma^2] + Var(N\mu) = \mathbb{E}N\sigma^2 + Var(N)\mu^2 \end{aligned}$$

Example. The sum, Y, of a geometrically distributed number, N, of independent, exponentially distributed rv's, X_i, gives an example that arises in many situations. Let p and λ be the parameters of the two respective distributions,and observe that

$$\mathbb{E}[Y] = \mathbb{E}N\mathbb{E}X = \frac{1}{p} \cdot \frac{1}{\lambda}$$

and

$$Var(Y) = \mathbb{E}N \cdot Var(X) + (\mathbb{E}X)^2 Var(N) = \frac{1}{p} \cdot \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{\lambda^2 p^2}$$

■

In terms of mgf's of randomized sums, we have, with the earlier notation,

$$M_Y(s) = \mathbb{E}e^{sY} = \mathbb{E}[\mathbb{E}[e^{sY}|N]] = \mathbb{E}[(M_X(s))^N] = \sum_{n\geq 0}(M_X(s))^n p_N(n),$$

so replacing e^s in the defining formula for $M_N(s)$ by $M_X(s)$ gives the mgf for the randomized sum.

Example (continued). In the preceding example,

$$M_X(s) = \frac{\lambda}{\lambda - s}, \ M_N(s) = \frac{pe^s}{1-(1-p)e^s}.$$

and so, after routine calculation,

$$M_Y(s) = \frac{pM_X(s)}{1-(1-p)M_X(s)} = \frac{p\lambda}{p\lambda - s}$$

Thus, while a fixed sum of i.i.d. exponentials is not exponential, a geometrically distributed number of such exponentials is, and its parameter is the product of the parameters of the geometric and exponential distributions.

2.1.8 Covariance, Correlation

We have already been exposed to calculations of pdf's and moments of products of rv's, say X and Y, in the context of derived distributions. But the topic is of fundamental interest as a means of measuring the interaction or relationship between two rv's. The basic measure is the *covariance* which is the expected value of the product of the zero-mean, or centered versions of the rv's:

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

or, as is easily verified,

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

If this has the value 0, which will clearly be the case if X and Y are independent, then X and Y are said to be *uncorrelated.* While the concepts of independence and correlation are intimately related in practice, examples can be contrived to show that a correlation of 0 does not imply independence. One such example is a uniform pmf applied to the sample points (1,0), (0,1), (-1,0), (0,-1) for rv's X and Y. It is easy to see by inspection that $\mathbb{E}XY = \mathbb{E}X = \mathbb{E}Y = 0$ and hence $Cov(X, Y) = 0$. But X and Y are not independent: when one has a nonzero value the other has value 0.

The covariance arises naturally in computing the variance of a sum $X + Y$ of rv's. One obtains, by expanding the square of the sum and taking expectations,

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Commonly, in studying the relationship between two rv's, a further normalization (beyond centering) is introduced, one that we have already encountered. We normalize the rv's by their standard deviations, which then gives rv's with unit variance as well as 0 means. In so doing, we define the *correlation coefficient* of X and Y, or equivalently, the covariance of the normalized rv's $\hat{X} = (X - \mathbb{E}X)/\sigma_X$ and $\hat{Y} = (Y - \mathbb{E}Y)/\sigma_Y$,

$$\rho(X, Y) = Cov(\hat{X}, \hat{Y}) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

It is easy to see that, with the new normalization, ρ is now limited to values in the interval $[-1, 1]$. A proof starts with the observation that

$$0 \le Var(\hat{X} + \hat{Y}) = Var(\hat{X}) + Var(\hat{Y}) + 2Cov(\hat{X}, \hat{Y}) = 2[1 + \rho(X, Y)]$$

which implies that $\rho \ge -1$. Similar manipulations with the inequality $0 \le Var(\hat{X} + \hat{Y})$ prove that $\rho \le 1$.

The correlation coefficient measures the extent to which two rv's "track" one another, or somewhat more precisely, the extent to which they are linearly related. For, if X and Y are linearly related, e.g., $X = aY + b$, then one finds that ρ is $+1$ or -1 according as $a > 0$ or $a < 0$. We leave a proof of this as an exercise.

