MATH 5255 - Math Theory of Probability Lecture Notes 1

Libao Jin (ljin1@uwyo.edu)

April 25, 2018

1 Combinatorial Analysis

1.1 The Basic Principle of Counting

Theorem 1.1 (The basic principle of counting)**.** Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

Example 1.1 (Example 2a)**.** A small community consists of 10 women, each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible.

SOLUTION. From the basic principle that there are $10 \times 3 = 30$ possible choices. □

Theorem 1.2 (The generalized basic principle of counting)**.** If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiments; and if, for each of the first two experiments, there are n_3 possible outcomes of the third experiment; and if ..., then there is a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

Example 1.2 (Example 2e)**.** How many different 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers if repetition among letters or numbers were prohibited?

SOLUTION. By the generalized version of the basic principle, there would be $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$ possible license plates. □

1.2 Permutations

Theorem 1.3 (Permutations)**.** Suppose that we have n objects, by the basic principle of counting, there are $n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$ different permutations of the n objects.

Theorem 1.4 (Permutations among some indistinguishable objects)**.** Suppose that there are n objects, of which n_1 are alike, n_2 are alike, ..., n_r are alike, then there are

$$\frac{n!}{n_1! n_2! \cdots n_r!}$$

different permutations.

1.3 Combinations

Theorem 1.5 (Combinations)**.** Suppose that there are n objects, then the number of different groups of r objects is

$$\frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!},$$

where $n(n-1)(n-2)\cdots(n-r+1)$ represents the number of different ways that a group of r items could be selected from n items when the order of selection is relevant, and as each group of r items will be counted $r!$ times in this count.

Definition 1.1 (n choose k)**.** We define $\binom{n}{r}$, for $r \leq n$, by

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

and say $\binom{n}{r}$ represents the number of possible combinations of n objects taken r at a time.

Theorem 1.6 (Combinatorial identity)**.**

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.$$

Proof. Consider a group of n objects, and fix attention on some particular one of there objects – call it object 1. Now, there are $\binom{n-1}{r-1}$ groups of size r that contain object 1 (since each group is formed by selecting $r-1$ from the remaining $n-1$ objects). Also, there are $\binom{n-1}{r}$ groups of size r that do not contain object 1. As there is a total of $\binom{n}{r}$ group of size r. □

Theorem 1.7 (The binomial theorem)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

HINT: Proof by induction.

1.4 Multinomial Coefficients

Theorem 1.8 (Multinomial)**.** A set of n distinct items is to be divided into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$, where $\sum_{i=1}^{r} n_i = n$. There are

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

Definition 1.2 (Multinomial notation)**.** If $\sum_{i=1}^{r} n_i = n$, we define $\binom{n}{n_1, n_2, \ldots, n_r}$ by

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}.$$

Thus, $\binom{n}{n_1, n_2, \ldots, n_r}$ represents the number of possible divisions of n distinct objects into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$.

Theorem 1.9 (The multinomial theorem)**.**

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

That is, the sum is over all nonnegative integer-valued vectors $(n_1, n_2, \ldots, n_r)$ such that $n_1+n_2+\cdots+n_r = n$.

Proposition 1.1. There are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \ldots, x_r)$ satisfying the equation

$$x_1 + x_2 + \cdots + x_r = n \quad x_i > 0, i = 1, \ldots, r.$$

Proposition 1.2. There are $\binom{n+r-1}{r-1}$ distinct nonnegative integer-valued vectors $(x_1, x_2, \ldots, x_r)$ satisfying the equation

$$x_1 + x_2 + \cdots + x_r = n.$$

2 Axioms of Probability

2.1 Sample Space and Events

Definition 2.1 (Sample space and events)**.** The set of all possible outcomes of an experiment is known as the *sample space* of the experiment and is denoted by S. Any subset E of the sample space is known as an event. In other words, an event is a set consisting of possible outcomes of the experiment.

Definition 2.2 (Union and intersection)**.** The event consists of all outcomes that are either in E or in F or in both E and F is called the *union* of the event E and F, denoted by $E \cup F$. Similarly, for any two events E and F, the event consists of all outcomes that are both in E and F is called the *intersection* of E and F, denoted by $E \cap F$ or EF. If EF does not contain any outcomes, it is called the null event, denoted by $\emptyset$, then E and F are said to be mutually exclusive.

Definition 2.3 (Generalized union and intersection)**.** If $E_1, E_2, \ldots$ are events, then the union of these events, denoted by $\bigcup_{n=1}^{\infty} E_n$, is defined to be that event which consists of all outcomes that are in E_n for at least one value of $n = 1, 2, \ldots$. Similarly, the intersection of the events E_n, denoted by $\bigcap_{n=1}^{\infty} E_n$, is defined to be the event consisting of those outcomes which are in all of the events $E_n, n = 1, 2, 3, \ldots$.

Definition 2.4 (Complement, subset and superset)**.** For any event E, we define the new event $E^{\mathbb{C}}$, referred to as the *complement* of E, to consist of all outcomes in the sample space S that are not in E. That is, $E^{\mathbb{C}}$ will occur if and only if E does not occur. For any two events E and F, if all of the outcomes in E are also in F, then we say E is *contained* in F, or E is a *subset* of F, and write $E \subset F$ (or equivalently, $E \supset E$, which we sometimes say as F is a *superset* of E). If $E \subset F$ and $F \subset E$, then E and F are equal and write $E = F$.

Theorem 2.1 (Laws)**.**

(1) *Commutative laws*: $E \cup F = F \cup E$, $EF = FE$.

(2) *Associative laws*: $(E \cup F) \cup G = E \cup (F \cup G)$, $(EF)G = F(EG)$.

(3) *Distributive laws*: $(E \cup F)G = EG \cup FG$, $(EF) \cup G = (E \cup G) \cap (F \cup G)$.

Theorem 2.2 (DeMorgan's laws)**.**

$$\left(\bigcup_{i=1}^{n} E_i\right)^{\complement} = \bigcap_{i=1}^{n} E_i^{\complement},$$
$$\left(\bigcap_{i=1}^{n} E_i\right)^{\complement} = \bigcup_{i=1}^{n} E_i^{\complement}.$$

2.2 Axioms of Probability

Theorem 2.3 (Axioms of probability)**.** Consider an experiment whose sample space is S. For each event E of the sample space S, we assume that a number $P(E)$ is defined and satisfies the following axioms:

(1) $0 \le P(E) \le 1$.

(2) $P(S) = 1$.

(3) For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (that is, events for which $E_iE_j = \emptyset$ when $i \ne j$),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

We refer to $P(E)$ as the probability of the event E.

2.3 Some Simple Propositions

Proposition 2.1.

$$P(E^{\complement}) = 1 - P(E).$$

Proposition 2.2. If $E \subset F$, then $P(E) \le P(F)$.

Proposition 2.3.

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

Proposition 2.4.

$$\begin{aligned} P(E_1 \cup E_2 \cup \cdots \cup E_n) = &\sum_{i=1}^{n} P(E_i) - \sum_{i_1<i_2} P(E_{i_1}E_{i_2}) + \cdots \\ &+ (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(E_{i_1}E_{i_2}\cdots E_{i_r}) + \cdots \\ &+ (-1)^{n+1} P(E_1E_2\cdots E_n) \\ = &\sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1<\cdots<i_r} P(E_{i_1}\cdots E_{i_r}). \end{aligned}$$

2.4 Sample Spaces Having Equally Likely Outcomes

Definition 2.5 (Probability of equally likely events)**.** Consider an experiment whose sample space S is a finite set, say, $S = \{1, 2, \ldots, N\}$, assume that all outcomes in the sample space are equally like to occur, i.e.,

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\}),$$

which implies that

$$P(\{i\}) = \frac{1}{N}, i = 1, 2, \ldots, N.$$

Then for any event E, we have

$$P(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S}.$$

2.5 Probability As a Continuous Set Function

Definition 2.6 (Increasing/decreasing sequence)**.** A sequence events $\{E_n, n \geq 1\}$ is said to be an increasing sequence if

$$E_1 \subset E_2 \subset \cdots E_n \subset E_{n+1} \subset \cdots,$$

whereas it is said to be a decreasing sequence if

$$E_1 \supset E_2 \supset \cdots E_n \supset E_{n+1} \supset \cdots.$$

Definition 2.7. It $\{E_n, n \geq 1\}$ is an increasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$, by

$$\lim_{n\to\infty} E_n = \bigcup_{i=1}^{\infty} E_i.$$

Similarly, if $\{E_n, n \geq 1\}$ is a decreasing sequence of events, we define $\lim_{n\to\infty} E_n$ by

$$\lim_{n\to\infty} E_n = \bigcap_{i=1}^{\infty} E_i.$$

Proposition 2.5. If $\{E_n, n \geq 1\}$ is either an increasing or a decreasing sequence of events, then

$$\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n).$$

3 Conditional Probability and Independence

3.1 Conditional Probabilities

Definition 3.1 (i)**.** For two events E and F, if $P(F) > 0$, then

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Theorem 3.1 (The multiplication rule)**.**

$$P(E_1E_2E_3\cdots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\cdots P(E_n|E_1\cdots E_{n-1}).$$

Theorem 3.2 (Bayes's formula)**.**

$$\begin{aligned} P(E) &= P(EF) + P(EF^{\complement}) \\ &= P(E|F)P(F) + P(E|F^{\complement})P(F^{\complement}) \\ &= P(E|F)P(F) + P(E|F^{\complement})(1 - P(F)). \end{aligned}$$

Definition 3.2. The odds of an event A are defined by

$$\frac{P(A)}{P(A^{\complement})} = \frac{P(A)}{1 - P(A)}.$$

That is, the odds of an event A tell how much more likely it is that the event A occurs than it is that it does not occur.

