1 Graduate Probability

This chapter gives a comprehensive and usable treatment of what is considered core graduate probability. It gives definitions, the main results, and examples. It can be used to teach an independent graduate course on probability, and it also serves the purpose of providing the essential background in probability, without which serious statistical theory cannot be understood. The chapter also comes with many exercises for additional practice and review. For inspiring introduction to probability, we recommend Feller (1968, 1971).

1.1 Sample Spaces and Events

Probability is a universally accepted tool for expressing degrees of confidence or doubt about some proposition in the presence of incomplete information or uncertainty. By convention, probabilities are calibrated to a scale of 0 to 1; assigning something a zero probability amounts to expressing the belief that we consider it impossible, while assigning a probability of one amounts to considering it a certainty. Most propositions fall somewhere in between.

Treatment of probability theory starts with the consideration of a *sample space*. The sample space is the set of all possible outcomes in some physical experiment. For example, if a coin is tossed twice and after each toss the face that shows is recorded, then the possible outcomes of this particular coin tossing experiment, say ξ are HH, HT, TH, TT, with H denoting the occurrence of heads and T denoting the occurrence of tails. We call

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

the sample space of the experiment ξ.

We instinctively understand what an experiment means. An experiment is a physical enterprise that can, in principle, be repeated infinitely many times independently. For example,

ξ = Choose a number between 1 and 10 and record the value of the chosen number,

ξ = Toss a coin three times and record the sequence of outcomes,

ξ = Arrange five people in a line up for taking a picture,

ξ = Distribute 52 cards in a deck of cards to four players so that each player gets 13 cards,

ξ = Count the number of calls you receive on your cell phone on a given day,

ξ = Measure someone's blood pressure are all activities that can, in principle, be repeated and are experiments. Notice that for each of these experiments, the ultimate outcome is uncertain until the experiment has actually been performed. For example, in the first experiment above, the number that ultimately gets chosen could be any of $1, 2, \cdots, 10$. The set of all these possible outcomes constitutes the sample space of the experiment. Individual possible outcomes are called the *sample points* of the experiment.

In general, a sample space is a general set Ω, finite or infinite. An easy example where the sample space Ω is infinite is to toss a coin until the first time heads show up and record the number of the trial at which the first head showed up. In this case, the sample space Ω is the *countably infinite* set

$$\Omega = \{1, 2, 3, \cdots\}.$$

Sample spaces can also be *uncountably infinite*; for example, consider the experiment of choosing a number *at random* from the interval $[0, 1]$. The sample space of this experiment is $\Omega = [0, 1]$. In this case, Ω is an uncountably infinite set. In all cases, individual elements of a sample space will be denoted as ω. The first task is to define *events* and to explain what is the meaning of the probability of an event.

Definition 1.1. Let Ω be the sample space of an experiment ξ. Then any subset A of Ω, including the empty set ϕ and the entire sample space Ω is called an *event*. Events may contain even one single sample point ω, in which case the event is a *singleton set* $\{\omega\}$.

1.2 Set Theory Notation and Axioms of Probability

Set theory notation will be essential in our treatment of events, because events are sets of sample points. So, at this stage, it might be useful to recall the following common set theory notation:

Given two subsets A, B of a set Ω,
$A^c =$ set of points of Ω not in A
$A \cap B =$ set of points of Ω which are in both A and B
$A \cup B =$ set of points of Ω which are in at least one of A, B
$A \triangle B =$ set of points of Ω which are in exactly one of A, B
$A - A \cap B =$ set of points of Ω which are in A but not in B.
If Ω is the sample space of some experiment, and A, B are events in that experiment, then the probabilistic meaning of this notation would be as follows:
Given two events, A, B in some experiment,
$A^c =$ A does not happen
$A \cap B =$ Both A, B happen; the notation AB is also sometimes used to mean $A \cap B$
$A \cup B =$ At least one of A, B happens
$A \triangle B =$ Exactly one of A, B happens
$A - A \cap B =$ A happens, but B does not

Example 1.1. This is to help interpret events of various types using the symbols of set operation. This becomes useful for calculating probabilities by setting up the events in

set theory notation and then use a suitable rule or formula. For example,

$$\text{At least one of A, B, C} = A \cup B \cup C;$$

$$\text{Each of A, B, C} = A \cap B \cap C;$$

$$\text{A, but not B or C} = A \cap B^c \cap C^c;$$

$$\text{A, and exactly one of B, C} = A \cap (B \triangle C) = (A \cap B \cap C^c) \cup (A \cap C \cap B^c);$$

$$\text{None of A, B, C} = A^c \cap B^c \cap C^c.$$

It is also useful to recall the following elementary facts about set operations.

Proposition $a) A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
$b) A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
$c) (A \cup B)^c = A^c \cap B^c$;
$d) (A \cap B)^c = A^c \cup B^c$.
Here is a definition of what counts as a legitimate probability on events.

Definition 1.2. Given a sample space Ω, a probability or a *probability measure* on Ω is a function P on subsets of Ω such that,

$$(a) P(A) \geq 0 \,\text{for any} A \subseteq \Omega;$$

$$(b) P(\Omega) = 1;$$

$$(c)\,\text{Given disjoint subsets}\, A_1, A_2, \cdots \text{ of } \Omega, P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Property (c) is known as *countable additivity*. Note that it is not something that can be proved, but it is like an assumption or an *axiom*. In our experience, we have seen that operating as if the assumption is correct leads to useful and credible answers in many problems, and so we accept it as a reasonable assumption. One important point is that finite additivity is subsumed in countable additivity, i.e., if there are some finite number m of disjoint subsets $A_1, A_2, \cdots, A_m$ of Ω, then $P(\cup_{i=1}^{m} A_i) = \sum_{i=1}^{m} P(A_i)$. Also, it is useful to note that the last two conditions in the definition of a probability measure imply that $P(\phi)$, the probability of the empty set or the *null event*, is zero.
One notational convention is that strictly speaking, for an event which is just a singleton set $\{\omega\}$, we should write $P(\{\omega\})$ to denote its probability. But to reduce clutter, we will simply use the more convenient notation $P(\omega)$.

One pleasant consequence of the axiom of countable additivity is the following basic result.

Theorem 1.1. Let $A_1 \supset A_2 \supset A_3 \supset \cdots$ be an infinite family of subsets of a sample space Ω such that $A_n \downarrow A$. Then, $P(A_n) \to P(A)$ as $n \to \infty$.

Proof: On taking the complements, $B_i = A_i^c, i \geq 1, B = A^c$, the result is equivalent to showing that if $B_1 \subset B_2 \subset B_3 \cdots, B_n \uparrow B$, then $P(B_n) \to P(B)$.

Decompose B_n for a fixed n into disjoint sets as $B_n = \cup_{i=1}^n (B_i - B_{i-1})$, where $B_0 = \phi$, and the difference notation $B_i - B_{i-1}$ means $B_i \cap B_{i-1}^c$. Therefore,

$$P(B_n) = \sum_{i=1}^{n} P(B_i - B_{i-1})$$

$$\Rightarrow \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i - B_{i-1}) = \sum_{i=1}^{\infty} P(B_i - B_{i-1}) = P(B),$$

as $\cup_{i=1}^{\infty} (B_i - B_{i-1}) = B$.

Next, the concept of equally likely sample points is a very fundamental one.

Definition 1.3. Let Ω be a finite sample space consisting of N sample points. We say that the sample points are *equally likely* if $P(\omega) = \frac{1}{N}$ for each sample point ω.

An immediate consequence, due to the addivity axiom, is the following useful formula.

Proposition Let Ω be a finite sample space consisting of N equally likely sample points. Let A be any event and suppose A contains n distinct sample points. Then

$$P(A) = \frac{n}{N} = \frac{\text{Number of sample points favorable to } A}{\text{Total number of sample points}}.$$

Example 1.2. (With and Without Replacement) Consider the experiment ξ where two numbers are chosen simultaneously *at random* from $\{0, 1, 2, \cdots, 9\}$. Since the numbers are chosen simultaneously, by implication they must be different; such sampling is called *without replacement* sampling. Probabilistically, without replacement sampling is also the same as drawing the two numbers, one at a time, with the restriction that the same number cannot be chosen twice. If the numbers are chosen one after the other, and the second number could be equal to the first number, then the sampling is called *with replacement* sampling. In this example, we consider without replacement sampling. Consider the events

$$A = \text{ The first chosen number is even;}$$

$$B = \text{ The second chosen number is even;}$$

$$C = \text{ Both numbers are even;}$$

$$D = \text{ At least one of the two numbers is even.}$$

The sample space $\Omega = \{01, 02, 03, \cdots, 96, 97, 98\}$ has $10 \times 9 = 90$ sample points. Suppose, due to the random or unbiased selection of the two numbers, we assign an equal probability,

$\frac{1}{90}$, for selecting any of the 90 possible pairs. Event A is favored by the sample points $\{01, 02, \cdots, 88, 89\}$; thus, A is favored by $5 \times 9 = 45$ sample points and so, $P(A) = 45/90 = .5$. Similarly, $P(B)$ is also .5. Event C is favored by those sample points which are in both A and B, i.e., in set theory notation, $C = A \cap B$. By direct listing, $A \cap B = \{02, 04, \cdots, 86, 88\}$; there are $5 \times 4 = 20$ such sample points, and so $P(C) = P(A \cap B) = 20/90 = 2/9$. On the other hand, event D is favored by those sample points which favor A or B or perhaps both; i.e., D is favored by sample points which favor at least one of A, B. In set theory notation $D = A \cup B$, and by direct listing, it is verified that $P(D) = P(A \cup B) = 70/90 = 7/9$. We notice that the collection of sample points which favor at least one of A, B can be found by writing the sample points in A, then writing on the sample points in B, and eventually taking out those sample points which got written twice, i.e., the sample points in $A \cap B$. So, we should have $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/2 - 2/9 = 7/9$, which is what we found by direct listing. Indeed, this is a general rule.

Addition Rule For any two events $A, B, P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

In more complicated experiments, it might be difficult or even impossible to manually list all the sample points. For example, if you toss a coin 20 times, the total number of sample points would be $2^{20} = 1048576$, which is larger than a million. Obviously we do not want to calculate probabilities in such an example by manual listing and manual counting.
Some facts about counting and basic combinatorics will be repeatedly useful in complex experiments. So it is useful to summarize them before we start using them.

Proposition. (a) Number of ways to linearly arrange n distinct objects when the order of arrangement matters $= n!$;
(b) Number of ways to choose r distinct objects from n distinct objects when the order of selection is important $= n(n-1)\cdots(n-r+1)$;
(c) Number of ways to choose r distinct objects from n distinct objects when the order of selection is not important $= \binom{n}{r} = \frac{n!}{r!(n-r)!}$;
(d) Number of ways to choose r objects from n distinct objects if the same object could be chosen repeatedly $= n^r$;
(e) Number of ways to distribute n distinct objects into k distinct categories when the order in which the distributions are made is not important, and n_i objects are to be allocated to the ith category $= \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots n_{k-1}}{n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}$.

Example 1.3. (**The shoe problem**). Suppose there are five pairs of shoes in a closet and four shoes are taken out at random. What is the probability that among the four that are taken out, there is at least one complete pair ?

The total number of sample points is $\binom{10}{4} = 210$. Since selection was done completely at random, we assume that all sample points are equally likely. At least one complete pair

would mean two complete pairs, or exactly one complete pair and two other nonconforming shoes. Two complete pairs can be chosen in $\binom{5}{2} = 10$ ways. Exactly one complete pair can be chosen in $\binom{5}{1}\binom{4}{2} \times 2 \times 2 = 120$ ways. The $\binom{5}{1}$ term is for choosing the pair that is complete; the $\binom{4}{2}$ term is for choosing two incomplete pairs, and then from each incomplete pair, one chooses the left or the right shoe. Thus, the probability that there will be at least one complete pair among the four shoes chosen is $(10 + 120)/210 = 13/21 = .62$.

Example 1.4. (**Five card Poker**). In five card poker, a player is given 5 cards from a full deck of 52 cards at random. Various named hands of varying degrees of rarity exist. In particular, we want to calculate the probabilities of $A =$ *two pairs* and $B =$ a *flush*. Two pairs is a hand with 2 cards each of 2 different denominations and the fifth card of some other denomination; a flush is a hand with 5 cards of the same suit, but the cards cannot be of denominations in a sequence.

Then, $P(A) = \binom{13}{2}[\binom{4}{2}]^2\binom{44}{1}/\binom{52}{5} = .04754$.

To find $P(B)$, note that there are ten ways to select five cards from a suit such that the cards are in a sequence, namely, $\{A, 2, 3, 4, 5\}, \{2, 3, 4, 5, 6\}, \cdots, \{10, J, Q, K, A\}$, and so, $P(B) = \binom{4}{1}\left(\binom{13}{5} - 10\right)/\binom{52}{5} = .00197$.

A major result in combinatorial probability is the *inclusion-exclusion formula*, which says the following.

Theorem 1.2. Let $A_1, A_2, \cdots, A_n$ be n general events. Let

$$S_1 = \sum_{i=1}^{n} P(A_i);\ S_2 = \sum_{1 \le i < j \le n} P(A_i \cap A_j);\ S_3 = \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k);\ \cdots$$

Then,

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots$$

$$+(-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

$$= S_1 - S_2 + S_3 - \cdots + (-1)^{n+1} S_n.$$

Proof: The theorem is proved by induction. Suppose it is known to be true for $n - 1$ general events $A_1, A_2, \cdots, A_{n-1}$. Define $A = \cup_{i=1}^{n-1} A_i$ and $B = A_n$. Then, by the addition rule for two events,

$$P(\cup_{i=1}^{n} A_i) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \sum_{i=1}^{n-1} P(A_i) - \sum_{1 \le i < j \le n-1} P(A_i \cap A_j) + \cdots + (-1)^n P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) + P(A_n) - P(U_{i=1}^{n-1}(A_i \cap A_n)).$$

Applying the inclusion exclusion formula to the $(n-1)$ events $A_i \cap A_n, i = 1, 2, \cdots, n-1$, and rearranging terms, the expression in the Theorem follows.

Example 1.5. (Missing Suits in a Bridge Hand). Consider a specific player, say North, in a Bridge game. We want to calculate the probability that North's hand is void in at least one suit. Towards this, denote the suits as 1, 2, 3, 4 and let A_i = North's hand is void in suit i.

Then, by the inclusion exclusion formula,

$$P(\text{North's hand is void in at least one suit})$$

$$= P(A_1 \cup A_2 \cup A_3 \cup A_4)$$

$= 4\binom{39}{13}/\binom{52}{13} - 6\binom{26}{13}/\binom{52}{13} + 4\binom{13}{13}/\binom{52}{13} = .051$, which is small, but not very small.

The inclusion exclusion formula can be hard to apply exactly, because the quantities S_j for large indices j can be difficult to calculate. However, fortunately, the inclusion exclusion formula leads to bounds in both directions for the probability of the union of n general events. We have the following series of bounds.

Theorem 1.3. (Bonferroni Bounds) Given n events $A_1, A_2, \cdots, A_n$, let $p_n = P(\cup_{i=1}^n A_i)$. Then,

$$p_n \leq S_1; p_n \geq S_1 - S_2; p_n \leq S_1 - S_2 + S_3; \ \cdots.$$

In addition,

$$P(\cap_{i=1}^n A_i) \geq 1 - \sum_{i=1}^n P(A_i^c).$$

Example 1.6. Suppose each of 10 events $A_1, A_2, \cdots A_{10}$ has probability .95 or more; thus, $P(A_i^c) \leq .05$ for each i. From the last Bonferroni bound given above, $P(\cap_{i=1}^n A_i) \geq 1 - 10 \times .05 = .5$. Each A_i, by itself, has a 95% probability or more of occuring. But that does not mean that with a high probability, all ten events will occur. What kinds of probability assurances can we provide that indeed all ten events will occur? The bound we just derived says that we can be at least 50% sure that all ten events will occur. This is typically rather crude, but these bounds are sometimes used by statisticians to make overall accuracy statements of their inference when they have made simultaneously a number of inferences.

1.3 Conditional Probability and Independence

Both conditional probability and independence are fundamental concepts for probabilists and statisticians alike. Conditional probabilities correspond to updating one's beliefs when new information becomes available. Independence corresponds to irrelevance of a piece of new information, even when it is made available. Additionally, the assumption of independence can and does significantly simplify development, mathematical analysis, and justification of tools and procedures.

Definition 1.4. Let A, B be general events with respect to some sample space Ω, and suppose $P(A) > 0$. The conditional probability of B given A is defined as

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

Some immediate consequences of the definition of a conditional probability are the following.

Theorem 1.4. (a) **(Multiplicative Formula).** For any two events A, B such that $P(A) > 0$, one has $P(A \cap B) = P(A)P(B|A)$;

(b) For any two events A, B such that $0 < P(A) < 1$, one has $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$;

(c) **(Total Probability Formula).** If $A_1, A_2, \cdots, A_k$ form a *partition* of the sample space Ω, i.e., $A_i \cap A_j = \phi$ for all $i \neq j$, and $\cup_{i=1}^k A_i = \Omega$, and if $0 < P(A_i) < 1$ for all i, then,

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

(d) **(Hierarchical Multiplicative Formula).** Let $A_1, A_2, \cdots, A_k$ be k general events in a sample space Ω. Then,

$$P(A_1 \cap A_2 \cap \cdots ... \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap A_2 \cap \cdots ... \cap A_{k-1}).$$

Example 1.7. A certain item is produced in a factory on one of three machines, A, B, and C. The percentages of items produced on machines A, B, and C are respectively $50\%, 30\%$, and 20%, and $4\%, 2\%$, and 4% of their products are defective. We want to know what percentage of all copies of this item are defective.

By defining A_1, A_2, A_3 as the events that a randomly selected item was produced by machine A, B, C respectively, and defining D as the event that it is a defective item, by the total probability formula, $P(D) = \sum_{i=1}^{3} P(D|A_i)P(A_i) = .04 \times .5 + .02 \times .3 + .04 \times .2 = .034$, i.e., 3.4% of all copies of the item produced in the factory are defective.

Example 1.8. One of two urns has a red and b black balls, and the other has c red and d black balls. One ball is chosen at random from each urn, and then one of these two balls is chosen at random. What is the probability that this ball is red?

If each ball selected from the two urns is red, then the final ball is definitely red. If one of those two balls is red, then the final ball is red with probability $1/2$. If none of those two balls is red, then the final ball cannot be red.

Thus, $P(\text{The final ball is red}) = a/(a+b) \times c/(c+d) + 1/2 \times \Big[a/(a+b) \times d/(c+d) + b/(a+b) \times c/(c+d) \Big] = \frac{2ac+ad+bc}{2(a+b)(c+d)}$.

As an example, suppose $a = 99, b = 1, c = 1, d = 1$. Then $\frac{2ac+ad+bc}{2(a+b)(c+d)} = .745$. Although the total percentage of red balls in the two urns is more than 98%, the chance that the final ball selected would be red is just about 75%.

Example 1.9. (An Example Open to Interpretation). This example was given to us in lecture when this author was a student of Dev Basu.

Mrs. Smith has two children. On a visit to the Smith household, you request a glass of water, and a boy brings it over. What is the probability that Mrs. Smith's other child is a boy?

Some people give the answer that this probability is 1/2. Others argue that with obvious notation, the sample space of the experiment is $\Omega = \{BB, BG, GB, GG\}$, and the required probability is $\frac{P(BB)}{P\{BB,BG,GB\}} = 1/3$.

Actually, the question does not have a unique answer, because the experiment to choose the child to carry the glass of water has not been specified. For instance, if Mrs. Smith will always send a girl child with the water if she has a girl, then the correct answer to the question is neither 1/2, nor 1/3, but 1!

Suppose Mrs. Smith chooses one of the two children to carry the glass of water at random if both children are of the same sex, and chooses the male child with probability p if the children are of different sex. Then,

$$P(\text{The other child is a boy}|\text{The chosen child is a boy}) = \frac{1/4}{1/4 + p/2} = 1/(2p+1).$$

If $p = .5$, then this is 1/2. Otherwise, the answer depends on Mrs. Smith's state of mind.

Next, we introduce the key concept of independence. Independence of events corresponds to lack of probabilistic information in one event A about some other event B; i.e., even if knowledge that some event A has occured was available, it would not cause us to modify the chances of the event B.

Definition 1.5. Two events A, B are called independent if $P(B|A) = P(B) \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$.
In applications, we often have to deal with more than two events simultaneously. But we may still want to know if they are independent. Fortunately, the concept of independence extends in a natural way to any number of events.

Definition 1.6. A collection of events $A_1, A_2, \cdots, A_n$ are said to be *mutually independent* (or just independent) if for each $k, 1 \le k \le n$, and any k of the events, $A_{i_1}, \cdots, A_{i_k}, P(A_{i_1} \cap \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$. They are called *pairwise independent* if this property holds for $k = 2$.

Example 1.10. Suppose a fair die is rolled twice and let A, B be the events that the sum of the two rolls is 7 and that the first roll is j, where j is any given number $1, 2, \cdots$, or 6.

Then $P(A) = 6/36 = 1/6, P(B) = 1/6$, and $P(A \cap B) = 1/36$. So A, B are independent events.

Now, change the event A to the event that the sum of the two rolls is 8. Then, A, B are not necessarily independent events. Why? For instance, with $j = 1, P(A|B) = 0$, but the unconditional probability $P(A)$ is not zero. Therefore, A and B cannot be independent events.

Example 1.11. (Lotteries). Although many people buy lottery tickets out of an expectation of good luck, probabilistically speaking, buying lottery tickets is usually a waste of money. Here is an example. Suppose in a weekly state lottery, five of the numbers $00, 01, \cdots, 49$ are selected without replacement at random, and someone to hold exactly those numbers wins the lottery. Then, the probability that someone holding one ticket will be the winner in a given week is $\frac{1}{\binom{50}{5}} = 4.72 \times 10^{-7}$. Suppose this person buys a ticket every week for 40 years. Then, the probability that he will win the lottery on at least one week is $1 - (1 - 4.72 \times 10^{-7})^{52 \times 40} = .00098 < .001$, still a very small probability. We assumed in this calculation that the weekly lotteries are all mutually independent, a reasonable assumption. The calculation will fall apart if we did not make this independence assumption.

Example 1.12. (An Interesting Example due to Emanuel Parzen). Consider two dice, with the side probabilities being $p_j, 1 \le j \le 6$ for the first die, and $q_j, 1 \le j \le 6$ for the second die. That is, we are just assuming that these are two arbitrarily loaded dice. The question is whether we can choose p_j, q_j in any way whatsoever such that the sum of the numbers obtained on tossing the dice once each has an equal probability of being any of $2, 3, \cdots 12$. The answer, interestingly, is that we cannot choose p_j, q_j in any way at all to make this happen. Here is a sketch of the proof.

Suppose we could. Then, since the sums of 2 and 12 will have an equal probability, we must have $p_1 q_1 = p_6 q_6 \Rightarrow q_1 = p_6 q_6 / p_1$. It follows, after some algebra, on using this that $(p_1 - p_6)/(q_1 - q_6) \le 0 \Rightarrow p_1 q_1 + p_6 q_6 \le p_1 q_6 + p_6 q_1$. But this means that

$$P(\text{The sum is 7}) \ge p_1 q_6 + p_6 q_1 \ge p_1 q_1 + p_6 q_6 = P(\text{The sum is 2 or 12}),$$

a contradiction, because by assumption P(The sum is 2 or 12) is supposed to be twice the probability that the sum is 7. Hence, we cannot construct two dice in any way to make the sum have an equal probability of taking the values $2, 3, \cdots 12$.

It is not uncommon to see the conditional probabilities $P(A|B)$ and $P(B|A)$ to be confused with each other. Suppose in some group of lung cancer patients, we see a large percentage of smokers. If we define B to be the event that a person is a smoker, and A to be the event that a person has lung cancer, then all we can conclude is that in our group of people $P(B|A)$ is large. But we cannot conclude from just this information that smoking

increases the chance of lung cancer, i.e., that $P(A|B)$ is large. In order to calculate a conditional probability $P(A|B)$ when we know the *other* conditional probability $P(B|A)$, a simple formula known as *Bayes' theorem* is useful. Here is a statement of a general version of Bayes' theorem.

Theorem 1.5. (Bayes' Theorem) Let $\{A_1, A_2, \cdots, A_m\}$ be a partition of a sample space Ω. Let B be some fixed event. Then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}.$$

Proof: By definition of conditional probability, $P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$, by the multiplicative formula, and the total probability formula.

Example 1.13. (Inaccurate Blood Tests). A certain blood test for a disease gives a positive result 90% of the times among patients having the disease. But it also gives a positive result 25% of the times among people who do not have the disease. It is believed that 30% of the population has this disease. What is the probability that a person with a positive test result indeed has the disease?

In medical terminology, the 90% value is called the *sensitivity* of the test, and $100-25 = 75\%$ is called the *specificity* of the test. Often, the sensitivity and the specificity would be somewhat higher than what they are in this example.

Define

$$A = \text{The person has the disease;}$$

$$B = \text{The blood test gives a positive result for the person.}$$

Then, by Bayes theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{.9 \times .3}{.9 \times .3 + .25 \times (1 - .3)} = .607.$$

Before the test was given, the physician had the *apriori* probability of 30% that the person has the disease. After a blood test came out positive, the physician has the *posterior* probability of 60.7% that the person has the disease.

Example 1.14. (Multiple choice exams). Suppose that the questions in a multiple choice exam have five alternatives each, of which a student has to pick one as the correct alternative. A student either knows the truly correct alternative with probability .7, or he randomly picks one of the five alternatives as his choice. Suppose a particular problem was answered correctly. We want to know what is the probability that the student really knew the correct answer.

Define

$$A = \text{The student knew the correct answer,}$$

$$B = \text{The student answered the question correctly.}$$

We want to compute $P(A|B)$. By Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{1 \times .7}{1 \times .7 + .2 \times .3} = .921.$$

Before the student answered the question, our probability that he would know the correct answer to the question was .7; but once he answered it correctly, the posterior probability that he knew the correct answer increases to .921. This is exactly what Bayes' theorem does; it updates our *prior* belief to the *posterior* belief, when new evidence becomes available.

1.4 Integer Valued and Discrete Random Variables

In some sense, the entire subject of probability and statistics is about distributions of random variables. Random variables, as the very name suggests, are quantities that vary, over time, or from individual to individual, and the reason for the variability is some underlying random process. Depending on exactly how an underlying experiment ξ ends, the random variable takes different values. In other words, the value of the random variable is determined by the sample point ω that prevails, when the underlying experiment ξ is actually conducted. We cannot apriori know the value of the random variable, because we do not know apriori which sample point ω will prevail when the experiment ξ is conducted. We try to understand the behavior of a random variable by analyzing the probability structure of that underlying random experiment.
Random variables, like probabilities, originated in gambling. Therefore, the random variables that come to us *more naturally*, are integer valued random variables; e.g., the sum of the two rolls when a die is rolled twice. Integer valued random variables are special cases of what are known as discrete random random variables. Discrete or not, a common mathematical definition of all random variables is the following.

Definition 1.7. Let Ω be a sample space corresponding to some experiment ξ and let $X : \Omega \to \mathcal{R}$ be a function from the sample space to the real line. Then X is called a *random variable*.

Discrete random variables are those that take a finite or a countably infinite number of possible values. In particular, all integer valued random variables are discrete. From the point of view of understanding the behavior of a random variable, the important thing is to know the probabilities with which X takes its different possible values.

Definition 1.8. Let $X : \Omega \to \mathcal{R}$ be a discrete random variable taking a finite or countably infinite number of values $x_1, x_2, x_3, \cdots$. The probability distribution or the *probability mass function* (pmf) of X is the function $p(x) = P(X = x), x = x_1, x_2, x_3, \cdots$, and $p(x) = 0$,

otherwise.

It is common to not explicitly mention the phrase "$p(x) = 0$ otherwise", and we will generally follow this convention. Some authors use the phrase *mass function* instead of *probability mass function*.

For any pmf, one must have $p(x) \geq 0$ for any x, and $\sum_i p(x_i) = 1$. Any function satisfying these two properties for some set of numbers $x_1, x_2, x_3, \cdots$ is a valid pmf.

1.4.1 CDF and Independence

A second important definition is that of a *cumulative distribution function* (CDF). The CDF gives the probability that a random variable X is less than or equal to any given number x. It is important to understand that the notion of a CDF is universal to all random variables; it is not limited to only the discrete ones.

Definition 1.9. The *cumulative distribution function* (CDF) of a random variable X is the function $F(x) = P(X \leq x), x \in \mathcal{R}$.

Definition 1.10. Let X have the CDF $F(x)$. Any number m such that $P(X \leq m) \geq .5$, and also $P(X \geq m) \geq .5$ is called a median of F, or equivalently, a median of X.

Remark: The median of a random variable *need not be* unique.

Example 1.15. Let ξ be the experiment of tossing a fair die twice and let X be the number of heads obtained. Then X takes the possible values $x_1 = 0, x_2 = 1, x_3 = 2$. Also, $P(X = 0) = P(TT) = 1/4; P(X = 1) = P(\{HT, TH\}) = 1/2$; and $P(X = 2) = P(HH) = 1/4$. We then have the following pmf for X:

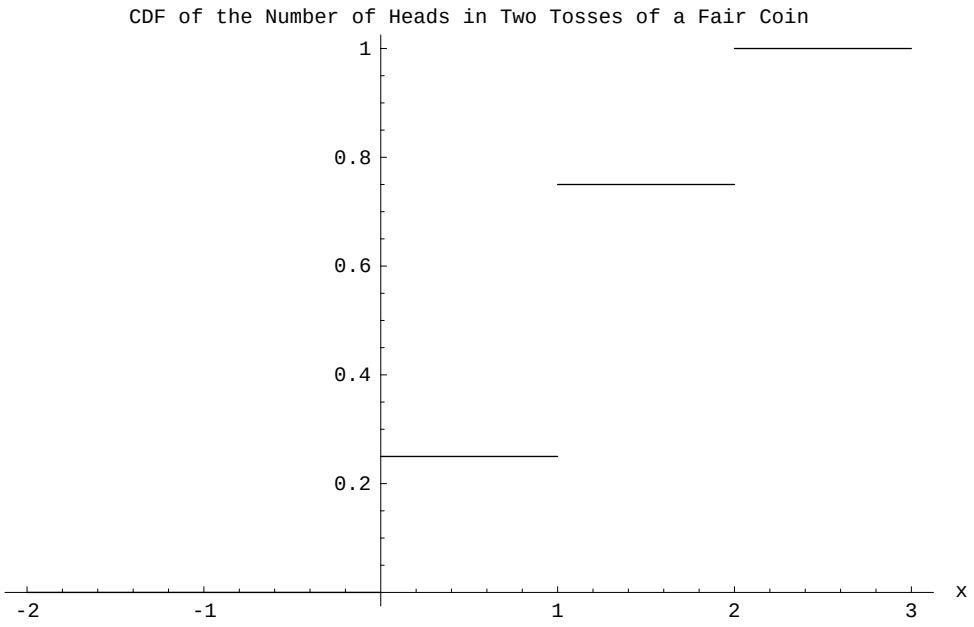
x	0	1	2
p(x)	.25	.5	.25

As regards the CDF of X, since X does not take any negative values, the CDF $F(x)$ is zero for any negative x. However, X takes the $x = 0$ value with a positive probability, namely .25. Thus, as soon as x reaches the zero value, the CDF $F(x)$ takes a jump and becomes equal to .25. Then, between $x = 0$ and $x = 1$, X does not take any other values, and no new probability is accumulated. So the CDF stays stuck at the .25 value until x reaches the value $x = 1$, and now it takes another jump of size .5, which is the probability that $X = 1$. The next jump is at $x = 2$, when the CDF takes another jump of size .25, and thereafter the CDF takes no further jumps. In symbols, the CDF $F(x)$ in this example is the jump function

$$F(x) = 0 \text{ if } x < 0;$$

$$= .25 \text{ if } 0 \leq x < 1;$$

$$= .75 \text{ if } 1 \leq x < 2;$$



$= 1$ if $x \geq 2$.

A plot of the CDF is helpful for understanding and is given below.

It is clear that because this CDF increases by jumps, it does not attain all values between 0 and 1. For example, there is no x at which $F(x) = .5$.

The CDF of any random variable satisfies a set of properties. Conversely, any function satisfying these properties is a valid CDF, i.e., it will be the CDF of some appropriately chosen random variable. These properties are given in the next result.

Theorem 1.6. A function $F(x)$ is the CDF of some real valued random variable X if and only if it satisfies all of the following properties:

$$(a)\, 0 \leq F(x) \leq 1 \ \forall x \in \mathcal{R};$$

$$(b)\, F(x) \to 0 \,\text{as}\, x \to -\infty, \,\text{and}\, F(x) \to 1 \,\text{as}\, x \to \infty;$$

$$(c)\, \text{Given any real number}\, a, F(x) \downarrow F(a) \,\text{as}\, x \downarrow a;$$

$$(d)\, \text{Given any two real numbers}\, x, y, x < y, F(x) \leq F(y).$$

Property (c) is called *continuity from the right*, or simply right continuity. It is clear that a CDF need not be continuous from the left; indeed, for discrete random variables, the CDF has a jump at the values of the random variable, and at the jump points, the CDF is not left continuous. More precisely, one has the following result.

Proposition Let $F(x)$ be the CDF of some random variable X. Then, for any x,

$$(a)\, P(X = x) = F(x) - \lim_{y \uparrow x} F(y) = F(x) - F(x-),$$

including those points x for which $P(X = x) = 0$.

(b) $P(X \geq x) = P(X > x) + P(X = x) = (1 - F(x)) + (F(x) - F(x-)) = 1 - F(x-)$.

Example 1.16. (Bridge). Consider the random variable

$$X = \text{ Number of aces in North's hand in a Bridge game.}$$

Clearly, X can take any of the values $x = 0, 1, 2, 3, 4$. If $X = x$, then the other $13 - x$ cards in North's hand must be non-ace cards. Thus, the pmf of X is

$$P(X = x) = \frac{\binom{4}{x}\binom{48}{13-x}}{\binom{52}{13}}, x = 0, 1, 2, 3, 4.$$

In decimals, the pmf of X is:

x	0	1	2	3	4
p(x)	.304	.439	.213	.041	.003

The CDF of X is a jump function, taking jumps at the values $0, 1, 2, 3, 4$, namely the possible values of X. The CDF is

$$F(x) = 0 \text{ if } x < 0;$$

$$= .304 \text{ if } 0 \leq x < 1;$$

$$= .743 \text{ if } 1 \leq x < 2;$$

$$= .956 \text{ if } 2 \leq x < 3;$$

$$= .997 \text{ if } 3 \leq x < 4;$$

$$= 1 \text{ if } x \geq 4.$$

Example 1.17. (Indicator Variables). Consider the experiment of rolling a fair die twice and now define a random variable Y as follows:

$$Y = 1 \text{ if the sum of the two rolls X is an even number;}$$

$$Y = 0 \text{ if the sum of the two rolls X is an odd number.}$$

If we let A be the event that X is an even number, then $Y = 1$ if A happens, and $Y = 0$ if A does not happen. Such random variables are called *indicator random variables* and are immensely useful in mathematical calculations in many complex situations.

Definition 1.11. Let A be any event in a sample space Ω. The *Indicator random variable* for A is defined as

$$I_A = 1 \text{ if A happens;}$$

$I_A = 0$ if A does not happen.

Thus, the distribution of an indicator variable is simply $P(I_A = 1) = P(A); P(I_A = 0) = 1 - P(A)$.

An indicator variable is also called a *Bernoulli variable* with parameter p, where p is just $P(A)$. We will later see examples of uses of indicator variables in calculation of *expectations*.

In applications, we are sometimes interested in the distribution of a function, say $g(X)$, of a basic random variable X. In the discrete case, the distribution of a function is found in the obvious way.

Proposition (Function of a Random Variable). Let X be a discrete random variable and $Y = g(X)$ a real valued function of X. Then, $P(Y = y) = \sum_{x:\, g(x)=y} p(x)$.

Example 1.18. Suppose X has the pmf $p(x) = \frac{c}{1+x^2}, x = 0, \pm 1, \pm 2, \pm 3$. Suppose we want to find the distribution of two functions of X:

$$Y = g(X) = X^3; Z = h(X) = \sin(\frac{\pi}{2}X).$$

First, the constant c must be explicitly evaluated. By directly summing the values,

$$\sum_x p(x) = \frac{13c}{5} \Rightarrow c = \frac{5}{13}.$$

Note that $g(X)$ is a one-to-one function of X, but $h(X)$ is not one-to-one. The values of Y are $0, \pm 1, \pm 8, \pm 27$. For example, $P(Y = 0) = P(X = 0) = c = 5/13; P(Y = 1) = P(X = 1) = c/2 = 5/26$, etc. In general, for $y = 0, \pm 1, \pm 8, \pm 27, P(Y = y) = P(X = y^{1/3}) = \frac{c}{1+y^{2/3}}$, with $c = 5/13$.