Scatterplots of samples of two rv's are standard graphics for suggesting the presence of correlation. For example, suppose we have two measures, each say integers from 0 and 100, ranking car driving skill of people and the condition of the cars they drive. For some sample of drivers, mark a point (place a dot) at the x, y coordinates (driver skill, car condition) for each driver. In the "cloud" of points obtained, one will surely see some positive correlation in the sense that the points plotted are approximated well by a straight line; points will deviate somewhat from a linear fit, but not by very much. In these cases, $|\rho|$ will be relatively close to 1. Lack of correlation is perceived in such plots as an apparently random scattering of points.

3 Limit Theorems

Let $X_1, X_2, \ldots$ be independent samples from a given distribution $F_X(x)$ (in the same sense that throws of a die give independent samples from a uniform distribution on $\{1, 2, 3, 4, 5, 6\}$). Suppose F_X has both a mean and variance, and denote them by μ and σ^2. The laws of large numbers both imply that, for any given $\epsilon > 0$, the probability that the sample mean $(X_1 + \cdots + X_n)/n$ deviates from μ by more than ϵ tends to 0. That is,

$$P\left(\frac{X_1 + \cdots + X_n}{n} - \mu > \epsilon\right) \to 0$$

as $n \to \infty$. Very roughly, as the number of samples grows the weak law, as above, allows for occasional large deviations, but the stronger version does not. The strong law states that, as $n \to \infty$, $(X_1 + \cdots + X_n)/n \to \mu$ with probability one.[11] We give a fundamental application of the strong law by defining, for a given sequence of independent trials of some experiment, the indicator rv's X_i which take the value 1 if a given event A occurs and the value 0 otherwise. By the strong law of large numbers $(X_1 + \cdots + X_n)/n \to \mathbb{E}X = P(A)$, and hence the limiting proportion of time that event A occurs is just $P(A)$, which confirms our initial motivation for the formal concept of probability.

The weak law is easy to prove and relies on two inequalities that are interesting and useful in their own right. The first is Markov's inequality:

If X is a nonnegative rv, then for any a, $P(X \geq a) \leq \mathbb{E}X/a$.

To prove this assertion, define the indicator function $I_A(X) = 1$, if $A = \{X \geq a\}$ holds, and $= 0$ otherwise, and note that $aI_A(X) \leq X$, so $\mathbb{E}I_A(X) \leq \mathbb{E}X/a$. But since $\mathbb{E}I_A(X) = P(X \geq a)$, we have our desired result.

We can be more precise if X has a known variance σ^2, as shown by Chebyshev's inequality:

[11] Roughly, for given ϵ, this means that in an infinite set of sufficiently long sample sequences, there will be at most a finite number of exceptions to the assertion that $(X_1 + \cdots + X_n/n) - \mu \leq \epsilon$.

Let X be an rv with mean μ and variance σ^2. Then for any $x > 0$, the probability of deviating from the mean by at least x is bounded by

$$P(|X - \mu| \geq x) \leq \sigma^2/x^2.$$

To prove the bound, use the Markov inequality with $a = x^2$ and write, since $(X - \mu)^2$ is nonnegative,

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

These bounds can be quite useful when probabilities are not known precisely but the first and/or second moment is. But because of their generality (weakness of the assumptions), the bounds tend to be rather coarse.

We now prove the weak law given earlier under the assumption that the rv's have the variance σ^2. Since

$$\mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu, \quad Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

then by Chebyshev's inequality

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

and the weak law follows.

In fact, the existence of a second moment is not needed for the laws of large numbers, but a proof of this fact is beyond our scope.

The next limit law, the central limit theorem, gives a much stronger assertion if in fact a variance does exist; in particular, it gives estimates of the probabilities of deviations from the mean. The theorem is stated in terms of the standardized, i.e., centered and normalized-variance, version of $S_n = X_i + \cdots + X_n$, which is given by

$$\hat{S_n} = \frac{(S_n - n\mu)}{\sqrt{n\sigma^2}}$$

which clearly has mean 0 and variance 1. In these terms, the central limit theorem states that, for any fixed x

$$P(\hat{S_n} \leq x) \to \Phi(x)$$

as $n \to \infty$.

Before proving the mgf version of this theorem, we give a couple of examples and make a few comments.