Theorem 3.3. Consider a hypothesis H that is true with probability $P(H)$, and suppose that new evidence E is introduced. Then the conditional probabilities, given the evidence E, that H is true and that H is not true are respectively given by

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}, \quad P(H^{\complement}|E) = \frac{P(E|H^{\complement})P(H^{\complement})}{P(E)}.$$

Therefore, the new odds after the evidence E has been introduced are

$$\frac{P(H|E)}{P(H^{\complement}|E)} = \frac{P(H)}{P(H^{\complement})}\frac{P(E|H)}{P(E|H^{\complement})}.$$

Proposition 3.1.

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

3.2 Independent Events

Definition 3.3. Two events E and F are said to be *independent* if $P(EF) = P(E)P(F)$ holds. Two events E and F that are not independent are said to be dependent.

Proposition 3.2. If E and F are independent, then so are E and $F^{\complement}$.

Definition 3.4. Three events E, F, and G are said to be independent if

$$\begin{aligned} P(EFG) &= P(E)P(F)P(G). \\ P(EF) &= P(E)P(F). \\ P(EG) &= P(E)P(G). \\ P(FG) &= P(F)P(G). \end{aligned}$$

Remark 3.1.

(a) Whereas the preceding argument established a condition on n and k that guarantees the existence of a coloring scheme satisfying the desired property, it gives no information about how to obtain such a scheme (although one possibility would be simply to choose the colors at random, check to see if the resulting coloring satisfies the property, and repeat the procedure until it does).

(b) The method of introducing probability into a problem whose statement is purely deterministic has been called the probabilistic method. Other examples of this method are given in Theoretical exercise and examples.

Proposition 3.3.

(a) $0 \le P(E|F) \le 1$.

(b) $P(S|F) = 1$.

(c) If $E_i, i = 1, 2, \ldots$, are mutually exclusive events, then

$$P\left(\bigcup_{1}^{\infty} E_i|F\right) = \sum_{1}^{\infty} P(E_i|F).$$

4 Random Variables

4.1 Random Variables

Definition 4.1. The quantities of interest, or, more formally, the real-valued functions defined on the sample space, are known as *random variables*.

4.2 Discrete Random Variables

Definition 4.2. A random variable that take can take on at most a countable number of possible value is said to be discrete. For a discrete random variable X, we define the *probability mass function* $p(a)$ of X by

$$p(a) = P(X = a).$$

The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

$$\begin{cases} p(x_i) \geq 0 & \text{for } i = 1, 2, \ldots \\ p(x) = 0 & \text{for all other values of } x. \end{cases}$$

Since X must take on one of the values x_i, we have

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

4.3 Expected Value

Definition 4.3. If X is a discrete random variable having a probability mass function $p(x)$, then the *expectation*, or the *expected value*, of X, denoted by $E[X]$, is defined by

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

In words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

4.4 Expectation of a Function of a Random Variable

Proposition 4.1. If X is a discrete random variable that takes on one of the values $x_i, i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g,

$$E[g(x)] = \sum_i g(x_i)p(x_i).$$

If a and b are constants, then

$$E[aX + b] = aE[X] + b.$$

4.5 Variance

Definition 4.4. If X is a random variable with mean μ, then the variance of X, denoted by $\mathrm{Var}(X)$ is defined by

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

Proposition 4.2. For any constants a and b,

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

4.6 The Bernoulli and Binomial Random Variables

Definition 4.5. Suppose that a trial, or an experiment, whose outcome can be classified as either a *success* or a *failure* is performed. If we let $X = 1$ when the outcome is a success and $X = 0$ when it is a failure, then the probability mass function of X is given by

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

where $p, 0 \leq p \leq 1$, is the probability that the trial is a success.

Definition 4.6. A random variable X is said to be a *Bernoulli random variable* (after the Swiss mathematician James Bernoulli) if its probability mass function is given by the above equation for some $p \in (0, 1)$.

Definition 4.7. Suppose that n independent trials, each of which results in a success with probability p and in a failure with probability $1-p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a *binomial random variable* with parameters (n,p). Thus, a Bernoulli random variable is just a binomial random variable with parameters $(1,p)$. And the probability mass function of a binomial random variable having parameters (n,p) is given by

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

4.6.1 Properties of Binomial Random Variables

Proposition 4.3.

$$E[X^k] = \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}. \tag{1}$$

Proposition 4.4. Since

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} = \frac{(n-1)!n}{[(n-1)-(i-1)]!(i-1)!i} = \frac{n}{i}\binom{n-1}{i-1}.$$

Then (1) becomes

$$\begin{aligned} E[X^k] &= n\sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^i (1-p)^{n-i} \\ &= np\sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{(n-1)-(i-1)} \\ &= np\sum_{i=0}^{n-1} (i+1)^{k-1} \binom{n-1}{i} p^i (1-p)^{n-1-i)} \\ &= npE[(Y+1)^{k-1}], \end{aligned}$$

where Y is binomial random variable with parameters $n-1, p$.

Proposition 4.5. Setting $k=1$ in the preceding equation yields

$$E[X] = np.$$

That is, the expected number of success that occur in n independent trials when each is a success with probability p is equal to np. Then let $k=2$, we have

$$E[X^2] = npE[Y+1] = np[(n-1)p+1].$$

It follows that

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

Proposition 4.6. If X is a binomial random variable with parameters (n,p), where $0 < p < 1$, then as k goes from 0 to n, $P(X=k)$ first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n+1)p$.

4.6.2 Computing the Binomial Distribution Function

Definition 4.8. Suppose that X is binomial with parameters (n,p). The key to computing its distribution function

$$P\{X \le i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \quad i = 0, 1, \ldots, n,$$

is to utilize the following relationship between $P\{X=k+1\}$ and $P\{X=k\}$, where

$$P\{X=k+1\} = \frac{p}{1-p}\frac{n-k}{k+1}P\{X=k\}.$$

4.7 The Poisson Random Variable

Definition 4.9. A random variable X that takes on one of the values $0, 1, 2, \ldots$ is said to be a *Poisson random variable* with parameter λ if, for some $\lambda >$,

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

Equation (2) defines a probability mass function, since

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

Proposition 4.7. The Poisson random variable may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small enough so that np is of moderate size. To see this, suppose that X is a binomial random variable with parameters (n, p), and let $\lambda = np$. Then

$$\begin{aligned}
P\{X = i\} &= \binom{n}{i} p^i(1-p)^{n-i} \\
&= \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} \\
&= \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n-1)(n-2)\cdots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\left(1-\frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n-1)(n-2)\cdots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.
\end{aligned}$$

Given that n is large enough and p is small enough to make np moderate, then we have the following

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)(n-2)\cdots(n-i+1)}{n^i} \approx 1, \quad \left(1-\frac{\lambda}{n}\right)^i \approx 1.$$

It follows that

$$P\{X = i\} \approx \frac{\lambda^i}{i!}e^{-\lambda}.$$

In other words, if n independent trials, each of which results in a success with probability p, are performed, then when n is large and p is small enough to make np moderate, the number of success occurring is approximately a Poisson random variable with parameter $\lambda = np$. This value λ will usually be determined empirically.

Example 4.1. Some examples of random variables that generally obey the Poisson probability law are as follows:

(a) The number of misprints on a page (or a group of pages) of a book.

(b) The number of people in a community who survive to age 100.

(c) The number of wrong telephone numbers that are dialed in a day.

(d) The number of packages of dog biscuits sold in a particular store each day.

(e) The number of customers entering a post office on a given day.

(f) The number of vacancies occurring during a year in the federal judicial system.

(g) The number of α-particles discharged in a fixed period of time from some radioactive material.

Proposition 4.8. The expected value and the variance of a Poisson random variable.

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\ &= \sum_{i=1}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \\ &= \lambda. \end{aligned}$$

To determine its variance, we first compute $E[X^2]$:

$$\begin{aligned} E[X^2] &= \sum_{i=0}^{\infty} \frac{i^2e^{-\lambda}\lambda^i}{i!} \\ &= \sum_{i=1}^{\infty} \frac{i^2e^{-\lambda}\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{ie^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{(i-1+1)e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda \sum_{i=0}^{\infty} \frac{(i+1)e^{-\lambda}\lambda^{i}}{i!} \\ &= \lambda \left(\sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} + \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} \right) \\ &= \lambda(\lambda+1). \end{aligned}$$

It follows that

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

Definition 4.10 (Poisson Paradigm)**.** Consider n events, with p_i equal to the probability that event i occurs, $i = 1, \ldots, n$. If all the p_i are "small" and the trials are either independent or at most "weakly dependent", then the number of these events that occur approximately has a Poisson distribution with mean $\sum_{i=1}^{n} p_i$.