However, $Z = h(X)$ is not a one-to-one function of X. The possible values of Z are as follows:

x	$h(x)$
-3	1
-2	0
-1	-1
0	0
1	1
2	0
3	-1

So, for example, $P(Z = 0) = P(X = -2) + P(X = 0) + P(X = 2) = \frac{7}{5}c = 7/13$. The pmf of $Z = h(X)$ is:

z	-1	0	1
$P(Z=z)$	3/13	7/13	3/13

Notice that Z has a *symmetric distribution*, i.e., Z and $-Z$ have the same pmf. This is not a coincidence. This is because X itself has a symmetric distribution and $Z = h(X)$ is an *odd function* of X. This is generally true.

Proposition Suppose X has a distribution symmetric about zero, i.e., $P(X = x) = P(X = -x)$ for any x. Let $h(x)$ be an odd function, i.e., $h(-x) = -h(x)$ for any x. Then $Z = h(X)$ also has a distribution symmetric about zero.

A key concept in probability is that of independence of a collection of random variables. The collection could be finite or infinite. In the infinite case, we want each finite subcollection of the random variables to be independent. The definition of independence of a finite collection is as follows.

Definition 1.12. Let $X_1, X_2, \cdots, X_k$ be $k \geq 2$ discrete random variables defined on the same sample space Ω. We say that $X_1, X_2, \cdots, X_k$ are *independent* if $P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_k = x_k)$, $\forall\, x_1, x_2, \cdots, x_k$.

It follows from the definition of independence of random variables that if X_1, X_2 are independent, then any function of X_1 and any function of X_2 are also independent. In fact, we have a more general result.

Theorem 1.7. Let $X_1, X_2, \cdots, X_k$ be $k \geq 2$ discrete random variables, and suppose they are independent. Let $U = f(X_1, X_2, \cdots, X_i)$ be some function of $X_1, X_2, \cdots, X_i$, and $V = g(X_{i+1}, \cdots, X_k)$ be some function of $X_{i+1}, \cdots, X_k$. Then, U and V are independent.

This result is true of *any types of random variables* $X_1, X_2, \cdots, X_k$, not just discrete ones.

Example 1.19. (Two Simple Illustrations). Consider the experiment of tossing a fair coin (or any coin) four times. Suppose X_1 is the number of heads in the first two tosses, and X_2 is the number of heads in the last two tosses. Then, it is intuitively clear that X_1, X_2 are independent, because the first two tosses have no information regarding the last two tosses. The independence can be easily mathematically verified by using the definition of independence.

Next, consider the experiment of drawing 13 cards at random from a deck of 52 cards. Suppose X_1 is the number of aces and X_2 is the number of clubs among the 13 cards. Then, X_1, X_2 are not independent. For example, $P(X_1 = 4, X_2 = 0) = 0$, but $P(X_1 = 4)$ and $P(X_2 = 0)$ are both > 0, and so $P(X_1 = 4)P(X_2 = 0) > 0$. So, X_1, X_2 cannot be independent.

A common notation of wide use in probability and statistics is now introduced.

If $X_1, X_2, \cdots, X_k$ are independent, and moreover have the same CDF, say F, then we say that $X_1, X_2, \cdots, X_k$ are iid (or IID) and write $X_1, X_2, \cdots, X_k \overset{iid}{\sim} F$. The abbreviation iid (IID) means independent and identically distributed.

1.4.2 Expectation and Moments

By definition, a random variable takes different values on different occasions. It is natural to want to know what value does it take on an average. Averaging is a very primitive concept. A simple average of just the possible values of the random variable will be misleading, because some values may have so little probability that they are relatively inconsequential. The average or the mean value, also called the expected value of a random variable is a weighted average of the different values of X, weighted according to how important the value is. Here is the definition.

Definition 1.13. Let X be a discrete random variable. We say that the *expected value* of X exists if $\sum_i |x_i| p(x_i) < \infty$, in which case the expected value is defined as

$$\mu = E(X) = \sum_i x_i p(x_i).$$

For notational convenience, we simply write $\sum_x xp(x)$ instead of $\sum_i x_i p(x_i)$.

The expected value is also known as *the expectation or the mean* of X.

If the set of possible values of X is infinite, then the infinite sum $\sum_x xp(x)$ can take different values on rearranging the terms of the infinite series unless $\sum_x |x| p(x) < \infty$. So, as a matter of definition, we have to include the qualification that $\sum_x |x| p(x) < \infty$.

If the sample space Ω of the underlying experiment is finite or countably infinite, then we can also calculate the expectation by averaging directly over the sample space.

Proposition (Change of Variable Formula). Suppose the sample space Ω is finite or countably infinite and X is a discrete random variable with expectation μ. Then,

$$\mu = \sum_x xp(x) = \sum_\omega X(\omega) P(\omega),$$

where $P(\omega)$ is the probability of the sample point ω.

Proof: $\sum_\omega X(\omega)P(\omega) = \sum_x \sum_{\omega: X(\omega)=x} X(\omega)P(\omega) = \sum_x x \sum_{\omega: X(\omega)=x} P(\omega) = \sum_x xp(x)$.

Important Point In applications we are often interested in more than one variable at the same time. To be specific, consider two discrete random variables X, Y defined on a common sample space Ω. Then, we could construct new random variables out of X and Y; for example, $XY, X+Y, X^2+Y^2$, etc. We can then talk of their expectations as well. Here is a general definition of expectation of a function of more than one random variable.

Definition 1.14. Let $X_1, X_2, \cdots, X_n$ be n discrete random variables, all defined on a common sample space Ω, with a finite or a countably infinite number of sample points. We say that the expectation of a function $g(X_1, X_2, \cdots, X_n)$ exists if $\sum_\omega |g(X_1(\omega), X_2(\omega), \cdots, X_n(\omega))|P(\omega) < \infty$, in which case, the expected value of $g(X_1, X_2, \cdots, X_n)$ is defined as

$$E[g(X_1, X_2, \cdots, X_n)] = \sum_\omega g(X_1(\omega), X_2(\omega), \cdots, X_n(\omega))P(\omega).$$

The next few results summarize the most fundamental properties of expectations.

Proposition (a) If there exists a finite constant c such that $P(X = c) = 1$, then $E(X) = c$.

(b) If X, Y are random variables defined on the same sample space Ω with finite expectations, and if $P(X \leq Y) = 1$, then $E(X) \leq E(Y)$.

(c) If X has a finite expectation, and if $P(X \geq c) = 1$, then $E(X) \geq c$. If $P(X \leq c) = 1$, then $E(X) \leq c$.

Proposition (Linearity of Expectations). Let $X_1, X_2, \cdots, X_n$ be random variables defined on the same sample space Ω, and $c_1, c_2, \cdots, c_n$ any real valued constants. Then, provided $E(X_i)$ exists for every X_i,

$$E(\sum_{i=1}^{n} c_i X_i) = \sum_{i=1}^{n} c_i E(X_i);$$

in particular, $E(cX) = cE(X)$ and $E(X_1 + X_2) = E(X_1) + E(X_2)$, whenever the expectations exist.

Proof: We assume that the sample space Ω is finite or countably infinite. Then, by the change of variable formula,

$$E(\sum_{i=1}^{k} c_i X_i) = \sum_\omega [\sum_{i=1}^{k} c_i X_i(\omega)]P(\omega)$$

$$= \sum_\omega [\sum_{i=1}^{k} c_i X_i(\omega)P(\omega)] = \sum_{i=1}^{k} \sum_\omega [c_i X_i(\omega)P(\omega)]$$

$$= \sum_{i=1}^{k} c_i \sum_\omega [X_i(\omega)P(\omega)] = \sum_{i=1}^{k} c_i E(X_i).$$

The following fact also follows easily from the definition of the pmf of a function of a random variable. The result says that the expectation of a function of a random variable X can be calculated directly using the pmf of X itself, without having to calculate the pmf of the function.

Proposition (Expectation of a Function). Let X be a discrete random variable on a sample space Ω with a finite or countable number of sample points, and $Y = g(X)$ a

function of X. Then,

$$E(Y) = \sum_{\omega} Y(\omega)P(w) = \sum_{x} g(x)p(x),$$

provided $E(Y)$ exists.

Caution: If $g(X)$ is a linear function of X, then, of course, $E(g(X)) = g(E(X))$. But, in general, the two things are not equal. For example, $E(X^2)$ is not the same as $(E(X))^2$; indeed, $E(X^2) > (E(X))^2$ for any random variable X which is not a constant.
A very important property of independent random variables is the following factorization result on expectations.

Theorem 1.8. Suppose $X_1, X_2, \cdots X_n$ are independent random variables. Then, provided each expectation exists,

$$E(X_1X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

Let us now see some more illustrative examples.

Example 1.20. (Dice Sum). Let X be the sum of the two rolls when a fair die is rolled twice. The pmf of X is $p(2) = p(12) = 1/36; p(3) = p(11) = 2/36; p(4) = p(10) = 3/36; p(5) = p(9) = 4/36; p(6) = p(8) = 5/36; p(7) = 6/36$. Therefore, $E(X) = 2 \times 1/36 + 3 \times 2/36 + 4 \times 3/36 + \cdots + 12 \times 1/36 = 7$. This can also be seen by letting $X_1 =$ The face obtained on the first roll, $X_2 =$ The face obtained on the second roll, and by using $E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$.

Let us now make this problem harder. Suppose that a fair die is rolled ten times and X is the sum of all ten rolls. The pmf of X is no longer so simple; it will be cumbersome to write it down. But, if we let $X_i =$ The face obtained on the ith roll, it is still true by the linearity of expectations that $E(X) = E(X_1 + X_2 + \cdots + X_{10}) = E(X_1) + E(X_2) + \cdots + E(X_{10}) = 3.5 \times 10 = 35$. We can easily compute the expectation, although the pmf would be difficult to write down.

Example 1.21. (A Random Variable without a Finite Expectation). Let X take the positive integers $1, 2, 3, \cdots$ as its values with the pmf $p(x) = P(X = x) = \frac{1}{x(x+1)}, x = 1, 2, 3, \cdots$. This is a valid pmf, because obviously $\frac{1}{x(x+1)} > 0$ for any $x = 1, 2, 3, \cdots$, and also the infinite series $\sum_{x=1}^{\infty} \frac{1}{x(x+1)}$ sums to 1, a fact from calculus. Now, $E(X) = \sum_{x=1}^{\infty} xp(x) = \sum_{x=1}^{\infty} x\frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \sum_{x=2}^{\infty} \frac{1}{x} = \infty$, also a fact from calculus.

This example shows that not all random variables have a finite expectation. Here, the reason for the infiniteness of $E(X)$ is that X takes large integer values x with probabilities $p(x)$ that are not adequately small. The large values are realized sufficiently often that on the average X becomes larger than any given finite number.

The zero-one nature of indicator random variables is extremely useful for calculating expectations of certain integer valued random variables, whose distributions are sometimes so complicated that it would be difficult to find their expectations directly from definition. We describe the technique and some illustrations of it below.

Proposition Let X be an integer valued random variable such that it can be represented as $X = \sum_{i=1}^{m} c_i I_{A_i}$ for some m, constants $c_1, c_2, \cdots, c_m$, and suitable events $A_1, A_2, \cdots, A_m$. Then, $E(X) = \sum_{i=1}^{m} c_i P(A_i)$.

Proof: $E(X) = E[\sum_{i=1}^{m} c_i I_{A_i}] = \sum_{i=1}^{m} c_i E[I_{A_i}] = \sum_{i=1}^{m} c_i P(A_i)$.

Example 1.22. (Coin Tosses). Suppose a coin which has probability p of show ing heads in any single toss is tossed n times, and let X denote the number of times in the n tosses that a head is obtained. Then, $X = \sum_{i=1}^{n} I_{A_i}$, where A_i is the event that a head is obtained in the ith toss. Therefore, $E(X) = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} p = np$.

Example 1.23. (The Matching Problem). Suppose n entities, say $1, 2, \cdots, n$ are linearly arranged at random in the locations marked as $1, 2, \cdots, n$. Suppose after rearrangement, the number at location i is $\pi(i)$. We want to study the number of matches defined as $X = \text{Number of locations } i \text{ such that } \pi(i) = i$.

We use the indicator variable method to find the expected number of matches. Towards this, define $A_i = \text{There is a match at location } i$. Then $X = \sum_{i=1}^{n} I_{A_i}$. Now, for any $i, P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$, and therefore we have the quite elegant result that whatever be $n, E(X) = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} \frac{1}{n} = n \times \frac{1}{n} = 1$. Direct verification of this would require calculation of the pmf of the number of matches for any given n. If we do, then we will find on algebra that $\sum_{k=1}^{n} kP(X = k) = 1$ for any n.

Another useful technique for calculating expectations of nonnegative integer valued random variables is based on the CDF of the random varaible, rather than directly the pmf. This method is useful when calculating probabilities of the form $P(X > x)$ is logically more straightforward than directly calculating $P(X = x)$. Here is the expectation formula based on the tail CDF.

Theorem 1.9. (Tailsum Formula) Let X take values $0, 1, 2, \cdots$. Then,

$$E(X) = \sum_{n=0}^{\infty} P(X > n).$$

Example 1.24. (Waiting Time to See the First Head). Suppose a coin with probability p for heads is tossed until a head is obtained for the first time. How many tosses will it take on an average to see the first head? Let X denote the number of heads necessary to obtain the very first head. Then $X > n$ simply means that the first n tosses have all

produced tails. Therefore,

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

X is called a *Geometric random variable with parameter* p. If the coin is fair, this says that on an average, the first head will be seen at the second toss.

Example 1.25. (Family Planning). Suppose a couple will have children until they have at least one child of each sex. How many children can they expect to have? Let X denote the childbirth at which they have a child of each sex for the first time. Suppose the probability that any particular childbirth will be a boy is p, and that all births are independent. Then,

$$P(X > n) = P(\text{The first } n \text{ children are all boys or all girls}) = p^n + (1-p)^n.$$

Therefore, $E(X) = 2 + \sum_{n=2}^{\infty} [p^n + (1-p)^n] = 2 + p^2/(1-p) + (1-p)^2/p = \frac{1}{p(1-p)} - 1$. If boys and girls are equally likely on any childbirth, then this says that a couple waiting to have a child of each sex can expect to have three children.

The expected value is calculated with the intention of understanding what is a typical value of a random variable. But two very different distributions can have exactly the same expected value. A common example is that of a return on an investment in a stock. Two stocks may have the same average return, but one may be much riskier than the other, in the sense that the variability in the return is much higher for that stock. In that case, most risk-averse individuals would prefer to invest in the stock with less variability. Measures of risk or variability are of course not unique. Some natural measures that come to mind are $E(|X-\mu|)$, known as the *mean absolute deviation*, or $P(|X-\mu| > k)$ for some suitable k. However, neither of these two is the most common measure of variability. The most common measure is the *standard deviation* of a random variable.

Definition 1.15. Let a random variable X have a finite mean μ. The *variance* of X is defined as

$$\sigma^2 = E[(X-\mu)^2],$$

and the *standard deviation* of X is defined as $\sigma = \sqrt{\sigma^2}$.

It is easy to prove that $\sigma^2 < \infty$ if and only if $E(X^2)$, the *second moment* of X, is finite. It is not uncommon to mistake the standard deviation for the mean absolute deviation; but they are not the same. In fact, an inequality always holds.

Proposition

$\sigma \geq E(|X-\mu|)$, and σ is strictly greater unless X is a constant random variable, namely, $P(X = \mu) = 1$.

We list some basic properties of the variance of a random variable.

Proposition

(a) $\text{Var}(cX) = c^2\text{Var}(X)$ for any real c;

(b) $\text{Var}(X + k) = \text{Var}(X)$ for any real k;

(c) $\text{Var}(X) \geq 0$ for any random variable X, and equals zero only if $P(X = c) = 1$ for some real constant c;

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

The quantity $E(X^2)$ is called the *second moment* of X. The definition of a general moment is as follows.

Definition 1.16. Let X be a random variable, and $k \geq 1$ a positive integer. Then $E(X^k)$ is called the *kth moment of* X, and $E(X^{-k})$ is called the *kth inverse moment of* X, provided they exist.

We therefore have the following realtionships involving moments and the variance:

Variance = Second Moment - (First Moment)^2;
Second Moment = Variance + (First Moment)^2.

Statisticians often use the third moment around the mean as a measure of lack of symmetry in the distribution of a random variable. The point is that if a random variable X has a symmetric distribution, and has a finite mean μ, then all odd moments around the mean, namely, $E[(X - \mu)^{2k+1}]$ will be zero, if the moment exists. In particular, $E[(X - \mu)^3]$ will be zero. Likewise, statisticians also use the fourth moment around the mean as a measure of how spiky the distribution is around the mean. To make these indices independent of the choice of unit of measurement, e.g., inches or cms, they use certain scaled measures of asymmetry and peakedness. Here are the definitions.

Definition 1.17. (a) Let X be a random variable with $E[|X|^3] < \infty$. The *skewness* of X is defined as

$$\beta = \frac{E[(X - \mu)^3]}{\sigma^3}.$$

(b) Suppose X is a random variable with $E[X^4] < \infty$. The *kurtosis* of X is defined as

$$\gamma = \frac{E[(X - \mu)^4]}{\sigma^4} - 3.$$

The skewness β is zero for symmetric distributions, but the converse need not be true. The kurtosis γ is necessarily ≥ -2, but can be arbitrarily large, with spikier distributions generally having a larger kurtosis. But a very good interpretation of γ is not really available. We will later see that $\gamma = 0$ for all *normal distributions*; hence the motivation for subtracting 3 in the definition of γ.

Example 1.26. (Variance of Dice Sum). Let X be the sum of two independent rolls of a fair die. Then, from the pmf of X previously derived, $E(X) = 7$, and $E(X^2) = 2^2 \times 1/36 + 3^2 \times 2/36 + 4^2 \times 3/36 + \cdots + 12^2 \times 1/36 = 329/6$, and therefore $\text{Var}(X) = E(X^2) - \mu^2 = 329/6 - 49 = \frac{35}{6} = 5.83$, and the standard deviation is $\sigma = \sqrt{5.83} = 2.415$.

Example 1.27. (Variance in the Matching Problem). Let again X be the number of locations at which a match occurs when n numbers, say, $1, 2, \cdots, n$ are rearranged in a random order. We have previously seen by use of the indicator variable method that $E(X) = 1$, whatever be n. We now use the indicator variable method to also calculate the variance.

Towards this, define again $A_i =$ There is a match at location i. Then, $X = \sum_{i=1}^n I_{A_i}$. We first find the second moment of X.

Now, $X^2 = (\sum_{i=1}^n I_{A_i})^2 = (\sum_{i=1}^n [I_{A_i}]^2 + 2\sum_{1 \le i < j \le n} I_{A_i} I_{A_j}) = (\sum_{i=1}^n I_{A_i} + 2\sum_{1 \le i < j \le n} I_{A_i} I_{A_j})$.

Therefore,

$$E(X^2) = \sum_{i=1}^n P(A_i) + 2 \sum_{1 \le i < j \le n} P(A_i \cap A_j).$$

For any $i, P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$, and for all $i, j, i < j, P(A_i \cap A_j) = \frac{(n-2!}{n!} = \frac{1}{n(n-1)}$.

Therefore,

$$E(X^2) = n \times \frac{1}{n} + 2\binom{n}{2}\frac{1}{n(n-1)} = 1 + 1 = 2.$$

Therefore, $\text{Var}(X) = E(X^2) - [E(X)]^2 = 2 - 1 = 1$, regardless of the value of n.

To summarize, in the matching problem, regardless of the value of n, the mean and the variance of the number of matches are both 1. We will later see that this property of equality of the mean and variance is true for a well known distribution called the *Poisson distribution*, and in fact, *a Poisson distribution does provide an extremely accurate approximation* for the exact pmf of the number of matches.

Example 1.28. (A Random Variable with an Infinite Variance). If a random variable has a finite variance, then it can be shown that it must have a finite mean. This example shows that the converse need not be true.

Let X be a discrete random variable with the pmf $P(X = x) = \frac{c}{x(x+1)(x+2)}, x = 1, 2, 3, \cdots$, where the normalizing constant $c = 4$. The expected value of X is

$$E(X) = \sum_{x=1}^{\infty} x \times \frac{4}{x(x+1)(x+2)} = 4\sum_{x=1}^{\infty} \frac{1}{(x+1)(x+2)} = 4 \times 1/2 = 2.$$

Therefore, by direct verification, X has a finite expectation. Let us now examine the second moment of X.

$$E(X^2) = \sum_{x=1}^{\infty} x^2 \times \frac{4}{x(x+1)(x+2)} = 4\sum_{x=1}^{\infty} x \times \frac{1}{(x+1)(x+2)} = \infty,$$

because the series $\sum_{x=1}^{\infty} x \times \frac{1}{(x+1)(x+2)}$ is not finitely summable, a fact from calculus. Since $E(X^2)$ is infinite, but $E(X)$ is finite, $\sigma^2 = E(X^2) - [E(X)]^2$ must also be infinite.

If a collection of random variables are independent, then just like the expectation, the variance also adds up. Precisely, one has the following very useful fact.

Theorem 1.10. Let $X_1, X_2, \cdots, X_n$ be n independent random variables. Then,

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n).$$

Proof: It is enough to prove this result for $n = 2$, because we can then prove the general case by induction. By definition, writing $E(X_i) = \mu_i$,

$$\text{Var}(X_1 + X_2) = E[(X_1 + X_2) - E(X_1 + X_2)]^2 = E[(X_1 + X_2) - (\mu_1 + \mu_2)]^2$$

$$= E[(X_1 - \mu_1) + (X_2 - \mu_2)]^2 = E[(X_1 - \mu_1)^2] + E[(X_2 - \mu_2)^2] + 2E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

$= \text{Var}(X_1) + \text{Var}(X_2) + 0$, because $E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[(X_1 - \mu_1)]E[(X_2 - \mu_2)] = 0 \times 0 = 0$, by virtue of the independence of X_1, X_2. This proves the result.

An important corollary of this result is the following variance formula for the mean, $\bar{X}$, of n independent and identically distributed (iid) random variables.

Corollary 1.1. Let $X_1, X_2, \cdots, X_n$ be independent random variables with a common variance $\sigma^2 < \infty$. Let $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. Then $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

1.4.3 Stochastically Larger

We understand what does it mean to say that a real number x is smaller than another real number y. There is a useful extension of this notion to a pair of random variables X, Y. For instance, if $X \sim N(5, 1)$ and $Y \sim N(10, 1)$ then intuitively we feel that in a practical sampling experiment, observed values on Y would tend to be larger than observed values on X. But, of course, we cannot guarantee that any observed value on Y would be larger than any observed value on X. So the correct notion for declaring one random variable as larger than another random variable must involve probabilities. Here is the definition.

Definition 1.18. Let X, Y be two real valued random variables. We say that Y is *stochastically larger* than X if $P(Y > u) \geq P(X > u)$ for all real numbers u.
In other words, in a real experiment, for any number u that you may choose, a larger percentage of the Y values would be to the right of u, compared to the percentage of the X values that are to the right of u,
Note that the definition given above is equivalent to saying that the CDFs of X, Y satisfy the ordering $F_X(u) = P(X \leq u) \geq F_Y(u) = P(Y \leq u)$ for all real numbers u.

Example 1.29. (Stochastic Ordering of Normal Random Variables). Suppose $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$, where $\mu_2 > \mu_1$. Then, for any given u,

$$P(X \leq u) = \Phi(\frac{u-\mu_1}{\sigma}) > \Phi(\frac{u-\mu_2}{\sigma}) = P(Y \leq u);$$

in the above, we use that $\frac{u-\mu_1}{\sigma} > \frac{u-\mu_2}{\sigma}$ since $\mu_1 < \mu_2$, and therefore $\Phi(\frac{u-\mu_1}{\sigma}) > \Phi(\frac{u-\mu_2}{\sigma})$, as Φ is an increasing function on the real line. So, Y is stochastically larger than X. The proof will not work if X, Y have unequal variances.
You will notice that this proof carries over to any location parameter family, and is not limited to the normal case.

Example 1.30. (Stochastic Ordering of χ^2 Random Variables). Suppose $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$, where $0 < m < n$. There is a neat way to prove that Y is stochastically larger than X. Write $Y = X + Z$, where $Z \sim \chi^2(n-m)$; this can always be done, and the meaning of it is that the distribution of Y is the same as the distribution of $X + Z$. Since Z is a positive random variable (being a chi square!), for any given u,

$$P(Y > u) = P(X + Z > u) > P(X > u),$$

which proves that Y is stochastically larger than X. The following intuitively plausible result is actually true.
Proposition Suppose Y is stochastically larger than X and $g(z)$ is a nondecreasing real valued function on the real line. Then,
(a) $g(Y)$ is stochastically larger than $g(X)$.
(b) $E[g(Y)] \geq E[g(X)]$.
(c) If X, Y are nonnegative random variables, then $E(Y^\alpha) \geq E(X^\alpha)$ for all $\alpha \geq 0$.

1.5 Inequalities

The mean and the variance, together, have earned the status of being the two most common summaries of a distribution. A relevant question is whether μ, σ are useful summaries of the distribution of a random variable. The answer is a qualified yes. The inequalities below suggest that knowing just the values of μ, σ, it is in fact possible to say something useful about the full distribution.

Theorem 1.11. a) **(Chebyshev's Inequality).** Suppose $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, assumed to be finite. Let k be any positive number. Then

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2};$$

b) **(Markov's Inequality).** Suppose X takes only nonnegative values , and suppose $E(X) = \mu$, assumed to be finite. Let c be any postive number. Then,

$$P(X \geq c) \leq \frac{\mu}{c};$$

The virtue of these two inequalities is that they make no restrictive assumptions on the random variable X. Whenever μ, σ are finite, Chebyshev's inequality is applicable, and whenever μ, is finite, Markov's inequality applies, provided the random variable is nonnegative. However, the universal nature of these inequalities also makes them typically quite conservative. Inequalities better than Chebyshev's or Markov's inequality are available under additional restrictions on the distribution of the underlying random variable X. You will see them in the supplementary section at the end of this chapter.

Although Chebyshev's inequality usually gives conservative estimates for tail probabilities, it does imply a major result in probability theory in a special case.

Theorem 1.12. (Weak Law of Large Numbers). Let $X_1, X_2, \cdots$ be iid random variables, with $E(X_i) = \mu, \text{Var}(X_i) = \sigma^2 < \infty$. Then, for any $\epsilon > 0, P(|\bar{X} - \mu| > \epsilon) \to 0$, as $n \to \infty$.

Proof: By Chebyshev's inequality, and our previously observed fact that $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$,

$$P(|\bar{X} - \mu| > \epsilon) = P(|\bar{X} - \mu| > \frac{\sqrt{n}\epsilon}{\sigma}\frac{\sigma}{\sqrt{n}}) \leq \frac{1}{(\frac{\sqrt{n}\epsilon}{\sigma})^2} = \frac{\sigma^2}{n\epsilon^2}$$

$\to 0$, as $n \to \infty$.

The interpretation of this result is that if we take a large sample from a population, then most of the times our sample mean will be numerically close to the population mean. Occasionally, just by bad luck, even though we have taken a large sample, the sample mean will not be close enough to the population mean. But such bad luck will occur only occasionally, i.e., with a small probability.

There is a stronger version of the weak law of large numbers, which says that in fact, with certainty, $\bar{X}$ will converge to μ as $n \to \infty$. The precise mathematical statement is that

$$P(\lim_{n\to\infty} \bar{X} = \mu) = 1.$$

The only condition needed is that $E(|X_i|)$ should be finite. This is called the *strong law of large numbers.*

The area of probability inequalities is an extremely rich and diverse area. The reason for it is that inequalities are tremendously useful in giving approximate answers when the exact answer to a problem, or a calculation, is very hard or perhaps even impossible to obtain. We will periodically present and illustrate inequalities over the rest of the chapter. Some really basic inequalities based on moments are presented in the next theorem.

Theorem 1.13. (a) **(Cauchy-Schwarz Inequality)** Let X, Y be two random variables such that $E(X^2)$ and $E(Y^2)$ are finite. Then,

$$E(|XY|) \leq \sqrt{E(X^2)}\sqrt{E(Y^2)};$$

(b) **(Hölder's Inequality)** Let X, Y be two random variables, and $1 < p < \infty$ a real number such that $E(|X|^p) < \infty$. Let $q = \frac{p}{p-1}$, and suppose $E(|Y|^q) < \infty$. Then,

$$E(|XY|) \leq [E(|X|^p)]^{\frac{1}{p}}[E(|Y|^q)]^{\frac{1}{q}};$$

(c) **(Minkowski's Inequality)** Let X, Y be two random variables, and $p \geq 1$ a real number such that $E(|X|^p), E(|Y|^p) < \infty$. Then,

$$[E(|X+Y|^p)]^{\frac{1}{p}} \leq [E(|X|^p)]^{\frac{1}{p}} + [E(|Y|^p)]^{\frac{1}{p}},$$

and, in particular, if $E(|X|), E(|Y|)$ are both finite, then,

$$E(|X+Y|) \leq E(|X|) + E(|Y|),$$

known as the **Triangular Inequality**.

1.6 Generating and Moment Generating Functions

Studying distributions of random variables and their basic quantitative properties, such as expressions for moments, occupy a central role in both statistics and probability. It turns out that a function called the probability generating function is often a very useful mathematical tool in studying distributions of random variables. The moment generating function, which is related to the probability generating function, is also extremely useful as a mathematical tool in numerous problems.

Definition 1.19. The *probability generating function* (pgf), also called simply the *generating function*, of a nonnegative integer valued random variable X is defined as $G(s) = G_X(s) = E(s^X) = \sum_{x=0}^{\infty} s^x P(X = x)$, provided the expectation is finite.

In this definition, 0^0 is to be understood as being equal to 1. Note that $G(s)$ is always finite for $|s| \leq 1$, but it could be finite over a larger interval, depending on the specific random variable X.

Some of the most important properties of generating functions are the following.

Theorem 1.14. a) Suppose $G(s)$ is finite in some open interval containing the origin. Then, $G(s)$ is infinitely differentiable in that open interval, and $P(X = k) = \frac{G^{(k)}(0)}{k!}, k \geq 0$, where $G^{(0)}(0)$ means $G(0)$;

b) If $\lim_{s\uparrow 1} G^{(k)}(s)$ is finite, then $E[X(X-1)\cdots(X-k+1)]$ exists and is finite, and $G^{(k)}(1) = \lim_{s\uparrow 1} G^{(k)}(s) = E[X(X-1)\cdots(X-k+1)]$.

c) Let $X_1, X_2, \cdots, X_n$ be independent random variables, with generating functions $G_1(s), G_2(s), \cdots, G_n(s)$. Then the generating function of $X_1 + X_2 + \cdots + X_n$ equals

$$G_{X_1+X_2+\cdots+X_n}(s) = \prod_{i=1}^{n} G_i(s).$$

$E[X(X-1)\cdots(X-k+1)]$ is called the kth *factorial moment* of X. The kth factorial moment of X exists if and only if the kth moment $E(X^k)$ exists.

One reason that the generating function is useful as a tool is its *distribution determining property*, in the following sense.

Theorem 1.15. Let $G(s)$ and $H(s)$ be the generating functions of two random variables X, Y. If $G(s) = H(s)$ in any nonempty open interval, then X, Y have the same distribution.

Proof: Let $P(X = n) = p_n, P(Y = n) = q_n, n \geq 0$. Then, $G(s) = \sum_{n=0}^{\infty} s^n p_n$, and $H(s) = \sum_{n=0}^{\infty} s^n q_n$. If there is a nonempty open interval in which $\sum_{n=0}^{\infty} s^n p_n = \sum_{n=0}^{\infty} s^n q_n$, then from the theory of power series, $p_n = q_n \, \forall n \geq 0$, and therefore X, Y have the same distribution.

Summarizing, then, one can find from the generating function of a nonnegative integer valued random variable X, the pmf of X, and every moment of X, including the moments that are infinite.

Example 1.31. (Discrete Uniform Distribution). Suppose X has the discrete uniform distribution on $\{1, 2, \cdots, n\}$. Then, its generating function is

$$G(s) = E[s^X] = \sum_{x=1}^{n} s^x P(X = x) = \frac{1}{n} \sum_{x=1}^{n} s^x$$

$$= \frac{s(s^n - 1)}{n(s-1)},$$

by summing the geometric series $\sum_{x=1}^{n} s^x$. As a check, if we differentiate $G(s)$ once, we get

$$G'(s) = \frac{1 + s^n[n(s-1) - 1]}{n(s-1)^2}.$$

On applying L'Hospital's rule, we get that $G'(1) = \frac{n+1}{2}$, which, therefore is the mean of X.

We have defined the probability generating function only for nonnegative integer valued random variables. The moment generating function is usually discussed in the context of

general random variables, not necessarily integer valued, or discrete. The two functions are connected. Here is the formal definition.

Definition 1.20. Let X be a real valued random variable. The moment generating function (mgf) of X is defined as

$$\psi_X(t) = \psi(t) = E[e^{tX}],$$

whenever the expectation is finite.

Note that the mgf $\psi(t)$ of a random variable X *always* exists and is finite if $t = 0$, and $\psi(0) = 1$. It may or may not exist when $t \neq 0$. If it does exist for t in a nonempty open interval containing zero, then many properties of X can be derived by using the mgf $\psi(t)$; it is an extremely useful tool. If X is a nonnegative integer valued random variable, then writing s^X as $e^{X \log s}$, it follows that the (probability) generating function $G(s)$ is equal to $\psi(\log s)$, whenever$G(s) < \infty$. Thus, the two generating functions, namely the probability generating function, and the moment generating function are connected.

The following theorem explains the name of a moment generating function.

Theorem 1.16. (a) Suppose the mgf $\psi(t)$ of a random variable X is finite in some open interval containing zero. Then, $\psi(t)$ is infinitely differentiable in that open interval, and for any $k \geq 1$,

$$E(X^k) = \psi^{(k)}(0).$$

(b) **(Distribution Determining Property).** If $\psi_1(t), \psi_2(t)$ are the mgfs of two random variables X, Y, and if $\psi_1(t) = \psi_2(t)$ in some nonempty open interval containing zero, then X, Y have the same distribution.

(c) If $X_1, X_2, \cdots, X_n$ are independent random variables, and if each X_i has a mgf $\psi_i(t)$, existing in some open interval around zero, then $X_1 + X_2 + \cdots + X_n$ also has a mgf in that open interval, and

$$\psi_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} \psi_i(t).$$

Explicit calculation of mgfs of various interesting random variables is done later in this chapter; we only give one example below right now.

Closely related to the moments of a random variable are *central moments* of a random variable.

Definition 1.21. Let a random variable X have a finite jth moment for some specified $j \geq 1$. The jth *central moment* of X is defined as $\mu_j = E[(X - \mu)^j]$, where $\mu = E(X)$.

Remark: Note that, $\mu_1 = E(X - \mu) = 0$, and $\mu_2 = E(X - \mu)^2 = \sigma^2$, the variance of X. If X has a distribution *symmetric about zero*, then every odd order central moment, $E[(X - \mu)^{2k+1}]$ is easily proved to be zero, provided it exists.

Example 1.32. Suppose X is the sum of two rolls of a fair die. Then, X can be written as $X = X_1 + X_2$, where X_1, X_2 are the numbers obtained on the two rolls respectively. The mgf of each of X_1, X_2 is obtained from the general mgf for a discrete uniform distribution worked out above, using $n = 6$. By part (c) of the preceding theorem, we have,

$$\psi_X(t) = [\frac{e^t(e^{6t} - 1)}{6(e^t - 1)}]^2 = \frac{e^{2t}(e^{6t} - 1)^2}{36(e^t - 1)^2}.$$

1.7 Standard Discrete Distributions

A few special discrete distributions arise very frquently in applications. Either the underlying probability mechanism of a problem is such that one of these distributions is truly the correct distribution for that problem, or the problem may be such that one of these distributions is a very good choice to model that problem. The special distributions we present are the discrete uniform, Binomial, the geometric, the negative binomial, the hypergeometric, and the Poisson.

The Discrete Uniform Distribution The discrete uniform distribution represents a finite number of equally likely values. The simplest real life example is the face obtained when a fair die is rolled once. It can also occur in some other physical phenomena, particularly when the number of possible values is small, and the scientist feels that that they are just equally likely. If we let the values of the random variable be $1, 2, \cdots, n$, then the pmf of the discrete uniform distribution is $p(x) = \frac{1}{n}, x = 1, 2, \cdots, n$, We sometimes write $X \sim Unif\{1, 2, \cdots, n\}$.