Examples. Recall that, using mgf's, we saw that a sum of independent Poisson rv's was Poisson distributed. It follows from the CLT, therefore, that under an appropriate normalization, the standard normal can be exhibited as a limit of the Poisson distribution. Similarly, we introduced Pascal's distribution as the pmf of the waiting time for the n-th success in a sequence of Bernoulli trials. Since waiting times were sums of independent geometric random variables, it follows that the normal law can be used as a large-n approximation for Pascal's law. ■

In "unnormalized" terms, the CLT says that a normal rv with mean $n\mu$ and variance $n\sigma^2$ is a good approximation of the distribution of S_n for n large enough. It is instructive to test experimentally how big n has to be before the normal gives a good approximation of the sum S_n. You can do this numerically quite easily using Matlab, or by simulations when exact results are infeasible, taking as examples a uniform distribution and an exponential distribution for the X_i's. Most people are surprised at how fast the convergence to the normal law takes place, i.e., how small n can be and still have a good approximation, even when dealing with distributions that are arguably not well served by the CLT. An even better feel can be gotten by graphing the distribution of S_n superimposed on the appropriate normal approximation.

To prove the mgf version of the theorem, we show that the mgf of $\hat{S}_n$ converges to the mgf of the standard normal, $e^{s^2/2}$. For simplicity, but no loss in generality, we assume that the X_i have mean 0 and variance 1 so that $\hat{S}_n = S_n/\sqrt{n}$ has mean 0 and variance n. Expand the mgf for X_i as a power series in s, assuming it's well behaved in a neighborhood of the origin, to get

$$M_X(s) = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \cdots$$

and so, after raising to the n-th power and taking the logarithm, we can write for $M_{\hat{S}_n}(s) = \mathbb{E}e^{sS_n/\sqrt{n}} = M_{S_n}(s/\sqrt{n})$

$$\ln M_{\hat{S}_n}(s) = n\ln(\alpha_0 + \alpha_1\frac{s}{\sqrt{n}} + \alpha_2\frac{s^2}{n} + O(\frac{s^3}{n^{3/2}}))$$

where, in this instance, the notation $O(g(x))$ just refers to a function proportional to $g(x)$, and so grows no faster (and decreases no slower) than $g(x)$. By the fact that $M_{S_n}(0) = 1$, $M'_{S_n}(0) = \mathbb{E}S_n = 0$, and $M''(0) = \mathbb{E}S_n^2 = Var(S_n) = nVar(X)$, we get $\alpha_0 = 1,\ \alpha_1 = 0$, and $\alpha_2 = 1/(2n)$, and hence

$$\ln M_{\hat{S}_n}(s) = \ln(1 + \frac{s^2}{2n} + O(\frac{s^3}{n^{3/2}}))$$

Expanding the logarithm as we did before (see p. 14), we get $\ln M_{\hat{S}_n}(s) \sim s^2/2$ as $n \to \infty$, and hence

$$\lim_{n\to\infty} M_{\hat{S}_n}(s) = e^{s^2/2}$$

which is the mgf for the standard normal, as claimed.

Example. Electra, a circuits person, collects n independent measurements of the current, $X_i,\ 1 \le i \le n$, at the output of some circuit. Electra knows that X_i's have an error given by a variance of 4 milliamps about the mean, but she doesn't know the mean, say d, which she will estimate by taking the empirical mean of the n measurements. How large should n be to be 95% sure that she has the right mean to within ±.5 milliamps?

The sum of the measurements has mean nd and standard deviation $\sqrt{4n}$, so the rv

$$Z_n = \frac{\sum_{i=1}^n X_i - nd}{2\sqrt{n}}$$

is approximately a standard normal. Electra wants a value of n such that

$$P\left(-.5 \le \frac{\sum_{i=1}^n X_i}{n} \le .5\right) \ge .95$$

or equivalently, a value of n such that Z_n falls in the interval $[-.5\sqrt{n}/2, .5\sqrt{n}/2]$ with probability at least .95. Using the symmetry of the normal distribution, Electra must compute an n such that

$$\Phi(\sqrt{n}/4) - \Phi(-\sqrt{n}/4) = 2\Phi(\sqrt{n}/4) - 1 \ge .95$$

or $\Phi(\sqrt{n}/4) \ge .975$. From tables for the standard normal, Electra finds that $\sqrt{n}/4 \ge 1.96$ or $n \ge 62$ will suffice.