4.7.1 Computing the Poisson Distribution Function

Definition 4.11. If X is Poisson with parameter λ, then

$$\frac{P\{X = i+1\}}{P\{X = i\}} = \frac{e^{\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^i/i!} = \frac{\lambda}{i+1}. \tag{3}$$

Then starting with $P\{X = 0\} = e^{-\lambda}$, we can use (3) to compute successively

$$\begin{aligned} P\{X = 1\} &= \lambda P\{X = 0\}, \\ P\{X = 2\} &= \frac{\lambda}{2} P\{X = 1\}, \end{aligned}$$

$$\vdots$$

$$P\{X = i+1\} = \frac{\lambda}{i+1} P\{X = i\}.$$

4.8 Other Discrete Probability Distributions

4.8.1 The Geometric Random Variable

Definition 4.12. Suppose that independent trials, each having a probability p, $0 < p < 1$, of being a success, are performed until a success occurs. If we let X equal the number of trials required, then

$$P\{X = n\} = (1-p)^{n-1} p, n = 1, 2, \ldots. \tag{4}$$

Equation (4) follows, in order for X to equal n, it is necessary and sufficient that the first $n-1$ trials are failures and the nth trial is a success. Equation (4) then follows, since the outcomes of the successive trials are assumed to be independent. Since

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

It follows that, with probability 1, a success will eventually occur. Any random variable X whose probability mass function is given by (4) is said to be *geometric* random variable with parameter p.

4.8.2 The Negative Binomial Random Variable

Definition 4.13. Suppose that independent trials, each having probability p, $0 < p < 1$, of being a success are performed until a total of r successes is accumulated. If we let X equal the number of trials required, then

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r}, n = r, r+1, \ldots. \tag{5}$$

Equation (5) follows because, in order for the rth success to occur at the nth trial, there must be $r-1$ successes in the first $n-1$ trials and the nth trial must be a success. The probability of the first event is

$$\binom{n-1}{r-1} p^{r-1} (1-p)^{n-r}.$$

and the probability of the second is p; thus, by independence, Equation (5) is established. To verify that a total of r successes must eventually be accumulated, either we can prove analytically that

$$\sum_{n=r}^{\infty} P\{X = n\} = \sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = 1.$$

Any random variable X whose probability mass function is given by (**??**) is said to be a *negative binomial* random variable with parameter (r, p). Note that a geometric random variable is just a negative binomial with parameter $(1, p)$.

4.8.3 The Hypergeometric Random Variable

Definition 4.14. Suppose that a sample of size n is to be chosen randomly (without replacement) from an urn containing N balls, of which m are white and $N-m$ are black. If we let X denote the number of white balls selected, then

$$P\{X = i\} = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, i = 0, 1, \ldots, n \tag{6}$$

A random variable X whose probability mass function is given by Equation (6) for some values of n, N, m is said to be a *hypergeometric* random variable.

Remark 4.1. Although we have written the hypergeometric probability mass function with i going from 0 to n, $P\{X = i\}$ will actually be 0, unless i satisfies the inequalities $n - (N - m) \leq i \leq \min(n, m)$. However, (6) is always valid because of our convention that $\binom{r}{k}$ is equal to 0 when either $k < 0$ or $r < k$.

4.8.4 The Zeta (or Zipf) Distribution

Definition 4.15. A random variable is said to have a zeta (sometimes called the Zipf) distribution if its probability mass function is given by

$$P\{X = k\} = \frac{C}{k^{\alpha+1}}, k = 1, 2, \ldots,$$

for some value of $\alpha > 0$. Since the sum of the foregoing probabilities must equal 1, it follows that

$$C = \left[\sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^{\alpha+1}\right]^{-1}.$$

The zeta distribution owes its name to the fact that the function

$$\zeta(s) = 1 + \left(\frac{1}{2}\right)^s + \left(\frac{1}{3}\right)^s + \cdots + \left(\frac{1}{k}\right)^s + \cdots$$

is known in mathematical discipline as the Riemann zeta function (after German mathematician G.F.B. Riemann). The zeta distribution was used by the Italian economist V. Pareto to describe the distribution of family incomes in a given country. However, it was G.K. Zipf who applied zeta distribution to a wide variety of problems in different areas and, in doing so, popularized its use.

4.9 Expected Value of Sums of Random Variables

Proposition 4.9. For a random variable X, let $X(s)$ denote the value of X when $s \in S$ is the outcome of the experiment. Now, if X and Y are both random variables, then so is their sum. That is, $Z = X + Y$ is also a random variable. Moreover, $Z(s) = X(s) + Y(s)$.

Proposition 4.10.

$$E[X] = \sum_{s \in S} X(s)p(s).$$

For random variables $X_1, X_2, \ldots, X_n$,

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

4.10 Properties of the Cumulative Distribution Function

Definition 4.16. Recall that, for the distribution function F of $X, F(b)$ denotes the probability that the random variable X takes on a value that is less than or equal to b. Following are some properties of the cumulative distribution function (c.d.f.) F:

1. F is a non-descending function; that is, if $a < b$, then $F(a) \leq F(b)$.
2. $\lim_{b \to \infty} F(b) = 1$.
3. $\lim_{b \to -\infty} F(b) = 0$.
4. F is right continuous. That is, for any b and any decreasing sequence $b_n, n \geq 1$, that converges to b, $\lim_{n \to \infty} F(b_n) = F(b)$.

5 Continuous Random Variables

5.1 Introduction

Definition 5.1 (Continuous random variables)**.** X is said to be a continuous random variable if there exists a nonnegative function f, defined for all real $x \in (-\infty, \infty)$, having the property that, for any set B of real numbers,

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

where the function f is called the *probability density function* of the random variable X. Note: the probability that a continuous random variable will assume any fixed value is zero. Hence, for a continuous random variable,

$$P\{X < a\} = P\{X \le a\} = F(a) = \int_{-\infty}^{a} f(x)\,dx.$$

Proposition 5.1. The relationship between the cumulative distribution F and the probability density f is expressed by

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

Differentiating both sides of the preceding equation yields

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

That is, the density is the derivative of the cumulative distribution function. A somewhat more intuitive interpretation of the density function maybe obtained from the following equation

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

as follows

$$P\left\{a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\right\} = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)\,dx \approx \varepsilon f(x),$$

when ε is small and when $f(\cdot)$ is continuous at $x = a$. In other words, the probability that X will be contained in an interval of length ε around the point a is approximately $\varepsilon f(a)$. From this result we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

5.2 Expectation and Variance of Continuous Random Variables

Definition 5.2. If X is a continuous random variable having probability density function $f(x)$, then, because

$$f(x)\,dx \approx P\{x \le X \le x + dx\} \text{ for } dx \text{ small},$$

it is easy to see that the analogous definition is to define the expected value of X by

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

Proposition 5.2. If X is a continuous random variable with probability density function $f(x)$, then, for any real-valued function g,

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

Lemma 5.1. For a nonnegative random variable Y,

$$E[Y] = \int_0^{\infty} P\{Y > y\}\,dy.$$

If a and b are constants, then

$$E[aX + b] = aE[x] + b.$$

Definition 5.3. If X is a random variable with expected value μ, then the variance of X is defined (for any type of random variable) by

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

5.3 The Uniform Random Variable

Definition 5.4. A random variable is said to be *uniformly* distributed over the interval $(0, 1)$ if its probability density function is given by

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

In general, X is said to be a uniform random variable on the interval (α, β) if the probability density function of X is given by

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

it follows from the preceding equation that the distribution function of a uniform random variable on the interval (α, β) is given by

$$F(a) = \begin{cases} 0 & a \le \alpha, \\ \frac{a - \alpha}{\beta - \alpha} & \alpha < a < \beta, \\ 1 & a \ge \beta. \end{cases}$$

5.4 Normal Random Variables

Definition 5.5. X is said to be a normal random variable or simply that X is normally distributed, with parameters μ and σ^2 if the density of X is given by

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

This density function is a bell-shaped curve that is symmetric about μ.

Proposition 5.3. If X is normally distributed with parameters μ and σ^2, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$.

Proposition 5.4. If X is normally distributed with parameters μ and σ^2, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters 0 and 1, which is said to be a *standard*, or a *unit*, normal random variable.