The Binomial Distribution The binomial distribution represents a sequence of independent coin tossing experiments. Suppose a coin with probability $p, 0 < p < 1$ for heads in a single trial is tossed independently a prespecified number of times, say n times, $n \geq 1$. Let X be the number of times in the n tosses that a head is obtained. Then the pmf of X is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \cdots, n,$$

the $\binom{n}{x}$ term giving the choice of the x tosses out of the n tosses in which the heads occur. Coin tossing, of course, is just an artifact. Suppose a trial can result in only one of two outcomes, called a *success*(S) or a *failure*(F), the probability of obtaining a success being p in any trial. Such a trial is called a *Bernoulli trial*. Suppose a Bernoulli trial is repeated independently a prespecified number of times, say n times, Let X be the number of times in the n trials that a success is obtained. Then X has the pmf given above, and we say that X has a *Binomial distribution with parameters n and p*, and write $X \sim Bin(n, p)$.

The Geometric Distribution Suppose a coin with probability $p, 0 < p < 1$, for heads

in a single trial is repeatedly tossed until a head is obtained for the first time. Assume that the tosses are independent. Let X be the number of the toss at which the very first head is obtained. Then the pmf of X is:

$$P(X = x) = p(1-p)^{x-1}, x = 1, 2, 3, \cdots.$$

We say that X has a *geometric distribution with parameter* p, and we will write $X \sim Geo(p)$. A geometric distribution measures a waiting time for the first success in a sequence of independent Bernoulli trials, each with the same success probability p, i.e., the coin cannot change from one toss to another.

The Negative Binomial Distribution The negative binomial distribution is a generalization of a geometric distribution, when we repeatedly toss a coin with probability p for heads, independently, until a total number of r heads has been obtained, where r is some fixed integer ≥ 1. The case $r = 1$ corresponds to the geometric distribution. Let X be the number of the first toss at which the rth success is obtained. Then the pmf of X is:

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \cdots,$$

the term $\binom{x-1}{r-1}$ simply giving the choice of the $r-1$ tosses among the first $x-1$ tosses where the first $r-1$ heads were obtained. We say that X has a *negative binomial distribution with parameters* r *and* p, and we will write $X \sim NB(r, p)$.

The Hypergeometric Distribution The hypergeometric distribution also represents the number of successes in a prespecified number of Bernoulli trials, but the trials happen to be dependent. A typical example is that of a finite population in which there are in all N objects, of which some D are of type I, and the other $N - D$ are of type II. A *without replacement sample* of size $n, 1 \leq n < N$ is chosen at random from the population. Thus, the selected sampling units are necessarily different. Let X be the number of units or individuals of type I among the n units chosen. Then the pmf of X is:

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},$$

$n - N + D \leq x \leq D$. We say that such an X has a *Hypergeometric distribution with parameters* n, D, N, and we will write $X \sim Hypergeo(n, D, N)$.

The Poisson Distribution The Poisson distribution is perhaps the most used and useful distribution for modelling nonnegative integer valued random variables.
The pmf of a *Poisson distribution with parameter* λ is:

$$P(X = x), \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \cdots;$$

by using the power series expansion of $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$, it follows that this is indeed a valid pmf.
Three specific situations where a Poisson distribution is almost routinely adopted as a model are the following:

(A) The number of times a specific event happens in a specified period of time; e.g., the number of phone calls received by someone over a 24 hour period.

B) The number of times a specific event or phenomenon is observed in a specified amount of area or volume; e.g. the number of bacteria of a certain kind in one liter of a sample of water, or the number of misprints per page of a book, etc.

(C) The number of times a success is obtained when a Bernoulli trial with success probability p is repeated independently n times, with p being small and n being large, such that the product np has a *moderate value*, say between .5 and 10. Thus, although the *true* distribution is a binomial, a Poisson distribution is used as an effective and convenient *approximation*.

We now present the most important properties of these special discrete distributions.

Theorem 1.17. Let $X \sim Unif\{1, 2, \cdots, n\}$. Then,

$$\mu = E(X) = \frac{n+1}{2}; \sigma^2 = \text{Var}(X) = \frac{n^2-1}{12}; E(X-\mu)^3 = 0; E(X-\mu)^4 = \frac{(3n^2-7)(n^2-1)}{240}.$$

Theorem 1.18. Let $X \sim Bin(n, p)$. Then,

$$(a)\, \mu = E(X) = np; \sigma^2 = \text{Var}(X) = np(1-p);$$

$$(b)\, \text{The mgf of } X \text{ equals } \psi(t) = (pe^t + 1 - p)^n \text{ at any } t;$$

$$(c)\, E[(X-\mu)^3] = np(1 - 3p + 2p^2);$$

$$(d)\, E[(X-\mu)^4] = np(1-p)[1 + 3(n-2)p(1-p)].$$

Theorem 1.19. (a) Let $X \sim Geo(p)$. Let $q = 1 - p$. Then,

$$E(X) = \frac{1}{p}; \text{Var}(X) = \frac{q}{p^2}.$$

(b) Let $X \sim NB(r, p), r \geq 1$. Then,

$$E(X) = \frac{r}{p}; \text{Var}(X) = \frac{rq}{p^2};$$

Furthermore, the mgf and the (probability) generating function of X equal

$$\psi(t) = (\frac{pe^t}{1 - qe^t})^r, t < \log(\frac{1}{q});$$

$$G(s) = (\frac{ps}{1-qs})^r, s < \frac{1}{q}.$$

Theorem 1.20. Let $X \sim Hypergeo(n, D, N)$, and let $p = \frac{D}{N}$. Then,

$$E(X) = np; \text{Var}(X) = np(1-p)(\frac{N-n}{N-1}).$$

Theorem 1.21. Let $X \sim Poi(\lambda)$. Then,

$$(a)\, E(X) = \text{Var(X)} = \lambda;$$

$$(b)\, E(X-\lambda)^3 = \lambda; E(X-\lambda)^4 = 3\lambda^2 + \lambda;$$

(c) The mgf of X equals

$$\psi(t) = e^{\lambda(e^t - 1)}.$$

Let us now see some illustrative examples.

Example 1.33. (Guessing on a Multiple Choice Exam). A multiple choice test with 20 questions has 5 possible answers for each question. A completely unprepared student picks the answer for each question at random and independently. Suppose X is the number of questions that the student answers correctly.

We identify each question with a Bernoulli trial and a correct answer as a success. Since there are 20 questions and the student picks his answer at random from five choices, $X \sim Bin(n, p)$, with $n = 20, p = \frac{1}{5} = .2$. We can now answer any question we want about X.
For example,

$$P(\text{The student gets every answer wrong}) = P(X = 0) = .8^{20} = .0115,$$

while,

$$P(\text{The student gets every answer right}) = P(X = 20) = .2^{20} = 1.05 \times 10^{-14},$$

a near impossibility. Suppose the instructor has decided that it will take at least 13 correct answers to pass this test. Then,

$$P(\text{The student will pass}) = \sum_{x=13}^{20} \binom{20}{x} .2^x .8^{20-x} = .000015,$$

still a very very small probability.

Example 1.34. Suppose a fair coin is tossed $n = 2m$ times. What is the probability that the number of heads obtained will be an even number?
Since $X =$ The number of heads $\sim Bin(2m, \frac{1}{2})$, we want to find:

$$P(X = 0) + P(X = 2) + \cdots + P(X = 2m) = \sum_{x=0}^{m} \binom{2m}{2x} / 2^{2m} = 2^{2m-1}/2^{2m}$$

$= \frac{1}{2}$, on using the identity that for any n,

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1}.$$

Thus, with a fair coin, the chances of getting an even number of heads in an even number of tosses are $\frac{1}{2}$. The same is true also if the number of tosses is odd, and is proved similarly.

Example 1.35. (Lack of Memory of Geometric Distribution). Let $X \sim Geo(p)$, and suppose m, n are given positive integers. Then, X has the interesting property

$$P(X > m + n | X > n) = P(X > m).$$

That is, suppose you are waiting for some event to happen for the first time. You have tried, say, 20 times, and you still have not succeeded. You may feel that it is due anytime now. But the chance that it will take another ten tries is the same as what it would be if you just started, and *forget* that you have been patient for a long time and have already tried very hard for a success.
The proof is simple. Indeed,

$$P(X > m + n | X > n) = \frac{P(X > m + n)}{P(X > n)} = \frac{\sum_{x>m+n} p(1-p)^{x-1}}{\sum_{x>n} p(1-p)^{x-1}}$$

$$= \frac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m = P(X > m).$$

Example 1.36. Suppose a door to door salesman makes an actual sale in 25% of the visits he makes. He is supposed to make at least two sales per day. How many visits should he plan on making to be 90% sure of making at least two sales?

Let X be the first visit at which the second sale is made. Then, $X \sim NB(r, p)$ with $r = 2, p = .25$. Therefore, X has the pmf $P(X = x) = (x-1)(.25)^2(.75)^{x-2}, x = 2, 3, \cdots$. Summing, for any given $k, P(X > k) = \sum_{x=k+1}^{\infty}(x-1)(.25)^2(.75)^{x-2} = \frac{k+3}{3}(3/4)^k$ (try to derive this). We want $\frac{k+3}{3}(3/4)^k \leq .1$. By directly computing this, we find that $P(X > 15) < .1$, but $P(X > 14) > .1$. So the salesman should plan on making 15 visits.

Example 1.37. (A Classic Example: Capture-Recapture). An ingenious use of the Hypergeometric distribution in estimating the size of a finite population is the *capture-recapture* method. It was originally used for estimating the total number of fish in a body

of water, such as a pond. Let N be the number of fish in the pond. In this method, a certain number of fish, say D of them are initially captured and tagged with a safe mark or identification device, and are returned to the water. Then, a second sample of n fish is recaptured from the water. Assuming that the fish population has not changed in any way in the intervening time, and that the initially captured fish remixed with the fish population homogeneously, the number of fish in the second sample, say X, which bear the mark is a hypergeometric random variable, namely, $X \sim Hypergeo(n, D, N)$. We know that the expected value of a hypergeometric random variable is $n\frac{D}{N}$. If we set, as a formalism, $X = n\frac{D}{N}$. and *solve* for N, we get $N = \frac{nD}{X}$. This is an estimate of the total number of fish in the pond. Although the idea is extremely original, this estimate can run into various kinds of difficulties, if for example the first catch of fish cluster around after being returned, or *hide*, or if the fish population has changed between the two catches due to death or birth, and of course if X turns out to be zero. Modifications of this estimate (known as the *Petersen estimate*) are widely used in wildlife estimation, census, and by the government for estimating tax frauds and number of people inflicted with some infection.

Example 1.38. (Events over Time). April receives three phone calls at her home on the average per day. On what percentage of days, does she receive no phone calls; more than five phone calls?

Because the number of calls received in a 24 hour period counts the occurrences of an event in a fixed time period, we model $X =$ Number of calls received by April on one day as a Poisson random variable with mean 3. Then,

$$P(X = 0) = e^{-3} = .0498; P(X > 5) = 1 - P(X \leq 5) = 1 - \sum_{x=0}^{5} e^{-3}3^x/x!$$
$$= 1 - .9161 = .0839.$$

Thus, she receives no calls on 4.98% of the days and she receives more than five calls on 8.39% of the days. *It is important to understand that X has only been modeled as a Poisson random variable, and other models could also be reasonable.*

Example 1.39. (Events over an Area). Suppose a 14 inch circular pizza has been baked with 20 pieces of barbecued chicken. At a party, you were served a $4 \times \times 4 \times 2$ (in inches) triangular slice. What is the probability that you got at least one piece of chicken?

The area of a circle of radius 7 is $\pi \times 7^2 = 153.94$. The area of a triangular slice of side lengths 4, 4, and 2 inches is $\sqrt{s(s-a)(s-b)(s-c)} = \sqrt{5 \times 1 \times 1 \times 3} = \sqrt{15} = 3.87$, where a, b, c are the three side lengths and $s = (a + b + c)/2$. Therefore, we model X, the number of pieces of chicken in the triangular slice as $X \sim Poi(\lambda)$, where $\lambda = 20 \times 3.87/153.94 = .503$. Using the Poisson pmf,

$$P(X \geq 1) = 1 - e^{-.503} = .395.$$

Example 1.40. (Gamma Ray Bursts). Gamma ray bursts are thought to be the most intense electromagnetic events observed in the sky, and they typically last a few seconds. While they are on, their intense brightness covers up any other gamma ray source in the sky. They occur at the rate of about one episode per day. It was initially thought that they are events within the Milky Way galaxy. But most astronomers now believe that to be not true, or not entirely true.

The 2000th gamma ray burst since 1991 was detected at the end of 1997 at NASA's Compton Gamma Ray Observatory. Are these data compatible with a model of Poisson distributed number of bursts with a rate of one per day?

Using a model of homogeneously distributed events, the number of bursts in a seven year period is $Poi(\lambda)$ with $\lambda = 7 \times 365 \times 1 = 2555$. The observed number of bursts is 2000, less than the expected number of bursts. But is it so much less that the postulated model will be in question? To assess this, we calculate $P(X \leq 2000)$, the probability that we could observe an observation as deviant from the expected as we did, just by chance. Statisticians call such a deviation probability a *P value*. The P value, then equals,

$$P(X \leq 2000) = \sum_{x=0}^{2000} \frac{e^{-2555}(2555)^x}{x!};$$

due to the large values of λ and the range of the summation, directly summing this is not recommended. But the sum can be approximated by using various other indirect means, including a theorem, known as the *central limit theorem*, which we will later discuss in detail. The approximate P value can be seen to be extremely small, virtually zero. So, the chance of such a deviant observation, if the Poisson model at the rate of one per day was correct, is very very small. One would doubt the model in such a case. The bursts may not occur at a homogeneous rate of one per day.

Example 1.41. (A Hierarchical Model with a Poisson Base). Suppose a chick lays a $Poi(\lambda)$ number of eggs in some specified period of time, say a month. Each egg has a probability p of actually developing. We want to find the distribution of the number of eggs that actually develop during that period of time.

Let $X \sim Poi(\lambda)$ denote the number of eggs the chick lays, and Y the number of eggs that develop. For example,

$$P(Y = 0) = \sum_{x=0}^{\infty} P(Y = 0|X = x)P(X = x) = \sum_{x=0}^{\infty} (1-p)^x \frac{e^{-\lambda}\lambda^x}{x!}$$

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

In general,

$$P(Y=y)=\sum_{x=y}^{\infty}\binom{x}{y}p^y(1-p)^{x-y}\frac{e^{-\lambda}\lambda^x}{x!}$$

$$=\frac{(p/(1-p))^y}{y!}e^{-\lambda}\sum_{x=y}^{\infty}\frac{1}{(x-y)!}(1-p)^x\lambda^x$$

$$=\frac{(p/(1-p))^y}{y!}e^{-\lambda}(\lambda(1-p))^y\sum_{n=0}^{\infty}\frac{(\lambda(1-p))^n}{n!},$$

on writing $n=x-y$ in the summation,

$$=\frac{(\lambda p)^y}{y!}e^{-\lambda}e^{\lambda(1-p)}=\frac{e^{-\lambda p}(\lambda p)^y}{y!},$$

and so, we recognize that $Y\sim Poi(\lambda p)$. What is interesting here is that the distribution still remains Poisson, under assumptions that seem to be very realistic physically.
Problems that should truly be modeled as hypergeometric distribution problems are often analyzed as if they were binomial distribution problems. That is, the fact that samples have been taken without replacement is ignored, and one pretends as if the successive draws are independent. When does it not matter that the dependence between the trials is ignored? Intuitively, we would think that if the population size N was large, and neither D nor $N-D$ was small, the trials would act like they are independent trials. The following theorem justifies this intuition.

Theorem 1.22. (Convergence of Hypergeometric to Binomial). Let $X=X_N\sim Hypergeo(n,D,N)$, where $D=D_N$ and N are such that $N\to\infty,\frac{D}{N}\to p, 0<p<1$. Then, for any fixed n, and for any fixed x,

$$P(X=x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}\to\binom{n}{x}p^x(1-p)^{n-x},$$

as $N\to\infty$.

This is proved by using *Stirling's approximation* (which says that as $k\to\infty$, $k!\sim e^{-k}k^{k+1/2}\sqrt{2\pi}$), for each factorial term in $P(X=x)$, and then doing some algebra.

1.7.1 Poisson Approximation

A Binomial random variable is the sum of n indicator variables. When the expectation of these indicator variables, namely p is small, and the number of summands n is large, the Poisson distribution provides a good approximation to the binomial. The Poisson distribution can also sometimes serve as a good approximation when the indicators are independent, but have different expectations p_i, or when the indicator variables have some weak dependence.

Theorem 1.23. Let $X_n \sim Bin(n, p_n), n \geq 1$. Suppose $np_n \to \lambda, 0 < \lambda < \infty$, as $n \to \infty$, Let $Y \sim Poi(\lambda)$. Then, for any given $k, 0 \leq k < \infty$,

$$P(X_n = k) \to P(Y = k),$$

as $n \to \infty$.

Proof: For ease of explanation, let us first consider the case $k = 0$. We have,

$$P(X_n = 0) = (1-p)^n = (1 - \frac{np}{n})^n \sim (1 - \frac{\lambda}{n})^n \sim e^{-\lambda}.$$

Note that we did not actually prove the claimed fact that $(1 - \frac{np}{n})^n \sim (1 - \frac{\lambda}{n})^n$; but it is true, and is not hard to prove.
Now consider $k = 1$. We have,

$$P(X_n = 1) = np(1-p)^{n-1} = (np)(1-p)^n \frac{1}{1-p} \sim \lambda(e^{-\lambda})(1)$$

$$= \lambda e^{-\lambda}.$$

The same technique works for any k. Indeed, for a general k,

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{k!}[n(n-1)\cdots(n-k+1)]p^k(1-p)^n[\frac{1}{(1-p)^k}]$$

$$= \frac{1}{k!} n^k [1 \frac{n-1}{n} \cdots \frac{n-k+1}{n}] p^k (1-p)^n [\frac{1}{(1-p)^k}]$$

$$= \frac{1}{k!} (np)^k [1 \frac{n-1}{n} \cdots \frac{n-k+1}{n}] (1-p)^n [\frac{1}{(1-p)^k}]$$

$$\sim \frac{1}{k!} (\lambda)^k [1] e^{-\lambda} [1] = \frac{e^{-\lambda}\lambda^k}{k!},$$

which is what the theorem says.

Example 1.42. (An Insurance Example). Suppose 5000 clients are each insured for one million dollars against fire damage in a coastal property. Each residence has a 1 in 10000 chance of being damaged by fire in a 12 month period. How likely is it that the insurance company has to pay out as much as 3 million dollars (4 million dollars) in fire damage claims in one year?

If X is the number of claims made during a year, then $X \sim Bin(n, p)$ with $n = 5000$ and $p = 1/10000$. We assume that no one makes more than one claim and that the clients are independent. Then we can approximate the distribution of X by $Poi(np) = Poi(.5)$. We need:

$$P(X \geq 3) = 1 - P(X \leq 2) \approx 1 - (1 + .5 + .5^2/2)e^{-.5} = .014,$$

and

$$P(X \geq 4) = 1 - P(X \leq 3) \approx 1 - (1 + .5 + .5^2/2 + .5^3/6)e^{-.5} = .002.$$

These two calculations are done above by using the Poisson approximation, namely $\frac{e^{-.5}.5^k}{k!}$, for $P(X = k)$. The insurance company is quite safe being prepared for 3 million dollars in payout, and very safe being prepared for 4 million dollars.

Example 1.43. (Poisson Approximation in the Birthday Problem). In the classic *birthday problem*, n unrelated people gather around and we want to know if there is at least one pair of individuals with the same birthday. Defining $I_{i,j}$ as the indicator of the event that individuals i, j have the same birthday, we have

$$X = \text{Number of different pairs of people who share a common birthday}$$

$$= \sum_{1 \leq i < j \leq n} I_{i,j}.$$

Each $I_{i,j} \sim Ber(p)$, where $p = 1/365$. Note, however, that the $I_{i,j}$ are definitely not independent. Now, the expected value of X is $\lambda = \binom{n}{2}/365$. This is moderate ($> .5$) if $n \geq 20$. So, a Poisson approximation may be accurate when n is about 20 or more.

If we use a Poisson approximation when $n = 23$, we get

$$P(X > 0) \approx 1 - e^{-\binom{23}{2}/365} = 1 - e^{-.693151} = .500002,$$

which is almost exactly equal to the true value of the probability that there will be a pair of people with the same birthday in a group of 23 people.

1.7.2 Distribution of Sums

Sums of random variables arise very naturally in practical applications. For example, the revenue over a year is the sum of the monthly revenues; the time taken to finish a test with ten problems is the sum of the times taken to finish the individual problems, etc. Sometimes we can reasonably assume that the various random variables being added are independent. Thus, the following general question is an important one:
Suppose $X_1, X_2, \cdots, X_k$ are k independent random variables, and suppose we know the distributions of the individual X_i. What is the distribution of the sum $X_1 + X_2 + \cdots + X_k$? In general, this is a very difficult question. Interestingly, if the individual X_i have one of the distinguished distributions we have discussed in this chapter, then their sum is also often a distribution of that same type.

Theorem 1.24. a) Suppose $X_1, X_2, \cdots, X_k$ are k independent Binomial random variables, with $X_i \sim Bin(n_i, p)$. Then $X_1 + X_2 + \cdots + X_k \sim Bin(n_1 + n_2 + \cdots + n_k, p)$;

b) Suppose $X_1, X_2, \cdots, X_k$ are k independent Negative Binomial random variables, with $X_i \sim NB(r_i, p)$. Then $X_1 + X_2 + \cdots + X_k \sim NB(r_1 + r_2 + \cdots + r_k, p)$;

c) Suppose $X_1, X_2, \cdots, X_k$ are k independent Poisson random variables, with $X_i \sim Poi(\lambda_i)$. Then $X_1 + X_2 + \cdots + X_k \sim Poi(\lambda_1 + \lambda_2 + \cdots \lambda_k)$.

Proof: Each of the three parts can be proved by various means. One possibility is to directly attack the problem. Alternatively, the results can also be proved by using generating functions or mgfs. It is useful to see a proof using both methods, and so, we show both methods of proof in the Poisson case. The proof for the other two cases is exactly the same and will be omitted.

First, note that it is enough to only consider the case $k = 2$, because then the general case follows by induction. We denote X_1, X_2 as X, Y for notational simplicity. Then,

$$P(X+Y=z) = \sum_{x=0}^{z} P(X=x, Y=z-x) = \sum_{x=0}^{z} P(X=x)P(Y=z-x)$$

$$= \sum_{x=0}^{z} \frac{e^{-\lambda_1}\lambda_1^x}{x!}\frac{e^{-\lambda_2}\lambda_2^{z-x}}{(z-x)!}$$

$$= e^{-(\lambda_1+\lambda_2)}\lambda_2^z \times \sum_{x=0}^{z} \frac{(\lambda_1/\lambda_2)^x}{x!(z-x)!}$$

$$= e^{-(\lambda_1+\lambda_2)}\frac{\lambda_2^z}{z!}\sum_{x=0}^{z}\binom{z}{x}(\lambda_1/\lambda_2)^x$$

$$= e^{-(\lambda_1+\lambda_2)}\frac{\lambda_2^z}{z!}(1+\lambda_1/\lambda_2)^z = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^z}{z!},$$

as was required to prove.

The second method uses the formula for the mgf of a Poisson distribution. Since X, Y are both Poisson and they are independent, the mgf of $X + Y$ is

$$\psi_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)}$$

$$= e^{(\lambda_1+\lambda_2)(e^t-1)},$$

which agrees with the mgf of the $Poi(\lambda_1+\lambda_2)$ distribution, and therefore by the distribution determining property of mgfs, the distribution of $X + Y$ must be $Poi(\lambda_1 + \lambda_2)$.

The calculation used in each of these two methods of proof is a useful calculation, and it is important to be familiar with each method.

Example 1.44. Suppose $X \sim Poi(1), Y \sim Poi(5)$, and $Z \sim Poi(10)$, and suppose X, Y, Z are independent. We want to find $P(X + Y + Z \geq 20)$.

By the previous theorem, $X + Y + Z \sim Poi(16)$, and therefore

$$P(X + Y + Z \geq 20) = 1 - P(X + Y + Z \leq 19) = 1 - \sum_{x=0}^{19} \frac{e^{-16}16^x}{x!}$$

$$= 1 - .8122 = .1878.$$

In the absence of the result that $X + Y + Z \sim Poi(16)$, computing this probability would call for enumeration of all the ways that $X + Y + Z$ could be 19 or smaller and adding up those probabilities. Clearly, it would be a much more laborious calculation.

1.8 Continuous Random Variables

Discrete random variables serve as good examples to develop probabilistic intuition, but they do not account for all the random variables that one studies in theory and in applications. We now introduce the so called *continuous random variables*, which typically take all values in some nonempty interval, e.g., the unit interval, or the entire real line, etc. The right probabilistic paradigm for continuous variables cannot be pmfs. Instead of pmfs, we operate with a density function for the variable. The density function fully describes the distribution.

Definition 1.22. Let X be a real valued random variable taking values in $\mathcal{R}$, the real line. A function $f(x)$ is called the *density function* or the *probability density function* (pdf) of X if

$$\text{For all } a, b, -\infty < a \leq b < \infty, P(a \leq X \leq b) = \int_a^b f(x)dx;$$

in particular, for a function $f(x)$ to be a density function of some random variable, it must satisfy:

$$f(x) \geq 0 \, \forall x \in \mathcal{R}; \int_{-\infty}^{\infty} f(x)dx = 1.$$

The statement that $P(a \leq X \leq b) = \int_a^b f(x)dx$ is the same as saying that if we plot the density function $f(x)$, then the area under the graph between a and b will give the probability that X is between a and b, while the statement that $\int_{-\infty}^{\infty} f(x)dx = 1$ is the same as saying that the area under the entire graph must be one.

The density function $f(x)$ can in principle be used to calculate the probability that the random variable X belongs to a general set A, not just an interval. Indeed, $P(X \in A) = \int_A f(x)dx$.

Caution Integrals over completely general sets A in the real line are not defined. To make this completely rigorous, one has to use measure theory and concepts of a *Lebesgue integral*. We will, however, generally only want to calculate $P(X \in A)$ for sets A that are countable union of intervals. For such sets, defining the integral $\int_A f(x)dx$ would not be

a problem and we can proceed as if we are just calculating ordinary integrals.
The definition of the CDF (cumulative distribution function) remains the same as before.

Definition 1.23. Let X be a continuous random variable with a pdf $f(x)$. Then the CDF of X is defined as

$$F(x) = P(X \leq x) = P(X < x) = \int_{-\infty}^{x} f(t)dt.$$

Remark: At any point x_0 at which $f(x)$ is continuous, the CDF $F(x)$ is differentiable, and $F'(x_0) = f(x_0)$. In particular, if $f(x)$ is continuous everywhere, then $F'(x) = f(x)$ at all x.

Again, to be strictly rigorous, one really needs to say in the above sentence that $F'(x) = f(x)$ at *almost all* x, a concept in measure theory.

Example 1.45. (Density vs CDF). Consider the functions

$$f(x) = 1, \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1];$$

$$f(x) = 3x^2, \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1];$$

$$f(x) = 6x(1-x), \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1];$$

$$f(x) = \frac{1}{\sqrt{x(1-x)}}, \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1];$$

$$f(x) = 4x^2 - \frac{2}{3}x, \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1].$$

We want to verify which, if any, of these functions is a valid density function.

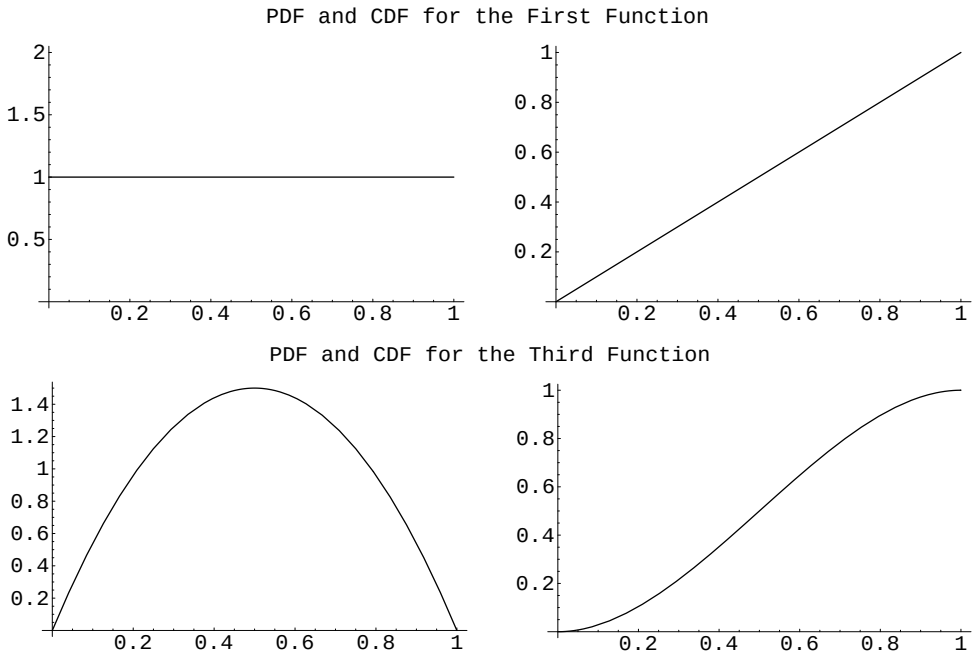
The first four functions are all clearly nonnegative; however, the last function in our list is negative if $4x^2 < \frac{2}{3}x$ (if $x < \frac{1}{6}$), and therefore it is not a valid pdf. Thus, we only need to verify if the first four functions integrate to one. Note that each function is zero when $x \notin [0,1]$, and so, $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 f(x)dx$. So we need to verify whether $\int_0^1 f(x)dx = 1$ for the first four functions.

For the first two functions, it is immediately verified that $\int_0^1 f(x)dx = 1$. For the third function,

$$\int_0^1 6x(1-x)dx = 6\int_0^1 x(1-x)dx = 6[\int_0^1 xdx - \int_0^1 x^2dx] = 6[\frac{1}{2} - \frac{1}{3}] = 1;$$

for the fourth function,

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}}dx = \int_0^{\pi/2} \frac{1}{\sin t \cos t} 2\sin t \cos t dt = \int_0^{\pi/2} 2dt = \pi,$$



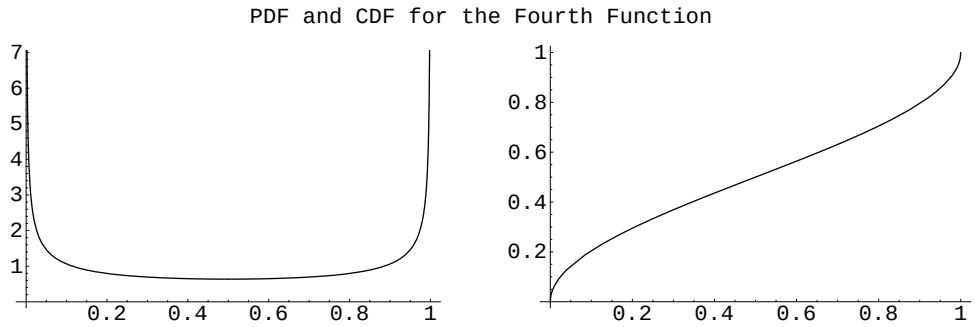
on making the substitution $x = \sin^2 t$. Since the function integrates to π, rather than to one, it is not a valid pdf; however, if we consider instead the function

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \text{ if } 0 \leq x \leq 1; 0 \text{ if } x \notin [0,1],$$

then it is both nonnegative and integrates to one, and so it will be a valid pdf. The constant $c = \frac{1}{\pi}$ is called a *normalizing constant*.

It is instructive to see a plot of these functions on $[0,1]$ to appreciate that density functions can take a variety of shapes. The only restriction being that they should be nonnegative and should integrate to one, there are no shape restrictions on a density function in general. For example, of our four functions, one is a constant, one is increasing, one is symmetric, first increasing and then decreasing, and the fourth one is shaped like a cereal bowl, being unbounded as $x \to 0, 1$. The density function which is constantly equal to 1 in the interval $[0,1]$ is known as the *Uniform density on* $[0,1]$. The word *uniform* suggests that we assign uniformly the same importance to every value in $[0,1]$. We can analogously define a uniform density on *any* bounded interval $[a,b]$; again, the density is constant throughout the interval $[a,b]$. If a random variable X is uniformly distributed on a bounded interval $[a,b]$, we write $X \sim U[a,b]$.

Side by side, for three of these density functions, we also plot the CDF. Note that the CDF is always a smooth, nondecreasing function, starting at zero when $x = 0$, and ending at one when $x = 1$. Unlike the density functions, the CDF has a certain uniformity in shape.



Example 1.46. (From CDF to PDF and Median). Consider the function $F(x) = 0,$ if $x < 0; = 1 - e^{-x}$ if $0 \leq x < \infty$. This is a nonnegative nondecreasing function, goes to one as $x \to \infty$, is continuous at any real number x, and is also differentiable at any x except $x = 0$. Thus, it is the CDF of a continuous random variable, and the PDF can be obtained by the relation $f(x) = F'(x) = e^{-x}, 0 < x < \infty$, and $f(x) = F'(x) = 0, x < 0$. At $x = 0$, $F(x)$ is *not* differentiable. But we can define the PDF in any manner we like at one specific point; so to be specific, we will write our PDF as

$$f(x) = e^{-x} \text{ if } 0 \leq x < \infty;$$

$$= 0, \text{ if } x < 0.$$

This density is called *the standard Exponential density* and is enormously important in practical applications.

From the formula for the CDF, we see that $F(m) = .5 \Rightarrow 1 - e^{-m} = .5 \Rightarrow e^{-m} = .5 \Rightarrow m = \log 2 = .693$. Thus, we have established that the standard Exponential density has median $\log 2 = .693$.

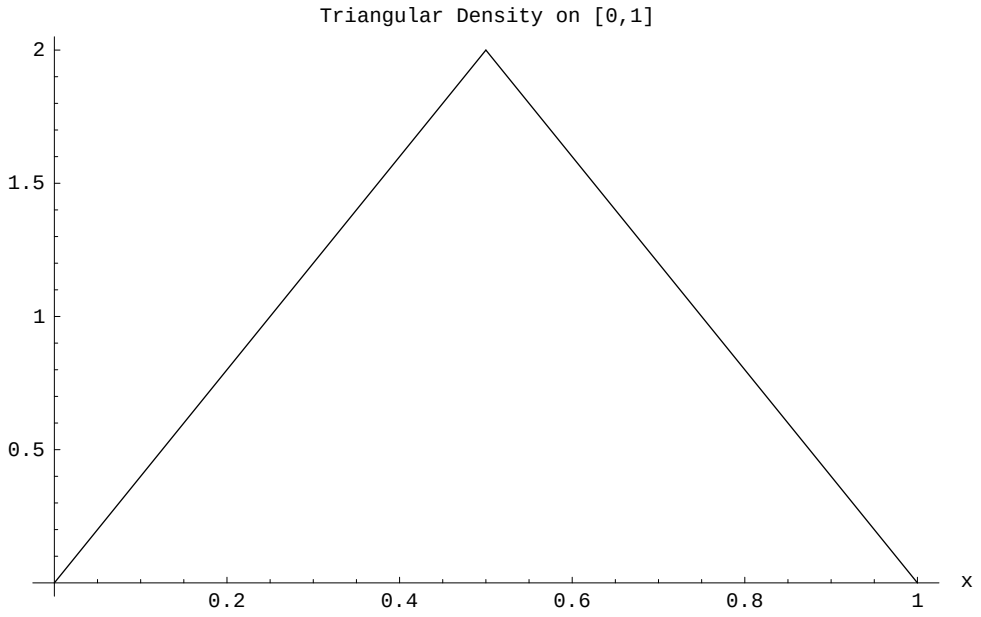
In general, given a number p, there can be infinitely many values x such that $F(x) = p$. Any such value splits the distribution into two parts, $100p\%$ of the probability below it, and $100(1-p)\%$ above. Such a value is called the pth quantile or percentile of F. However, in order to give a prescription for choosing a unique value when there is more than one x at which $F(x) = p$, the following definition is adopted.

Definition 1.24. Let X have the CDF $F(x)$. Let $0 < p < 1$. The pth *quantile* or the pth *percentile* of X is defined to be the first x such that $F(x) \geq p$:

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

The function $F^{-1}(p)$ is also sometimes denoted as $Q(p)$ and is called the *quantile function of F or X*.