Exercise. For large x an excellent approximation to the tail of the standard normal is[12]

$$1 - \Phi(x) = \int_x^\infty \frac{e^{-y^2/2}dy}{\sqrt{2\pi}} \approx \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$$

Use this estimate to answer the following question. 1,359,672 boys and 1,285,088 girls were born in Sweden between 1867 and 1890. Is this assertion consistent with (reasonably probable under) the assumption that both sexes are equally likely?

[12] To prove that the estimate is an upper bound, verify by differentiation (using Leibniz's rule in the case of the integral) that

$$\frac{e^{-x^2/2}}{x\sqrt{2\pi}} = \int_x^\infty e^{-y^2/2}\left\{1 + \frac{1}{y^2}\right\} dy$$

then notice that the integrand is larger than the integrand in the integral defining $1 - \Phi(x)$. To prove a corresponding lower bound proceed similarly to show that

$$\frac{e^{-x^2/2}}{\sqrt{2\pi}}\left\{\frac{1}{x} - \frac{1}{x^3}\right\} = \int_x^\infty \frac{e^{-y^2/2}}{\sqrt{2\pi}}\left\{1 - \frac{3}{y^4}\right\} dy$$

4 Random processes

4.1 Bernoulli and Poisson processes

We have had much to say about sequences of Bernoulli trials. These can be thought of as *Bernoulli processes* in discrete time by the simple device of partitioning time into units or slots, say of duration δ, with successive trials being made in successive slots. With applications in mind, we also change terminology: A success at the n-th trial becomes an 'arrival' occurring at time n (the n-th time slot), and a failure is a null event (nothing happens). By a limit process that we've already illustrated, we can convert the Bernoulli process in discrete time into a Poisson process in continuous time. The properties and statistics of the discrete time process all have their parallels in the continuous process; these parallels are listed below as a means of introducing the Poisson process.

1. The times between arrivals in discrete time are geometric whereas they are exponential in continuous time. Of course, an arrival of the former process requires a time slot for its occurrence, but it becomes an instantaneous event in the continuous process. Just as the independent, geometrically distributed times between successes can be used as a definition of the Bernoulli process, so can the independent, exponentially distributed times between arrivals define the Poisson process.

2. The probability of a Poisson arrival in $[t, t + \Delta t]$ is $\lambda \Delta t + o(\Delta t)$, where λ is a *rate parameter* of the process corresponding to p in the Bernoulli process, and where $o(\Delta t)$ means "of smaller order of magnitude than Δt" (i.e., if $g(\Delta t) = o(\Delta t)$ then $g(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$). This probability is independent of the events in all intervals disjoint from $[t, t+\Delta t]$. Note that this property replicates the Bernoulli trial mechanism at the differential level with an (asymptotically negligible) error term that tends to 0 at a faster rate in the continuous limit than does the size Δt of the differential interval.

3. The number of arrivals in time $n\delta$ has a binomial distribution with mean pn for the Bernoulli process and a Poisson distribution with mean $\lambda t = np$ in the corresponding Poisson process. The latter is obtained in the limit $n \to \infty$, $\delta \to 0$, with $n\delta$ fixed at t and $p \to 0$ with p/δ held fixed at λ. The numbers of arrivals in disjoint intervals are independent. Since we can take the disjoint intervals to be adjacent, this gives a new proof that the sum of Poisson rv's is also a Poisson rv.

   We should be able to get this result for the Poisson process from the preceding property as well, in which we work directly with the continuous process. This goes as follows. Let $N(t)$, often called a *counting* process, denote the number of arrivals in $[0, t]$. Write, for $i \geq 1$,

   $$P(N(t + \Delta t) = i) = P(N(t) = i - 1)\lambda\Delta t + P(N(t) = i)(1 - \lambda\Delta t) + o(\Delta t)$$

and for $i = 0$ just omit the first term on the right-hand side. Now bring $P(N(t) = i)$ over to the left-hand side and divide through by Δt to get, for $i \geq 1$,

$$\frac{P(N(t+\Delta t) = i) - P(N(t) = i)}{\Delta t} = \lambda P(N(t) = i-1) - \lambda P(N(t) = i) + \frac{o(\Delta t)}{\Delta t}$$

Let $p_i(t) := P(N(t) = i)$ and take the limit $\Delta t \to 0$ to get the differential equation

$$p_i'(t) = \lambda p_{i-1}(t) - \lambda p_i(t), i \geq 1,$$

with $p_0'(t) = -\lambda p_0(t)$ which, by inspection, has the solution $p_0(t) = e^{-\lambda t}$. It is easy to see that the differential equation is satisfied by the Poisson pmf

$$p_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}$$

which is what we had set out to verify.[13]

Property 1 is also readily seen to be implied by this property. For, the probability of no arrivals in time t since the last arrival is $e^{-\lambda t}$ according to the Poisson law. But this is just the tail of the interarrival-time distribution, which must then be the exponential.