Proposition 5.5. If X is a normal random variable with parameters μ and σ^2, then

$$E[X] = \mu, \quad \text{Var}(X) = \sigma^2.$$

Definition 5.6. It is customary to denote the cumulative distribution function of a standard normal random variable by $\Phi(x)$. That is,

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

Proposition 5.6. By the symmetry, we have

$$\Phi(-x) = 1 - \Phi(x), \quad -\infty < x < \infty.$$

That is,

$$P\{Z \le -x\} = P\{Z > x\}, \quad -\infty < x < \infty.$$

Since $Z = (X - \mu)/\sigma$ is a standard normal random variable whenever X is normally distributed with parameters μ an σ^2, it follows that the distribution function of X can be expressed as

$$F_X(a) = P\{X \le a\} = P\left(\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right).$$

5.4.1 The Normal Approximation to the Binomial Distribution

Theorem 5.1 (The DeMoivre-Laplace limit theorem)**.** If S_n denotes the number of successes that occur when n independent trials, each resulting in a success with probability p, are performed, then, for any $a < b$,

$$P\left\{a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right\} \to \Phi(b) - \Phi(a) \text{ as } n \to \infty.$$

5.5 Exponential Random Variables

Definition 5.7. A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

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

is said to be an *exponential* random variable (or, more simply, is said to be exponentially distributed) with parameter λ. The cumulative distribution function $F(a)$ of an exponential random variable is given by

$$F(a) = P\{X \le a\} = \int_0^a \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda a}, \quad a \ge 0.$$

Proposition 5.7. Let X be an exponential random variable with parameter λ. Then

$$\begin{aligned} E[X^n] &= \int_{-\infty}^{\infty} x^n f(x)\, dx \\ &= \int_0^{\infty} x^n \lambda e^{-\lambda x}\, dx \\ &= -x^n e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} n x^{n-1}\, dx \\ &= \frac{n}{\lambda}\int_0^{\infty} x^{n-1}\lambda e^{-\lambda x}\, dx \\ &= \frac{n}{\lambda} E[X^{n-1}]. \end{aligned}$$

Let $n = 1$ and $n = 2$, follows that $E[X] = 1/\lambda$ and $E[X^2] = 2/\lambda^2$. Then

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = 1/\lambda^2.$$

Definition 5.8. A nonnegative random variable X is said to be *memoryless* if

$$P\{X > s + t | X > t\} = P\{X > s\}, \forall s, t \ge 0.$$

It turns out that the not long is the exponential distribution memoryless, but it is also the unique distribution possessing this property.

Definition 5.9. A variation of the exponential distribution is the distribution of a random variable that is equally likely to be either positive or negative and whose absolute value is exponentially distributed with parameter λ, $\lambda \ge 0$. Such a random variable is said to have a *Laplace* distribution, and its density is given by

$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}, \quad -\infty < x < \infty.$$

Its distribution function is given by

$$F(x) = \begin{cases} \frac{1}{2}\int_{-\infty}^{x} \lambda e^{\lambda x}\,dx & x < 0, \\ \frac{1}{2}\int_{-\infty}^{0} \lambda e^{\lambda x}\,dx + \frac{1}{2}\int_{0}^{x} \lambda e^{-\lambda x}\,dx & x > 0, \end{cases} = \begin{cases} \frac{1}{2}e^{\lambda x} & x < 0, \\ 1 - \frac{1}{2}e^{-\lambda x} & x > 0. \end{cases}$$

5.5.1 Hazard Rate Functions

Definition 5.10. Consider a positive continuous random variable X that we interpret as being the lifetime of some item. Let X have distribution function F and density f. The hazard rate (sometimes called the *failure rate*) function $\lambda(t)$ of F is defined by

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)}, \quad \text{where } \overline{F} = 1 - F.$$

To interpret $\lambda(t)$, suppose that the item has survived for a time t and we desire the probability that it will not survive for an additional time dt. That is, consider $P\{X \in (t, t+dt) | X > t\}$. Now,

$$\begin{aligned} P\{X \in (t, t+dt) | X > t\} &= \frac{P\{X \in (t, t+dt), X > t\}}{P\{X > t\}} \\ &= \frac{P\{X \in (t, t+dt)\}}{P\{X > t\}} \\ &\approx \frac{f(t)}{\overline{F}(t)}\,dt. \end{aligned}$$

Thus, $\lambda(t)$ represents the conditional probability intensity that a t-unit-old item will fail. Suppose now that the lifetime distribution is exponential. Then, by the memoryless property, it follows that the distribution of remaining life for a t-year-old item is the same as that for a new item. Hence $\lambda(t)$ should be constant. In fact, this checks out, since

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda}} = \lambda.$$

Thus, the failure rate function for the exponential distribution is constant. The parameter λ is often referred to as the *rate* of the distribution. It turns out that the failure rate function $\lambda(t)$ uniquely determines the distribution F. To prove this, note that, by definition,

$$\lambda(t) = \frac{dF(t)/dt}{1 - F(t)}.$$

Integrating both sides yields

$$\log(1 - F(t)) = -\int_0^t \lambda(\tau)\,d\tau + k \implies 1 - F(t) = e^k \exp\left\{-\int_0^t \lambda(\tau)\,d\tau\right\}.$$

Letting $t = 0$ shows that $k = 0$; thus,

$$F(t) = 1 - \exp\left\{-\int_0^t \lambda(\tau)\,d\tau\right\}.$$

Hence, a distribution function of a positive continuous random variable has a linear hazard rate function – that is, if

$$\lambda(t) = a + bt,$$

then its distribution function is given by

$$F(t) = 1 - e^{-at - bt^2/2},$$

and differentiation yields its density, namely,

$$f(t) = (a + bt)e^{-(at + bt^2/2)}, \quad t \geq 0.$$

When $a = 0$, the preceding equation is known as the *Rayleigh density function*.

5.6 Other Continuous Distributions

5.6.1 The Gamma Distribution

Definition 5.11. A random variable is said to be a gamma distribution with parameter $(\alpha, \lambda), \lambda > 0, \alpha > 0$, if its distribution function is given by

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

where $\Gamma(\alpha)$, called the *gamma function*, is defined as

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy.$$

Integration of $\Gamma(\alpha)$ by parts yields

$$\begin{aligned} \Gamma(\alpha) &= -e^y y^{\alpha-1}\Big|_0^\infty + \int_0^\infty e^{-y}(\alpha-1)y^{\alpha-2}\, dy \\ &= (\alpha-1)\int_0^\infty e^{-y} y^{\alpha-2}\, dy \\ &= (\alpha-1)\Gamma(\alpha-1). \end{aligned}$$

Since $\Gamma(1) = \int_0^\infty e^{-x}\, dx = 1$, it follows that, for integral values of n,

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

When α is a positive integer, say, $\alpha = n$, the gamma distribution with parameters (α, λ) often arises, in practice as the distribution of the amount of time one has to wait unitl a toal of n events has occurred. More specifically, if events are occurring randomly and in accordance with the three axioms, then it turns out that the amount of time one has to wait until a total of n events has occurred will be a gamma random variable with parameters (n, λ). To prove this, let T_n denote the time at which the nth event occurs, and note that T_n is less than or equal to t if and only if the number of events that have occurred by time t is at least n. That is, with $N(t)$ equal to the number of events in $[0, t]$,

$$P\{T_n \leq t\} = P\{N(t) \geq n\} = \sum_{j=n}^\infty P\{N(t) = j\} = \sum_{j=n}^\infty \frac{e^{-\lambda t}(\lambda t)^j}{j!},$$

where the final identity follows because the number of events in $[0, t]$ has a Poisson distribution with parameter λt. Differentiation of the preceding now yields the density function T_n:

$$f(t) = \sum_{j=n}^\infty \frac{e^{-\lambda t} j(\lambda t)^{j-1}\lambda}{j!} - \sum_{j=n}^\infty \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!} = \sum_{j=n}^\infty \frac{\lambda e^{-\lambda t}(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^\infty \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!} = \frac{\lambda e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}.$$

Hence, T_n has the gamma distribution with parameters (n, λ). (This distribution is often referred to in the literature as the *n-Erlang distribution.*) Note that when $n = 1$, this distribution reduces to the exponential distribution. The gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$, n a positive integer, is called the χ_n^2 (chi-squared) distribution with n degrees of freedom. The chi-squared distribution often arises in practice as the distribution of the error involved in attempting to hit a target in n-dimensional space when each coordinate error is normally distributed.

5.6.2 The Weibull Distribution

Definition 5.12. A random variable whose cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & x \leq \nu, \\ 1 - \exp\left\{-\left(\frac{x-\nu}{\alpha}\right)^\beta\right\} & x > \nu, \end{cases}$$

is said to be a *Weibull random variable* with parameters ν, α and β. Differentiation yields the density:

$$f(x) = \begin{cases} 0 & x \le \nu, \\ \frac{\beta}{\alpha}\left(\frac{x-\nu}{\alpha}\right)^{\beta} \exp\left\{-\left(\frac{x-\nu}{\alpha}\right)^{\beta}\right\} & x > \nu. \end{cases}$$

The Weibull distribution is widely used in engineering practice due to its versatility. It was originally proposed for the interpretation of fatigue data, but now its use has been extended to many other engineering problems. In particular, it is widely used in the field of life phenomena as the distribution of the lifetime of some object, especially when the "weakest link" model is appropriate for the object. That is, consider an object consisting of many parts, and suppose that the object experiences death (failure) when any of its parts fail. It has been shown (both theoretically and empirically) that under these condition a Weibull distribution provides a close approximation to the distribution of lifetime of the item.

5.6.3 The Cauchy Distribution

Definition 5.13. A random variable is said to be a Cauchy distribution with parameter $\theta, -\infty < \theta < \infty$, if its density is given by

$$f(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty.$$

5.6.4 The Beta Distribution

Definition 5.14. A random variable is said to be to have a beta distribution if its density is given by

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1, \\ 0 & \text{otherwise}, \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx.$$

The beta distribution can be used to model a random phenomenon whose set of possible values is some finite interval $[c,d]$ – which, by letting c denote the origin and taking $d-c$ as a unit measurement, can be transformed into the interval $[0,1]$. When $a=b$, the beta density is symmetric about $1/2$, giving more and more weight to regions about $1/2$ as the common value a increases. When $b>a$, the density is skewed to the left (in the sense that smaller values become more likely); and it is skewed to the right when $a>b$. The relationship

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

can be shown to exist between

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx.$$

and the gamma function.