Two very familiar concepts in probability and statistics are those of *symmetry* and *unimodality*. Symmetry of a density function means that around some point, the density



has two halves which are exact mirror images of each other. Unimodality means that the density has just one peak point at some value. We give the formal definitions.

Definition 1.25. A density function $f(x)$ is called *symmetric* around a number M if $f(M+u) = f(M-u) \forall\, u > 0$. In particular, $f(x)$ is symmetric around zero if $f(u) = f(-u) \forall\, u > 0$.

Definition 1.26. A density function $f(x)$ is called *strictly unimodal* at (or around) a number M if $f(x)$ is increasing for $x < M$, and decreasing for $x > M$.

Example 1.47. (The Triangular Density). Consider the density function

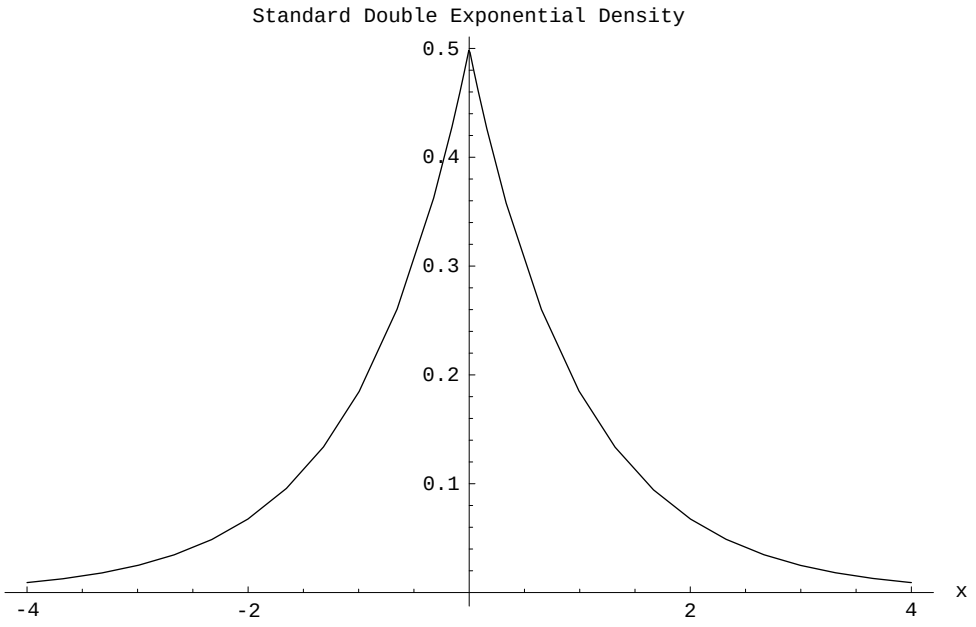
$$f(x) = cx, 0 \le x \le \frac{1}{2};$$

$$= c(1-x), \frac{1}{2} \le x \le 1,$$

where c is a normalizing constant. It is easily verified that $c = 4$. This density consists of two different linear segments on $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. A plot of this density looks like a triangle, and it is called the *triangular density* on $[0, 1]$. Note that it is symmetric and strictly unimodal.

Example 1.48. (The Double Exponential Density). We have previously seen the standard Exponential density on $[0, \infty)$ defined as $e^{-x}, x \ge 0$. We can extend this to the negative real numbers by writing $-x$ for x in the above formula; i.e., simply define the density to be e^x for $x \le 0$. Then, we have an overall function which equals

$$e^{-x} \text{ for } x \ge 0;$$



e^x for $x \leq 0$.

This function integrates to

$$\int_0^\infty e^{-x}dx + \int_{-\infty}^0 e^x dx = 1 + 1 = 2.$$

So, if we use a *normalizing constant* of $\frac{1}{2}$, then we get a valid density on the entire real line:

$$f(x) = \frac{1}{2}e^{-x} \text{ for } x \geq 0;$$

$$f(x) = \frac{1}{2}e^{x} \text{ for } x \leq 0.$$

The two lines can be combined into one formula as

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

This is the *standard Double Exponential density*, and is symmetric, unimodal, and has a *cusp* at $x = 0$; see the plot.

Example 1.49. (The Normal density). The double exponential density tapers off to zero at the linear exponential rate at both tails, i.e., as $x \to \pm\infty$. If we force the density to taper off at a quadratic exponential rate, then we will get a function like e^{-ax^2}, for some chosen $a > 0$. While this is obviously nonnegative, and also has a finite integral over the whole real line, it does not integrate to one. So we need a normalizing constant to make it a valid density function. Densities of this form are called *normal densities*, and occupy the central place among all distributions in the theory and practice

of probability and statistics. Gauss, while using the method of least squares for analyzing astronomical data, used the normal distribution to justify least squares methods; the normal distribution is also often called the *Gaussian distribution*, although de Moivre and Laplace both worked with it before Gauss. Physical data on many types of variables approximately fit a normal distribution. Theory of statistical methods is often the best understood when the underlying distribution is normal. The normal distributions have many unique properties not shared by any other distribution. Because of all these reasons, the normal density, also called *the bell curve*, is the most used, important, amd well studied distribution.

Let

$$f(x) = f(x|\mu, \sigma) = ce^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

where c is a normalizing constant. The normalizing constant can be proved to be equal to $\frac{1}{\sigma\sqrt{2\pi}}$. Thus, a normal density with parameters μ and σ is given by

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

We write $X \sim N(\mu, \sigma^2)$; we will see later that the two parameters μ and σ^2 are the mean and the variance of this distribution. Note that the $N(\mu, \sigma^2)$ density is a location-scale parameter density. If $\mu = 0$ and $\sigma = 1$, this simplifies to the formula $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$, and is universally denoted by the notation $\phi(x)$. It is called the *standard normal density*. The standard normal density, then, is:

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

Consequently, the CDF of the standard normal density is the function $\int_{-\infty}^{x} \phi(t)dt$. It is *not* possible to express the CDF in terms of the elementary functions. It is standard practice to denote it by using the notation $\Phi(x)$, and compute it using widely available tables or software for a given x, needed in a specific application.

The distribution of a continuous random variable is completely described if we describe either its density function, or its CDF. For flexible modelling, it is useful to know how to create new densities or new CDFs out of densities or CDFs that we have already thought of. This is similar to generating new functions out of old functions in calculus. The following theorem describes some standard methods to make new densities or CDFs out of already available ones.

Theorem 1.25. (a) **(Location-scale Shift).** Let $f(x)$ be any density function. Then, for any real number μ and any $\sigma > 0$,

$$g(x) = g_{\mu,\sigma}(x) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$$

is also a valid density function.

(b) **(Mixtures).** Let $f_1, f_2, \cdots, f_k$ be k densities for some $k, 2 \leq k < \infty$, and let $p_1, p_2, \cdots, p_k$ be k constants such that each $p_i \geq 0$, and $\sum_{i=1}^{k} p_i = 1$. Then,

$$f(x) = \sum_{i=1}^{k} p_i f_i(x)$$

is also a valid density function.

(c) **(Powers).** Let F be a CDF, and α any positive real number. Then,

$$G(x) = F^{\alpha}(x)$$

is also a valid CDF.

(d) **(Products).** Let $F_1, F_2, \cdots, F_k$ be k CDFs, for some $k, 2 \leq k < \infty$. Then,

$$G(x) = F_1(x)F_2(x)\cdots F_k(x)$$

is also a valid CDF.

Densities of the form $\sum_{i=1}^{k} p_i f_i(x)$ are called *mixture densities*, because they are formed by *mixing* $f_1, f_2, \cdots, f_k$ according to the weights $p_1, p_2, \cdots, p_k$. Mixture densities are greatly useful in generating densities of various shapes, and *tails*. See Chapter 2 for explicit examples.

1.9 Functions of a Continuous Random Variable

As for discrete random variables, we are often interested in the distribution of some function $g(X)$ of a continuous random variable X. For example, X could measure the input into some production process, and $g(X)$ could be a function that describes the output. For one-to-one functions $g(X)$, one has the following important formula.

Theorem 1.26. (The Jacobian Formula) Let X have a continuous pdf $f(x)$ and a CDF $F(x)$, and suppose $Y = g(X)$ is a strictly monotone function of X with a nonzero derivative. Then Y has the pdf

$$f_Y(y) = \frac{f(g^{-1}(y))}{|g'(g^{-1}(y))|},$$

where y belongs to the range of g.

Proof: Since $g(X)$ is strictly monotone, it has an inverse function. Suppose $g(X)$ is strictly increasing. Then,

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

$$= F(g^{-1}(y)).$$

On differentiating,

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

the proof for the strictly decreasing case is similar.

Example 1.50. (Simple Linear Transformations). Suppose X is any continuous random variable with a pdf $f(x)$ and let $Y = g(X)$ be the linear function (a location and scale change on X) $g(X) = a + bX, b \neq 0$. This is obviously a strictly monotone function, as $b \neq 0$. Take $b > 0$. Then the inverse function of g is $g^{-1}(y) = \frac{y-a}{b}$, and of course $g'(x) \equiv b$. Putting it all together, from the theorem above,

$$f_Y(y) = \frac{f(g^{-1}(y))}{|g'(g^{-1}(y))|} = \frac{1}{b}f(\frac{y-a}{b});$$

in general, whether b is positive or negative, the formula is:

$$f_Y(y) = \frac{1}{|b|}f(\frac{y-a}{b}).$$

Example 1.51. (The Cauchy Distribution). The Cauchy density, like the normal and the double exponential, is also symmetric and unimodal, but the properties are very different. It is such an atypical density that we often think of the Cauchy density first when we look for a counterexample to a conjecture. There is a very interesting way to obtain a Cauchy density from a uniform density by using the quantile transformation. We describe that derivation in this example.

Suppose a person holds a flashlight in his hand, and standing one foot away from an infinitely long wall, points the beam of light at a *random direction*. Here, by random direction, we mean that the point of landing of the light ray makes an angle X with the individual (considered to be a straight line one foot long), and this angle $X \sim U[-\pi/2, \pi/2]$. Let Y be the horizontal distance of the point at which the light lands from the person, with Y being considered negative if the light lands on the person's left, and it being considered positive if it lands on the person's right. vskip 2pt Then, by elementary trigonometry,

$$\tan(X) = \frac{Y}{1} \Rightarrow Y = \tan(X).$$

Now $g(X) = \tan X$ is a strictly monotone function of X, and the inverse function is $g^{-1}(y) = \arctan(y), -\infty < y < \infty$. Also, $g'(x) = 1 + \tan^2 x$. Putting it all together,

$$f_Y(y) = \frac{\frac{1}{\pi}}{1 + [\tan(\arctan y)]^2} = \frac{1}{\pi(1+y^2)}, -\infty < y < \infty.$$

This is the *standard Cauchy density*.

The Cauchy density is particularly notorious for its heavy tail. See Chapter 2.

Example 1.52. (From Exponential to Uniform). Suppose X has the standard Exponential density $f(x) = e^{-x}, x \geq 0$. Let $Y = g(X) = e^{-X}$. Again, $g(X)$ is a strictly monotone function, and the inverse function is found as follows:

$$g(x) = e^{-x} = y \Rightarrow x = -\log y = g^{-1}(y).$$

Also, $g'(x) = -e^{-x}$,

$$\Rightarrow f_Y(y) = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{e^{-(-\log y)}}{|e^{-(-\log y)}|}$$
$$= \frac{y}{y} = 1, 0 \leq y \leq 1.$$

We have thus proved that if X has a standard Exponential density, then $Y = e^{-X}$ is *uniformly distributed* on $[0, 1]$.

There is actually nothing special about choosing X to be the standard Exponential; the following important result says that what we saw in the above example is completely general for all continuous random variables.

Theorem 1.27. Let X have a continuous CDF $F(x)$. Consider the new random variables $Y = 1 - F(X)$ and $Z = F(X)$. Then both Y, Z are distributed as $U[0, 1]$.

It is useful to remember this result in informal notation:

$$F(X) = U, \text{ and } F^{-1}(U) = X.$$

The implication is a truly useful one. Suppose for purposes of computer experiments, we want to have computer simulated values of *some* random variable X which has *some* CDF F and the quantile function $Q = F^{-1}$. Then, all we need to do is to have the computer generate $U[0, 1]$ values, say $u_1, u_2, \cdots, u_n$, and use $x_1 = F^{-1}(u_1), x_2 = F^{-1}(u_2), \cdots, x_n = F^{-1}(u_n)$ as the set of simulated values for our random variable of interest, namely X. Thus, the problem can be reduced to simulation of uniform values, a simple task. The technique has so many uses, that there is a name for this particular function $Z = F^{-1}(U)$ of a uniform random variable U.

Definition 1.27. (Quantile Transformation). Let U be a $U[0, 1]$ random variab le and let $F(x)$ be a continuous CDF. Then the function of U defined as $X = F^{-1}(U)$ is called the *quantile transformation of* U, and it has exactly the CDF F.

Example 1.53. (An Interesting Function which is Not Strictly Monotone). Suppose X has the standard normal density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ on $(-\infty, \infty)$. We want to find the density of $Y = g(X) = X^2$. However, we immediately realize that X^2 is not a strictly monotone function on the whole real line (its graph is a parabola). Thus, the general

formula given above for densities of strictly monotone functions cannot be applied in this problem. We attack the problem directly. Thus,

$$P(Y \le y) = P(X^2 \le y) = P(X^2 \le y, X > 0) + P(X^2 \le y, X < 0)$$

$$= P(0 < X \le \sqrt{y}) + P(-\sqrt{y} \le X < 0)$$

$$= F(\sqrt{y}) - F(0) + [F(0) - F(-\sqrt{y})] = F(\sqrt{y}) - F(-\sqrt{y}),$$

where F is the CDF of X, i.e., the standard normal CDF.

Since we have obtained the CDF of Y, we now differentiate to get the pdf of Y:

$$f_Y(y) = \frac{d}{dy}[F(\sqrt{y}) - F(-\sqrt{y})] = \frac{f(\sqrt{y})}{2\sqrt{y}} - \frac{f(-\sqrt{y})}{-2\sqrt{y}}$$

(by use of the chain rule)

$$= \frac{f(\sqrt{y})}{2\sqrt{y}} + \frac{f(\sqrt{y})}{2\sqrt{y}}$$

(since f is symmetric around zero, i.e., $f(-u) = f(u)$ for any u)

$$= \frac{2f(\sqrt{y})}{2\sqrt{y}} = \frac{f(\sqrt{y})}{\sqrt{y}} = \frac{e^{-y/2}}{\sqrt{2\pi y}},$$

$y > 0$. This is a very special density in probability and statistics, and is called the *chi square density with one degree of freedom*. We have thus proved that the square of a standard normal random variable has a chi square distribution with one degree of freedom.

1.9.1 Expectation and Moments

For discrete random variables, expectation was seen to be equal to $\sum_x xP(X = x)$. Of course, for continuous random variables, the analogous sum $\sum_x xf(x)$ is not defined. The correct definition of expectation for continuous random variables replaces sums by integrals.

Definition 1.28. Let X be a continuous random variable with a pdf $f(x)$. We say that the expectation of X exists if $\int_{-\infty}^{\infty} |x| f(x)dx < \infty$, in which case the expectation, or the expected value, or the mean of X is defined as

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

Suppose X is a continuous random variable with a pdf $f(x)$ and $Y = g(X)$ is a function of X. If Y has a density, say $f_Y(y)$, then we can compute the expectation as $\int yf_Y(y)dy$, or as $\int g(x)f(x)dx$. Since Y need not always be a continuous random variable just because X is, it may not in general have a density $f_Y(y)$; but the second expression is always applicable and correct.

Theorem 1.28. Let X be a continuous random variable with pdf $f(x)$. Let $g(X)$ be a function of X. The expectation of $g(X)$ exists if and only if $\int_{-\infty}^{\infty} |g(x)|f(x)dx < \infty$, in which case the expectation of $g(X)$ is

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

The definitions of moments and the variance remain the same as in the discrete case.

Definition 1.29. Let X be a continuous random variable with pdf $f(x)$. Then the kth moment of X is defined to be $E(X^k), k \geq 1$. We say that the kth moment does not exist if $E(|X|^k) = \infty$.

Corollary 1.2. Suppose X is a continuous random variable with pdf $f(x)$. Then its variance, provided it exists, is equal to

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

One simple observation that saves calculations, but is sometimes overlooked, is the following fact; the proof of it merely uses the integration result that the integral of the product of an odd function and an even function on a symmetric interval is zero, if the integral exists.

Proposition Suppose X has a distribution symmetric around some number a, i.e., $X - a$ and $a - X$ have the same distribution. Then, $E[(X-a)^{2k+1}] = 0$, for any $k \geq 0$ for which the expectation $E[(X-a)^{2k+1}]$ exists.

For example, if X has a distribution symmetric about zero, then any odd moment, e.g., $E(X), E(X^3)$, etc., provided it exists, must be zero. There is no need to calculate it; it is automatically zero.

For the next example, we will need the definition of the *Gamma function*. It will be repeatedly necessary for us to work with the Gamma function in this text.

Definition 1.30. The Gamma function is defined as

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

In particular,

$$\Gamma(n) = (n-1)!, \text{ for any positive integer } n;$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \ \forall \alpha > 0;$$

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

Example 1.54. (Moments of Exponential). Let X have the standard Exponential density. Then, all its moments exist, and indeed,

$$E(X^n) = \int_0^\infty x^n e^{-x} dx = \Gamma(n+1) = n!.$$

In particular,

$$E(X) = 1; E(X^2) = 2,$$

and therefore, $\text{Var}(X) = E(X^2) - [E(X)]^2 = 2 - 1 = 1$. Thus, the standard Exponential density has the same mean and variance.

Example 1.55. (Absolute Value of a Standard Normal). This is often required in calculations in statistical theory. Let X have the standard normal distribution; we want to find $E(|X|)$. By definition,

$$E(|X|) = \int_{-\infty}^\infty |x| f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty |x| e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^\infty x e^{-x^2/2} dx$$

(since $|x|e^{-x^2/2}$ is an even function of x on $(-\infty, \infty)$)

$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty [\frac{d}{dx}(-e^{-x^2/2})] dx = \frac{2}{\sqrt{2\pi}} (-e^{-x^2/2})|_0^\infty$$

$$= \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}.$$

Example 1.56. (A Random Variable whose Expectation Does Not Exist). Consider the standard Cauchy random variable with the density $f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$. Recall that for $E(X)$ to exist, we must have $\int_{-\infty}^\infty |x| f(x) dx < \infty$. But,

$$\int_{-\infty}^\infty |x| f(x) dx = \frac{1}{\pi} \int_{-\infty}^\infty \frac{|x|}{1+x^2} dx \geq \frac{1}{\pi} \int_0^\infty \frac{x}{1+x^2} dx$$

$$\geq \frac{1}{\pi} \int_0^M \frac{x}{1+x^2} dx$$

(for any $M < \infty$)

$$= \frac{1}{2\pi} \log(1+M^2),$$

and on letting $M \to \infty$, we see that

$$\int_{-\infty}^\infty |x| f(x) dx = \infty.$$

Therefore the expectation of a standard Cauchy random variable, or synonymously, the expectation of a standard Cauchy distribution does not exist.

Example 1.57. (Moments of the Standard Normal). In contrast to the standard Cauchy, every moment of a standard normal variable exists. The basic reason is that the tail of the standard normal density is too thin. A formal proof is as follows:
Fix $k \geq 1$. Then,

$$|x|^k e^{-x^2/2} = |x|^k e^{-x^2/4} e^{-x^2/4} \leq C e^{-x^2/4},$$

where C is a finite constant such that $|x|^k e^{-x^2/4} \leq C$ for any real number x (such a constant C does exist). Therefore,

$$\int_{-\infty}^{\infty} |x|^k e^{-x^2/2} dx \leq C \int_{-\infty}^{\infty} e^{-x^2/4} dx < \infty.$$

Hence, by definition, for any $k \geq 1, E(X^k)$ exists.

Now, take k to be an odd integer, say $k = 2n+1, n \geq 0$. Then,

$$E(X^k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2n+1} e^{-x^2/2} dx = 0,$$

because x^{2n+1} is an *odd function* and $e^{-x^2/2}$ is an *even function.* Thus, every odd moment of the standard normal distribution is zero.

Next, take k to be an even integer, say $k = 2n, n \geq 1$. Then,

$$E(X^k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2n} e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^{2n} e^{-x^2/2} dx$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z^n e^{-z/2} \frac{1}{2\sqrt{z}} dz = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z^{n-1/2} e^{-z/2} dz,$$

on making the substitution $z = x^2$.

Now make a further substitution $u = \frac{z}{2}$. Then, we get,

$$E(X^{2n}) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} (2u)^{n-1/2} e^{-u} 2du = \frac{2^n}{\sqrt{\pi}} \int_0^{\infty} u^{n-1/2} e^{-u} du.$$

Now, we recognize $\int_0^{\infty} u^{n-1/2} e^{-u} du$ to be $\Gamma(n + \frac{1}{2})$, and so, we get the formula

$$E(X^{2n}) = \frac{2^n \Gamma(n + \frac{1}{2})}{\sqrt{\pi}}, n \geq 1.$$

By using the *Gamma duplication formula*

$$\Gamma(n + \frac{1}{2}) = \sqrt{\pi} 2^{1-2n} \frac{(2n-1)!}{(n-1)!},$$

this reduces to

$$E(X^{2n}) = \frac{(2n)!}{2^n n!}, n \geq 1.$$

1.9.2 Moments and the Tail of a CDF

We now describe methods to calculate moments of a random variable from its survival function, namely $\bar{F}(x) = 1 - F(x)$.

Theorem 1.29. (a) Let X be a nonnegative random variable and suppose $E(X)$ exists. Then $E(X) = \int_0^\infty \overline{F}(x)dx$.

(b) Let X be a nonnegative random variable and suppose $E(X^k)$ exists. Then

$$E(X^k) = \int_0^\infty (kx^{k-1})[1 - F(x)]dx.$$

(c) Let X be a general real valued random variable and suppose $E(X)$ exists. Then

$$E(X) = \int_0^\infty [1 - F(x)]dx - \int_{-\infty}^0 F(x)dx.$$

Example 1.58. (Expected Value of the Minimum of Several Uniform Variables). Suppose $X_1, X_2, \cdots, X_n$ are independent $U[0,1]$ random variables, and let $m_n = \min\{X_1, X_2, \cdots, X_n\}$ be their minimum. By virtue of the independence of $X_1, X_2, \cdots, X_n$,

$$P(m_n > x) = P(X_1 > x, X_2 > x, \cdots, X_n > x) = \prod_{i=1}^n P(X_i > x) = (1-x)^n, 0 < x < 1,$$

and, $P(m_n > x) = 0$ if $x \geq 1$. Therefore, by the above theorem,

$$E(m_n) = \int_0^\infty P(m_n > x)dx = \int_0^1 P(m_n > x)dx = \int_0^1 (1-x)^n dx$$

$$= \int_0^1 x^n dx = \frac{1}{n+1}.$$

1.10 Moment Generating Functions in the Continuous Case

The definition previously given of the moment generating function(mgf) of a random variable is completely general. We work out a few examples for some continuous random variables.

Example 1.59. (Moment Generating Function of Standard Exponential). Let X have the standard Exponential density. Then,

$$E(e^{tX}) = \int_0^\infty e^{tx}e^{-x}dx = \int_0^\infty e^{-(1-t)x}dx = \frac{1}{1-t},$$

if $t < 1$, and it equals $+\infty$ if $t \geq 1$. Thus, the mgf of the standard Exponential distribution is finite if and only if $t < 1$. So, the moments can be found by differentiating the mgf, namely, $E(X^n) = \psi^{(n)}(0)$. Now, at any $t < 1$, by direct differentiation, $\psi^{(n)}(t) = \frac{n!}{(1-t)^{n+1}} \Rightarrow E(X^n) = \psi^{(n)}(0) = n!$, a result we have derived before directly.

Example 1.60. (Moment Generating Function of Standard Normal). Let X have the standard normal density. Then,

$$E(e^{tX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \times e^{t^2/2}$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \times e^{t^2/2} = 1 \times e^{t^2/2} = e^{t^2/2},$$

because $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz$ is the integral of the standard normal density, and so must be equal to one.

We have therefore proved that the mgf of the standard normal distribution exists at any real t and equals $\psi(t) = e^{t^2/2}$.

The mgf is useful in deriving inequalities on probabilities of tail values of a random variable that have proved to be extremely useful in many problems in statistics and probability. In particular, these inequalities typically give much sharper bounds on the probability that a random variable would be far from its mean value than Chebyshev's inequality can give. Such probabilities are called *large deviation probabilities*. We present a particular large deviation inequality below.

Theorem 1.30. (Chernoff-Bernstein Inequality) Let X have the mgf $\psi(t)$, and assume that $\psi(t) < \infty$ for $t < t_0$ for some $t_0, 0 < t_0 \leq \infty$. Let $\kappa(t) = \log \psi(t)$, and for a real number x, define

$$I(x) = \sup_{0<t<t_0} [tx - \kappa(t)].$$

Then,

$$P(X \geq x) \leq e^{-I(x)}.$$

The function $I(x)$ is called the *rate function* corresponding to the distribution of X. See the chapter supplementary section for an example.

There are numerous other moment inequalities on positive and general real valued random variables. They have a variety of uses in theoretical calculations. We present a few fundamental moment inequalities that are special.

Theorem 1.31. (Jensen's Inequality). Let X be a random variable with a finite mean, and $g(x) : \mathcal{R} \to \mathcal{R}$ a convex function. Then $g(E(X)) \leq E(g(X))$.

Proof: Denote the mean of X by μ, and suppose that g has a finite derivative $g'(\mu)$ at μ. Now consider any $x > \mu$. By the convexity of g, $\frac{g(x)-g(\mu)}{x-\mu} \geq g'(\mu) \Rightarrow g(x) - g(\mu) \geq (x-\mu)g'(\mu)$. For $x < \mu$, $\frac{g(x)-g(\mu)}{x-\mu} \leq g'(\mu) \Rightarrow g(x) - g(\mu) \geq (x-\mu)g'(\mu)$. For $x = \mu, g(x) - g(\mu) = (x-\mu)g'(\mu)$. Since we have $g(x) - g(\mu) \geq (x-\mu)g'(\mu)\ \forall x$, by taking an expectation,

$$E[g(X) - g(\mu)] \geq E[(X-\mu)g'(\mu)] = 0$$

$$\Rightarrow g(\mu) \le E(g(X)).$$

When g does not have a finite derivative at μ, the proof uses the geometric property of a convex function that the chord line joining two points is always above the graph of the convex function between the two points. We will leave that case as an exercise.

Example 1.61. Let X be any random variable with a finite mean μ. Consider the function $g(x) = e^{ax}$, where a is a real number. Then, by the second derivative test, g is a convex function on the entire real line, and therefore, by Jensen's inequality

$$E(e^{aX}) \ge e^{a\mu}.$$

1.11 Some Special Continuous Distributions

A number of densities, by virtue of their popularity in modelling, or because of their special theoretical properties, are considered to be special. We discuss, when suitable, their moments, the form of the CDF, the mgf, shape and modal properties, and interesting inequalities. Classic references to standard continuous distributions are Johnson, Kotz, and Balakrishnan (1995), and Kendall and Stuart (2009); Everitt (1998) contains many unusual facts.

Definition 1.31. Let X have the pdf

$$f(x) = \frac{1}{b-a}, a \le x \le b,$$
$$= 0 \text{ otherwise},$$

where $-\infty < a < b < \infty$ are given real numbers.

Then we say that X has the uniform distribution on $[a, b]$ and write $X \sim U[a, b]$.

The basic properties of a uniform density are given next.

Theorem 1.32. (a) If $X \sim U[0, 1]$, then $a+(b-a)X \sim U[a, b]$, and if $X \sim U[a, b]$, then $\frac{X-a}{b-a} \sim U[0, 1]$.

(b) The CDF of the $U[a, b]$ distribution equals:

$$F(x) = 0, x < a;$$
$$= \frac{x-a}{b-a}, a \le x \le b;$$
$$= 1, x > b.$$

(c) The mgf of the $U[a, b]$ distribution equals $\psi(t) = \frac{e^{tb}-e^{ta}}{(b-a)t}$.

(d) The nth moment of the $U[a, b]$ distribution equals

$$E(X^n) = \frac{b^{n+1} - a^{n+1}}{(b-a)(n+1)}.$$

(e) The mean and the variance of the $U[a,b]$ distribution equal

$$\mu = \frac{a+b}{2}; \sigma^2 = \frac{(b-a)^2}{12}.$$

Example 1.62. A point is selected at random on the unit interval, dividing it into two pieces with total length 1. Find the probability that the ratio of the length of the shorter piece to the length of the longer piece is less than 1/4.

Let $X \sim U[0,1]$; we want $P(\frac{\min\{X,1-X\}}{\max\{X,1-X\}} < 1/4)$. This happens only if $X < 1/5$ or $> 4/5$. Therefore, the required probability is $P(X < 1/5)+P(X > 4/5) = 1/5+1/5 = 2/5$. We defined the standard exponential density in the previous section. We now introduce the general exponential density. Exponential densities are used to model waiting times, e.g., waiting times for an elevator or at a supermarket check out, or failure times, e.g., the time till the first failure of some equipment, or renewal times, e.g., time elapsed between successive earthquakes at a location, etc. The exponential density also has some very interesting theoretical properties.

Definition 1.32. A nonnegative random variable X has the Exponential distribution with parameter $\lambda > 0$ if it has the pdf $f(x) = \frac{1}{\lambda}e^{-x/\lambda}, x > 0$. We write $X \sim Exp(\lambda)$.

Here are the basic properties of an exponential density.

Theorem 1.33. Let $X \sim Exp(\lambda)$. Then,

$$(a)\ \frac{X}{\lambda} \sim Exp(1);$$

$$(b)\ \text{The CDF } F(x) = 1 - e^{-x/\lambda}, x > 0,$$

(and zero for $x \le 0$.)

$$(c)\ E(X^n) = \lambda^n n!, n \ge 1;$$

$$(d)\ \text{The mgf } \psi(t) = \frac{1}{1-\lambda t}, t < 1/\lambda.$$

Example 1.63. (Mean is Larger than Median for Exponential). Suppose $X \sim Exp(4)$. What is the probability that $X > 4$?

Since $X/4 \sim Exp(1)$,

$$P(X > 4) = P(X/4 > 1) = \int_1^\infty e^{-x}dx = e^{-1} = .3679,$$

quite a bit smaller than 50%. This implies that the median of the distribution has to be smaller than 4, where 4 is the mean. Indeed, the median is a number m such that $F(m) = \frac{1}{2}$ (the median is unique in this example) $\Rightarrow 1 - e^{-m/4} = \frac{1}{2} \Rightarrow m = 4\log 2 = 2.77$.

This phenomenon that the mean is larger than the median is quite typical of distributions which have a long right tail, as does the Exponential.

In general, if $X \sim Exp(\lambda)$, the median of X is $\lambda \log 2$.

Example 1.64. (Lack of Memory of the Exponential Distribution). The exponential densities have a lack of memory property similar to the one we established for the geometric distribution. Let $X \sim Exp(\lambda)$, and let s, t be positive numbers. The lack of memory property is that $P(X > s + t|X > s) = P(X > t)$. So, suppose that X is the waiting time for an elevator, and suppose that you have already waited $s = 3$ minutes. Then the probability that you have to wait another two minutes is the same as the probability that you would have to wait two minutes, if you just came. This is not true if the waiting time distribution is something other than an Exponential.

The proof of the property is simple;

$$P(X > s + t|X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-(s+t)/\lambda}}{e^{-s/\lambda}}$$
$$= e^{-t/\lambda} = P(X > t).$$

The Exponential density is decreasing on $[0, \infty)$. A generalization of the Exponential density with a mode usually at some strictly positive number m is the Gamma distribution. It includes the Exponential as a special case, and can be very skewed, or even almost a bell shaped density. We will later see that it also arises naturally, as the density of the sum of a number of independent Exponential random variables.

Example 1.65. (The Weibull Distribution). Suppose $X \sim Exp(1)$, and let $Y = X^\alpha$, where $\alpha > 0$ is a constant. Since this is a strictly monotone function with the inverse function $y^{1/\alpha}$, the density of Y is

$$f_Y(y) = \frac{f(y^{1/\alpha})}{|g'(y^{1/\alpha})|} = e^{-y^{1/\alpha}} \times \frac{1}{\alpha y^{(\alpha-1)/\alpha}}$$
$$= \frac{1}{\alpha} y^{(1-\alpha)/\alpha} e^{-y^{1/\alpha}}, y > 0.$$

This final answer can be made to look a little simpler by writing $\beta = \frac{1}{\alpha}$. If we do so, the density becomes

$$\beta y^{\beta-1} e^{-y^\beta}, y > 0.$$

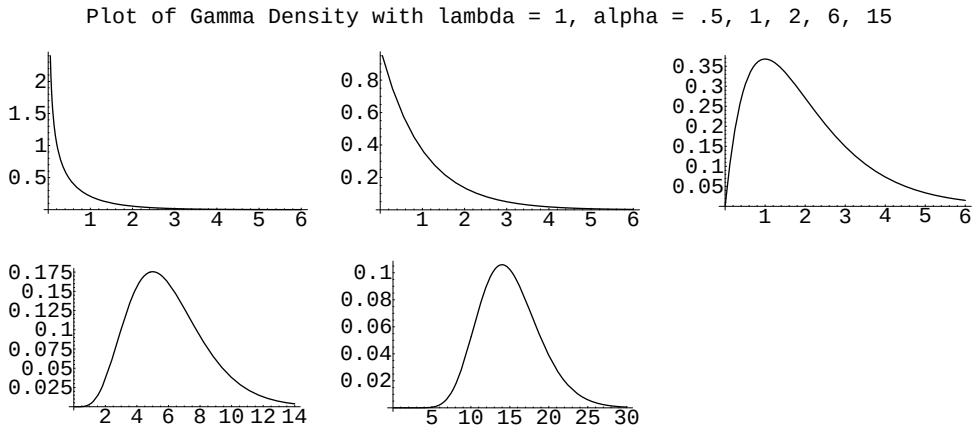
We can introduce an extra scale parameter akin to we do for the exponential case itself. In that case, we have the general two parameter Weibull density

$$f(y|\beta, \lambda) = \frac{\beta}{\lambda}(\frac{x}{\lambda})^{\beta-1} e^{-(\frac{x}{\lambda})^\beta}, y > 0.$$

This is the *Weibull density* with parameters β, λ.

Definition 1.33. A positive random variable X is said to have a Gamma distribution with shape parameter α and scale parameter λ if it has the pdf

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



we write $X \sim G(\alpha, \lambda)$. The Gamma density reduces to the Exponential density with mean λ when $\alpha = 1$; for $\alpha < 1$, the Gamma density is decreasing and unbounded, while for large α, it becomes nearly a bell shaped curve. A plot of some Gamma densities reveals these features.

The basic facts about a Gamma distribution are given in the following theorem.

Theorem 1.34. (a) The CDF of the $G(\alpha, \lambda)$ density is the *normalized incomplete Gamma function*

$$F(x) = \frac{\gamma(\alpha, x/\lambda)}{\Gamma(\alpha)},$$

where $\gamma(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1} dt$.

(b) The nth moment equals

$$E(X^n) = \lambda^n \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}, n \geq 1.$$

(c) The mgf equals

$$\psi(t) = (1 - \lambda t)^{-\alpha}, t < \frac{1}{\lambda}.$$

(d) The mean and the variance equal

$$\mu = \alpha\lambda; \sigma^2 = \alpha\lambda^2.$$

An important consequence of the mgf formula is the following result.

Corollary Suppose $X_1, X_2, \cdots, X_n$ are independent $Exp(\lambda)$ variables. Then $X_1 + X_2 + \cdots + X_n \sim G(n, \lambda)$.

Example 1.66. (The General Chi Square Distribution). We saw in the previous section that the distribution of the square of a standard normal variable is the chi square distribution with one degree of freedom. A natural question is what is the distribution of the sum of squares of several independent standard normal variables. Although we do

not yet have the technical tools necessary to derive this distribution, it turns out that this distribution is in fact a Gamma distribution. Precisely, if $X_1, X_2, \cdots, X_m$ are m independent standard normal variables, then $T = \sum_{i=1}^{m} X_i^2$ has a $G(\frac{m}{2}, 2)$ distribution, and therefore has the density

$$f_m(t) = \frac{e^{-t/2} t^{m/2-1}}{2^{m/2}\Gamma(\frac{m}{2})}, t > 0.$$

This is called the *chi square density with m degrees of freedom*, and arises in numerous contexts in statistics and probability. We write $T \sim \chi^2_m$. From the general formulas for the mean and variance of a Gamma distribution, we get that

$$\text{Mean of a } \chi^2_m \textit{ distribution } = m;$$

$$\text{Variance of a } \chi^2_m \textit{ distribution } = 2m.$$

The chi square density is rather skewed for small m, but becomes approximately bell shaped when m gets large; we have seen this for general Gamma densities.