4. Both processes exhibit the memoryless property: at any time t, the future of the process evolves independently of the past, and in particular, the time already transpired waiting for the next arrival. Formally, we proved earlier that if X denotes the interarrival time in progress and λ is the arrival rate, then, *independently of* t, $P(X - t \leq x | X > t) = 1 - e^{-\lambda x}$, which was the distribution looking forward from the last arrival t time units earlier.

5. In the discrete model we posed the question: Given that a single arrival took place within a given sequence of n time slots, what is the conditional pmf of the slot in which it occurred? The answer was the uniform law $1/n$, and we get the same answer for the analogous question in continuous time: Given the event A that exactly one arrival occurred in an interval of duration τ, what is the conditional density of the time T in that interval when it took place? We have $P(A) = \lambda\tau e^{-\lambda\tau}$, and the probability of an arrival in $[t, t+dt]$ and nowhere else in the interval of duration τ is $\lambda dte^{-\lambda\tau}$, so the density is again given by a uniform density,

$$f_{T|A}(t|A)dt = \frac{\lambda dte^{-\lambda\tau}}{\lambda\tau e^{-\lambda\tau}} = \frac{dt}{\tau}$$

More generally, given that there are exactly k arrivals in the interval of duration τ, their conditional distribution is that of k independent uniform

[13] To solve the differential equation directly, take the z-transform and then solve, by a simple integration, the resulting partial differential equation to get the z-transform of the Poisson distribution.

random draws from the interval. This is easily proved by induction on k using the properties of the Poisson process.

6. The waiting time W for the k-th arrival (success) in the Bernoulli process gave rise to the Pascal distribution (p. 10), and in the Poisson process it gives rise to the *Erlang-k* distribution as follows. The probability that the k-th arrival takes place in $[t, t+dt]$ is λdt (plus lower order terms) times the probability that $k-1$ arrivals took place in $[0, t]$, an independent event, as we have stated above. Thus, the pdf for W is as given in

$$f_W(t)dt = \lambda dt \cdot \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} = \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda t} dt$$

7. We have yet to speak of the *merging* and *splitting* of processes, but as these concepts are entirely analogous in the Bernoulli and Poisson processes, now is a good time to describe them. Suppose that, for each arrival in a Bernoulli process, we toss a biased coin (q is the probability of heads) and if it comes up heads we reject the arrival. The resulting process is still Bernoulli, as is easily seen, but with a new parameter $p(1-q)$; at each slot the probability of an arrival is $p(1-q)$ independently of the past. Similarly, the process of rejected arrivals is Bernoulli, but with a parameter pq.

   We do the same thing with arrivals in a Poisson process, tossing a coin to determine which of two split processes is to receive an arrival in the original process. It is easily proved that the two split processes are independent and Poisson, one with arrival rate $\lambda(1-q)$ and the other with arrival rate λq. (Use the fact that the sum of a geometrically distributed number of exponentials is exponential.)
   Exercise Describe the split processes when we transfer every k-th Poisson arrival to a new process. This gives us two arrival processes, one at rate $\lambda(k-1)/k$ and one at rate λ/k. Are the new processes Poisson processes? Explain. ■

   We merge two independent Bernoulli processes with parameters p and q into a single Bernoulli process with parameter $1-(1-p)(1-q)$ by creating an arrival in a slot of the new process whenever there is an arrival in the same slot of one or both of the merging processes. The merging of independent Poisson processes at rates λ_1, λ_2 is simply their superposition, which, as is easily proved, is a Poisson process at rate $\lambda_1 + \lambda_2$.