5.7 The Distribution of A Function of A Random Variable

Suppose that we known the distribution of X and want to find the distribution of $g(X)$. To do so, it is necessary to express the even that $g(X) \le y$ in terms of X being in some set.

Theorem 5.2. Let X be a continuous random variable having probability density function f_X. Suppose that $g(x)$ is a strictly monotonic (increasing or decreasing), differentiable (and thus continuous) function x. Then the random variable Y defined by $Y = g(X)$ has a probability density function given by

$$f_Y(y) = \begin{cases} f_X[g^{-1}(y)]\left|\frac{d}{dy}g^{-1}(y)\right| & \text{if } y = g(x) \text{ for some } x, \\ 0 & \text{if } y \ne g(x) \text{ for all } x, \end{cases}$$

where $g^{-1}(y)$ is defined to equal that value of x such that $g(x) = y$.

6 Jointly Distributed Random Variables

6.1 Joint Distribution Functions

Definition 6.1. For any two random variables X and Y, the *joint cumulative probability distribution function* of X and Y by

$$F(a,b) = P\{X \le a, Y \le b\}, \quad -\infty < a, b < \infty.$$

The distribution of X can be obtained from the joint distribution of X and Y as follows:

$$\begin{aligned} F_X(a) &= P\{X \le a\} \\ &= P\{X \le a, Y < \infty\} \\ &= P\left(\lim_{b\to\infty}\{X \le a, Y \le b\}\right) \\ &= \lim_{b\to\infty} P\{X \le a, Y \le b\} \\ &= \lim_{b\to\infty} F(a,b) \\ &= F(a,\infty). \end{aligned}$$

Similarly, $F_Y(b) = P\{Y \le b\} = \lim_{a\to\infty} F(a,b) = F(\infty, b)$. The distribution function F_X and F_Y are sometimes referred to as the *marginal* distributions of X and Y.

Proposition 6.1. The joint probability that X is greater than a and Y is greater than b is

$$\begin{aligned} P\{X > a, Y > b\} &= 1 - P(\{X > a, Y > b\}^{\complement}) \\ &= 1 - P(\{X > a\}^{\complement} \cup \{Y > b\}^{\complement}) \\ &= 1 - P(\{X \le a\} \cup \{Y \le b\}) \\ &= 1 - P\{X \le a\} - P\{Y \le b\} + P\{X \le a, Y \le b\} \\ &= 1 - F_X(a) - F_Y(b) + F(a,b). \end{aligned}$$

Proposition 6.2.

$$P\{a_1 < X \le a_2, b_1 < Y \le b_2\} = F(a_2,b_2) + F(a_1,b_1) - F(a_1,b_2) - F(a_2,b_1),$$

whenever $a_1 < a_2$, $b_1 < b_2$.

Proposition 6.3. When X and Y are both discrete random variables, it is convenient to define the *joint probability mass function* of X and Y by

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

The probability mass function of X can be obtained from $p(x,y)$ by

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

Similarly,

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

Definition 6.2. The random variables X and Y are *jointly continuous* if there exists a function $f(x,y)$, defined for all x and y, having the property that, for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane),

$$P\{(X,Y) \in C\} = \iint_{(x,y)\in C} f(x,y)\,dx\,dy.$$

The function $f(x,y)$ is called the *joint probability density function* of X and Y. If A and B are any sets of real numbers, then, by defining $C = \{(x,y) : x \in A, y \in B\}$, we have

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

Proposition 6.4. If X and Y are jointly continuous, they are individually continuous, and their probability density functions can be obtained as follows:

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

where

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

is thus the probability density function of X. Similarly, the probability density function of Y is given by

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

Definition 6.3. The joint probability distributions for n random variables in exactly the same manner as we did for $n = 2$. For instance, the joint cumulative probability distribution function $F(a_1, a_2, \ldots, a_n)$ of the n random variables $X_1, X_2, \ldots, X_n$ is defined by

$$F(a_1, a_2, \ldots, a_n) = P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\}.$$

Further, the n random variables are said to be *jointly continuous* if there exists a function $f(x_1, x_2, \ldots, x_n)$, called the *joint probability density function*, such that, for any set C in n-space,

$$P\{(X_1, X_2, \ldots, X_n) \in C\} = \iint \cdots \int_{(x_1,x_2,\ldots,x_n)\in C} f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2 \cdots dx_n.$$

In particular, for any n sets of real numbers $A_1, A, \ldots, A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} = \int_{A_n} \int_{A_{n-1}} \cdots \int_{A_1} f(x_1, x_2, \ldots, x_n)\,dx_1 \cdots dx_{n-1} dx_n.$$

6.2 Independent Random Variables

Definition 6.4. The random variables X and Y are said to be *independent* if, for any two sets of real numbers A and B,

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}. \tag{7}$$

In other words, X and Y are independent if, for all A and B, the events $E_A = \{X \in A\}$ and $F_B = \{Y \in B\}$ are independent.
It can be shown by using the three axioms of probability that (7) will follow if and only if, for all a, b,

$$P\{X \le a, Y \le b\} = P\{X \le a\}P\{Y \le b\}.$$

Hence, in terms of the joint distribution function F of X and Y, X and Y are independent if

$$F(a,b) = F_X(a)F_Y(b) \quad \text{for all } a, b.$$

When X and Y are discrete random variables, the condition of independence (7) is equivalent to

$$p(x,y) = p_X(x)p_Y(y) \quad \text{for all } x, y.$$

In the jointly continuous case, the condition of independence is equivalent to

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

Proposition 6.5. The continuous (discrete) random variables X and Y are independent if and only if their joint probability density (mass) function can be expressed as

$$f_{X,Y}(x,y) = h(x)g(y) \quad -\infty < x < \infty, -\infty < y < \infty.$$

Definition 6.5. In general, the n random variables $X_1, X_2, \ldots, X_n$ are said to be independent if, for all sets of real numbers $A_1, A_2, \ldots, A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\}.$$

As before, it can be shown that this condition is equivalent to

$$P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\} = \prod_{i=1}^{n} P\{X_i \le a_i\} \quad \text{for all } a_1, a_2, \ldots, a_n.$$

Finally, it can be shown an infinite collection of random variables is independent if every finite subcollection of them is independent.

6.3 Sums of Independent Random Variables

Definition 6.6. Suppose that X and Y are independent, continuous random variables having probability density functions f_X and f_Y. The cumulative distribution function of $X + Y$ is obtained as follows:

$$\begin{aligned} F_{X+Y}(a) &= P\{X + Y \le a\} \\ &= \iint_{x+y \le a} f_X(x) f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)\, dx f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\, dy. \end{aligned}$$

The cumulative distribution function F_{X+Y} is called the *convolution* of the distributions F_X and F_Y (the cumulative distribution functions of X and Y, respectively). By differentiating the above equation, we find that the probability density function f_{X+Y} of $X + Y$ is given by

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} F_X(a-y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy. \end{aligned}$$

6.3.1 Identically Distributed Uniform Random Variables

Example 6.1 (Sum of two independent uniform random variables)**.** If X and Y are independent random variables, both uniformly distributed on $(0, 1)$, the probability density of $X + Y$ is

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy = \int_0^1 f_X(a-y)\, dy = \begin{cases} a & 0 \le a \le 1, \\ 2-a & 1 < a < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Because of the shape of its density function, the random variable $X + Y$ is said to have a *triangular* distribution.

Proposition 6.6. Suppose that $X_1, X_2, \ldots, X_n$ are independent uniform$(0,1)$ random variables, and let

$$F_n(x) = P\{X_1 + \cdots + X_n \le x\}.$$

Whereas a general formula for $F_n(x)$ is messy, it has a particularly nice form when $x \le 1$. Indeed, we now use mathematical induction to prove that

$$F_n(x) = \frac{x^n}{n!}, \quad 0 \le x \le 1.$$

Because the proceeding equation is true for $n = 1$, assume that

$$F_{n-1}(x) = \frac{x^{n-1}}{(n-1)!}, \quad 0 \le x \le 1.$$

Now, writing

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n,$$

and using the fact that the X_i are all nonnegative, we see that, for $0 \le x \le 1$,

$$F_n(x) = \int_0^1 F_{n-1}(x-y) f_{X_n}(y)\, dy = \frac{1}{(n-1)!} \int_0^x (x-y)^{n-1}\, dy = \frac{x^n}{n!},$$

which completes the proof.

6.3.2 Gamma Random Variable

Proposition 6.7. It is known that the density of a gamma random variable has the form

$$f(y) = \frac{\lambda e^{-\lambda y} (\lambda y)^{t-1}}{\Gamma(t)}, \quad 0 < y < \infty.$$

Then if X and Y are independent gamma random variables with respective parameters (s, λ) and (t, λ), then $X + Y$ is a gamma random variable with parameters $(s + t, \lambda)$. It is now a simple matter to establish, by using the preceding proposition and induction, that if $X_i, i = 1, \ldots, n$ are independent gammma random variables with respective parameters $(t_i, \lambda), i = 1, \ldots, n$, then $\sum_{i=1}^{n} X_i$is gamma with parameters $\left(\sum_{i=1}^{n} t_i, \lambda\right)$.