One especially important context in which the chi square distribution arises is in consideration of the distribution of the *sample variance* for iid normal observations. The sample variance of a set of n random variables $X_1, X_2, \cdots, X_n$ is defined as $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X} = \frac{X_1+\cdots+X_n}{n}$ is the mean of $X_1, \cdots, X_n$. The name sample variance derives from the following property.

Theorem 1.35. Suppose $X_1, \cdots, X_n$ are independent with a common distribution F having a finite variance σ^2. Then, for any n, $E(s^2) = \sigma^2$.

Proof: First note the algebraic identity

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

Therefore,

$$E(s^2) = \frac{1}{n-1}E[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] = \frac{1}{n-1}[n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)] = \sigma^2.$$

If, in particular, $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, then $\frac{X_i - \bar{X}}{\sigma}$ are also normally distributed, each with mean zero. However, they are no longer independent. If we sum their squares, then the sum of the squares will still be distributed as a chi square, but there will be a loss of one degree of freedom, due to the fact that $X_i - \bar{X}$ are not independent, even though the X_i are independent.

We state this important fact formally.

Theorem 1.36. Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$. Then $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.

For continuous random variables that take values between 0 and 1, the most standard family of densities is the family of Beta densities. Their popularity is due to their analytic tractability, and due to the large variety of shapes that Beta densities can take when the parameter values change. It is a generalization of the $U[0,1]$ density. Beta densities, however, cannot have more than one mode in the open interval $(0,1)$.

Definition 1.34. X is said to have a Beta density with parameters α and β if it has the density

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1, \alpha, \beta > 0,$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. We write $X \sim Be(\alpha,\beta)$. An important point is that by its very notation, $\frac{1}{B(\alpha,\beta)}$ must be the normalizing constant of the function $x^{\alpha-1}(1-x)^{\beta-1}$; thus, another way to think of $B(\alpha,\beta)$ is that for any $\alpha, \beta > 0$,

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

This fact will be repeatedly useful in the following.

Theorem 1.37. Let $X \sim Be(\alpha,\beta)$.

(a) The CDF equals

$$F(x) = \frac{B_x(\alpha,\beta)}{B(\alpha,\beta)},$$

where $B_x(\alpha,\beta)$ is the *incomplete Beta function* $\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$.

(b) The nth moment equals

$$E(X^n) = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}.$$

(c) The mean and the variance equal

$$\mu = \frac{\alpha}{\alpha+\beta}; \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Example 1.67. (Square of a Beta). Suppose X has a Beta density. Then, X^2 also takes values in $[0,1]$, but it does not have a Beta density. To have a specific example, suppose $X \sim Be(7,7)$. Then the density of $Y = X^2$ is

$$f_Y(y) = \frac{f(\sqrt{y})}{2\sqrt{y}} = \frac{y^3(1-\sqrt{y})^6}{B(7,7)2\sqrt{y}}$$

$$= 6006y^{5/2}(1-\sqrt{y})^6, 0 \le y \le 1.$$

Clearly, this is not a Beta density.

In practical applications, certain types of random variables consistently exhibit a long right tail, in the sense that a lot of small values are mixed with a few large or excessively large values in the distributions of these random variables. Economic variables such as wealth typically manifest such heavy tail phenomena. Other examples include sizes of oil fields, insurance claims, stock market returns, river height in a flood, etc. The tails are sometimes so heavy that the random variable may not even have a finite mean. *Extreme value distributions* are common and increasingly useful models for such applications. A brief introduction to two specific extreme value distributions is provided next. These two distributions are the *Pareto* distribution and the *Gumbel* distribution. One peculiarity of semantics is that the Gumbel distribution is often called the *Gumbel law*.
A random variable X is said to have the Pareto density with parameters θ and α if it has the density

$$f(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, x \geq \theta > 0, \alpha > 0.$$

We write $X \sim Pa(\alpha, \theta)$. The density is monotone decreasing. It may or may not have a finite expectation, depending on the value of α. It never has a finite mgf in any nonempty interval containing zero. The basic facts about a Pareto density are given in the next result.

Theorem 1.38. Let $X \sim Pa(\alpha, \theta)$.
(a) The CDF of X equals

$$F(x) = 1 - (\frac{\theta}{x})^\alpha, x \geq \theta,$$

and zero for $x < \theta$.
(b) The nth moment exists if and only if $n < \alpha$, in which case

$$E(X^n) = \frac{\alpha\theta^n}{\alpha - n}.$$

(c) For $\alpha > 1$, the mean exists; for $\alpha > 2$, the variance exists. Furthermore, they equal

$$E(X) = \frac{\alpha\theta}{\alpha - 1}; \text{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}.$$

We next define the Gumbel law. A random variable X is said to have the Gumbel density with parameters μ, σ if it has the density

$$f(x) = \frac{1}{\sigma} e^{(-e^{-\frac{x-\mu}{\sigma}})} e^{-\frac{x-\mu}{\sigma}}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$$

If $\mu = 0$ and $\sigma = 1$, the density is called the *standard Gumbel density*. Thus, the standard Gumbel density has the formula $f(x) = e^{-e^{-x}} e^{-x}, -\infty < x < \infty$. The density converges

extremely fast (at a superexponential rate) at the left tail, but only at a regular exponential rate at the right tail. Its relation to the density of the maximum of a large number of independent normal variables makes it a special density in statistics and probability.
And, now the general normal density is formally introduced with its properties.

Definition 1.35. A random variable X is said to have a normal distribution with parameters μ and σ^2 if it has the density

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

where μ can be any real number, and $\sigma > 0$. We write $X \sim N(\mu, \sigma^2)$. If $X \sim N(0,1)$, we call it a *standard normal variable*.

The density of a standard normal variable is denoted as $\phi(x)$, and equals the function

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

and the CDF is denoted as $\Phi(x)$. Note that the standard normal density is symmetric and unimodal about zero. The general $N(\mu, \sigma^2)$ density is symmetric and unimodal about μ.

By definition of a CDF,

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)dz.$$

The CDF $\Phi(x)$ cannot be written in terms of the elementary functions, but can be computed at a given value x, and tables of the values of $\Phi(x)$ are widely available. For example, here are some selected values.

Example 1.68. (Standard Normal CDF at Selected Values).

x	$\Phi(x)$
-4	.00003
-3	.00135
-2	.02275
-1	.15866
0	.5
1	.84134
2	.97725
3	.99865
4	.99997

Here are the most basic properties of a normal distribution.

Theorem 1.39. (a) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, and if $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

In words, if X is any normal random variable, then its standardized version is always a standard normal variable.

(b) If $X \sim N(\mu, \sigma^2)$, then

$$P(X \le x) = \Phi(\frac{x-\mu}{\sigma}) \; \forall \, x.$$

In particular, $P(X \le \mu) = P(Z \le 0) = .5$, i.e., the median of X is μ.

(c) Every moment of any normal distribution exists, and the odd central moments $E[(X-\mu)^{2k+1}]$ are all zero.

(d) If $Z \sim N(0,1)$, then

$$E(Z^{2k}) = \frac{(2k)!}{2^k k!}, k \ge 1.$$

(e) The mgf of the $N(\mu, \sigma^2)$ distribution exists at all real t, and equals

$$\psi(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}.$$

(f) If $X \sim N(\mu, \sigma^2)$,

$$E(X) = \mu; \text{var}(X) = \sigma^2; E(X^3) = \mu^3 + 3\mu\sigma^2; E(X^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$$

An important consequence of part (b) of this theorem is the following result.

Corollary Let $X \sim N(\mu, \sigma^2)$, and let $0 < \alpha < 1$. Let $Z \sim N(0,1)$. Suppose x_α is the $(1-\alpha)$th quantile (also called percentile) of X, and z_α is the $(1-\alpha)$th quantile of Z. Then

$$x_\alpha = \mu + \sigma z_\alpha.$$

Example 1.69. (Using a Standard Normal Table). Suppose $X \sim N(5, 16)$; we want to know which number x has the property that $P(X \le x) = .95$.

This amounts to asking what is the 95th percentile of X. By the general formula for percentiles of a normal distribution, the 95th percentile of X equals

$$\sigma \times \text{ 95th percentile of standard normal } + \mu = 4 \times 1.645 + 5 = 11.58.$$

Now change the question: which number x has the property that $P(x \le X \le 9) = .68$? This means, by standardizing X to a standard normal,

$$\Phi(1) - \Phi(\frac{x-5}{4}) = .68 \Rightarrow \Phi(\frac{x-5}{4}) = \Phi(1) - .68$$

$$= .8413 - .68 = .1613.$$

By reading a standard normal table, $\Phi(-.99) = .1613$, and so,

$$\frac{x-5}{4} = -.99 \Rightarrow x = 1.04.$$

Example 1.70. (A Reliability Problem). Let X denote the length of time (in minutes) an automobile battery will continue to crank an engine. Assume that $X \sim N(10, 4)$. What is the probability that the battery will crank the engine longer than $10 + x$ minutes given that it is still cranking at 10 minutes?

We want to find

$$P(X > 10 + x \,|X > 10) = \frac{P(X > 10 + x)}{P(X > 10)} = \frac{P(Z > x/2)}{1/2}$$

$$= 2[1 - \Phi(\frac{x}{2})];$$

note that this is decreasing in x. If X had been exponentially distributed, then by the lack of memory property, this probability would have been $P(X > x) = e^{-x/10}$, assuming that the mean was still 10 minutes. But if the distribution is normal, we can no longer get an analytic expression for the probability; we only get an expression involving the standard normal CDF.

As a specific choice, if $x = 2$, then we get

$$P(X > 10 + x \,|X > 10) = 2[1 - \Phi(\frac{x}{2})] = 2[1 - \Phi(1)] = .3174.$$

Example 1.71. (lognormal Distribution). lognormal distributions are common models in studies of economic variables, such as income and wealth, because they can adequately describe the skewness that one sees in data on such variables. If $X \sim N(\mu, \sigma^2)$, then the distribution of $Y = e^X$ is called a *lognormal distribution with parameters* μ, σ^2. Note that the lognormal name can be confusing; a lognormal variable is *not* the logarithm of a normal variable. A better way to remember its meaning is *log is normal*.

Since $Y = e^X$ is a strictly monotone function of X, by the usual formula for the density of a monotone function, Y has the pdf

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}, y > 0;$$

this is called the lognormal density with parameters μ, σ^2. Since a lognormal variable is defined as e^X for a normal variable X, its mean and variance are easily found from the mgf of a normal variable. A simple calculation shows that

$$E(Y) = e^{\mu + \frac{\sigma^2}{2}}; \mathrm{Var}(Y) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}.$$

One of the main reasons for the popularity of the lognormal distribution is its skewness; the lognormal density is extremely skewed for large values of σ. The coefficient of skewness has the formula

$$\beta = (2 + e^{\sigma^2})\sqrt{e^{\sigma^2} - 1} \to \infty, \text{ as } \sigma \to \infty.$$

Note that the lognormal densities do not have a finite mgf at any $t > 0$, although all its moments are finite. It is also the only standard continuous distribution that is *not determined by its moments*. That is, there exist *other distributions besides the lognormal* all of whose moments exactly coincide with the moments of a given lognormal distribution. This is not true of any other distribution with a name that we have come across so far. For example, the normal and the Poisson distributions are all determined by their moments. A phenomenal fact in statistics and probability is that sums of many independent variables tend to be approximately normally distributed. A precise version of this is the *central limit theorem*, which we will study in the next section. What is interesting is that sums of *any number* of independent normal variables are *exactly* normally distributed. Here is the result.

Theorem 1.40. Let $X_1, X_2, \cdots, X_n, n \geq 2$ be independent random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $S_n = \sum_{i=1}^n X_i$. Then,

$$S_n \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2).$$

Proof: The quickest proof of this uses the mgf technique. Since the X_i are independent, the mgf of S_n is

$$\psi_{S_n}(t) = E(e^{tS_n}) = E(e^{tX_1} \cdots e^{tX_n})$$

$$= \prod_{i=1}^n E(e^{tX_i}) = \prod_{i=1}^n e^{t\mu_i + t^2\sigma_i^2/2} = e^{t(\sum_{i=1}^n \mu_i) + (t^2/2)\ (\sum_{i=1}^n \sigma_i^2)},$$

which agrees with the mgf of the $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ distribution, and therefore by the distribution determining property of mgfs, it follows that $S_n \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

An important consequence is the following result.

Corollary Suppose $X_i, 1 \leq i \leq n$ are independent, and each distributed as $N(\mu, \sigma^2)$. Then $\overline{X} = \frac{S_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$.

The theorem above implies that any linear function of independent normal variables is also normal, i.e.,

$$\sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2).$$

Example 1.72. Suppose $X \sim N(-1, 4), Y \sim N(1, 5)$, and suppose that X, Y are independent. We want to find the CDF of $X + Y$, and of $X - Y$.

By the above theorem,

$$X + Y \sim N(0, 9), \text{and } X - Y \sim N(-2, 9).$$

Therefore,

$$P(X+Y \le x) = \Phi(\frac{x}{3}), \text{and } \; P(X-Y \le x) = \Phi(\frac{x+2}{3}).$$

For example,

$$P(X+Y \le 3) = \Phi(1) = .8413, \text{and } \; P(X-Y \le 3) = \Phi(\frac{5}{3}) = .9525.$$

The standard normal CDF cannot be represented in terms of the elementary functions. But it arises in many mathematical problems to do with normal distributions. As such, it is important to know the behavior of the standard normal CDF $\Phi(x)$, especially for large x. The ratio

$$R(x) = \frac{1-\Phi(x)}{\phi(x)}$$

is called *Mills ratio*. Among many bounds and approximations for $R(x)$, we mention the following.

Theorem 1.41. (Mills ratio). (a) For $x > 0$,

$$\frac{x}{x^2+1} \le R(x) \le \frac{1}{x}.$$

(b) For $x > 0$,

$$\frac{1}{x} - \frac{1}{x^3} < R(x) < \frac{1}{x}.$$

See DasGupta (2008) for these and some additional bounds on $R(x)$.

1.12 Stein's Lemma

In 1981, Charles Stein gave a simple lemma for a normal distribution, and extended it to the case of a finite number of independent normal variables, which seems innocuous on its face, but has proved to be a really powerful tool in numerous areas of statistics. It has also had its technical influence on the area of Poisson approximations, which we briefly discussed in this chapter. We present the basic lemma, its extension to the case of several independent variables, and show some applications. It would not be possible to give more than just a small glimpse of the applications of Stein's lemma here; the applications are too varied. Regrettably, no comprehensive book or review of the various applications of Stein's lemma is available at this time. The original article is Stein (1981).

Theorem 1.42. (a) Let $X \sim N(\mu, \sigma^2)$, and suppose $g : \mathcal{R} \to \mathcal{R}$ is such that g is differentiable at all but at most a finite number of points, and

$$(i) \text{ For some } \lambda < 1, g(x)e^{-\frac{\lambda x^2}{2\sigma^2}} \to 0, \text{ as } x \to \pm\infty;$$

$$(ii) \; E[|g'(X)|] < \infty.$$

Then,

$$E[(X-\mu)g(X)] = \sigma^2 E[g'(X)].$$

(b) Let $X_1, X_2, \cdots, X_k$ be independent $N(\mu_i, \sigma^2)$ variables, and suppose $g : \mathcal{R}^k \to \mathcal{R}$ is such that g has a partial derivative with respect to each x_i at all but at most a finite number of points. Then,

$$E[(X_i-\mu_i)g(X_1, X_2, \cdots, X_k)] = \sigma^2 E[\frac{\partial}{\partial X_i} g(X_1, X_2, \cdots, X_k)].$$

Proof: We will prove part (a). By definition of expectation, and by using integration by parts,

$$E[(X-\mu)g(X)] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)g(x)e^{-(x-\mu)^2/(2\sigma^2)}dx = -\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x)[\frac{d}{dx}e^{-(x-\mu)^2/(2\sigma^2)}]dx$$

$$= -\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} g(x)e^{-(x-\mu)^2/(2\sigma^2)}|_{-\infty}^{\infty} + \sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x)e^{-(x-\mu)^2/(2\sigma^2)}dx$$

$$= -\sigma \frac{1}{\sqrt{2\pi}} g(x)e^{-\frac{\lambda x^2}{2\sigma^2}} e^{-\frac{(1-\lambda)x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}|_{-\infty}^{\infty} + \sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x)e^{-(x-\mu)^2/(2\sigma^2)}dx$$

$$= 0 + \sigma^2 E[g'(X)] = \sigma^2 E[g'(X)],$$

because by assumption

$$g(x)e^{-\frac{\lambda x^2}{2\sigma^2}} \to 0, \text{as } x \to \pm\infty,$$

and,

$$e^{-\frac{(1-\lambda)x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}} \text{ is uniformly bounded in } x.$$

The principal applications of Stein's lemma are in statistical theory; we will make plenty of uses of it in Chapters 7 and 14. Here is a simple application.

Example 1.73. (Application of Stein's Lemma). Let $X \sim N(\mu, 1)$ and k a positive integer. Take any function $g(x)$ such that g is differentiable and is bounded by an exponential function, $|g(x)| \le e^{a|x|}$ for some finite a. These conditions are enough for applying Stein's lemma, and we then get

$$E[X^{k+1}g(X)] = E[X(X^k g(X))] = E[(X-\mu+\mu)(X^k g(X))]$$

$$= E[(X-\mu)X^k g(X)] + +\mu E[X^k g(X)]$$

$$= E[kX^{k-1}g(X) + X^k g'(X)] + \mu E[X^k g(X)] = kE[X^{k-1}g(X)] + \mu E[X^k g(X)] + E[X^k g'(X)].$$

In particular, if we let $g(X) = 1$, then we get the recursion

$$E[X^{k+1}] = kE[X^{k-1}] + \mu E[X^k],$$

which allows one to recursively find the moments of X by starting from the first moment $E(X) = \mu$. This was pointed out by Charles Stein.

A famous *characterization of normal distributions* is that the mean and the variance of an iid sample taken from any normal distribution are independently distributed. Here is this famous result.

Theorem 1.43. (Independence of Sample Mean and Sample Variance) Suppose $X_1, X_2, \cdots, X_n$ are independent $N(\mu, \sigma^2)$ variables. Then, for any $n \geq 2, \bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent. Conversely, for given $n \geq 2$, if $X_1, X_2, \cdots, X_n$ are independent variables with some common distribution, and if $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent, then the common distribution of the X_i must be $N(\mu, \sigma^2)$ for some μ, σ^2.

1.13 Normal Approximations and Central Limit Theorem

Many of the special discrete and special continuous distributions that we have discussed can be well approximated by a normal distribution, for suitable configurations of their underlying parameters. Typically, the normal approximation works well when the parameter values are such that the skewness of the distribution is small. For example, Binomial distributions are well approximated by a normal when n is large and p is not too small or too large. Gamma distributions are well approximated by a normal when the shape parameter α is large. There is a unifying mathematical result here. The unifying mathematical result is one of the most important results in all of mathematics, and is called the *central limit theorem*. Roughly, the central limit theorem (CLT) says that sums of many independent random variables are often approximately normally distributed. The subject of central limit theorems is incredibly diverse. In this section, we present the basic or the *canonical central limit theorem*, and present its applications to certain problems that we are already familiar with. Among numerous excellent references on central limit theorems, we recommend Feller (1968, 1971) for lucid exposition and examples. The subject of central limit theorems also has a really interesting history; we recommend Le Cam (1986) and Stigler (1986) for reading some history of the central limit theorem. Careful and comprehensive mathematical treatment is available in Hall (1992) and Bhattacharya and Rao (1986).

Theorem 1.44. (Central Limit Theorem) For $n \geq 1$, let $X_1, X_2, \cdots, X_n$ be n independent random variables, each having *the same distribution*, and suppose this common distribution, say F, has a finite mean μ, and a finite variance σ^2. Let $S_n = X_1 + X_2 + \cdots + X_n, \bar{X} = \bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$. Then, as $n \to \infty$,

$$(a)\, P(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x) \to \Phi(x)\ \forall\, x \in \mathcal{R};$$

$$(b)\, P(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq x) \to \Phi(x)\ \forall\, x \in \mathcal{R}.$$

In words, for large n,

$$S_n \approx N(n\mu, n\sigma^2);$$

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n}).$$

Example 1.74. (A Motivating Example). Consider a binomial random variable X with parameters n and p; we will fix $p = .1$ and see the effect of increasing n on the pmf of X. Recall that the binomial pmf has the formula $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \cdots, n$. Using this formula, with $n = 10, 20, 50$, and 100, we have computed and plotted the pmf of X in the form of a *histogram*, which is a system of rectangles with the height of the rectangle corresponding to a specific x value equal to (or proportional to) the probability of that x value.

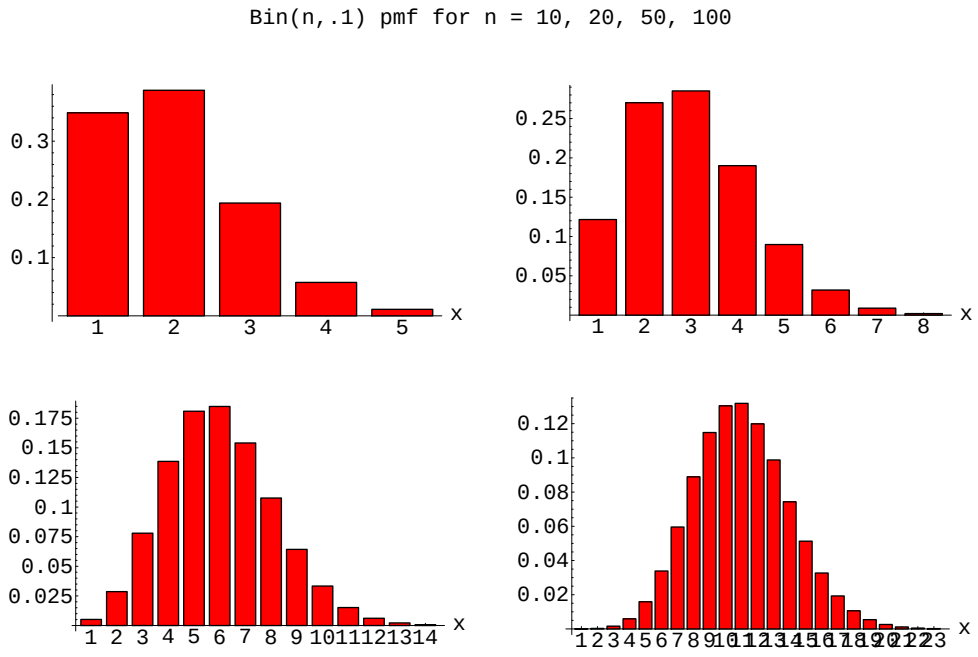
We see that the histogram is rather skewed for the smallest n, namely $n = 10$. As n increases, the histogram gets less skewed, and for the largest value, $n = 100$, the histogram looks like a bell shaped histogram, centered at 10 and 11, resembling a normal density curve.

What is the explanation? The formula for the coefficient of skewness of a binomial distribution is $\frac{1-2p}{\sqrt{np(1-p)}}$, which goes to zero, as $n \to \infty$, for any fixed p. That is, the distribution is becoming nearly symmetric as n gets large, although it started out being very skewed when n was small. Indeed, it is true in general, that the $Bin(n, p)$ distribution can be well approximated by the $N(np, np(1-p))$ distribution for any fixed p, when n is large. If p is near .5, a normal looking histogram will be produced even for n as small as 20; if p is closer to zero or one, a larger n is necessary to produce a normal looking histogram. We will see this empirical illustration borne out by a theorem below.

The proof of this theorem in the generality stated here requires use of certain sophisticated tools in probability theory that we have not discussed. We will prove the theorem under the more restrictive condition that the underlying distribution F has a finite mgf in some open interval containing zero.

We recall the following notation to be used in the proof:
If a sequence of numbers $a_n \to 0$ as $n \to \infty$, we write $a_n = o(1)$. If a_n, b_n are two sequences of numbers, and $\frac{a_n}{b_n} \to 0$ as $n \to \infty$, we write $a_n = o(b_n)$. For example, $a_n = o(n^{-1})$ means that not only does $a_n \to 0$ as $n \to \infty$, even $na_n \to 0$ as $n \to \infty$.
We will also need a fact about mgfs in this proof; we state this fact below.

Lemma (Continuity Theorem on MGFs) Let Y_n be a sequence of random variables with Y_n having a finite mgf $\psi_n(t)$ in some open interval $(-a, a)$ containing zero. If $\psi_n(t) \to e^{t^2/2}$ for each $t \in (-a, a)$ as $n \to \infty$, then $P(Y_n \leq x) \to \Phi(x)$ for all real numbers x, as $n \to \infty$.



Proof of CLT Part (b) of the theorem is in fact equivalent to part (a), and so we prove just part (a). A crucial algebraic simplification that we can make is that we may assume, without loss of generality, that $\mu = 0$ and $\sigma^2 = 1$. This is because if we define a sequence of new random variables $W_i = \frac{X_i - \mu}{\sigma}, i \geq 1$, then the W_i are also iid, and they have mean zero and variance one. Furthermore, $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^n W_i}{\sqrt{n}}, n \geq 1$. Thus, we have that $P(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x) \to \Phi(x)$ if and only if $P(\frac{\sum_{i=1}^n W_i}{\sqrt{n}} \leq x) \to \Phi(x)$.
Therefore, we go ahead and set $\mu = 0, \sigma^2 = 1$, and show that $P(\frac{S_n}{\sqrt{n}} \leq x) \to \Phi(x)$ as $n \to \infty$. We will prove this by appealing to the lemma concerning mgfs stated above.
Since the X_i are independent, the mgf of $Z_n := \frac{S_n}{\sqrt{n}}$ equals

$$\psi_{Z_n}(t) = E[e^{tZ_n}] = E[e^t(\frac{S_n}{\sqrt{n}})]$$

$$= E[e^{\sum_{i=1}^n [t(\frac{X_i}{\sqrt{n}})]}] = \prod_{i=1}^n E[e^{t\frac{X_i}{\sqrt{n}}}]$$

$$= (E[e^{\frac{t}{\sqrt{n}}X_1}])^n = \psi^n(\frac{t}{\sqrt{n}})$$

$$\Rightarrow \log \psi_{Z_n}(t) = n \log \psi(\frac{t}{\sqrt{n}})$$

$$= n \log[1 + \frac{t}{\sqrt{n}}\psi'(0) + \frac{t^2}{2n}\psi''(0) + o(n^{-1})]$$

(by a Taylor expansion of $\psi(\frac{t}{\sqrt{n}})$ around $t = 0$)

$$= n(\frac{t}{\sqrt{n}}\psi'(0) + \frac{t^2}{2n}\psi''(0) - \frac{t^2}{2n}(\psi'(0))^2 + o(n^{-1}))$$

(by expanding $\log(1+x)$ around $x=0$, which gives $\log(1+x) \approx x - \frac{x^2}{2}$ for $x \approx 0$)

$$= \frac{t^2}{2}[\psi''(0) - (\psi'(0))^2] + o(1) = \frac{t^2}{2} + o(1),$$

(since $\psi'(0) = \mu = 0$, and $\psi''(0) - (\psi'(0))^2 = \sigma^2 = 1$)

$$\Rightarrow \psi_{Z_n}(t) \to e^{t^2/2}.$$

This proves, by the Lemma that we stated above, that $P(Z_n \leq x) \to \Phi(x)$ for all x, as was needed.

A very important case in which the general central limit theorem applies is the binomial distribution. The CLT allows us to approximate clumsy binomial probabilities involving large factorials by simple and accurate normal approximations. We first give the exact result on normal approximation of the binomial.

Theorem 1.45. (de Moivre-Laplace Central Limit Theorem) Let $X = X_n \sim Bin(n,p)$. Then, for any fixed p and $x \in \mathcal{R}$,

$$P(\frac{X-np}{\sqrt{np(1-p)}} \leq x) \to \Phi(x),$$

as $n \to \infty$.

The demoivre-Laplace CLT tells us that if $X \sim Bin(n,p)$, then we can approximate the $\leq$ type probability $P(X \leq k)$ as

$$P(X \leq k) = P(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{k-np}{\sqrt{np(1-p)}})$$

$$\approx \Phi(\frac{k-np}{\sqrt{np(1-p)}}).$$

Note that, in applying the normal approximation in the binomial case, we are using a *continuous* distribution to approximate a discrete distribution taking only integer values. The quality of the approximation improves, sometimes dramatically, if we *fill up the gaps between the successive integers*. That is, *pretend* that an event of the form $X = x$ really corresponds to $x - \frac{1}{2} \leq X \leq x + \frac{1}{2}$. In that case, in order to approximate $P(X \leq k)$, we will in fact expand the domain of the eventto $k + \frac{1}{2}$, and approximate $P(X \leq k)$ as

$$P(X \leq k) \approx \Phi(\frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}}).$$

This adjusted normal approximation is called *normal approximation with a continuity correction*. Continuity correction should always be done while computing a normal approximation to a binomial probability. Here are the continuity corrected normal approxomation

formulas for easy reference:

$$P(X \le k) \approx \Phi(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}});$$

$$P(m \le X \le k) \approx \Phi(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}) - \Phi(\frac{m - \frac{1}{2} - np}{\sqrt{np(1-p)}}).$$

Example 1.75. (Coin Tossing). This is the simplest example of a normal approximation of binomial probabilities. We will solve a number of problems by applying the normal approximation method.

First, suppose a fair coin is tossed 100 times. What is the probability that we obtain between 45 and 55 heads? Denoting X as the number of heads obtained in 100 tosses, $X \sim Bin(n, p)$, with $n = 100, p = .5$. Therefore, by using the continuity corrected normal approximation,

$$P(45 \le X \le 55) \approx \Phi(\frac{55.5 - 50}{\sqrt{12.5}}) - \Phi(\frac{44.5 - -50}{\sqrt{12.5}})$$

$$= \Phi(1.56) - \Phi(-1.56) = .9406 - .0594 = .8812.$$

So, the probability that the percentage of heads is between 45% and 55% is high, but not really high if we toss the coin 100 times. Here is the next question; how many times do we need to toss a fair coin to be 99% sure that the percentage of heads will be between 45% and 55%? The percentage of heads is between 45% and 55% if and only if the number of heads is betwwen $.45n$ and $.55n$. Using the continuity corrected normal approximation, again, we want

$$.99 = \Phi(\frac{.55n + .5 - .5n}{\sqrt{.25n}}) - \Phi(\frac{.45n - .5 - .5n}{\sqrt{.25n}})$$

$$\Rightarrow .99 = 2\Phi(\frac{.55n + .5 - .5n}{\sqrt{.25n}}) - 1$$

(because, for any real number $x, \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$)

$$\Rightarrow \Phi(\frac{.55n + .5 - .5n}{\sqrt{.25n}}) = 995$$

$$\Rightarrow \Phi(\frac{.05n + .5}{\sqrt{.25n}}) = .995.$$

Now, from a standard normal table, we find that $\Phi(2.575) = .995$. Therefore, we equate

$$\frac{.05n + .5}{\sqrt{.25n}} = 2.575$$

$$\Rightarrow .05n + .5 = 2.575 \times .5\sqrt{n} = 1.2875\sqrt{n}.$$

Writing $\sqrt{n} = x$, we have here a quadratic equation $.05x^2 - 1.2875x + .5 = 0$ to solve. The root we want is $x = 25.71$, and squaring it gives $n \ge (25.71)^2 = 661.04$. Thus, an

approximate value of n such that in n tosses of a fair coin, the percentage of heads will be between 45% and 55% with a 99% probability is $n = 662$. Most people find that the value of n needed is higher than what they would have guessed.

Example 1.76. (Public Polling: Predicting the Correct Winner). Normal approximation to binomial probabilities is routinely used in designing polls on an issue, for example polls to predict a winner in an election. Suppose in an election, there are two candidates, A and B, and among *all voters*, 52% support A and 48% support B. A poll of 1400 voters is done; what is the probability that the poll will predict the correct winner?

Let X demote the number of respondents in the poll who favor A. The poll will predict the correct winner if $X > 700$. By using the continuity corrected normal approximation,

$$P(X > 700) = 1 - P(X \le 700) \approx 1 - \Phi(\frac{700.5 - 1400 \times .52}{\sqrt{1400 \times .52 \times .48}})$$

$$= 1 - \Phi(-1.5) = \Phi(1.5) = .9332.$$

As long as the spread between the candidates' support is sufficiently large, say 4% or more, a poll that uses about 1500 respondents will predict the correct winner with a high probability. *But it takes much larger polls to predict the correct spread accurately.*

Example 1.77. (Rounding Errors and CLT). Suppose n positive numbers are rounded to their nearest integers, and suppose the rounding errors e_i = (True value of X_i − Rounded value of X_i) are independently distributed as $U[-.5, .5]$. We want to find the probability that the total error is at most some number k in magnitude. An example would be a tax agency rounding off the exact refund amount to the nearest integer, in which case the total error would be the agency's loss or profit due to this rounding process.

From the general formulas for the mean and variance of a uniform distribution, each e_i has mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{12}$. Therefore, by the CLT, the total error $S_n = \sum_{i=1}^{n} e_i$ has the approximate normal distribution

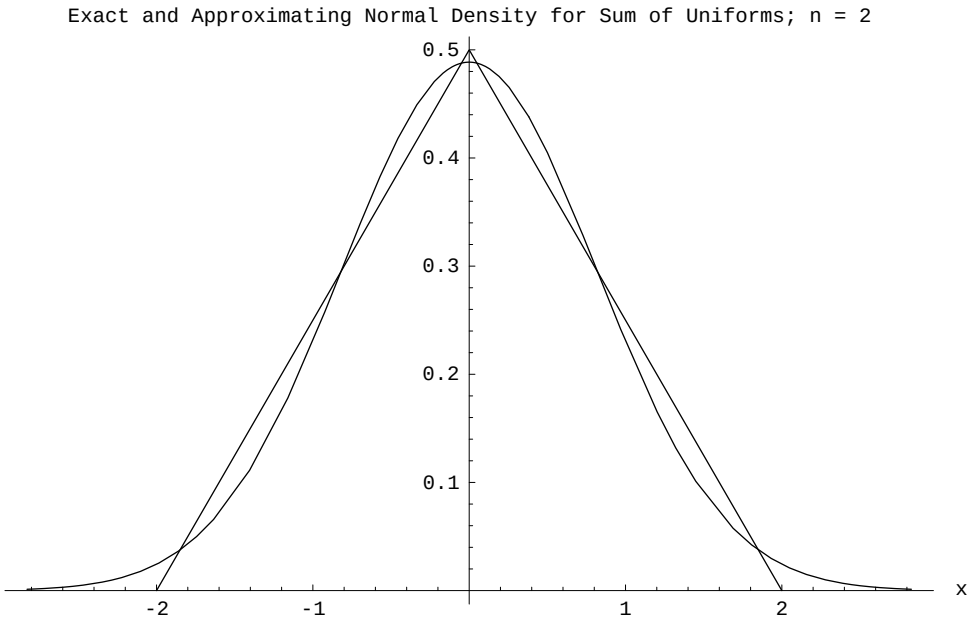
$$S_n \approx N(0, \frac{n}{12}).$$

For example, if $n = 1000$, then

$$P(|S_n| \le 20) = P(S_n \le 20) - P(S_n \le -20) = P(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{20}{\sqrt{\frac{n}{12}}}) - P(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{-20}{\sqrt{\frac{n}{12}}})$$

$$\approx \Phi(2.19) - \Phi(-2.19) = .9714.$$

We see from here that due to the cancellations of positive and negative errors, the tax agency is unlikely to lose or gain much money due to rounding.