4.2 Markov Chains

Recall the rv S_n representing the sum of i.i.d. rv's with values in $\{-1, 1\}$ in the random walk of midterm 2, where S_n was the position of a particle, say, after step n. The sequence $\{S_n,\ n \geq 1\}$ is a widely applicable special case of discrete-time *Markov chains.* Technically, the term 'chain' means nothing more than 'sequence.' The term 'discrete-time' refers to step-by-step evolution as in random walks or Bernoulli trials. The rv S_n is called a *state* in the parlance of MC's and the MC is said to be in state i at time n if $S_n = i$. At each *step* of the MC, it moves from a state S_n to a new state S_{n+1} obtained in the random-walk example simply by adding to the current state the value of the next step drawn from $\{-1, 1\}$. These MC's are *infinite-state* in the sense that there is no bound on their values.

The general case of Markov chains $\{X_n, n \geq 0\}$ of interest here is a sequence of random variables (states) with integer values, in which one-step transitions from any state to any other state are allowed. Each transition $X_n = i \to X_{n+1} = j$ has a *transition probability* $p_{i,j}$, which is independent of n and all $X_j,\ j < n$. The essential simplification of Bernoulli-type processes relative to MC's, is that in the former the steps are i.i.d. rv's independent of the states of the process, and in the latter the steps can depend on the current state. However, we emphasize the above *Markov property: the step made by an MC in any given state X_n depends only on X_n, and not on the history of the process leading up to X_n.* Our initial focus is on *finite* MC's, i.e., those with a finite set $\mathcal{S}$ of m states which we usually take to be the integers $1, \ldots, m$. Generating sequences of states, which we refer to as *paths*, is completely defined by an initial state, which is drawn from a given distribution, and an $m \times m$ *one-step transition matrix* $\mathcal{P} := \{p_{i,j}\}$, where

$$p_{i,j} = P(X_{n+1} = j | X_n = i), \quad i, j \in \mathcal{S}, \quad n \geq 0,$$

independently of all prior state transitions. We allow $p_{i,j} = 0$ in which case transitions are never made from state i to state j, and $p_{i,j} = 1$, in which case state j always follows state i. If $p_{i,i} = 1$, then if the MC ever gets into state i, it stays (is *absorbed*) there. Every row in $\mathcal{P}$ must sum to 1, $\sum_{j \in \mathcal{S}} p_{i,j} = 1$, since from any given state a transition to some state is obligatory, even if it is back to the given state itself.

The transition matrix $\mathcal{P}$ and the paths that can be generated from it are conveniently represented in the form of a directed graph, say G. Draw a node for each state in $\mathcal{P}$ and label it with the state it represents. Then draw a directed edge from node i to node j (i.e., an edge headed by an arrow pointing towards j) if and only if $p_{i,j} > 0$ in $\mathcal{P}$. Following sequences of directed edges in G traces the possible paths in the MC.

Example. Consider the bits received at a digital receiver, one per time unit, with 1's and 0's equally likely. The current state is 1 meaning the last bit transmission was successful and delivered a 0, or 2 meaning that the last bit was transmitted successfully and was a 1, or 3 meaning a bit was delivered in error in a transition starting in state 1, or 4 meaning that a bit was delivered in error in a transition starting in state 2. The probability of a bit being received in error is ϵ, and in that event, after a delay of one time unit, the bit is resent with the system in the same state it was in before the error occurred. Model this process by the transition matrix

$$\begin{bmatrix} \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & \epsilon & 0 \\ \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & 0 & \epsilon \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

■

The conditional probability that the MC moves along the path $i_1, \ldots i_n$ given that it starts in state i_0, is, by the Markov property, simply the product of the associated transition probabilities

$$P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n | X_0 = i_0) = p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},n},$$

consistent with the fact that an MC is completely defined by the distribution from which the initial state is drawn and the transition matrix $\mathcal{P}$. It is of obvious interest to know the distribution of the state of the MC after it has made n steps, and this is available from the n-step transition probabilities $p_{i,j}^{(n)} := P(X_n = j | X_0 = i)$. The following recurrence[14] is easily established for all i, j

$$(*) \qquad p_{i,j}^{(n)} = \sum_{k=1}^{m} p_{i,k}^{(n-1)} p_{k,j}, \quad n > 1$$

with $p_{i,j}^{(1)} \equiv p_{i,j}$.