Proposition 6.8. If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then $Y = \sum_{i=1}^{n} Z_i^2$ is said to have the *chi-squared* (sometimes seen as χ^2) distribution with n degrees of freedom. Let us compute the density function of Y. When $n = 1, Y = Z_1^2$, we see that its probability density function is given by

$$f_{Z^2}(y) = \frac{1}{2\sqrt{y}}[f_Z(\sqrt{y}) + f_Z(-\sqrt{y})] = \frac{1}{2\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-y/2} = \frac{1/2 e^{-y/2} (y/2)^{1/2-1}}{\sqrt{\pi}}$$

We can recognize the preceding as the gamma distribution with parameters $(1/2, 1/2)$. But since each Z_i^2 is gamma$(1/2, 1/2)$, it follows that the χ^2 distribution with n degrees of freedom is the gamma distribution with parameters $(n/2, 1/2)$ and hence has a probability density function given by

$$f_{\chi^2}(y) = \frac{\frac{1}{2} e^{-y/2} \left(\frac{y}{2}\right)^{n/2-1}}{\Gamma\left(\frac{n}{2}\right)} = \frac{e^{-y/2} y^{n/2-1}}{2^{n/2} \Gamma\left(\frac{n}{2}\right)}, \quad y > 0.$$

6.3.3 Normal Random Variables

Proposition 6.9. If $X_i, i = 1, \ldots, n$, are independent random variables that are normally distributed with respective parameters $\mu_i, \sigma_i^2, i = 1, \ldots, n$, then $\sum_{i=1}^{n} X_i$ is normally distributed with parameters $\sum_{i=1}^{n} \mu_i$ and $\sum_{i=1}^{n} \sigma_i^2$.

6.3.4 Poisson and Binomial Random Variables

Proposition 6.10. If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, the distribution of $X + Y$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Proposition 6.11. Let X and Y be independent binomial random variables with respective parameters (n, p) and (m, p). Then $X + Y$ has a binomial distribution with parameters $(n + m, p)$.

6.3.5 Geometric Random Variables

Proposition 6.12. Let $X_1, \ldots, X_n$ be independent geometric random variables, with X_i having parameters p_i for $i = 1, \ldots, n$. If all the p_i are distinct, then, for $k \geq n$,

$$P\{S_n = k\} = \sum_{i=1}^{n} p_i q_i^{k-1} \prod_{j \neq i} \frac{p_j}{p_j - p_i}.$$

6.4 Conditional Distributions: Discrete Case

Definition 6.7. If X and Y are discrete random variables, define the conditional probability mass function of X given that $Y = y$ by

$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= \frac{P\{X = x, Y = y\}}{P\{Y = y} \\ &= \frac{P\{X = x\}P\{Y = y\}}{P\{Y = y} \\ &= \frac{p(x, y)}{p_Y(y)}, \end{aligned}$$

for all values of y such that $p_Y(y) > 0$. Similarly, the conditional probability distribution function of X given that $Y = y$ is defined, for all y such that $p_Y(y) > 0$, by

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

6.5 Conditional Distributions: Continuous Case

Definition 6.8. If X and Y have a joint probability density function $f(x, y)$, then the conditional probability density function of X given that $Y = y$ is defined, for all values of y such that $f_Y(y) > 0$, by

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

To motivate this definition, multiply the left-hand side by dx and the right-hand side by $dxdy)/dy$ to obtain

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

The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable. That is, if X and Y are jointly continuous, then, for any set A,

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\, dx.$$

6.6 Order Statistics

Definition 6.9. Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed continuous random variable having a common density f and distribution function F. Define

$$\begin{aligned} X_{(1)} &= \text{ smallest of } X_1, X_2, \ldots, X_n, \\ X_{(2)} &= \text{ second smallest of } X_1, X_2, \ldots, X_n, \\ &\vdots \\ X_{(j)} &= j^{\text{th}} \text{ smallest of } X_1, X_2, \ldots, X_n, \\ &\vdots \\ X_{(n)} &= \text{ largest of } X_1, X_2, \ldots, X_n. \end{aligned}$$

The ordered values $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are known as the *order statistics* corresponding to the random variables $X_1, X_2, \ldots, X_n$. In other words, $X_{(1)}, \ldots, X_{(n)}$ are the ordered values of $X_1, \ldots, X_n$. The joint density function of the order statistics is obtained by noting that the order statistics $X_{(1)}, \ldots, X_{(n)}$ will take on the values $x_1 \le x_2 \le \cdots \le x_n$ if and only if, for some permutation $(i_1, i_2, \ldots, i_n)$ of $(1, 2, \ldots n)$,

$$X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}.$$

Since, for any permutation $(i_1, \ldots, i_n)$ of $(1, 2, \ldots, n)$,

$$\begin{aligned} P\left\{x_{i_1} - \frac{\varepsilon}{2} < X_1 < x_{i_1} + \frac{\varepsilon}{2}, \ldots, x_{i_n} - \frac{\varepsilon}{2} < X_n < x_{i_n} + \frac{\varepsilon}{2}\right\} &\approx \varepsilon^n f_{X_1,\ldots X_n}(x_{i_1}, \ldots, x_{i_n}) \\ &= \varepsilon^n f(x_{i_1}) \cdots f(x_{i_n}) \\ &= \varepsilon^n f(x_1) \cdots f(x_n), \end{aligned}$$

it follows that, for $x_1 < x_2 < \ldots < x_n$,

$$P\left\{x_1 - \frac{\varepsilon}{2} < X_1 < x_1 + \frac{\varepsilon}{2}, \ldots, x_n - \frac{\varepsilon}{2} < X_n < x_n + \frac{\varepsilon}{2}\right\} = n!\varepsilon^n f(x_1) \cdots f(x_n).$$

Dividing by ε^n and letting $\varepsilon \to 0$ yields

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f(x_1) \cdots f(x_n), \quad x_1 < x_2 < \cdots < x_n,$$

which argues that in order for the vector $\langle X_{(1)}, \ldots X_{(n)} \rangle$ to equal $\langle x_1, \ldots, x_n \rangle$, it is necessary and sufficient for $\langle X_1, \ldots, X_n \rangle$ to equal one of the $n!$ permutations of $\langle x_1, \ldots, x_n \rangle$. Since the probability (density) that $\langle X_1, \ldots, X_n \rangle$ equals any given permutation of $\langle x_1, \ldots, x_n \rangle$ is just $f(x_1) \cdots f(x_n)$.

6.7 Joint Probability Distribution of Functions of Random Variables

Definition 6.10. Let X_1 and X_2 be jointly continuous random variables with joint probability density function f_{X_1,X_2}. It is sometimes necessary to obtain the joint distribution of the random variables Y_1 and Y_2, which arise as function of X_1 and X_2. Specifically, suppose that $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2.
Assume that the functions g_1 and g_2 satisfy the following conditions:

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

2. The functions g_1 and g_2 have continuous partial derivatives at all points (x_1, x_2) and are such that the 2×2 determinant
$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} \equiv \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \neq 0$$
at all points (x_1, x_2).

Under these two conditions, it can be shown that the random variables Y_1 and Y_2 are jointly continuous with joint density function given by

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

where $x_1 = h_1(y_1,y_2), x_2 = h_2(y_1,y_2)$.

Definition 6.11. When the joint density function of the n random variables $X_1, X_2, \ldots, X_n$ is given and we want to compute the joint density function of $Y_1, Y_2, \ldots, Y_n$, where

$$Y_1 = g_1(X_1,\ldots,X_n), \quad Y_2 = g_2(X_1,\ldots,X_n), \ldots, \quad Y_n = g_n(X_1,\ldots,X_n),$$

the approach is the same – namely, we assume that the functions g_i have continuous partial derivatives and that the Jacobian determinant.

$$J(x_1,\ldots,x_n) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix} \neq 0,$$

at all points $(x_1,\ldots,x_n)$. Furthermore, we suppose that the equation $y_1 = g_1(x_1,\ldots,x_n), y_2 = g_2(x_1,\ldots,x_n), \ldots, y_n = g_n(x_1,\ldots,x_n)$ have a unique solution, say, $x_1 = h_1(y_1,\ldots,y_n), \ldots, x_n = h_n(y_1,\ldots,y_n)$. Under these assumptions, the joint density function of the random variables Y_i is given by

$$f_{Y_1,\ldots,Y_n}(y_1,y_2,\ldots,y_n) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)|J(x_1,\ldots,x_n)|^{-1},$$

where $x_i = h_i(y_1,\ldots,y_n)$.

6.8 Exchangeable Random Variables

Definition 6.12. The random variables $X_1, X_2, \ldots, X_n$ are said to be *exchangeable* if, for every permutation $i_1,\ldots,i_n$ of the integers $1,\ldots,n$,

$$P\{X_{i_1} \le x_1, X_{i_2} \le x_2, \ldots, X_{i_n} \le x_n\} = P\{X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\},$$

for all $x_1, x_2, \ldots, x_n$. That is, the n random variables are exchangeable if their joint distribution is the same to no matter in which order the variables are observed. Discrete random variables will be exchangeable if

$$P\{X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_n} = x_n\} = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\},$$

for all permutations $i_1,\ldots,i_n$ and all values $x_1,\ldots,x_n$. This is equivalent to stating that $p(x_1,x_2,\ldots,x_n) = P\{X_1 = x_1,\ldots,X_n = x_n\}$ is a symmetric function of the vector $(x_1,\ldots,x_n)$, which means that its value does not change when the values of the vector are permuted.