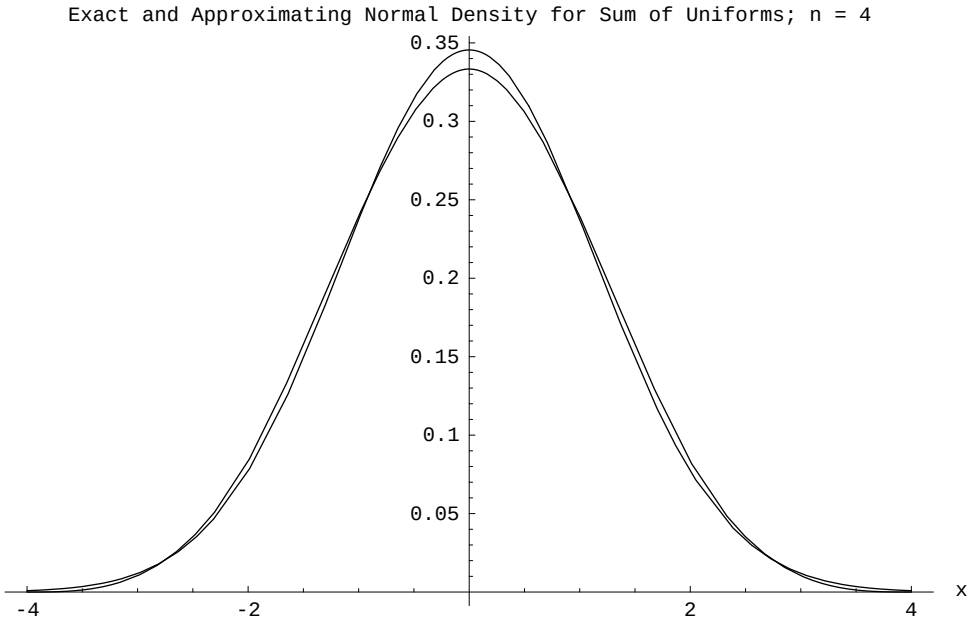
Example 1.78. (Sum of Uniforms). We can approximate the distribution of the sum of n independent uniforms on a general interval $[a, b]$ by a suitable normal distribution. However, it is interesting to ask what is the exact density of the sum of n independent uniforms on a general interval $[a, b]$. Since a uniform random variable on a general interval $[a, b]$ can be transformed to a uniform on the unit interval $[-1, 1]$ by a linear transformation and vice versa, we ask what is the exact density of the sum of n independent uniforms on $[-1, 1]$. We want to compare this exact density to a normal approximation for various values of n.

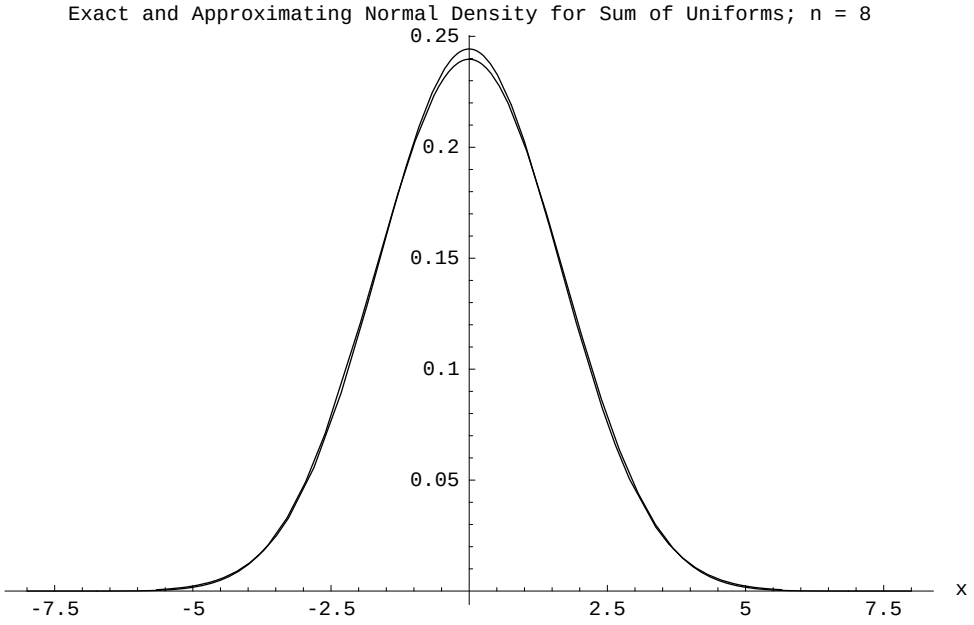
When $n = 2$, the density of the sum is a triangular density on $[-2, 2]$, which is a piecewise linear polynomial. In general, the density of the sum of n independent uniforms on $[-1, 1]$ is a piecewise polynomial of degree $n - 1$, there being n different arcs in the graph of the density. The exact formula is:

$$f_n(x) = \frac{1}{2^n(n-1)!} \sum_{k=0}^{\lfloor \frac{n+x}{2} \rfloor} (-1)^k \binom{n}{k} (n + x - 2k)^{n-1}, \text{ if } |x| \leq n;$$

see Feller (1971, pp 27).

On the other hand, the CLT approximates the density of the sum by the $N(0, \frac{n}{3})$ density. It would be interesting to compare plots of the exact and the approximating normal density for various n. We see from the plots that the normal approximation is already nearly exact when $n = 8$.


Exact and Approximating Normal Density for Sum of Uniforms; n = 4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
-4
-2
2
4
x


Exact and Approximating Normal Density for Sum of Uniforms; n = 8
0.25
0.2
0.15
0.1
0.05
-7.5
-5
-2.5
2.5
5
7.5
x

1.14 Normal Approximation to Poisson and Gamma

A Poisson variable with an integer parameter $\lambda = n$ can be thought of as the sum of n independent Poisson variables, each with mean 1. Likewise, a Gamma variable with parameters $\alpha = n$ and λ can be thought of as the sum of n independent Exponential variables, each with mean λ. So, in these two cases the CLT already implies that a normal approxmation to the Poisson and the Gamma holds, when n is large. However, *even if the Poisson parameter λ is not an integer, and even if the Gamma parameter α is not an integer*, if λ is large, or if α is large, *a normal approximation still holds*. These results can be proved directly, by using the mgf technique. Here are the normal approximation results for general Poisson and Gamma distributions.

Theorem 1.46. Let $X \sim Poisson(\lambda)$. Then

$$P(\frac{X-\lambda}{\sqrt{\lambda}} \le x) \to \Phi(x), \text{ as } \lambda \to \infty,$$

for any real number x.

Notationally, for large λ,

$$X \approx N(\lambda, \lambda).$$

Theorem 1.47. Let $X \sim G(\alpha, \lambda)$. Then, for every fixed λ,

$$P(\frac{X-\alpha\lambda}{\lambda\sqrt{\alpha}} \le x) \to \Phi(x), \text{ as } \alpha \to \infty,$$

for any real number x.

Notationally, for large α,

$$X \approx N(\alpha\lambda, \alpha\lambda^2).$$

Example 1.79. (Nuclear Accidents). Suppose the probability of having any nuclear accidents in any single nuclear plant during a given year is .0005, and that a country has 100 such nuclear plants. What is the probability that there will be at least six nuclear accidents in the country during the next 250 years?

Let X_{ij} be the number of accidents in the ith year in the jth plant. We assume that each X_{ij} has a common Poisson distribution. The parameter , say θ of this common Poisson distribution is determined from the equation $e^{-\theta} = 1 - .0005 = .9995 \Rightarrow \theta = -\log(.9995) = .0005$.

Assuming that these X_{ij} are all independent, the number of accidents T in the country during 250 years has a $Poi(\lambda)$ distribution, where $\lambda = \theta \times 100 \times 250 = .0005 \times 100 \times 250 = 12.5$. If we now do a normal approximation with continuity correction,

$$P(T \ge 6) \approx 1 - \Phi(\frac{5.5 - 12.5}{\sqrt{12.5}})$$

$$= 1 - \Phi(-1.98) = .9761.$$

So we see that although the chances of having any accidents in a particular plant in any particular year are small, collectively, and in the long run, the chances are high that there will be quite a few such accidents.

Example 1.80. (Diabetic Retinopathy and CLT). Diabetes is one of the main causes for development of an eye disease known as retinopathy, which causes damage to the blood vessels in the retina and growth of abnormal blood vessels, potentially causing loss of vision. The average time to develop retinopathy after onset of diabetes is 15 years, with a standard deviation of 4 years.

Suppose we let X be the time from onset of diabetes till development of retinopathy, and suppose we model it as $X \sim G(\alpha, \lambda)$. Then, we have

$$\alpha\lambda = 15; \lambda\sqrt{\alpha} = 4$$

$$\Rightarrow \sqrt{\alpha} = \frac{15}{4} = 3.75 \Rightarrow \alpha = 14.06, \lambda = 1.07.$$

Suppose we want to know what percentage of diabetes patients develop retinopathy within 20 years. Since $\alpha = 14.06$ is large, we can use a normal approximation:

$$P(X \leq 20) \approx \Phi(\frac{20 - 15}{4}) = \Phi(1.25) = .8944;$$

i.e., under the Gamma model, approximately 90% develop diabetic retinopathy within twenty years.

1.15 Bivariate Discrete Distributions

We have so far provided a detailed overview of distributions of one discrete or one continuous random variable in the previous sections. But often in applications, we are just naturally interested in two or more random variables simultaneously. We may be interested in them simultaneously because they provide information about each other, or because they arise simultaneously as part of the data in some scientific experiment. For instance, on a doctor's visit, the physician may check someone's blood pressure, pulse rate, blood cholesterol level, and blood sugar level, because together they give information about the general health of the patient. In such cases, it becomes essential to know how to operate with many random variables simultaneously. This is done by using *joint distributions*. Joint distributions naturally lead to considerations of *marginal and conditional distributions*. We will now study joint, marginal, and conditional distributions for discrete random variables in this section. The concepts of joint, marginal, and conditional distributions for continuous random variables are not different; but the techniques are mathematically more sophisticated. The continuous case will be treated in the next section.

1.15.1 Joint Distributions and Expectations of Functions

We present the fundamentals of joint distributions of two variables in this section. The concepts in the multivariate case are the same, although the technicalities are somewhat more involved. We will treat the multivariate case in a later section. The idea is that there is still an underlying experiment ξ, with an associated sample space Ω. But now, we have two or more random variables on the sample space Ω. Random variables being functions on the sample space Ω, we now have multiple functions, say $X(\omega), Y(\omega), \cdots$, etc. on Ω. We want to study their *joint behavior*.

Example 1.81. (Coin tossing). Consider the experiment ξ of tossing a fair coin three times. Let X be the number of heads among the first two tosses, and Y the number of heads among the last two tosses. If we consider X and Y *individually*, we realize immediately that they are each $Bin(2, .5)$ random variables. But the individual distributions hide part of the full story. For example, if we knew that X was 2, then that would imply that Y must be at least 1. Thus, their joint behavior cannot be fully understood from their individual distributions; we must study their *joint distribution*.

Here is what we mean by their joint distribution. The sample space Ω of this experiment is

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

Each sample point has an equal probability $\frac{1}{8}$. Denoting the sample points as $\omega_1, \omega_2, \cdots, \omega_8$, we see that if ω_1 prevails, then $X(\omega_1) = Y(\omega_1) = 2$. but if ω_2 prevails, then $X(\omega_2) = 2, Y(\omega_2) = 1$. The combinations of *all* possible values of (X, Y) are

$$(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2).$$

The joint distribution of (X, Y) provides the probability $p(x, y) = P(X = x, Y = y)$ for each such combination of possible values (x, y). Indeed, by direct counting using the eight equally likely sample points, we see that

$$p(0,0) = \frac{1}{8}, p(0,1) = \frac{1}{8}, p(0,2) = 0, p(1,0) = \frac{1}{8}, p(1,1) = \frac{1}{4};$$

$$p(1,2) = \frac{1}{8}, p(2,0) = 0, p(2,1) = \frac{1}{8}, p(2,2) = \frac{1}{8}.$$

For example, why is $p(0,1)\frac{1}{8}$? This is because the combination $(X = 0, Y = 1)$ is favored by only one sample point, namely TTH. It is convenient to present thes e nine different probabilities in the form of a table as follows.

	Y		
X	0	1	2
0	$\frac{1}{8}$	$\frac{1}{8}$	0
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$

Such a layout is a convenient way to present the joint distribution of two discrete random variables with a small number of values. The distribution itself is called *the joint pmf*; here is a formal definition. Def Let X, Y be two discrete random variables with respective sets of values $x_1, x_2, \cdots$, and $y_1, y_2, \cdots$, defined on a common sample space Ω. The *joint pmf* of X, Y is defined to be the function $p(x_i, y_j) = P(X = x_i, Y = y_j), i, j \geq 1$, and $p(x, y) = 0$ at any other point (x, y) in $\mathcal{R}^2$.

The requirements of a joint pmf are that

$$(i)\, p(x, y) \geq 0 \ \forall (x, y);$$

$$(ii) \sum_i \sum_j p(x_i, y_j) = 1.$$

Thus, if we write the joint pmf in the form of a table, then all entries should be nonnegative, and the sum of all the entries in the table should be one.

As in the case of a single variable, we can define a CDF for more than one variable also. For the case of two variables, here is the definition of a CDF.

Definition 1.36. Let X, Y be two discrete random variables, defined on a common sample space Ω. The joint CDF, or simply the CDF, of (X, Y) is a function $F : \mathcal{R}^2 \to [0, 1]$ defined as $F(x, y) = P(X \leq x, Y \leq y), x, y \in \mathcal{R}$.

Like the joint pmf, the CDF also characterizes the joint distribution of two discrete random variables. But it is not very convenient or even interesting to work with the CDF in the case of discrete random variables. It ismuch preferable to work with the pmf, when dealing with discrete random variables.

Example 1.82. (Maximum and Minimum in Dice Rolls). Suppose a fair die is rolled twice, and let X, Y be the larger and the smaller of the two rolls (note that X can be equal to Y). Each of X, Y takes the individual values $1, 2, \cdots, 6$, but we have necessarily $X \geq Y$. The sample space of this experiment is

$$\{11, 12, 13, \cdots, 64, 65, 66\}.$$

By direct counting, for example, $p(2, 1) = \frac{2}{36}$. Indeed, $p(x, y) = \frac{2}{36}$ for each $x, y = 1, 2, \cdots, 6, x > y$, and $p(x, y) = \frac{1}{36}$ for $x = y = 1, 2, \cdots, 6$. Here is how the joint pmf looks like in the form of a table:

	Y					
X	1	2	3	4	5	6
1	$\frac{1}{36}$	0	0	0	0	0
2	$\frac{1}{18}$	$\frac{1}{36}$	0	0	0	0
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0	0	0
4	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0	0
5	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0
6	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$

The individual pmfs of X, Y are easily recovered from the joint distribution. For example, $P(X = 1) = \sum_{y=1}^{6} P(X = 1, Y = y) = \frac{1}{36}$, and $P(X = 2) = \sum_{y=1}^{6} P(X = 2, Y = y) = \frac{1}{18} + \frac{1}{36} = \frac{1}{12}$, etc. The individual pmfs are obtained by summing the joint probabilities over all values of the other variable. They are:

x	1	2	3	4	5	6
$p_X(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

y	1	2	3	4	5	6
$p_Y(y)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

From the individual pmf of X, we can find the expectation of X. Indeed, $E(X) = 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + \cdots + 6 \times \frac{11}{36} = \frac{161}{36}$. Similarly, $E(Y) = \frac{91}{36}$.
The individual pmfs are called *marginal pmfs*, and here is the formal definition.

Definition 1.37. Let $p(x, y)$ be the joint pmf of (X, Y). The *marginal pmf* of a function $Z = g(X, Y)$ s defined as $p_Z(z) = \sum_{(x,y):g(x,y)=z} p(x, y)$. In particular,

$$p_X(x) = \sum_y p(x, y);\ p_Y(y) = \sum_x p(x, y),$$

and for any event A,

$$P(A) = \sum_{(x,y)\in A} p(x, y).$$

Example 1.83. (Dice Rolls Revisited). Consider again the example of two rolls of a fair die, and suppose X, Y are the larger and the smaller of the two rolls. We have just worked out the joint distribution of (X, Y). Suppose we want to find the distribution of the difference, $X - Y$. The possible values of $X - Y$ are $0, 1, \cdots, 5$, and we find $P(X - Y = k)$ by using the joint distribution of (X, Y):

$$P(X - Y = 0) = p(1, 1) + p(2, 2) + \cdots + p(6, 6) = \frac{1}{6};$$

$$P(X-Y=1) = p(2,1) + p(3,2) + \cdots + p(6,5) = \frac{5}{18};$$

$$P(X-Y=2) = p(3,1) + p(4,2) + p(5,3) + p(6,4) = \frac{2}{9};$$

$$P(X-Y=3) = p(4,1) + p(5,2) + p(6,3) = \frac{1}{6};$$

$$P(X-Y=4) = p(5,1) + p(6,2) = \frac{1}{9};$$

$$P(X-Y=5) = p(6,1) = \frac{1}{18}.$$

There is no way to find the distribution of $X-Y$ except by using the joint distribution of (X,Y).

Suppose now we want to also know the expected value of $X-Y$. Now that we have the distribution of $X-Y$ worked out, we can find the expectation by directly using the definition of expectation:

$$E(X-Y) = \sum_{k=0}^{5} kP(X-Y=k)$$

$$= \frac{5}{18} + \frac{4}{9} + \frac{1}{2} + \frac{4}{9} + \frac{5}{18} = \frac{35}{18}.$$

But, we can also use linearity of expectations and find $E(X-Y)$ as

$$E(X-Y) = E(X) - E(Y) = \frac{161}{36} - \frac{91}{36} = \frac{35}{18}$$

(see the previous example for $E(X), E(Y)$).

A third possible way to compute $E(X-Y)$ is to treat $X-Y$ as a function of (X,Y) and *use the joint pmf of* (X,Y) to find $E(X-Y)$ as $\sum_x \sum_y (x-y)p(x,y)$. In this particular example, this will be an unncessarily laborious calculation, because luckily we can find $E(X-Y)$ by other quicker means in this example, as we just saw. But in general, one has to resort to the joint pmf to calculate the expectation of a function of (X,Y). Here is the formal formula.

Theorem 1.48. (Expectation of a Function). Let (X,Y) have the joint pmf $p(x,y)$, and let $g(X,Y)$ be a function of (X,Y). We say that the expectation of $g(X,Y)$ exists if $\sum_x \sum_y |g(x,y)|p(x,y) < \infty$, in which case,

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

Example 1.84. Consider the example of three tosses of a fair coin, and let X, Y be the number of heads in the first two and the last two tosses, respectively. Let $g(X,Y) = |X-Y|$. We want to find the expectation of $g(X,Y)$. Because of the absolute value, we

cannot find this expectation from the marginal distributions of X and Y; we must use the joint pmf in this case.

Using the joint pmf of (X, Y) from Example 3.77,

$$E(|X - Y|) = \sum_{x=0}^{2} \sum_{y=0}^{2} |x - y| p(x, y)$$

$$= 1 \times [p(0,1) + p(1,0) + p(1,2) + p(2,1)] + 2 \times [p(0,2) + p(2,0)] = \frac{4}{8} = \frac{1}{2}.$$

1.15.2 Conditional Distributions and Conditional Expectations

Sometimes, we want to know what is the expected value of one of the variables, say X, if we knew the value of the other variable Y. For example, in the die tossing experiment above, what should we expect the larger of the two rolls to be if the smaller roll is known to be 2?

To answer this question, we have to find the probabilities of the various values of X, *conditional on* knowing that Y equals some given y, and then average by using these conditional probabilities. Here are the formal definitions.

Definition 1.38. (Conditional Distribution) Let (X, Y) have the joint pmf $p(x, y)$. The *conditional distribution of X given $Y = y$* is defined to be

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

and the *conditional expectation of X given $Y = y$* is defined to be

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

The conditional distribution of Y given $X = x$ and the conditional expectation of Y given $X = x$ are defined analogously, by switching the roles of X and Y in the above definitions.

We often casually write $E(X|y)$ to mean $E(X|Y = y)$.

Two easy facts that are nevertheless often useful are the following.

Proposition Let X, Y be random variables defined on a common sample space Ω. Then,

$$(a)\, E(g(Y)|Y = y) = g(y),\ \forall y,\ \text{and for any function}\, g;$$

$$(b)\, E(Xg(Y)|Y = y) = g(y)E(X|Y = y)\ \forall y,\ \text{and for any function} g.$$

Recall that two random variables are called independent if $P(X \le x, Y \le y) = P(X \le x)P(Y \le y)\ \forall\, x, y \in \mathcal{R}$. This is of course a correct definition; but in the case of discrete

random variables, it is more convenient to think of independence in terms of the pmf. The definition below puts together some equivalent definitions of independence of two discrete random variables,

Definition 1.39. (Independence) Let (X, Y) have the joint pmf $p(x, y)$. Then X, Y are said to be independent if

$$p(x|y) = p_X(x),\ \forall\, x, y \text{ such that } p_Y(y) > 0;$$

$$\Leftrightarrow p(y|x) = p_Y(y).\ \forall\, x, y \text{ such that } p_X(x) > 0;$$

$$\Leftrightarrow p(x, y) = p_X(x)p_Y(y), \forall\, x, y;$$

$$\Leftrightarrow P(X \le x, Y \le y) = P(X \le x)P(Y \le y)\ \forall\, x, y.$$

The third equivalent condition in the above list is usually the most convenient one to verify and use.

One more frequently useful fact about conditional expectations is the following.

Proposition Suppose X, Y are independent random variables. Then, for any function $g(X)$ such that the expectations below exist, and for any y,

$$E[g(X)|Y = y] = E[g(X)].$$

Example 1.85. In the experiment of three tosses of a fair coin, we have worked out the joint mass function of X, Y, where X is the number of heads in the first two tosses, and Y the number of heads in the last two tosses. Using this joint mass function, we now find

$$P(X = 0|Y = 0) = \frac{p(0,0)}{p_Y(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$

$$P(X = 1|Y = 0) = \frac{p(1,0)}{p_Y(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$

$$P(X = 2|Y = 0) = \frac{p(2,0)}{p_Y(0)} = \frac{0}{1/4} = 0.$$

That is, the conditional distribution of X given $Y = 0$ is a *two point distribution*, although X, by itself, takes three values.

We can also similarly find

$$P(Y = 0|X = 0) = \frac{p(0,0)}{p_X(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$

$$P(Y = 1|X = 0) = \frac{p(0,1)}{p_X(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$

$$P(Y = 2|X = 0) = \frac{p(0,2)}{p_X(0)} = \frac{0}{1/4} = 0.$$

Thus, the conditional distribution of Y given $X = 0$ is also a two point distribution, and in fact, as distributions, the two conditional distributions that we worked out in this example are the same.

Example 1.86. (Maximum and Minimum in Dice Rolls: Continued). In the experiment of two rolls of a fair die, we have worked out the joint distribution of X, Y, where X is the larger and Y the smaller of the two rolls. Using this joint distribution, we can now find the conditional distributions. For instance,

$$P(Y = 1|X = 1) = 1; P(Y = y|X = 1) = 0, \text{ if } y > 1;$$

$$P(Y = 1|X = 2) = \frac{1/18}{1/18 + 1/36} = \frac{2}{3};$$

$$P(Y = 2|X = 2) = \frac{1/36}{1/18 + 1/36} = \frac{1}{3};$$

$$P(Y = y|X = 2) = 0, \text{ if } y > 2;$$

$$P(Y = y|X = 6) = \frac{1/18}{5/18 + 1/36} = \frac{2}{11}, \text{ if } 1 \le y \le 5;$$

$$P(Y = 6|X = 6) = \frac{1/36}{5/18 + 1/36} = \frac{1}{11}.$$

Now consider the problem of finding the conditioonal expectations. We will find $E(X|Y = y)$ for various values of y.

By using the definition of $E(X|Y = y)$, we have, for example,

$$E(X|Y = 1) = \frac{1 \times \frac{1}{36} + \frac{1}{18}[2 + \cdots + 6]}{\frac{1}{36} + \frac{5}{18}} = \frac{41}{11} = 3.73;$$

as another example,

$$E(X|Y = 3) = \frac{3 \times \frac{1}{36} + \frac{1}{18} \times 15}{\frac{1}{36} + \frac{3}{18}} = \frac{33}{7} = 4.71;$$

and,

$$E(X|Y = 5) = \frac{5 \times \frac{1}{36} + 6 \times \frac{1}{18}}{\frac{1}{36} + \frac{1}{18}} = \frac{17}{3} = 5.77.$$

We notice that $E(X|Y = 5) > E(X|Y = 3) > E(X|Y = 1)$; in fact, it is true that $E(X|Y = y)$ is increasing in y in this example. This does make intuitive sense.

Example 1.87. (Conditional Expectation of Number of Aces). Consider again the example of number of aces X, Y in the hands of North and South in a Bridge game. We want to find $E(X|Y = y)$ for $y = 0, 1, 2, 3, 4$. Of these, note that $E(X|Y = 4) = 0$.

For the rest, from definition,

$$E(X|Y = y) = \frac{\sum_x xp(x, y)}{\sum_x p(x, y)}$$

$$= \frac{\sum_{x=0}^{4-y} xp(x, y)}{\sum_{x=0}^{4-y} p(x, y)},$$

where $p(x,y) = \frac{\binom{4}{x}\binom{48}{13-x}\binom{4-x}{y}\binom{35+x}{13-y}}{\binom{52}{13}\binom{39}{13}}$, by a simple counting argument.
For example,

$$E(X|Y=2) = \frac{0\times p(0,2) + 1\times p(1,2) + 2\times p(2,2)}{p(0,2)+p(1,2)+p(2,2)}$$
$$= \frac{.0974 + 2\times .0225}{.0936 + .0974 + .0225} = .67.$$

Note that the .67 value is actually $\frac{2}{3}$, and this makes intuitive sense. If South already has 2 aces, then the remaining 2 aces should be divided among East, West, and North equitably, which would give $E(X|Y=2)$ as $\frac{2}{3}$.

Just as in the case of a distribution of a single variable, we often also want a measure of variability in addition to a measure of average for conditional distributions. This motivates defining a *conditional variance*.

Definition 1.40. (Conditional Variance) Let (X,Y) have the joint pmf $p(x,y)$. Let $\mu_X(y) = E(X|Y=y)$. The *conditional variance* of X given $Y=y$ is defined to be

$$\mathrm{Var}(X|Y=y) = E[(X-\mu_X(y))^2|Y=y] = \sum_x (x-\mu_X(y))^2 p(x|y).$$

We often write casually $\mathrm{Var}(X|y)$ to mean $\mathrm{Var}(X|Y=y)$.

Example 1.88. (Conditional Variance in Dice experiment). We will work out the conditional variance of the maximum of two rolls of a die given the minimum. That is, suppose a fair die is rolled twice, and X,Y are the larger and the smaller of the two rolls; we want to compute $\mathrm{Var}(X|y)$.

For example, if $y=3$, then $\mu_X(y) = E(X|Y=y) = E(X|Y=3) = 4.71$ (see the previous example). Therefore,

$$\mathrm{Var}(X|y) = \sum_x (x-4.71)^2 p(x|3)$$

$$= \frac{(3-4.71)^2\times\frac{1}{36} + (4-4.71)^2\times\frac{1}{18} + (5-4.71)^2\times\frac{1}{18} + (6--4.71)^2\times\frac{1}{18}}{\frac{1}{36}+\frac{1}{18}+\frac{1}{18}+\frac{1}{18}} = 1.06.$$

To summarize, given that the minimum of two rolls of a fair die is 3, the expected value of the maximum is 4.71 and the variance of the maximum is 1.06.

These two values, $E(X|y)$ and $Var(X|y)$, change as we change the given value y. Thus, $E(X|y)$ and $Var(X|y)$ are functions of y, and for each separate y, a new calculation is needed. If X,Y happen to be independent, then of course whatever be y, $E(X|y) = E(X)$, and $Var(X|y) = Var(X)$.

The next result is an important one in many applications.

Theorem 1.49. (Poisson Conditional Distribution) Let X, Y be independent Poisson random variables, with means λ, μ. Then the conditional distribution of X given $X + Y = t$ is $Bin(t, p)$, where $p = \frac{\lambda}{\lambda+\mu}$.

Proof: Clearly, $P(X = x|X + Y = t) = 0 \, \forall x > t$. For $x \leq t$,

$$P(X = x|X + Y = t) = \frac{P(X = x, X + Y = t)}{P(X + Y = t)}$$

$$= \frac{P(X = x, Y = t - x)}{P(X + Y = t)}$$

$$= \frac{e^{-\lambda}\lambda^x}{x!} \frac{e^{-\mu}\mu^{t-x}}{(t-x)!} \frac{t!}{e^{-(\lambda+\mu)}(\lambda+\mu)^t}$$

(on using the fact that $X + Y \sim Poi(\lambda + \mu)$)

$$= \frac{t!}{x!(t-x)!} \frac{\lambda^x \mu^{t-x}}{(\lambda+\mu)^t}$$

$$= \binom{t}{x} (\frac{\lambda}{\lambda+\mu})^x (\frac{\mu}{\lambda+\mu})^{t-x},$$

which is the pmf of the $Bin(t, \frac{\lambda}{\lambda+\mu})$ distribution.

1.15.3 Using Conditioning to Evaluate Mean and Variance

Conditioning is often an extremely effective tool to calculate probabilities, means, and variances of random variables with a complex or clumsy joint distribution. Thus, in order to calculate the mean of a random variable X, it is sometimes greatly convenient to follow an *iterative process*, whereby we first evaluate the mean f X *after conditioning on the value* y *of some suitable random variable* Y, and then average over y. The random variable Y has to be chosen judiciously, but is often clear from the context of the specific problem. Here are the precise results on how this technique works; it is important to note that the next two results hold for any kinds of random variables, not just discrete ones.

Theorem 1.50. (Iterated Expectation Formula) Let X, Y be random variables defined on the same probability space Ω. Suppose $E(X)$ and $E(X|Y = y)$ exist for each y. Then,

$$E(X) = E_Y[E(X|Y = y)];$$

thus, in the discrete case,

$$E(X) = \sum_y \mu_X(y) p_Y(y),$$

where $\mu_X(y) = E(X|Y = y)$.

Proof: We prove this for the discrete case. By definition of conditional expectation,

$$\mu_X(y) = \frac{\sum_x xp(x,y)}{p_Y(y)}$$

$$\Rightarrow \sum_y \mu_X(y)p_Y(y) = \sum_y \sum_x xp(x,y) = \sum_x \sum_y xp(x,y)$$

$$= \sum_x x \sum_y p(x,y) = \sum_x xp_X(x) = E(X).$$

The corresponding variance calculation formula is the following.

Theorem 1.51. (Iterated Variance Formula) Let X, Y be random variables defined on the same probability space Ω. Suppose $\mathrm{Var}(X)$, $\mathrm{Var}(X|Y=y)$ exist for each y. Then,

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

Remark: These two formulas for iterated expectation and iterated variance are valid for all types of variables, not just the discrete ones. Thus, these same formulas will still hold when we discuss joint distributions for continuous random variables in the next section. Some operational formulas that one should be familiar with are summarized below:

Conditional Expectation and Variance Rules

$$E(g(X)|X=x) = g(x); E(g(X)h(Y)|Y=y) = h(y)E(g(X)|Y=y);$$

$E(g(X)|Y=y) = E(g(X))$ if X, Y are independent;

$$\mathrm{Var}(g(X)|X=x) = 0; \mathrm{Var}(g(X)h(Y)|Y=y) = h^2(y)\mathrm{Var}(g(X)|Y=y);$$

$\mathrm{Var}(g(X)|Y=y) = \mathrm{Var}(g(X))$ if X, Y are independent.

Example 1.89. Suppose in a certain population, 30% of couples have one child, 50% have two children, and 20% have three children. One family is picked at random from this population. What is the expected number of boys in this family?

Let Y denote the number of children in the family that was picked, and let X be the number of boys it has. Making the usual assumption of a childbirth being equally likely to be a boy or a girl,

$$E(X) = E_Y[E(X|Y=y)] = .3 \times .5 + .5 \times 1 + .2 \times 1.5 = .95.$$

Example 1.90. Suppose a chicken lays a Poisson number of eggs per week with mean λ. Each egg, independently of the others, has a probability p of fertilizing. We want to find the mean and the variance of the number of eggs fertilized in a week.

Let N denote the number of eggs hatched and X the number of eggs fertilized. Then, $N \sim Poi(\lambda)$, and given $N = n, X \sim Bin(n, p)$. Therefore,

$$E(X) = E_N[E(X|N = n)] = E_N[np] = p\lambda,$$

and,

$$\begin{aligned}\text{Var}(X) &= E_N[\text{Var}(X|N = n)] + \text{Var}_N(E(X|N = n) \\ &= E_N[np(1-p)] + \text{Var}_N(np) = \lambda p(1-p) + p^2\lambda = p\lambda.\end{aligned}$$

Interestingly, the number of eggs actually fertilized has the same mean and variance $p\lambda$; (can you see why?)

Sometimes, a formal generalization of the iterated expectation formula when a third variable Z is present is useful. It is particularly useful in hierarchical statistical modelling of distributions, where an ultimate marginal distribution for some X is constructed by first conditioning on a number of auxiliary variables, and then gradually unconditioning them. We state the more general iterated expectation formula; its proof is exactly similar to that of the usual iterated expectation formula.

Theorem 1.52. (Higher Order Iterated Expectation) Let X, Y, Z be random variables defined on the same sample space Ω. Assume that each conditional expectation below and the marginal expectation $E(X)$ exist. Then,

$$E(X) = E_Y[E_{Z|Y}\{E(X|Y = y, Z = z)\}].$$

1.15.4 Covariance and Correlation

We know that variance is additive for independent random variables; i.e., if $X_1, X_2, \cdots, X_n$ are independent random variables, then $\text{Var}(X_1+X_2+\cdots+X_n) = \text{Var}(X_1)+\cdots+\text{Var}(X_n)$. In particular, for two independent random variables $X, Y, \text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$. However, *in general*, variance is not additive. Let us do the general calculation for $\text{Var}(X+Y)$.

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

We thus have the extra term $2[E(XY) - E(X)E(Y)]$ in the expression for $\text{Var}(X+Y)$; of course, when X, Y are independent, $E(XY) = E(X)E(Y)$, and so the extra term drops out. But, in general, one has to keep the extra term. The quantity $E(XY) - E(X)E(Y)$ is called the *covariance* of X and Y.

Definition 1.41. (Covariance) Let X, Y be two random variables defined on a common sample space Ω, such that $E(XY), E(X), E(Y)$ all exist. The *covariance* of X and Y is defined as

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

Remark: Covariance is a measure of whether two random variables X, Y tend to increase or decrease together. If a larger value of X generally causes an increment in the value of Y, then often (but not always) they have a positive covariance. For example, taller people tend to weigh more than shorter people, and height and weight usually have a positive covariance.

Unfortunately, however, covariance can take arbitrary positive and arbitrary negative values. Therefore, by looking at its value in a particular problem, we cannot judge whether it is a large value or not. We cannot compare a covariance with a standard to judge if it is large or small. A renormalization of the covariance cures this problem, and calibrates it to a scale of -1 to $+1$. The renormalized quantity is the *correlation coefficient* or simply the *correlation* between X and Y.

Definition 1.42. (Correlation) Let X, Y be two random variables defined on a common sample space Ω, such that $\text{Var}(X)$, $\text{Var}(Y)$ are both finite. The *correlation* between X, Y is defined to be

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

Some important properties of covariance and correlation are put together in the next theorem.

Theorem 1.53. (Properties of Covariance and Correlation) Provided that the required variances and the covariances exist,

$$(a)\, \text{Cov}(X, c) = 0 \text{ for any } X \text{ and any constant } c;$$

$$(b)\, \text{Cov}(X, X) = \text{var}(X) \text{ for any } X;$$

$$(c)\, \text{Cov}(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j),$$

and in particular,

$$\text{Var}(aX + bY) = \text{Cov}(aX + bY, aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y),$$

and,

$$\text{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum \sum_{i<j=1}^{n} \text{Cov}(X_i, X_j);$$

(*d*) For any two independent random variables $X, Y, \text{Cov}(X, Y) = \rho_{X,Y} = 0$;

$$(e)\ \rho_{a+bX,c+dY} = sgn(bd)\rho_{X,Y},$$

where $sgn(bd) = 1$ if $bd > 0$, and $= -1$ if $bd < 0$.

(f) Whenever $\rho_{X,Y}$ is defined, $-1 \le \rho_{X,Y} \le 1$.

(g) $\rho_{X,Y} = 1$ if and only if for some a, some $b > 0, P(Y = a + bX) = 1$; $\rho_{X,Y} = -1$ if and only if for some a, some $b < 0, P(Y = a + bX) = 1$.

Example 1.91. (Correlation between Minimum and Maximum in Dice Rolls). Consider again the experiment of rolling a fair die twice, and let X, Y be the maximum and the minimum of the two rolls. We want to find the correlation between X, Y.