States can be characterized in various ways determined by when and how often they can be entered and exited. For the moment, let us suppose that all states in $\mathcal{S}$ are *communicating*, i.e., each state can be reached from every other state with positive probability. We do this so that we can quickly come to the fundamental questions of long-term behavior, where n is large in (*). The communicating states are *recurrent* in the sense that they will recur infinitely often and have finite expected times between successive recurrences. We need one other concept.

The state space $\mathcal{S}$, and hence the MC, can be *periodic* in the sense that it moves cyclically from one subset of $\mathcal{S}$ to another in a fixed sequence. Specifically, $\mathcal{S}$ can be partitioned into subsets $A_1, \ldots, A_k$ such that each one-step transition of the MC in $\mathcal{S}$ is, for some ℓ, $1 \leq \ell < k$, a transition from a state in A_ℓ to a state in $A_{\ell+1}$, or a transition from a state

[14] The recurrence specializes the Chapman-Kolmogorov characterization of Markov chains, with $p_{i,j}(m,n) := P(X_n = j | X_m = i)$,

$$p_{i,j}(m,n) = \sum_{k=1}^{m} p_{i,k}(m,r) p_{k,j}(r,n)$$

for any r satisfying $m < r < n$.

in A_k to a state in A_1. If no such partition exists for $\mathcal{S}$, then $\mathcal{S}$ and the MC are *aperiodic.* It is readily proved that $\mathcal{S}$ is aperiodic if there exists some state $s \in \mathcal{S}$ and some step count $n \geq 1$ such that every state in $\mathcal{S}$ can reach s in n steps. In the other direction, one can prove without difficulty that if $\mathcal{S}$ is aperiodic then there exists an n such that in n steps a transition can be made from any state in $\mathcal{S}$ to any other state in $\mathcal{S}$.

We have, in a simplified form, the following fundamental result of Markov chain theory.

If $\mathcal{S}$ is a finite set of communicating states, and is aperiodic, then the n-step transition probabilities converge to a distribution

$$\lim_{n\to\infty} p_{i,j}^{(n)} = \pi_j > 0, \text{ forall } j$$

which, as can be seen, is independent of the initial state i, and is the unique solution to the limit of the recurrence in (*)

$$\pi_j = \sum_{k=1}^{m} p_{k,j}\pi_k$$

with

$$\sum_{k=1}^{m} \pi_k = 1$$

Such MC's (and their states) are said to be *ergodic* and their distributions $\{\pi_j\}$ are variously called *stationary*, *invariant* (under $\mathcal{P}$), or *steady-state* distributions.

In general, not all recurrent states of $\mathcal{S}$ need communicate; rather they can be broken down into a number of *recurrence classes*: If A is a subset of $\mathcal{S}$ such that (a) all states in A communicate with each other, and (b) no state outside A is reachable from a state in A, then A is a recurrent class of states; A is said to be a *closed set.* There may be a positive probability of never entering the class A, but once entered, the MC remains there. In that case, every state in A will be entered infinitely often, and the recurrence times will have finite means.

A *transient* state is not recurrent: although there is no bound in general on the number of times it can be visited in paths of a MC, the expected number of visits is finite. An *absorbing state j* is a recurrent class containing just the one state j. An MC must have at least one recurrent class; it may or may not have transient states. The MC is *irreducible* if it has exactly one recurrent class. Note that aperiodicity specializes to recurrence classes: some classes may be aperiodic and some not. In the fundamental result of finite Markov chains above, we could have allowed up to $m-1$ transient states (a nonempty recurrent class is needed), and the result would have been the same except that for the stipulation that $p_{i,j}^{(n)} \to 0$, and hence $\pi_j = 0$, for transient states S_j. Finally, if in the fundamental result, 'aperiodic' had been replaced by 'periodic', a stationary distribution $\{\pi\}$ solving the balance equations would still have existed, even though periodicities prevent the convergence of the state probabilities at time n; the π_j may still be interpreted as the fractions of time spent in states j.

Convergence rates are of great importance as they measure the approximation in models of stationary behavior. For finite MC's convergence is exponentially (more precisely geometrically) fast in the sense that there exists a $\rho < 1$ such that $|p_{i,j}^{(n)} - \pi_j| = O(\rho^n)$, where the hidden constants do not depend on i, j. Thus, these approximations are typically very good.