Proposition 6.13. If $X_1, X_2, \ldots, X_n$ are exchangeable, then each X_i has the same probability distribution. For instance, if X and Y are exchangeable discrete random variables, then

$$P\{X = x\} = \sum_y P\{X = x, Y = y\} = \sum_y P\{X = y, Y = x\} = P\{Y = x\}.$$

7 Properties of Expectation

7.1 Introduction

Proposition 7.1. Since $E[X]$ is a weighted average of the possible values of X, it follows that if X must lie between a and b, then so must its expected value. That is, if

$$P\{a \le X \le b\} = 1,$$

then

$$a \le E[X] \le b.$$

7.2 Expectation of Sums of Random Variables

Proposition 7.2. If X and Y have a joint probability mass function $p(x, y)$, then

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

If X and Y have a joint probability density function $f(x, y)$, then

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

Proposition 7.3. Suppose that $E[X]$ and $E[Y]$ are both finite and let $g(X, Y) = X + Y$. Then, in the continuous case,

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y)\, dy\, dx + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} xf_X(x)\, dx + \int_{-\infty}^{\infty} yf_Y(y)\, dy \\ &= E[X] + E[Y]. \end{aligned}$$

The same result holds in general; thus, whenever $E[X_i], i = 1, 2, \ldots, n$ are finite,

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n].$$

7.2.1 Obtaining Bounds from Expectation via the Probabilistic Method

Proposition 7.4. The probabilistic method is a technique for analyzing the properties of the elements of a set by introducing on the set and then studying an element chosen according to those probabilities. Let f be a function on the elements of a finite set $\mathcal{S}$, and suppose that we are interested in

$$m = \max_{s \in \mathcal{S}} f(s).$$

A useful lower bounds for m can often be obtained by letting $\mathcal{S}$ be a random element of $\mathcal{S}$ for which the expected value of $f(\mathcal{S})$ is computable and then noting that $m \geq f(\mathcal{S})$ implies that

$$m \geq E[f(\mathcal{S})],$$

with strict inequality if $f(\mathcal{S})$ is not a constant random variable. That is, $E[f(\mathcal{S})]$ is a lower bound on the maximum value.

7.2.2 The Maximum-Minimums Identity

Proposition 7.5. For arbitrary numbers $x_i, i = 1, \ldots, n$,

$$\max_i x_i = \sum_i x_i - \sum_{i<j} \min(x_i, x_j) + \sum_{i<j<k} \min(x_i, x_j, x_k) + \cdots + (-1)^{n+1} \min(x_1, \ldots, x_n).$$

Proposition 7.6. For any random variables $X_1, \ldots, X_n$,

$$\max_i X_i = \sum_i X_i - \sum_{i<j} \min(X_i, X_j) + \cdots + (-1)^{n+1} \min(X_1, \ldots, X_n).$$

Taking expectations of both sides of this equality yields the following relationship between the expected value of the maximum and those of the partial minimums:

$$E\left[\max_i X_i\right] = \sum_i E[X_i] - \sum_{i<j} E[\min(X_i, X_j)] + \cdots + (-1)^{n+1} E[\min(X_1, \ldots, X_n)].$$

7.3 Moments of the Number of Events That Occur

Proposition 7.7. For given events $A_1, \ldots, A_n$, find $E[X]$, where X is the number of the number of these events that occur. The solution then involved defining an indicator I_i for event A_i such that

$$I_i = \begin{cases} 1, & \text{if } A_i \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Because

$$X = \sum_{i=1}^{n} I_i,$$

we obtained the result

$$E[X] = E\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(A_i).$$

Now suppose we are interested in the number of *pairs* of events that occur. Because $I_i I_j$ will equal 1 if both A_i and A_j occur, and will equal 0 otherwise, it follows that the number of pairs is equal to $\sum_{i<j} I_i I_j$. But because X is the number of events that occur, it also follows that the number of pairs of events that occur is $\binom{X}{2}$. Consequently,

$$\binom{X}{2} = \sum_{i<j} I_i I_j.$$

where there are $\binom{n}{2}$ terms in the summation. Taking expectations yields

$$E\left[\binom{X}{2}\right] = \sum_{i<j} E[I_i I_j] = \sum_{i<j} P(A_i A_j),$$

or

$$E\left[\frac{X(X-1)}{2}\right] = \sum_{i<j} P(A_i A_j),$$

giving that

$$E[X^2] - E[X] = 2\sum_{i<j} P(A_i A_j),$$

which yields $E[X^2]$, and thus $\mathrm{Var}(X) = E[X^2] - (E[X])^2$. Moreover, by considering the number of distinct subsets of k events that all occur, we see that

$$\binom{X}{k} = \sum_{i_1 < i_2 < \cdots < i_k} I_{i_1} I_{i_2} \cdots I_{i_k}.$$

Taking expectations gives the identity

$$E\left[\binom{X}{k}\right] = \sum_{i_1 < i_2 < \cdots < i_k} E[I_{i_1} I_{i_2} \cdots I_{i_k}] = \sum_{i_1 < i_2 < \cdots < i_k} P(A_{i_1} A_{i_2} \cdots A_{i_k}).$$

7.4 Covariance, Variance of Sums, and Correlations

Proposition 7.8. If X and Y are independent, then, for any functions h and g,

$$E[g(X)h(Y)] = E[g(X)]E[h(X)].$$

Definition 7.1. The covariance between X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

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

Proposition 7.9. (i) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

(ii) $Cov(X,X) = \mathrm{Var}(X)$.

(iii) $\mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$.

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

Proposition 7.10.

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

Definition 7.2. The correlation of two random variables X and Y, denoted by $\rho(X,Y)$, is defined, as long as $\mathrm{Var}(X)\,\mathrm{Var}(Y)$ is positive, by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

It can be shown that

$$-1 \le \rho(X,Y) \le 1.$$

7.5 Conditional Expectation

7.5.1 Definitions

Definition 7.3. If X and Y are jointly discrete random variables, then the conditional probability mass function of X, given that $Y = y$, is defined, for all y such that $P\{Y = y\} > 0$, by

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

It is therefore natural to define, in this case, the conditional expectation of X given that $Y = y$, for all values of y such that $p_Y(y) > 0$, by

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

Similarly, if X and Y are jointly continuous with a joint probability density function $f(x,y)$, then the conditional probability density of X, given that $Y = y$, is defined, for all values of y such that $f_Y(y) > 0$, by

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

It is natural, in this case, to define the conditional expectation of X, given that $Y = y$, by

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

We also have

$$E[g(X)|Y = y] = \begin{cases} \sum_x g(x) p_{X|Y}(x|y) & \text{in the discrete case,} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\,dx & \text{in the continuous case.} \end{cases}$$

and

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

7.5.2 Computing Expectation by Conditioning

Proposition 7.11.

$$E[X] = E[E[X|Y]].$$

If Y is a discrete random variable, then the above equation states that

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

whereas if Y is continuous with density $f_Y(y)$, then the equation states

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

7.5.3 Conditional Variance

Definition 7.4. The conditional variance of X given that $Y = y$ is defined by

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

Proposition 7.12. The conditional variance formula

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

7.6 Conditional Expectation and Prediction

Proposition 7.13.

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2].$$

7.7 Moment Generating Functions

Definition 7.5. The moment generating function $M(t)$ of the random variable X is defined for all real values of t by

$$M(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx}p(x) & \text{if } X \text{ is discrete with mass function } p(x), \\ \int_{-\infty}^{\infty} e^{tx}f(x)\,dx & \text{if } X \text{ continuous with density } f(x). \end{cases}$$

We call $M(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $M(t)$ and then evaluate the result at $t = 0$. For example,

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

In general, the nth derivative of $M(t)$ is given by

$$M^n(t) = E[X^n e^{tX}], \quad n \geq 1,$$

implying that

$$M^n(0) = E[X^n], \quad n \geq 1.$$

Proposition 7.14.