The joint distribution of (X, Y) was worked out in Example 3.78. From the joint distribution,

$$E(XY) = 1/36 + 2/18 + 4/36 + 3/18 + 6/18 + 9/36 + \cdots + 30/18 + 36/36 = 49/4.$$

The marginal pmfs of X, Y are also known to us. From the marginal pmfs, by direct calculation, $E(X) = 161/36, E(Y) = 91/36, \text{Var}(X) = \text{Var}(Y) = 2555/1296$. Therefore,

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

$$= \frac{49/4 - 161/36 \ \times 91/36}{2555/1296} = \frac{35}{73} = .48.$$

The correlation between the maximum and the minimum is in fact positive for *any* number of rolls of a die, although the correlation will converge to zero when the number of rolls converges to ∞.

Example 1.92. (Correlation in the Chicken-Eggs Example). Consider again the example of a chicken laying a Poisson number of eggs, N, with mean λ, and each egg fertilizing, independently of others, with probability p. If X is the number of eggs actually fertilized, we want to find the correlation between the number of eggs laid and the number fertlized, i.e., the correlation between X and N.

First,

$$E(XN) = E_N[E(XN|N = n)] = E_N[nE(X|N = n)]$$

$$= E_N[n^2 p] = p(\lambda + \lambda^2).$$

Next, from our previous calculations, $E(X) = p\lambda, E(N) = \lambda,\ \text{var}(X) = p\lambda, \text{var}(N) = \lambda$. Therefore,

$$\rho_{X,N} = \frac{E(XN) - E(X)E(N)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(N)}}$$

$$= \frac{p(\lambda + \lambda^2) - p\lambda^2}{\sqrt{p\lambda}\sqrt{\lambda}} = \sqrt{p}.$$

Thus, the correlation goes up with the fertility rate of the eggs.

Example 1.93. (Zero Correlation Does Not Mean Independence). If X, Y are independent, then necessarily $\text{Cov}(X, Y) = 0$, and hence the correlation is also zero. The converse is not true. Take a three valued random variable X with the pmf $P(X = \pm 1) = p, P(X = 0) = 1 - 2p, 0 < p < \frac{1}{2}$. Let the other variable Y be $Y = X^2$. Then, $E(XY) = E(X^3) = 0$, and $E(X)E(Y) = 0$, because $E(X) = 0$. Therefore, $\text{Cov}(X, Y) = 0$. But X, Y are certainly not independent; e.g. $P(Y = 0|X = 0) = 1$, but $P(Y = 0) = 1 - 2p \neq 0$.

Indeed, if X has a distribution symmetric around zero, and if X has three finite moments, then X and X^2 always have a zero correlation, although they are not independent.

Example 1.94. (Best Linear Predictor). Suppose X and Y are two jointly distributed random variables, and either by necessity, or by omission, the variable Y was not observed. But X was observed, and there may be some information in the X value about Y. The problem is to predict Y by using X. Linear predictors, because of their functional simplicity, are appealing. The mathematical problem is to choose the *best linear predictor* $a + bX$ of Y, where best is defined as the predictor that minimizes *the mean squared error* $E[Y - (a + bX)]^2$. We will see that the answer has something to do with the covariance between X and Y.

By breaking the square, $R(a, b) =$

$$E[Y - (a + bX)]^2 = a^2 + b^2E(X^2) + 2abE(X) - 2aE(Y) - 2bE(XY) + E(Y^2).$$

To minimize this with respect to a, b, we partially differentiate $R(a, b)$ with respect to a, b, and set the derivatives equal to zero:

$$\frac{\partial}{\partial a}R(a, b) = 2a + 2bE(X) - 2E(Y) = 0$$

$$\Leftrightarrow a + bE(X) = E(Y);$$

$$\frac{\partial}{\partial b}R(a, b) = 2bE(X^2) + 2aE(X) - 2E(XY) = 0$$

$$\Leftrightarrow aE(X) + bE(X^2) = E(XY).$$

Simultaneously solving this two equations, we get

$$b = \frac{E(XY) - E(X)E(Y)}{\text{Var}(X)}, a = E(Y) - \frac{E(XY) - E(X)E(Y)}{\text{Var}(X)}E(X).$$

These values do minimize $R(a, b)$ by an easy application of the second derivative test. So, the best linear predictor of Y based on X is

$$\text{Best Linear Predictor of } Y = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)}E(X) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}X$$

$$= E(Y) + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}[X - E(X)].$$

The best linear predictor is also known as *the regression line of* Y *on* X. It is of widespread use in statistics.

1.16 Multivariate Case

The extension of the concepts for the bivariate discrete case to the multivariate discrete case is straightforward. We will give the appropriate definitions and an important example, namely that of the *multinomial distribution*, an extension of the binomial distribution.

Definition 1.43. Let $X_1, X_2, \cdots, X_n$ be discrete random variables defined on a common sample space Ω, with X_i taking values in some countable set $\mathcal{X}_i$. The *joint pmf* of $(X_1, X_2, \cdots, X_n)$ is defined as $p(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n), x_i \in \mathcal{X}_i$, and zero otherwise..

Definition 1.44. Let $X_1, X_2, \cdots, X_n$ be random variables defined on a common sample space Ω. The *joint CDF* of $X_1, X_2, \cdots, X_n$ is defined as $F(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n), x_1, x_2, \cdots, x_n \in \mathcal{R}$.

The requirements of a joint pmf are the usual:

$$(i)\, p(x_1, x_2, \cdots, x_n) \geq 0 \;\forall\, x_1, x_2, \cdots, x_n \in \mathcal{R};$$

$$(ii) \sum_{x_1 \in \mathcal{X}_1, \cdots, x_n \in \mathcal{X}_n} p(x_1, x_2, \cdots, x_n) = 1.$$

1.16.1 Joint MGF

Analogous to the case of one random variable, we can define the joint mgf for several random variables. The definition is the same for all types of random variables, discrete or continuous, or other mixed types. As in the one dimensional case, the joint mgf of several random variables is also a very useful tool. First, we repeat the definition of expectation of a function of several random random variables.

Definition 1.45. Let $X_1, X_2, \cdots, X_n$ be discrete random variables defined on a common sample space Ω, with X_i taking values in some countable set $\mathcal{X}_i$. Let the joint pmf of $X_1, X_2, \cdots, X_n$ be $p(x_1, \cdots, x_n)$. Let $g(x_1, \cdots, x_n)$ be a real valued function of n variables. We say that $E[g(X_1, X_2, \cdots, X_n)]$ exists if

$\sum_{x_1 \in \mathcal{X}_1, \cdots, x_n \in \mathcal{X}_n} |g(x_1, \cdots, x_n)| p(x_1, \cdots, x_n) < \infty$, in which case, the expectation is defined as

$$E[g(X_1, X_2, \cdots, X_n)] = \sum_{x_1 \in \mathcal{X}_1, \cdots, x_n \in \mathcal{X}_n} g(x_1, \cdots, x_n) p(x_1, \cdots, x_n).$$

A corresponding definition when $X_1, X_2, \cdots, X_n$ are all continuous random variables will be given in the next section.

Definition 1.46. Let $X_1, X_2, \cdots, X_n$ be n random variables defined on a common sample space Ω, The *joint moment generating function* of $X_1, X_2, \cdots, X_n$ is defined to be

$$\psi(t_1, t_2, \cdots, t_n) = E[e^{t_1X_1+t_2X_2+\cdots+t_nX_n}] = E[e^{\mathbf{t}'\mathbf{X}}],$$

provided the expectation exists, and where $\mathbf{t}'\mathbf{X}$ denotes the inner product of the vectors $\mathbf{t} = (t_1, \cdots, t_n), \mathbf{X} = (X_1, \cdots, X_n)$.

Note that the joint moment generating function (mgf) always exists at the origin, namely, $\mathbf{t} = (0, \cdots, 0)$, and equals one at that point. It may or may not exist at other points $\mathbf{t}$. If it does exist in a nonempty rectangle containing the origin, then many important characteristics of the joint distribution of $X_1, X_2, \cdots, X_n$ can be derived by using the joint mgf. As in the one dimensional case, it is a very useful tool. Here is the moment generation property of a joint mgf.

Theorem 1.54. Suppose $\psi(t_1, t_2, \cdots, t_n)$ exists in a nonempty open rectangle containing the origin $\mathbf{t} = \mathbf{0}$. Then a partial derivative of $\psi(t_1, t_2, \cdots, t_n)$ of every order with respect to each t_i exists in that open rectangle, and furthermore,

$$E(X_1^{k_1}X_2^{k_2}\cdots X_n^{k_n}) = \frac{\partial^{k_1+k_2+\cdots+k_n}}{\partial t_1^{k_1}\cdots\partial t_n^{k_n}}\psi(t_1, t_2, \cdots, t_n)|t_1 = 0, t_2 = 0, \cdots, t_n = 0.$$

A corollary of this result is sometimes useful in determining the covariance between two random variables.

Corollary Let X, Y have a joint mgf in some open rectangle around the origin $(0, 0)$. Then,

$$\mathrm{Cov}(X, Y) = \frac{\partial^2}{\partial t_1 \partial t_2}\psi(t_1, t_2)|_{0,0} - (\frac{\partial}{\partial t_1}\psi(t_1, t_2)|_{0,0})(\frac{\partial}{\partial t_2}\psi(t_1, t_2)|_{0,0}).$$

We also have the distribution determining property, as in the one dimensional case.

Theorem 1.55. Suppose $(X_1, X_2, \cdots, X_n)$ and $(Y_1, Y_2, \cdots, Y_n)$ are two sets of jointly distributed random variables, such that their mgfs $\psi_{\mathbf{X}}(t_1, t_2, \cdots, t_n)$ and $\psi_{\mathbf{Y}}(t_1, t_2, \cdots, t_n)$ exist and coincide in some nonempty open rectangle containing the origin. Then $(X_1, X_2, \cdots, X_n)$ and $(Y_1, Y_2, \cdots, Y_n)$ have the same joint distribution.

Remark: It is important to note that the last two theorems are not limited to discrete random variables; they are valid for general random variables.

1.16.2 Multinomial Distribution

One of the most important multivariate discrete distributions is the multinomial distribution. The multinomial distribution corresponds to n balls being distributed to k cells,

independently, with each ball having the probability p_i of being dropped into the ith cell. The random variables under consideration are $X_1, X_2, \cdots, X_k$, where X_i is the number of balls that get dropped into the ith cell. Then their joint pmf is the *multinomial pmf* defined below.

Definition 1.47. A multivariate random vector $(X_1, X_2, \cdots, X_k)$ is said to have a multinomial distribution with parameters $n, p_1, p_2, \cdots, p_k$ if it has the pmf

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, x_i \geq 0, \sum_{i=1}^{k} x_i = n,$$

$p_i \geq 0, \sum_{i=1}^{k} p_i = 1$.

We write $(X_1, X_2, \cdots, X_k) \sim Mult(n, p_1, \cdots, p_k)$ to denote a random vector with a multinomial distribution.

Example 1.95. (Bridge). Consider a Bridge game with four players, North, South, East, and West. We want to find the probability that North and South together have two or more aces. Let X_i denote the number of aces in the hands of player i, i = 1, 2, 3, 4; we let i = 1, 2 mean North and South. Then, we want to find $P(X_1 + X_2 \geq 2)$.

The joint distribution of (X_1, X_2, X_3, X_4) is $Mult(4, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (think of each ace as a ball, and the four players as cells). Then, $(X_1 + X_2, X_3 + X_4) \sim Mult(4, \frac{1}{2}, \frac{1}{2})$. Therefore,

$$\begin{aligned} P(X_1 + X_2 \geq 2) &= \frac{4!}{2!2!}(\frac{1}{2})^4 + \frac{4!}{3!1!}(\frac{1}{2})^4 + \frac{4!}{4!0!}(\frac{1}{2})^4 \\ &= \frac{11}{16}. \end{aligned}$$

Important formulas and facts about the multinomial distribution are given in the next theorem.

Theorem 1.56. Let $(X_1, X_2, \cdots, X_k) \sim Mult(n, p_1, p_2, \cdots, p_k)$. Then,

$$(a)\, E(X_i) = np_i; \mathrm{Var}(X_i) = np_i(1 - p_i);$$

$$(b)\, \forall\, i, X_i \sim Bin(n, p_i);$$

$$(c)\, \mathrm{Cov}(X_i, X_j) = -np_ip_j,\ \forall i \neq j;$$

$$(d)\, \rho_{X_i, X_j} = -\sqrt{\frac{p_ip_j}{(1 - p_i)(1 - p_j)}},\ \forall i \neq j;$$

$$(e)\, E[e^{t_1X_1 + \cdots + t_kX_k}] = (p_1e^{t_1} + p_2e^{t_2} + \cdots + p_ke^{t_k})^n.$$

1.17 Multidimensional Densities

Similar to several discrete random variables, we are frequently interested in applications in studying several continuous random variables simultaneously. And similar to the case of one continuous random variable, again we do not talk of pmfs of several continuous variables, but of a pdf, jointly for all the continuous random variables. The joint density function completely characterizes the joint distribution of the full set of continuous random variables. We refer to the entire set of random variables as a random vector. Both the calculation aspects, as well as the applications aspects of multidimensional density functions are generally sophisticated. As such, acquiring skills in using and operating with multidimensional densities is among the most important skills one needs to have in probability and also in statistics. The general concepts and calculations are discussed in this section. Some special multidimensional densities are introduced separately in the later sections.

1.17.1 Joint Density Function and Its Role

Exactly as in the one dimensional case, it is important to note the following points:

(a) The joint density function of all the variables does not equal the probability of a specific point in the multidimensional space; *the probability of any specific point is still zero.*

(b) The joint density function reflects the relative importance of a particular point. Thus, the probability that the variables together belong to a small set around a specific point, say $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is roughly equal to the volume of that set multiplied by the density function at the specific point $\mathbf{x}$. *This volume interpretation for probabilities is useful for intuitive understanding of distributions of multidimensional continuous random variables.*

(c) For a general set A in the multidimensional space, the probability that the random vector $\mathbf{X}$ belongs to A is obtained by integrating the joint density function over the set A. These are all just the most natural extensions of the corresponding one dimensional facts to the present multidimensional case. We now formally define a joint density function.

Definition 1.48. Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be an n-dimensional random vector, taking values in $\mathcal{R}^n$, for some $n, 1 < n < \infty$. We say that $f(x_1, x_2, \cdots, x_n)$ is the *joint density* or simply the density of $\mathbf{X}$ if for all $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots, b_n, -\infty < a_i \leq b_i < \infty$,

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \cdots, a_n \leq X_n \leq b_n)$$

$$= \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n.$$

In order that a function $f : \mathcal{R}^n \to \mathcal{R}$ be a density function of some n-dimensional random vector, it is necessary and sufficient that

$$(i) f(x_1, x_2, \cdots, x_n) \geq 0 \ \forall (x_1, x_2, \cdots, x_n) \in \mathcal{R}^n;$$

$$(ii) \int_{\mathcal{R}^n} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n = 1.$$

The definition of the joint CDF is the same as what was given in the discrete case. But now the joint CDF is an integral of the density rather than a sum. Here is the precise definition.

Definition 1.49. Let $\mathbf{X}$ be a n-dimensional random vector with the density function $f(x_1, x_2, \cdots, x_n)$. The *joint CDF* or simply the CDF of $\mathbf{X}$ is defined as

$$F(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \cdots, t_n) dt_1 \cdots dt_n.$$

As in the one dimensional case, *both the CDF and the density completely specify the distribution of a continuous random vector* and one can be obtained from the other. We know how to obtain the CDF from the density; the reverse relation is that (for *almost all* $(x_1, x_2, \cdots, x_n)$),

$$f(x_1, x_2, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, x_2, \cdots, x_n).$$

Analogous to the case of several discrete variables, the marginal densities are obtained by integrating out (instead of summing) all the other variables. In fact, all lower dimensional marginals are obtained that way. The precise statement is the following.

Proposition Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be a continuous random vector with a joint density $f(x_1, x_2, \cdots, x_n)$. Let $1 \leq p < n$. Then the *marginal joint density* of $(X_1, X_2, \cdots, X_p)$ is given by

$$f_{1,2,\cdots,p}(x_1, x_2, \cdots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_{p+1} \cdots dx_n.$$

At this stage, it is useful to give a characterization of independence of a set of n continuous random variables by using the density function.

Proposition Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be a continuous random vector with a joint density $f(x_1, x_2, \cdots, x_n)$. Then, $X_1, X_2, \cdots, X_n$ are independent if and only if the joint density factorizes as

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

where $f_i(x_i)$ is the marginal density function of X_i.
Let us see some examples.

Example 1.96. (Bivariate Uniform). Consider the function

$$f(x,y) = 1 \text{ if } 0 \le x \le 1, 0 \le y \le 1;$$

$$= 0 \text{ otherwise.}$$

Clearly, f is always nonnegative, and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_0^1\int_0^1 f(x,y)dxdy$$

$$= \int_0^1\int_0^1 dxdy = 1.$$

Therefore, f is a valid bivariate density function. The marginal density of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$= \int_0^1 f(x,y)dy = \int_0^1 dy = 1,$$

if $0 \le x \le 1$, and zero otherwise. Thus, marginally, $X \sim U[0,1]$, and similarly, marginally, $Y \sim U[0,1]$. Furthermore, clearly, for *all* x, y the joint density $f(x,y)$ factorizes as $f(x,y) = f_1(x)f_2(y)$, and so X, Y are independent too. The joint density $f(x,y)$ of this example is called *the bivariate uniform density.* It gives the constant density of one to all points (x,y) in the unit square $[0,1] \times [0,1]$ and zero density outside of the unit square. The bivariate uniform, therefore, is the same as putting two independent $U[0,1]$ variables together as a bivariate vector.

Example 1.97. (Uniform in a Triangle). Consider the function

$$f(x,y) = c, \text{ if } x, y \ge 0, x + y \le 1,$$

$$= 0 \text{ otherwise.}$$

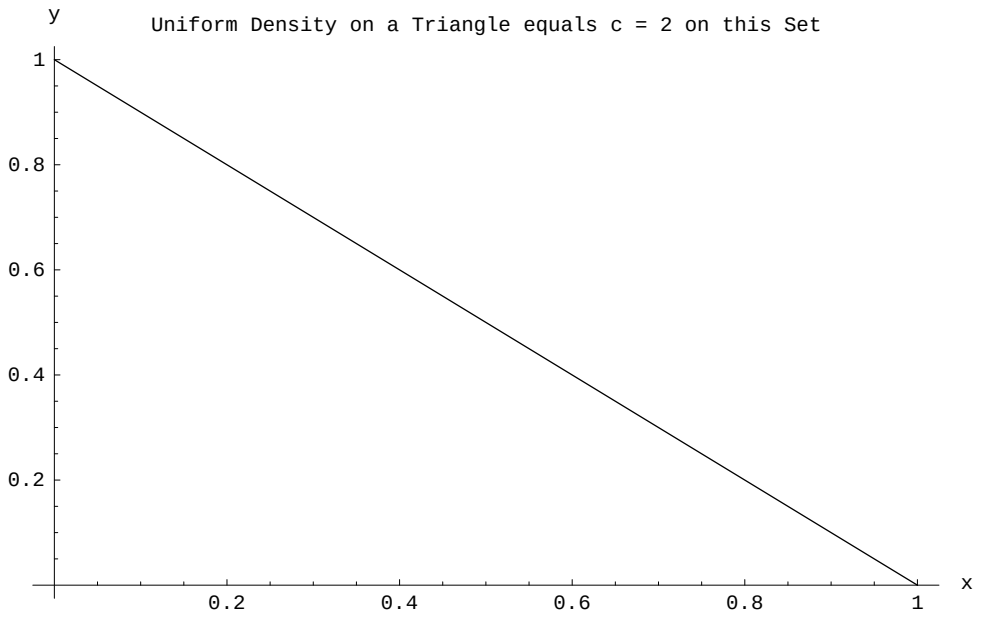
The set of points $x, y \ge 0, x+y \le 1$ form a triangle in the plane with vertices at $(0,0), (1,0)$ and $(0,1)$; thus, it is just half the unit square. The normalizing constant c is easily evaluated:

$$1 = \int_{x,y:x,y\ge 0,x+y\le 1} cdxdy$$

$$= \int_0^1\int_0^{1-y} cdxdy$$

$$= c\int_0^1 (1-y)dy$$

$$= \frac{c}{2}$$



$$\Rightarrow c = 2.$$

The marginal density of X is

$$f_1(x) = \int_0^{1-x} 2dy = 2(1-x), 0 \leq x \leq 1.$$

Similarly, the marginal density of Y is

$$f_2(y) = 2(1-y), 0 \leq y \leq 1.$$

Contrary to the previous example, X, Y are *not independent* now. There are many ways to see this. For example,

$$P(X > \frac{1}{2} | Y > \frac{1}{2}) = 0.$$

But, $P(X > \frac{1}{2}) = \int_{\frac{1}{2}}^1 2(1-x)dx = \frac{1}{4} \neq 0$. So, X, Y cannot be independent. We can also see that the joint density $f(x,y)$ does not factorize as the product of the marginal densities, and so X, Y cannot be independent.

Example 1.98. (Nonuniform Joint Density with Uniform Marginals). Let (X, Y) have the joint density function $f(x,y) = c - 2(c-1)(x+y-2xy), x, y \in [0,1], 0 < c < 2$. This is nonnegative in the unit square, as can be seen by considering the cases $c < 1, c = 1, c > 1$ separately. Also,

$$\int_0^1 \int_0^1 f(x,y)dxdy$$

$$= c - 2(c-1) \int_0^1 \int_0^1 (x + y - 2xy)dxdy$$

$$= c - 2(c-1)\int_0^1 (\frac{1}{2} + y - y)dy = c - (c-1) = 1.$$

Now, the marginal density of X is

$$f_1(x) = \int_0^1 f(x,y)dy$$

$$= c - 2(c-1)[x + \frac{1}{2} - x] = 1.$$

Similarly, the marginal density of Y is also the constant function 1. So each marginal is uniform, although the joint density is not uniform if $c \neq 1$.

Example 1.99. (Uniform Distribution in a Circle). Suppose C denotes the unit circle in the plane:

$$C = \{(x,y) : x^2 + y^2 \leq 1\}.$$

We pick a point (X,Y) at random from C; what that means is that (X,Y) has the density

$$f(x,y) = c, \text{ if } (x,y) \in C,$$

and is zero otherwise. Since

$$\int_C f(x,y)dxdy = c\int_C dxdy = c \times \text{ Area of } C = c\pi = 1,$$

we have that the normalizing constant $c = \frac{1}{\pi}$. Let us find the marginal densities. First,

$$f_1(x) = \int_{y:x^2+y^2\leq 1} \frac{1}{\pi}dy = \frac{1}{\pi}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy$$

$$= \frac{2\sqrt{1-x^2}}{\pi}, -1 \leq x \leq 1.$$

Since the joint density $f(x,y)$ is symmetric between x, y, i.e., $f(x,y) = f(y,x)$, Y has the same marginal density as X, i.e.,

$$f_2(y) = \frac{2\sqrt{1-y^2}}{\pi}, -1 \leq y \leq 1.$$

Since $f(x,y) \neq f_1(x)f_2(y)$, X, Y are not independent. Note that if X, Y has a joint uniform density in the unit square, we have found them to be independent; but now, when they have a uniform density in the unit circle, we find them to be not indpendent. In fact, the following general rule holds;

Suppose a joint density $f(x,y)$ can be written in a form $g(x)h(y), (x,y) \in S$, and $f(x,y)$ zero otherwise. Then, X, Y are independent if and only if S is a rectangle (including squares).

Example 1.100. (Using the Density to Calculate a Probability). Suppose (X, Y) has the joint density $f(x, y) = 6xy^2, x, y \geq 0, x + y \leq 1$. Thus, this is yet another density on the triangle with vertices at $(0, 0), (1, 0)$, and $(0, 1)$. We want to find $P(X + Y < \frac{1}{2})$. By definition,

$$\begin{aligned} P(X + Y < \frac{1}{2}) &= \int_{(x,y);x,y\geq 0,x+y<\frac{1}{2}} 6xy^2 dxdy \\ &= 6\int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-y} xy^2 dxdy \\ &= 6\int_0^{\frac{1}{2}} y^2 \frac{(\frac{1}{2}-y)^2}{2} dy \\ &= 3\int_0^{\frac{1}{2}} y^2 (\frac{1}{2}-y)^2 dy \\ &= 3 \times \frac{1}{960} = \frac{1}{320}. \end{aligned}$$

This example gives an elementary illustration of the need to work out the limits of the iterated integrals carefully while using a joint density to calculate the probability of some event. In fact, properly finding the limits of the iterated integrals is the part that requires the greatest care, when working with joint densities.

1.17.2 Expectation of Functions

Expectations for multidimensional densities are defined analogously to the one dimensional case. Here is the definition.

Definition 1.50. Let $(X_1, X_2, \cdots, X_n)$ have a joint density function $f(x_1, x_2, \cdots, x_n)$, and let $g(x_1, x_2, \cdots, x_n)$ be a real valued function of $x_1, x_2, \cdots, x_n$. We say that the expectation of $g(X_1, X_2, \cdots, X_n)$ exists if

$$\int_{\mathcal{R}^n} |g(x_1, x_2, \cdots, x_n)| f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n < \infty,$$

in which case the expected value of $g(X_1, X_2, \cdots, X_n)$ is defined as

$$E[g(X_1, X_2, \cdots, X_n)] = \int_{\mathcal{R}^n} g(x_1, x_2, \cdots, x_n) f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n.$$

Remark: It is clear from the definition that the expectation of each individual X_i can be evaluated by either interpreting X_i as a function of the full vector $(X_1, X_2, \cdots, X_n)$, or by simply using the marginal density $f_i(x)$ of X_i; that is,

$$\begin{aligned} E(X_i) &= \int_{\mathcal{R}^n} x_i f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\ &= \int_{-\infty}^{\infty} x f_i(x) dx. \end{aligned}$$

A similar comment applies to any function $h(X_i)$ of just X_i alone. All the properties of expectations that we have previously established, for example, linearity of expectations, continue to hold in the multidimensional case. Thus,

$$E[ag(X_1, X_2, \cdots, X_n) + bh(X_1, X_2, \cdots, X_n)]$$

$$= aE[g(X_1, X_2, \cdots, X_n)] + bE[h(X_1, X_2, \cdots, X_n)].$$

We work out some examples.

Example 1.101. (Bivariate Uniform). Two numbers X, Y are picked independently at random from $[0, 1]$. What is the expected distance between them?

Thus, if X, Y are independent $U[0, 1]$, we want to compute $E(|X - Y|)$, which is

$$E(|X - Y|) = \int_0^1 \int_0^1 |x - y| dx dy$$

$$= \int_0^1 [\int_0^y (y - x) dx + \int_y^1 (x - y) dx] dy$$

$$= \int_0^1 [(y^2 - \frac{y^2}{2}) + (\frac{1 - y^2}{2} - y(1 - y))] dy$$

$$= \int_0^1 [\frac{1}{2} - y + y^2] dy$$

$$= \frac{1}{2} - \frac{1}{2} + \frac{1}{3} = \frac{1}{3}.$$

Example 1.102. (Independent Exponentials). Suppose X, Y are independently distributed as $Exp(\lambda), Exp(\mu)$ respectively. We want to find the expectation of the minimum of X and Y. The calculation below requires patience, but is not otherwise difficult.

Denote $W = \min\{X, Y\}$. Then,

$$E(W) = \int_0^\infty \int_0^\infty \min\{x, y\} \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dx dy$$

$$= \int_0^\infty \int_0^y x \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dx dy + \int_0^\infty \int_y^\infty y \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dx dy$$

$$= \int_0^\infty \frac{1}{\mu} e^{-y/\mu} [\int_0^y x \frac{1}{\lambda} e^{-x/\lambda} dx] dy$$

$$+ \int_0^\infty \frac{1}{\mu} e^{-y/\mu} [\int_y^\infty y \frac{1}{\lambda} e^{-x/\lambda} dx] dy$$

$$= \int_0^\infty \frac{1}{\mu} e^{-y/\mu} [\lambda - \lambda e^{-y/\lambda} - y e^{-y/\lambda}] dy$$

$$+ \int_0^\infty \frac{1}{\mu} e^{-y/\mu} y e^{-y/\lambda} dy$$

(on integrating the x integral in the first term by parts)

$$= \frac{\lambda\mu^2}{(\lambda+\mu)^2} + \frac{\mu\lambda^2}{(\lambda+\mu)^2}$$

(once again, by integration by parts)

$$= \frac{\lambda\mu}{\lambda+\mu} = \frac{1}{\frac{1}{\lambda}+\frac{1}{\mu}},$$

a very pretty result.

Example 1.103. (Use of Polar Coordinates). Suppose a point (x, y) is picked at random from inside the unit circle. We want to find its expected distance from the center of the circle.

Thus, let (X, Y) have the joint density

$$f(x, y) = \frac{1}{\pi}, x^2 + y^2 \leq 1,$$

and zero otherwise.

We will find $E[\sqrt{X^2+Y^2}]$. By definition,

$$E[\sqrt{X^2+Y^2}]$$

$$= \frac{1}{\pi}\int_{(x,y):x^2+y^2\leq 1} \sqrt{x^2+y^2}dxdy.$$

It is now very useful to make a transformation by using the polar coordinates

$$x = r\cos\theta, y = r\sin\theta,$$

with $dxdy = rdrd\theta$. Therefore,

$$E[\sqrt{X^2+Y^2}] = \frac{1}{\pi}\int_{(x,y):x^2+y^2\leq 1} \sqrt{x^2+y^2}dxdy$$

$$= \frac{1}{\pi}\int_0^1\int_{-\pi}^{\pi} r^2 d\theta dr$$

$$= 2\int_0^1 r^2 dr = \frac{2}{3}.$$

We will later see in various calculations that transformation to polar and spherical coordiantes often simplifies the integrations involved.

1.18 Bivariate Normal

The bivariate normal density is one of the most important densities for two jointly distributed continuous random variables, just like the univariate normal density is for one continuous variable. Many correlated random variables across applied and social sciences are approximately distributed as a bivariate normal. A typical example is the joint distribution of two size variables, such as height and weight.

Definition 1.51. The function $f(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, -\infty < x, y < \infty$ is called the *bivariate standard normal density.*

Clearly, we see that $f(x,y) = \phi(x)\phi(y)\ \forall\, x, y$. Therefore, the bivariate standard normal distribution corresponds to a pair of independent standard normal variables X, Y. If we make a linear transformation

$$U = \mu_1 + \sigma_1 X$$

$$V = \mu_2 + \sigma_2[\rho X + \sqrt{1-\rho^2}Y],$$

then we get the general *five parameter bivariate normal density*, with means μ_1, μ_2, standard deviations σ_1, σ_2, and correlation $\rho_{U,V} = \rho$; here, $-1 < \rho < 1$.

Definition 1.52. The density of the five parameter bivariate normal distribution is

$$f(u,v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}]},$$

$-\infty < u, v < \infty$.

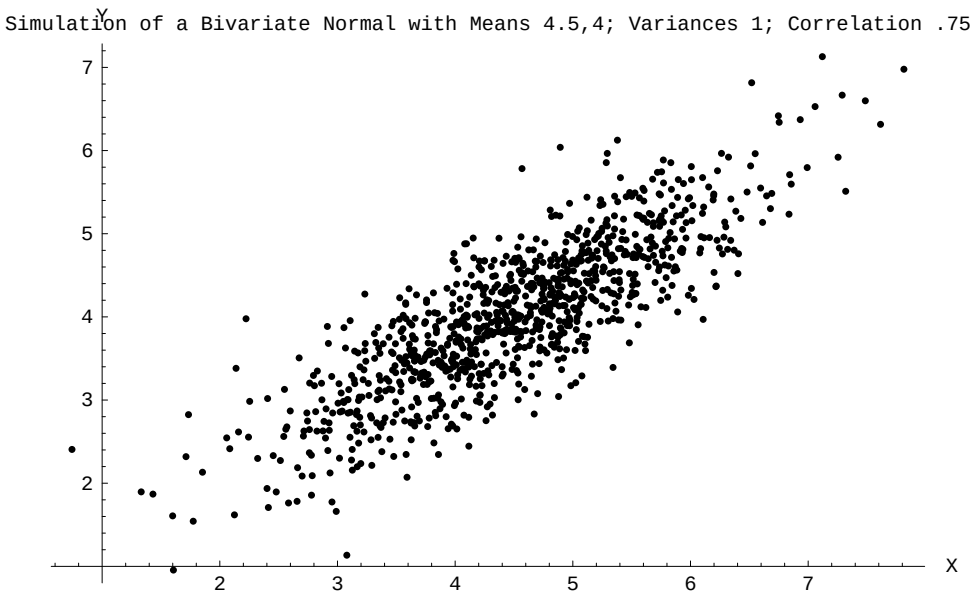
If $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$, then the bivariate normal density has just the parameter ρ, and it is denoted as $SBVN(\rho)$.

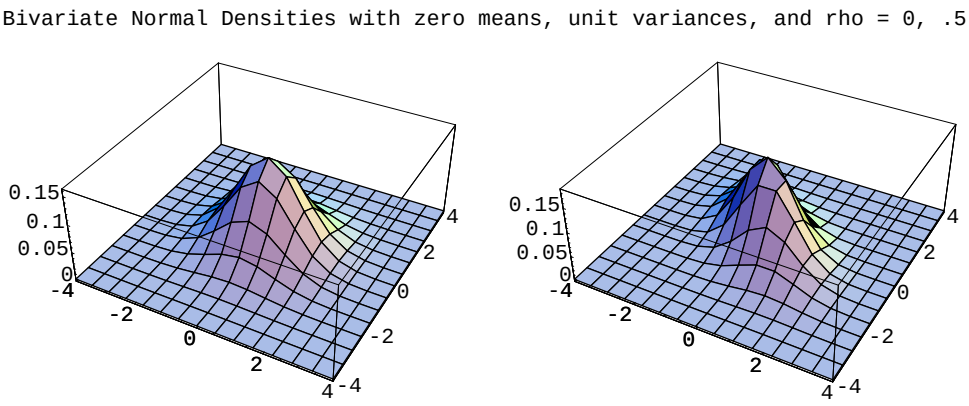
If we sample observations from a general bivariate normal distribution, and plot the data points as points in the plane, then they would roughly plot out to an elliptical shape. The reason for this approximate elliptical shape is that the exponent in the formula for the density function is a quadratic form in the variables. A plot is given here of a simulation of 1000 values from a bivariate normal distribution. The roughly elliptical shape is clear. It is also seen in the plot that the center of the point cloud is quite close to the true means of the variables, which were chosen to be $\mu_1 = 4.5, \mu_2 = 4$.

From the representation we have given above of the general bivariate normal vector (U, V) in terms of independent standard normals X, Y, it follows that

$$E(UV) = \rho\sigma_1\sigma_2 + \mu_1\mu_2$$

$$\Rightarrow \mathrm{Cov}(U,V) = \rho\sigma_1\sigma_2.$$





The symmetric matrix with the variances as diagonal entries and the covariance as the off diagonal entry is called the *variance covariance matrix*, or the *dispersion matrix*, or sometimes simply the *covariance matrix* of (U, V). Thus, the covariance matrix of (U, V) is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

A plot of the $SBVN(\rho)$ density is provided here for $\rho = 0, .5$; the zero correlation case corresponds to independence. We see from the plots that the bivariate density has a unique peak at the mean point $(0, 0)$ and falls off from that point like a mound. The higher the correlation, the more the density concentrates near a plane. In the limiting case, when $\rho = \pm 1$, the density becomes fully concentrated on a plane, and we call it a *singular bivariate normal*.

When $\rho = 0$, the bivariate normal density does factorize into the product of the two marginal densities. Therefore, if $\rho = 0$, then U, V are actually independent, and so, in that

case, $P(U > \mu_1, V > \mu_2) = P(\text{Each variable is larger than its mean value}) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. When the parameters are general, one has the following classic formula.

Theorem 1.57. (A Classic Bivariate Normal Formula) Let (U, V) have the five parameter bivariate normal density with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Then,

$$P(U > \mu_1, V > \mu_2) = P(U < \mu_1, V < \mu_2)$$
$$= \frac{1}{4} + \frac{\arcsin \rho}{2\pi}$$

Another important property of a bivariate normal distribution is the following result.

Theorem 1.58. Let (U, V) have a general five parameter bivariate normal distribution. Then, any linear function $aU + bV$ of (U, V) is normally distributed:

$$aU + bV \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2).$$

In particular, each of U, V is marginally normally distributed:

$$U \sim N(\mu_1, \sigma_1^2), V \sim N(\mu_2, \sigma_2^2).$$

If $\rho = 0$, then U, V are *independent* with $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ marginal distributions.