	$p(x)$	$M(t)$	Mean	Variance
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$(pe^t+1-p)^n$	np	$np(1-p)$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\exp[\lambda(e^t-1)]$	λ	λ
Geometric(p)	$p(1-p)^{x-1}$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial(r,p)	$\binom{n-1}{r-1}p^r(1-p)^{n-r}$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

7.7.1 Joint Moment Generating Functions

It is also possible to define the joint moment generating function of two or more random variables. This is done as follows: For any n random variables $X_1,\dots,X_n$, the joint momment generating function, $M(t_1,\dots,t_n)$, is defined, for all real values of $t_1,\dots,t_n$, by

$$M(t_1,\dots,t_n)=E[e^{t_1X_1+\cdots+t_nX_n}].$$

The individual moment generating functions can be obtained from $M(t_1,\dots,t_n)$ by letting all but one of the t_j's be 0. That is,

$$M_{X_i}(t)=E[e^{tX_i}]=M(0,\dots,0,t,0,\dots,0),$$

where the t is in the ith place. It can be proven that the joint moment generating functoin $M(t_1,\dots,t_n)$ uniquely determines the joint distribution of $X_1,\dots,X_n$. This result can then be used to prove that the n random vairables $X_1,\dots,X_n$ are independent if and only if

$$M(t_1,\dots,t_n)=M_{X_1}(t_1)\cdots M_{X_n}(t_n).$$

7.8 Additional Properties of Normal Random Variables

7.8.1 The Multivariate Normal Distribution

Let $Z_1,\dots,Z_n$ be a set of n independent unit normal random variables. If, for some constants $a_{ij}, 1\le i\le m, 1\le j\le n$, and $\mu_i, 1\le i\le m$,

$$\begin{aligned}
X_1&=a_{11}Z_1+\cdots+a_{1n}Z_n+\mu_1\\
X_2&=a_{21}Z_1+\cdots+a_{2n}Z_n+\mu_2\\
&\vdots\\
X_i&=a_{i1}Z_1+\cdots+a_{in}Z_n+\mu_i\\
&\vdots\\
X_m&=a_{m1}Z_1+\cdots+a_{mn}Z_n+\mu_m,
\end{aligned}$$

then the random vairables $X_1,\dots,X_m$ are said to have a multivariate normal distribution. From the fact that the sum of independent normal random variables is itself a normal random variable, it follows that each X_i is a normal random variable with mean and variance given, respectively, by

$$\begin{aligned}
E[X_i]&=\mu_i\\
\mathrm{Var}(X_i)&=\sum_{j=1}^{n}a_{ij}^2.
\end{aligned}$$

Let us now consider

$$M(t_1,\ldots,t_m) = E[\exp\{t_1X_1 + \cdots + t_mX_m\}],$$

the joint moment generating function of $X_1,\ldots,X_m$. The first thing to note is that since $\sum_{i=1}^{m} t_iX_i$ is itself a linear combination of the independent normal random variables $Z_1,\ldots,Z_n$, it is also normally distributied. Its mean and variance are

$$E[\sum_{i=1}^{m} t_iX_i] = \sum_{i=1}^{m} t_i\mu_i,$$

and

$$\text{Var}\left(\sum_{i=1}^{m} t_iX_i\right) = \text{Cov}\left(\sum_{i=1}^{m} t_iX_i, \sum_{j=1}^{m} t_jX_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} t_it_j\,\text{Cov}(X_i,X_j).$$

Now, if Y is a normal random variable with mean μ and variance σ^2, then

$$E[e^Y] = M_Y(t)|_{t=1} = e^{\mu+\sigma^2/2}.$$

Thus,

$$M(t_1,\ldots,t_m) = \exp\left\{\sum_{i=1}^{m} t_i\mu_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} t_it_jCov(X_i,X_j)\right\},$$

which show that the joint distribution of $X_1,\ldots,X_m$ is completely determined from a knowledge of the values of $E[X_i]$ and $\text{Cov}(X_i,X_j), i,j = 1,\ldots,m$. It can be shown that when $m = 2$, the multivariate normal distribution reduces to the bivariate normal.

7.8.2 The Joint Distribution of the Sample Mean and Sample Variance

Let $X_1,\ldots,X_m$ be independent normal random variables, each with mean μ and variance σ^2. Let $\overline{X} = \sum_{i=1}^{n} X_i/n$ denote their sample mean. Since the sum of independent normal random variables is also a normal random variable, it follows that $\overline{X}$ is a normal random variable with expected value μ and variance σ^2/n. Now recall that

$$\text{Cov}(\overline{X}, X_i - \overline{X}) = 0, i = 1,\ldots,n.$$

Also, note that since $\overline{X}, X_1-\overline{X}, X_2-\overline{X},\ldots,X_n-\overline{X}$ are all linear combinations of the independent standard normals $(X_i-\mu)/\sigma, i = 1,\ldots,n$, it follows that $\overline{X}, X_i - \overline{X}, i = 1,\ldots,n$ has a joint distribution that is multivariate normal. If we let Y be a normal random variable, with mean μ and variance σ^2/n, that is independent of the $X_i, i = 1,\ldots,n$, then $Y, X_i - \overline{X}, i = 1,\ldots,n$ also has a multivariate normal distribution and, indeed, because of the above equation, has the same expected values and covariance as the random variables $\overline{X}, X_i - \overline{X}, i = 1,\ldots,n$. BUt since multivariate normal distribution is determinded completely by its expected values and covariance, it follows that $Y, X_i-\overline{X}, i = 1,\ldots,n$ and $\overline{X}, X_i-\overline{X}, i = 1,\ldots n$ have the same joint distribution, thus showing that $\overline{X}$ is independent of the sequence of deviations $X_i-\overline{X}, i = 1,\ldots,n$. Since $\overline{X}$ is independent of the sequence of deviation $X_i-\overline{X}, i = 1,\ldots,n$, it is also independent of the sample variance $S^2 = \sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$.

Since we already know that $\overline{X}$ is normal with mean μ and variance σ^2/n, it remains only to determine the distribution of S^2. To accomplish this, recall, the algebraic identity

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

Upon dividing the preceding equation by σ^2, we obtain

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2.$$

Now,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

is the sum of the squares of n independent standard normal random variables and so is a chi-squared random variable with n degrees of freedom. Hence, its moment generating functoin is $(1-2t)^{-n/2}$. Also, because

$$\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

is the squre of a standard normal variable, it is a chi-squared random variable with 1 degree of freedom, and so has moment generating function $(1-2t)^{-1/2}$. Now we have seen previously that the two random variables on the left side are independent. Hence, as the moment generating function of the sum of independent random variables is equal to the product of their individual moment generating functions, we have

$$E[e^{t(n-1)S^2/\sigma^2}](1-2t)^{-1/2}=(1-2t)^{-n/2},$$

or

$$E[e^{t(n-1)S^2/\sigma^2}]=(1-2t)^{-(n-1)/2}.$$

But as $(1-2t)^{-(n-1)/2}$ is the moment generating function of a chi-squared random variable with $n-1$ degrees of freedom, we can conclude, since the moement generating function uniquely determines the distribution of the random variable, it follows that that is the distribution of $(n-1)S^2/\sigma^2$.

Proposition 7.15. If $X_1,\ldots,X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then the sample mean $\overline{X}$ and the sample variance S^2 are independent. $\overline{X}$ is a normal random variable with mean μ and variance σ^2/n; $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degrees of freedom.

7.9 General Definition of Expectation

Expectation for random variables that are neither discrete nor continuous. Let X be a Bernoulli random variable with parmeter $p=1/2$, and let Y be a uniformly distributed random variable over the interval $[0,1]$. Furthermore, suppose that X and Y are independent, and define the new random variable W by

$$W=\begin{cases}X & \text{if } X=1,\\ Y & \text{if } X\neq 1.\end{cases}$$

Clearly, W is neither a discrete (since its set of possible values, $[0,1]$, is uncountable) nor a continuous \$(since $P\{W=1\}=1/2$) random variable.
In order to define the expectation of an arbitary random variable, we require the notion of a Stieltjes intergral. Before defining this intergral, let us recall that, for any function g, $\int_a^b g(x)\,dx$ is defined by

$$\int_a^b g(x)\,dx=\lim\sum_{i=1}^{n}g(x_i)(x_i-x_{i-1}),$$

where the limit is taken over all $a=x_0<x_1<x_2<\cdots<x_n=b$ as $n\to\infty$ and where $\max_{i=1,\ldots,n}(x_i-x_{i-1})\to 0$.
For any distribution function F, we define the Stieltjes integral of the nonnegative function g over the interval $[a,b]$ by

$$\int_a^b g(x)\,dF(x)=\lim\sum_{i=1}^{n}g(x_i)[F(x_i)-F(x_{i-1})],$$

where, as beofore, the limit is taken over all $a=x_0<x_1<\cdots<x_n=b$ as $n\to\infty$ and where $\max_{i=1,\ldots,n}(x_i-x_{i-1})\to 0$. Further, we define the Stieltjes integral over the whole real line by

$$\int_{-\infty}^{\infty}g(x)\,dF(x)=\lim_{\substack{a\to-\infty\\ b\to+\infty}}\int_a^b g(x)\,dF(x).$$

Finally, if g is not a nonnegative function, we define g^+ and g^- by

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

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

Because $g(x) = g^+(x) - g^- g(x)$ and g^+ and g^- are both nonnegative functions, it is natural to define

$$\int_{-\infty}^{\infty} g(x)\, dF(x) = \int_{-\infty}^{\infty} g^+\, dF(x) - \int_{-\infty}^{\infty} g^-\, dF(x),$$

and we say that $\int_{-\infty}^{\infty} g(x)\, dF(x)$ exists as long as $\int_{-\infty}^{\infty} g^+(x)\, dF(x)$ and $\int_{-\infty}^{\infty} g^-(x)\, dF(x)$ are not both equal to $+\infty$.
If X is an arbitrary random variable having cumulative distribution F, we define the expected value of X by

$$E[X] = \int_{-\infty}^{\infty} x\, dF(x).$$

It can be shown that if X is a discrete random variable with mass function $p(x)$, then

$$\int_{-\infty}^{\infty} x dF(x) = \sum_{x:p(x)>0} xp(x),$$

whereas if X is a continuous random variable with density function $f(x)$, then

$$\int_{-\infty}^{\infty} x\, dF(x) = \int_{-\infty}^{\infty} xf(x)\, dx.$$