Proof: First note that $E(aU + bV) = a\mu_1 + b\mu_2$ by linearity of expectations, and $\text{var}(aU + bV) = a^2\text{var}(U) + b^2\text{var}(V) + 2ab\text{Cov}(U, V)$ by the general formula for the variance of a linear combination of two jointly distributed random variables. But $\text{var}(U) = \sigma_1^2, \text{var}(V) = \sigma_2^2$, and $\text{Cov}(U, V) = \rho\sigma_1\sigma_2$. Therefore, $\text{var}(aU + bV) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2$.

Therefore, we only have to prove that $aU + bV$ is normally distributed. For this, we use our representation of U, V in terms of a pair of independent standard normal variables X, Y:

$$U = \mu_1 + \sigma_1 X$$
$$V = \mu_2 + \sigma_2[\rho X + \sqrt{1 - \rho^2}Y].$$

Multiplying the equations by a, b and adding, we get the representation

$$aU + bV = a\mu_1 + b\mu_2 + [a\sigma_1 X + b\sigma_2\rho X + b\sigma_2\sqrt{1 - \rho^2}Y]$$
$$= a\mu_1 + b\mu_2 + [(a\sigma_1 + b\sigma_2\rho)X + b\sigma_2\sqrt{1 - \rho^2}Y].$$

That is, $aU + bV$ can be represented as a linear function $cX + dY + k$ of two independent standard normal variables X, Y, and so $aU + bV$ is necessarily normally distributed

In fact, a result stronger than the previous theorem holds. What is true is that *any two* linear functions of U, V will again be distributed as a bivariate normal. Here is the stronger result.

Theorem 1.59. Let (U, V) have a general five parameter bivariate normal distribution. Let $Z = aU + bV, W = cU + dV$ be two linear functions, such that $ad - bc \neq 0$. Then, (Z, W) also has a bivariate normal distribution, with parameters

$$E(Z) = a\mu_1 + b\mu_2, E(W) = c\mu_1 + d\mu_2;$$

$$\text{Var}(Z) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2;$$

$$\text{Var}(W) = c^2\sigma_1^2 + d^2\sigma_2^2 + 2cd\rho\sigma_1\sigma_2;$$

$$\rho_{Z,W} = \frac{ac\sigma_1^2 + bd\sigma_2^2 + (ad + bc)\rho\sigma_1\sigma_2}{\sqrt{\text{Var}(Z)\text{Var}(W)}}.$$

Example 1.104. (Normal Marginals Do Not Guarantee Joint Normal). Although joint bivariate normality of two random variables implies that each variable must be marginally a univariate normal, the converse is in general not true.

Let $Z \sim N(0, 1)$, and let U be a two valued random variable with the pmf $P(U = \pm 1) = \frac{1}{2}$. Take U and Z to be independent. Define now, $X = U|Z|$, and $Y = Z$.

Then, each of X, Y has a standard normal distribution. That X has a standard normal distribution is easily seen in many ways, for example, by just evaluating its CDF. Take $x > 0$; then,

$$\begin{aligned} P(X \leq x) &= P(X \leq x|U = -1) \times \frac{1}{2} + P(X \leq x|U = 1) \times \frac{1}{2} \\ &= 1 \times \frac{1}{2} + P(|Z| \leq x) \times \frac{1}{2} \\ &= \frac{1}{2} + \frac{1}{2} \times [2\Phi(x) - 1] = \Phi(x). \end{aligned}$$

But, jointly, X, Y cannot be bivariate normal, because $X^2 = U^2Z^2 = Z^2 = Y^2$ with probability one. That is, the joint distribution of (X, Y) *lives on just the two lines* $y = \pm x$, and so is certainly not bivariate normal.

1.19 Conditional Densities and Expectations

The conditional distribution for continuous random variables is defined analogously to the discrete case, with pmfs replaced by densities. The formal definitions are as follows.

Definition 1.53. (Conditional Density) Let (X, Y) have a joint density $f(x, y)$. The *conditional density* of X given $Y = y$ is defined as

$$f(x|y) = f(x|Y = y) = \frac{f(x, y)}{f_Y(y)}, \ \forall y \text{ such that } f_Y(y) > 0.$$

The *conditional expectation* of X given $Y = y$ is defined as

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

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

$\forall y$ such that $f_Y(y) > 0$.

For fixed x, the conditional expectation $E(X|y) = \mu_X(y)$ is a number. As we vary y, we can think of $E(X|y)$ as a function of y. The corresponding function of Y is written as $E(X|Y)$ and is a random variable. It is very important to keep this notational distinction in mind.

The conditional density of Y given $X = x$ and the conditional expectation of Y given $X = x$ are defined analogously. That is, for instance,

$$f(y|x) = \frac{f(x,y)}{f_X(x)}, \ \forall x \text{ such that } f_X(x) > 0.$$

An important relationship connecting the two conditional densities is the following result.

Theorem 1.60. (Bayes' Theorem for Conditional Densities) Let (X, Y) have a joint density $f(x, y)$. Then, $\forall x, y,$ such that $f_X(x) > 0, f_Y(y) > 0,$

$$f(y|x) = \frac{f(x|y)f_Y(y)}{f_X(x)}.$$

Proof:

$$\frac{f(x|y)f_Y(y)}{f_X(x)} = \frac{\frac{f(x,y)}{f_Y(y)}f_Y(y)}{f_X(x)}$$

$$= \frac{f(x,y)}{f_X(x)} = f(y|x).$$

Thus, we can convert one conditional density to the other one by using Bayes' theorem; note the similarity to Bayes' theorem for two events A, B.

Definition 1.54. (Conditional Variance). Let (X, Y) have a joint density $f(x, y)$. The *conditional variance* of X given $Y = y$ is defined as

$$\text{Var}(X|y) = \text{Var}(X|Y = y) = \frac{\int_{-\infty}^{\infty}(x - \mu_X(y))^2 f(x,y)dx}{\int_{-\infty}^{\infty} f(x,y)dx},$$

$\forall y$ such that $f_Y(y) > 0$, where $\mu_X(y)$ denotes $E(X|y)$.

Remark: All the facts and properties about conditional pmfs and conditional expectations that were presented before for discrete random variables, continue to hold verbatim in the continuous case, with densities replacing the pmfs in their statements. In particular,

the iterated expectation and variance formula, and all the rules about conditional expectations and variance hold in the continuous case. An important optimizing property of the conditional expectation is that the best predictor of Y based on X among all possible predictors is the conditional expectation of Y given X. Here is the exact result.

Proposition (Best Predictor) Let (X, Y) be jointly distributed random variables (of any kind). Suppose $E(Y^2) < \infty$. Then $E_{X,Y}[(Y - E(Y|X))^2] \leq E_{X,Y}[(Y - g(X))^2]$, for any function $g(X)$. Here, the notation $E_{X<Y}$ stands for expectation with respect to the joint distribution of X, Y.

We will now see a number of examples.

Example 1.105. (Uniform in a Triangle). Consider the joint density

$$f(x, y) = 2, \text{ if } x, y \geq 0, x + y \leq 1.$$

By using the results derived in Example 3.2,

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{1}{1 - y},$$

if $0 \leq x \leq 1 - y$, and is zero otherwise. Thus, we have the interesting conclusion that given $Y = y, X$ is distributed uniformly in $[0, 1 - y]$. Consequently,

$$E(X|y) = \frac{1 - y}{2}, \ \forall y, 0 < y < 1.$$

Also, the conditional variance of X given $Y = y$ is, by the general variance formula for uniform distributions,

$$\text{Var}(X|y) = \frac{(1 - y)^2}{12}.$$

Example 1.106. (Uniform Distribution in a Circle). Let (X, Y) have a uniform density in the unit circle, $f(x, y) = \frac{1}{\pi}$, $x^2 + y^2 \leq 1$, We will find the conditional expectation of X given $Y = y$. First, the conditional density is

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-y^2}}{\pi}} = \frac{1}{2\sqrt{1 - y^2}},$$

$-\sqrt{1 - y^2} \leq x \leq \sqrt{1 - y^2}$.

Thus, we have the interesting result that the conditional density of X given $Y = y$ is uniform on $[-\sqrt{1 - y^2}, \sqrt{1 - y^2}]$. It being an interval symmetric about zero, we have in addition the result that for any $y, E(X|Y = y) = 0$.

Let us now find the conditional variance. Since the conditional distribution of X given $Y = y$ is uniform on $[-\sqrt{1 - y^2}, \sqrt{1 - y^2}]$, by the general variance formula for uniform distributions,

$$\text{Var}(X|y) = \frac{(2\sqrt{1 - y^2})^2}{12} = \frac{1 - y^2}{3}.$$

Thus, the conditional variance decreases as y moves away from zero, which makes sense intuitively, because as y moves away from zero, the line segment in which x varies becomes smaller.

Example 1.107. ($E(X|Y=y)$ exists for any y, but $E(X)$ does not). Consider the setup of the preceding example once again, i.e., $X \sim f(x)$, and given $X = x, Y \sim U[0,x]$. Suppose $f(x) = \frac{1}{x^2}, x \geq 1$. Then the marginal expectation $E(X)$ does not exist, because $\int_1^\infty x\frac{1}{x^2}dx = \int_1^\infty \frac{1}{x}dx$ diverges.
However, from the general formula in the preceding example,

$$E(X|Y=y) = \frac{1-F(y)}{\int_y^\infty \frac{f(x)}{x}dx} = \frac{\frac{1}{y}}{\frac{1}{2y^2}} = 2y,$$

and thus, $E(X|Y=y)$ exists for every y.

Example 1.108. (Using Conditioning to Evaluate Probabilities). We described the iterated expectation technique in the last chapter to calculate expectations. It turns out that it is in fact a really useful way also to calculate probabilities. The reason for it is that the probability of any event A is also the expectation of $X = I_A$, and so, by the iterated expectation technique, we can calculate $P(A)$ as

$$P(A) = E(I_A) = E(X) = E_Y[E(X|Y=y)] = E_Y[P(A|Y=y)],$$

by using a conditioning variable Y judiciously. The choice of the conditioning variable Y is usually clear from the particular context. Here is an example.

Example 1.109. Let X, Y be independent $U[0,1]$ random variables. Then $Z = XY$ also takes values in $[0,1]$, and suppose we want to find an expression for $P(Z \leq z)$. We can do this by using the iterated expectation technique:

$$\begin{aligned} P(XY \leq z) &= E[I_{XY\leq z}] = E_Y[E(I_{XY\leq z}|Y=y)] \\ &= E_Y[E(I_{Xy\leq z}|Y=y)] = E_Y[E(I_{X\leq \frac{z}{y}}|Y=y)] \\ &= E_Y[E(I_{X\leq \frac{z}{y}})] \end{aligned}$$

(because X and Y are independent)

$$= E_Y[P(X \leq \frac{z}{y})].$$

Now, note that $P(X \leq \frac{z}{y})$ is $\frac{z}{y}$ if $\frac{z}{y} \leq 1 \Leftrightarrow y \geq z$, and $P(X \leq \frac{z}{y}) = 1$ if $y < z$. Therefore,

$$\begin{aligned} E_Y[P(X \leq \frac{z}{y})] &= \int_0^z 1dy + \int_z^1 \frac{z}{y}dy \\ &= z - z\log z, \end{aligned}$$

$0 < z \leq 1$. So, the final answer to our problem is $P(XY \leq z) = z - z\log z, 0 < z \leq 1$.

Example 1.110. (A Two Stage Experiment). Suppose X is a positive random variable with density $f(x)$, and given $X = x$, a number Y is chosen at random between 0 and x. Suppose, however, that you are *only* told the value of Y, and the x value is kept hidden from you. What is your guess for x?
The formulation of the problem is:

$$X \sim f(x); Y|X = x \sim U[0, x]; \text{ we want to find } E(X|Y = y).$$

To find $E(X|Y = y)$, our first task would be to find $f(x|y)$, the conditional density of X given $Y = y$. This is, by its definition,

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(y|x)f(x)}{f_Y(y)}$$

$$= \frac{\frac{1}{x} I_{\{x \geq y\}} f(x)}{\int_y^\infty \frac{1}{x} f(x)dx}.$$

Therefore,

$$E(X|Y = y) = \int_y^\infty x f(x|y)dx$$

$$= \frac{\int_y^\infty x \frac{1}{x} f(x)dx}{\int_y^\infty \frac{1}{x} f(x)dx} = \frac{1 - F(y)}{\int_y^\infty \frac{1}{x} f(x)dx},$$

where F denotes the CDF of X.

1.19.1 Bivariate Normal Conditional Distributions

Suppose (X, Y) have a joint bivariate normal distribution. A very important property of the bivariate normal is that *each conditional distribution*, the distribution of Y given $X = x$, and that of X given $Y = y$ is a univariate normal, for any x, and any y. This really helps in easily computing conditional probabilities involving one variable, when the other variable is held fixed at some specific value.

Theorem 1.61. Let (X, Y) have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Then,

$$(a)\, X|Y = y \sim N(\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2));$$

$$(b)\, Y|X = x \sim N(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)).$$

In particular, the conditional expectations of X given $Y = y$ and of Y given $X = x$ are linear functions of y and x respectively:

$$E(X|Y = y) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2);$$

$$E(Y|X = x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1),$$

and the variance of each conditional distribution is a constant, and does not depend on the conditioning values x or y.

Remark: We see here that the conditional expectation is linear in the bivariate normal case. Specifically, take $E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$. Previously, we have seen in Section 3.15 that the conditional expectation $E(Y|X)$ is, in general, the best predictor of Y based on X. Now we see that the conditional expectation is a linear predictor in the bivariate normal case, and it is the best predictor and therefore, also the best linear predictor. We thus have the very pretty result that in the bivariate normal case, the best linear predictor and the best overall predictor are the same.

Example 1.111. Suppose incomes of husbands and wives in a population are bivariate normal with means 75 and 60 (in thousands of dollars), standard deviations 20 each, and a correlation of .75. We want to know in what percentage of those families where the wife earns 80,000 dollars, the family income exceeds 175,000 dollars.

Denote the income of the husband and the wife by X and Y. Then, we want to find $P(X+Y > 175|Y=80)$. By the above theorem. $X|Y=80 \sim N(75+.75(80-60), 400(1-.75^2)) = N(90, 175)$. Therefore,

$$\begin{aligned} P(X+Y>175|Y=80) &= P(X>95|Y=80) \\ &= P(Z > \frac{95-90}{\sqrt{175}}) = P(Z>.38) \\ &= .3520, \end{aligned}$$

where Z denotes a standard normal variable.

Example 1.112. (Galton's Observation: Regression to the Mean). This example is similar to the previous example, but makes an interesting different point. It is often found that students who get a very good grade on the first midterm, do not do as well on the second midterm. We can try to explain it by doing a bivariate normal calculation.

Denote the grade on the first midterm by X, that on the second midterm by Y, and suppose X, Y are jointly bivariate normal with means 70, standard deviations 10, and a correlation .7. Suppose a student scored 90 on the first midterm. What are the chances that he will get a lower grade on the second midterm?

This is

$$\begin{aligned} P(Y<X|X=90) &= P(Y<90|X=90) \\ &= P(Z < \frac{90-84}{\sqrt{51}}) = P(Z<.84) \\ &= .7995, \end{aligned}$$

where Z is a standard normal variable, and we have used the fact that $Y|X=90 \sim N(70+.7(90-70), 100(1-.7^2)) = N(84, 51)$.

Among the numerous characterizations of the bivariate normal distribution, there is one that stands out in its elegance and in being useful. We state this characterization below.

Theorem 1.62. (Characterization) Let X, Y be jointly distributed random variables. X, Y are jointly bivariate normal if and only if every linear combination $aX + bY$ has a univariate normal distribution.

1.19.2 Computing Bivariate Normal Probabilities

There are many approximations to the CDF of a general bivariate normal distribution. The most accurate ones are too complex for quick use. The relatively simple approximations are not computationally accurate for all configurations of the arguments and the parameters. Keeping a balance between simplicity and accuracy, we present here two approximations.

Mee-Owen Approximation Let (X, Y) have the general five parameter bivariate normal distribution. Then,

$$P(X \leq \mu_1 + h\sigma_1, Y \leq \mu_2 + k\sigma_2) \approx \Phi(h)\Phi(\frac{k - c}{\tau}),$$

where $c = -\rho\frac{\phi(h)}{\Phi(h)}, \tau^2 = 1 + \rho hc - c^2$.

Cox-Wermuth Approximation

$$P(X \geq \mu_1 + h\sigma_1, Y \geq \mu_2 + k\sigma_2) \approx \Phi(-h)\Phi(\frac{\rho\mu(h) - k}{\sqrt{1 - \rho^2}}),$$

where $\mu(h) = \frac{\phi(h)}{1-\Phi(h)}$.

1.20 Convolutions in General

Definition 1.55. Let X, Y be independent random variables. The distribution of their sum $X + Y$ is called the *convolution* of X and Y.

Remark: Usually, we study convolutions of two continuous or two discrete random variables. But, in principle, one could be continuous and the other discrete.

Example 1.113. Suppose X, Y have a joint density function $f(x, y)$, and suppose we want to find the density of their sum, namely $X + Y$. Denote the conditional density of X given $Y = y$ by $f_{X|y}(x|y)$ and the conditional CDF, namely, $P(X \leq u|Y = y)$ by $F_{X|Y}(u)$. Then, by the iterated expectation formula,

$$P(X + Y \leq z) = E[I_{X+Y\leq z}]$$

$$= E_Y[E(I_{X+Y\leq z}|Y = y)] = E_Y[E(I_{X+y\leq z}|Y = y)]$$

$$= E_Y[P(X \le z - y|Y = y)]$$
$$= E_Y[F_{X|Y}(z - y)] = \int_{-\infty}^{\infty} F_{X|Y}(z - y) f_Y(y) dy.$$

In particular, if X and Y are independent, then the conditional CDF $F_{X|Y}(u)$ will be the same as the marginal CDF $F_X(u)$ of X. In this case, the expression above simplifies to

$$P(X + Y \le z) = \int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy.$$

The density of $X + Y$ can be obtained by differentiating the CDF of $X + Y$:

$$f_{X+Y}(z) = \frac{d}{dz} P(X + Y \le z)$$
$$= \frac{d}{dz} \int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy$$
$$= \int_{-\infty}^{\infty} [\frac{d}{dz} F_X(z - y) f_Y(y)] dy$$

(assuming that the derivative can be carried inside the integral)

$$= \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy.$$

Indeed, this is the general formula for the density of the sum of two real valued independent continuous random variables.

Theorem 1.63. Let X, Y be independent real valued random variables with densities $f_X(x), f_Y(y)$ respectively. Let $Z = X + Y$ be the sum of X and Y. Then, the density of the convolution is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy.$$

More generally, if X, Y are not necessarily independent, and have joint density $f(x, y)$, then $Z = X + Y$ has the density

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X|Y}(z - y) f_Y(y) dy.$$

Definition 1.56. If X, Y are independent continuous random variables with a common density $f(x)$, then the density of the convolution is denoted as $f * f$. In general, if $X_1, X_2, \cdots, X_n$ are n independent continuous random variables with a common density $f(x)$, then the density of their sum $X_1 + X_2 + \cdots + X_n$ is called the *n- fold convolution of* f and is denoted as $f^{*(n)}$.

Example 1.114. (Sum of Exponentials). Suppose X, Y are independent $Exp(\lambda)$ variables, and we want to find the density of $Z = X + Y$. By the convolution formula, for $z > 0$,

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\lambda} e^{-\frac{z-y}{\lambda}} I_{y<z} \frac{1}{\lambda} e^{-\frac{y}{\lambda}} I_{y>0} dy$$

$$= \frac{1}{\lambda^2}\int_0^z e^{-\frac{z}{\lambda}}dy$$
$$= \frac{ze^{-\frac{z}{\lambda}}}{\lambda^2},$$

which is the density of a Gamma distribution with parameters 2 and λ. Note that it would be possible to prove this also by using mgfs, but the proof here is a direct proof.

Example 1.115. (Difference of Exponentials). Let U, V be independent standard Exponentials. We want to find the density of $Z = U - V$. Writing $X = U$, and $Y = -V$, we notice that $Z = X + Y$, and X, Y are still independent. However, now Y is a *negative Exponential*, and so has density $f_Y(y) = e^y Iy < 0$. It is also important to note that Z can now take any real value, positive or negative. Substituting into the formula for the convolution density,

$$f_Z(z) = \int_{-\infty}^{\infty} e^{-(z-y)}(Iy < z)e^y(Iy < 0)dy.$$

Now, first consider $z > 0$. Then this last expression becomes

$$f_Z(z) = \int_{-\infty}^{0} e^{-(z-y)}e^y dy = e^{-z}\int_{-\infty}^{0} e^{2y}dy$$
$$= \frac{1}{2}e^{-z}.$$

On the other hand, for $z < 0$, the convolution formula becomes

$$f_Z(z) = \int_{-\infty}^{z} e^{-(z-y)}e^y dy = e^{-z}\int_{-\infty}^{z} e^{2y}dy$$
$$= e^{-z}\frac{1}{2}e^{2z} = \frac{1}{2}e^z.$$

Combining the two cases, we can write the single formula

$$f_Z(z) = \frac{1}{2}e^{-|z|}, -\infty < z < \infty,$$

i.e., if X, Y are independent standard Exponentials, then the difference $X - Y$ has a standard double exponential density. This representation of the double exponential is often useful. Also note that although the standard exponential distribution is obviously not symmetric, the distribution of the difference of two independent exponentials is symmetric. This is a useful technique for symmetrizing a random variable.

Definition 1.57. (Symmetrization of a Random Variable) Let X_1, X_2 be independent random variables with a common distribution F. Then $X_s = X_1 - X_2$ is called *the symmetrization of F or symmetrization of X_1.*

If X_1 is a continuous random variable with density $f(x)$, then its symmetrization has the density

$$f_s(z) = \int_{-\infty}^{\infty} f(z + y)f(y)dy.$$

Example 1.116. (Sums of Cauchy Variables). Let X, Y be independent standard Cauchy random variables with the common density function $f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$. Then, the density of the convolution is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \int_{-\infty}^{\infty} f(z-y) f(y) dy$$

$$= \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{(1+(z-y)^2)(1+y^2)} dy$$

$$= \frac{1}{\pi^2} \frac{2\pi}{4+z^2}$$

(on a partial fraction expansion of $\frac{1}{(1+(z-y)^2)(1+y^2)}$)

$$= \frac{2}{\pi(4+z^2)}.$$

Therefore, the density of $W = \frac{Z}{2} = \frac{X+Y}{2}$ would be $\frac{1}{\pi(1+w^2)}$, which is, remarkably, the same standard Cauchy density that we had started with. *By using more advanced techniques, it can be shown that if $X_1, X_2, \cdots, X_n$ are independent standard Cauchy variables, then for any $n \geq 2$, their average $\bar{X}$ also has the standard Cauchy distribution.*

Example 1.117. (Normal-Poisson Convolution). Here is an example of the convolution of one continuous and one discrete random variable. Let $X \sim N(0,1)$ and $Y \sim Poi(\lambda)$. Then their sum $Z = X + Y$ is still continuous, and has the density

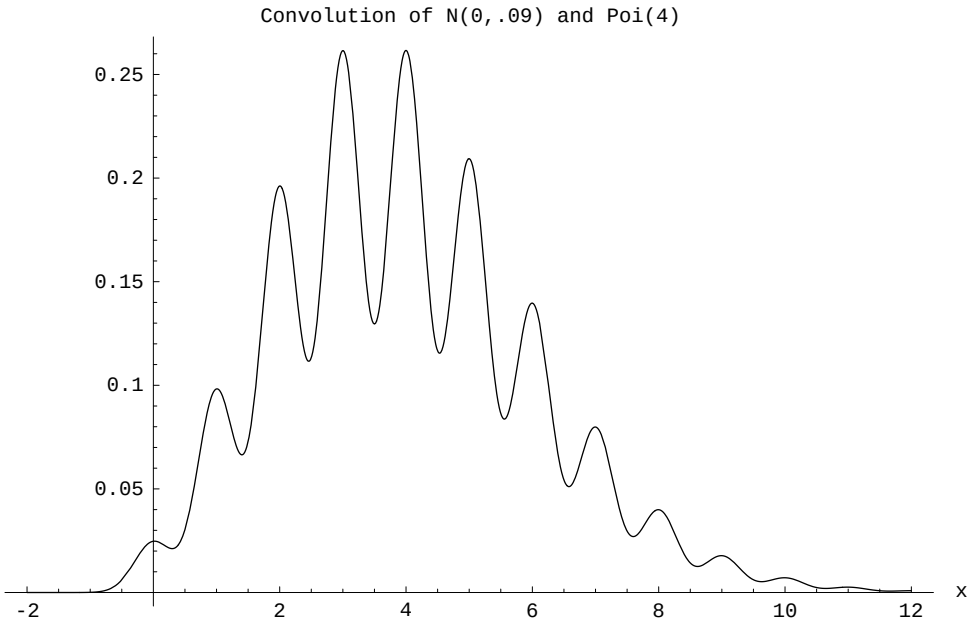
$$f_Z(z) = \sum_{y=0}^{\infty} \phi(z-y) \frac{e^{-\lambda} \lambda^y}{y!}$$

More generally, if $X \sim N(0, \sigma^2)$, and $Y \sim Poi(\lambda)$. Then the density of the sum is

$$f_Z(z) = \frac{1}{\sigma} \sum_{y=0}^{\infty} \phi(\frac{z-y}{\sigma}) \frac{e^{-\lambda} \lambda^y}{y!}$$

This is not expressible in terms of the elementary functions. However, it is interesting to plot the density. The plot below shows an unconventional density for the convolution with multiple local maxima and shoulders.

For purposes of summary and easy reference, we list some convolutions of common types below.



Distribution of Summands	Distribution of Sum
$X_i \sim Bin(n_i, p)$	$Bin(\sum n_i, p)$
$X_i \sim Poi(\lambda_i)$	$Poi(\sum \lambda_i)$
$X_i \sim NB(r_i, p)$	$NB(\sum r_i, p)$
$X_i \sim Exp(\lambda)$	$Gamma(n, \lambda)$
$X_i \sim N(\mu_i, \sigma_i^2)$	$N(\sum \mu_i, \sum \sigma_i^2)$
$X_i \sim C(\mu_i, \sigma_i^2)$	$C(\sum \mu_i, (\sum \sigma_i)^2)$
$X_i \sim U[a, b]$	See Chapter 12

1.21 Products and Quotients and the t and F Distribution

Suppose X, Y are two random variables. Then two other functions that arise naturally in many applications are the product XY, and the quotient $\frac{X}{Y}$. Following exactly the same technique as for convolutions, one can find the density of each of XY and $\frac{X}{Y}$. More precisely, one first finds the CDF by using the iterated expectation technique, exactly as we did for convolutions, and then differentiates the CDF to obtain the density. Here are the density formulas; they are extremely important and useful.

Theorem 1.64. Let X, Y be continuous random variables with a joint density $f(x, y)$. Let $U = XY, V = \frac{X}{Y}$. Then the densities of U, V are given by

$$f_U(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x, \frac{u}{x}) dx;$$

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy.$$

Example 1.118. (Ratio of Standard Normals). The distribution of the ratio of two independent standard normal variables is an interesting one; we show now that it is in fact a standard Cauchy distribution. Indeed, by applying the general formula, the density of the quotient $V = \frac{X}{Y}$ is

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} |y| f(vy, y) dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |y| e^{-\frac{y^2}{2}(1+v^2)} dy \\ &= \frac{1}{\pi} \int_0^{\infty} y e^{-\frac{y^2}{2}(1+v^2)} dy \\ &= \frac{1}{\pi(1+v^2)}, -\infty < v < \infty, \end{aligned}$$

by making the substitution $t = \sqrt{1+v^2}y$ in the integral on the last line. This proves that the quotient has a standard Cauchy distribution.

It is important to note that zero means for the normal variables are essential for this result. If either X or Y has a nonzero mean, the quotient has a complicated distribution, and is definitely not Cauchy.

Example 1.119. (The F Distribution). Let $X \sim G(\alpha, 1), Y \sim G(\beta, 1)$, and suppose X, Y are independent. The distribution of the ratio $R = \frac{X/\alpha}{Y/\beta}$ arises in statistics in many contexts and is called *an F distribution*. We derive the explicit form of the density here.

First, we will find the density of $\frac{X}{Y}$, from which the density of $R = \frac{X/\alpha}{Y/\beta}$ will follow easily. Again, by applying the general formula for the density of a quotient, the density of the quotient $V = \frac{X}{Y}$ is

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} |y| f(vy, y) dy \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{\infty} y e^{-y(1+v)} (vy)^{\alpha-1} y^{\beta-1} dy \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1} \int_0^{\infty} e^{-y(1+v)} y^{\alpha+\beta-1} dy \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1} \frac{\Gamma(\alpha+\beta)}{(1+v)^{\alpha+\beta}} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{v^{\alpha-1}}{(1+v)^{\alpha+\beta}}, 0 < v < \infty. \end{aligned}$$

To complete the example, notice now that $R = cV$, where $c = \frac{\beta}{\alpha}$. Therefore, the density of R is immediately obtained from the density of V. Indeed,

$$f_R(r) = \frac{1}{c} f_V(\frac{r}{c}),$$

where f_V is the function we just derived above. If we simplify $f_R(r)$, we get the final expression

$$f_R(r) = \frac{(\frac{\beta}{\alpha})^{\beta} r^{\alpha-1}}{B(\alpha,\beta)(r+\frac{\beta}{\alpha})^{\alpha+\beta}}, r > 0.$$

This is the F-density with parameters α, β; it is common in statistics to refer to 2α and 2β as the degrees of freedom of the distribution.

Example 1.120. (The Student t Distribution). Once again, the t distribution is one that arises frequently in statistics. Suppose, $X \sim N(0,1), Z \sim \chi^2_m$, and suppose X, Z are independent. Let $Y = \sqrt{\frac{Z}{m}}$. Then the distribution of the quotient $V = \frac{X}{Y}$ is called the *the t distribution with m degrees of freedom*. We derive its density in this example.

Recall that Z has the density $\frac{e^{-z/2} z^{m/2-1}}{2^{m/2}\Gamma(\frac{m}{2})}, z > 0$. Therefore, $Y = \sqrt{\frac{Z}{m}}$ has the density

$$f_Y(y) = \frac{m^{m/2} e^{-my^2/2} y^{m-1}}{2^{m/2\,-1}\Gamma(\frac{m}{2})}, y > 0.$$

Since, by hypothesis, X and Z are independent, it follows that X and Y are also independent, and so their joint density $f(x,y)$ is just the product of the marginal densities of X and Y.

Once again, by applying our general formula for the density of a quotient,

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy$$

$$= \frac{m^{m/2}}{\sqrt{2\pi} 2^{m/2\,-1}\Gamma(\frac{m}{2})} \int_0^{\infty} y e^{-v^2y^2/2} e^{-my^2/2} y^{m-1} dy$$

$$= \frac{m^{m/2}}{\sqrt{2\pi} 2^{m/2\,-1}\Gamma(\frac{m}{2})} \int_0^{\infty} e^{-(v^2+m)y^2/2} y^m dy$$

$$= \frac{m^{m/2}}{\sqrt{2\pi} 2^{m/2\,-1}\Gamma(\frac{m}{2})} \times \frac{\Gamma(\frac{m+1}{2}) 2^{(m-1)/2}}{(m+v^2)^{(m+1)/2}}$$

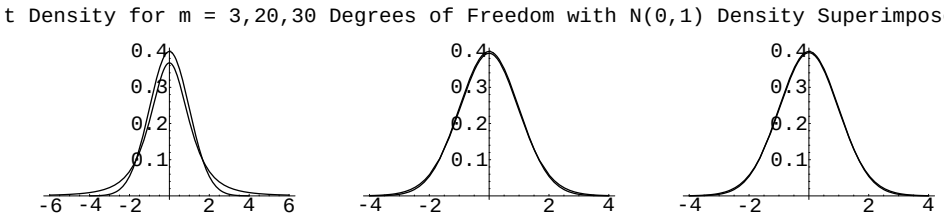
$$= \frac{m^{m/2}\Gamma(\frac{m+1}{2})}{\sqrt{\pi}\Gamma(\frac{m}{2})} \frac{1}{(m+v^2)^{(m+1)/2}}$$

$$= \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})(1+\frac{v^2}{m})^{(m+1)/2}}, -\infty < v < \infty.$$

This is the density of *the Student t distribution with m degrees of freedom*.

Note that when the degree of freedom $m = 1$, this becomes just the standard Cauchy density. The t distribution was first derived in 1908 by William Gossett under the pseudonym *Student*. The distribution was later named the *Student t distribution* by Ronald Fisher.

The t density, just like the standard normal, is symmetric and unimodal around zero,



although with tails much heavier than that of the standard normal for small values of m. However, as $m \to \infty$, the density converges pointwise to the standard normal density, and then the t and the standard normal density look almost the same. We give a plot of the t density for a few degrees of freedom, and of the standard normal density for a visual comparison.

1.22 Transformations

The simple technique that we used in the previous section to derive the density of a sum or a product does not extend to functions of a more complex nature. Consider the simple case of just two continuous variables X, Y with some joint density $f(x, y)$, and suppose we want to find the density of some function $U = g(X, Y)$. Then, the general technique is to pair up U with *another* function $V = h(X, Y)$, and first obtain the joint CDF of (U, V) from the joint CDF of (X, Y). The pairing up has to be done carefully; i.e., only some judicious choices of V will work in a given example. Having found the joint CDF of (U, V), by differentiation one finds the joint density of (U, V), and then finally integrates v out to obtain the density of just U. Fortunately, this agenda does work out, because the transformation from (X, Y) to (U, V) can be treated as just a change of variable in manipulation with double integrals, and calculus tells us how to find double integrals by making suitable changes of variables (i.e., substitutions). Indeed, the method works out for *any* number of jointly distributed variables, $X_1, X_2, \cdots, X_n$, and a function $U = g(X_1, X_2, \cdots, X_n)$, and the reason it works out is that the whole method is just a change of variables in manipulating a multivariate integral.

Here is the theorem on density of a multivariate transformation, a major theorem in multivariate distribution theory. It is really nothing but the change of variable theorem of multivariate calculus.

Theorem 1.65. (Multivariate Jacobian Formula) Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ have the joint density function $f(x_1, x_2, \cdots, x_n)$, such that there is an open set $S \subseteq \mathcal{R}^n$ with $P(\mathbf{X} \in S) = 1$. Suppose $u_i = g_i(x_1, x_2, \cdots, x_n), 1 \leq i \leq n$ are n real valued functions of $x_1, x_2, \cdots, x_n$ such that

(a) $(x_1, x_2, \cdots, x_n) \to (g_1(x_1, x_2, \cdots, x_n), \cdots, g_n(x_1, x_2, \cdots, x_n))$ is a one-to-one function of $(x_1, x_2, \cdots, x_n)$ on S with range space T;

(b) The inverse functions $x_i = h_i(u_1, u_2, \cdots, u_n), 1 \leq i \leq n,$ are continuously differentiable on T with respect to each u_j;

(c) The Jacobian determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix}$$

is nonzero.

Then the joint density of $(U_1, U_2, \cdots, U_n)$ is given by

$$f_{U_1,U_2,\cdots,U_n}(u_1, u_2, \cdots, u_n) = f(h_1(u_1, u_2, \cdots, u_n), \cdots, h_n(u_1, u_2, \cdots, u_n)) \times |J|,$$

where $|J|$ denotes the absolute value of the Jacobian determinant J, and the notation f on the right hand side means the original joint density of $(X_1, X_2, \cdots, X_n)$.

1.22.1 Applications of Jacobian Formula

We will now see a number of examples of its applications to finding the density of interesting transformations. We emphasize that quite often *only one of the functions* $u_i = g_i(x_1, x_2, \cdots, x_n)$ *will be provided, whose density function we want to find. But unless that function is a really simple one, its density cannot be found directly, without invoking the Jacobian theorem given here. It is necessaary to make up the remaining* $(n-1)$ *functions, and then obtain their joint density by using this Jacobian theorem. Finally, one would integrate out all these other coordinates to get the density function of just* u_i. *The other* $(n-1)$ *functions need to be found judiciously.*

Example 1.121. (A Relation Between Exponential and Uniform). Let X, Y be independent standard exponentials, and define $U = \frac{X}{X+Y}$. We want to find the density of U. We have to pair it up with another function V in order to use the Jacobian theorem. We choose $V = X + Y$. We have here a one-to-one function for $x > 0, y > 0$. Indeed, the inverse functions are

$$x = x(u, v) = uv; y = y(u, v) = v - uv = v(1 - u).$$

The partial derivatives of the inverse functions are

$$\frac{\partial x}{\partial u} = v; \frac{\partial x}{\partial v} = u; \frac{\partial y}{\partial u} = -v; \frac{\partial y}{\partial v} = 1 - u.$$

Thus, the Jacobian determinant equals $J = v(1 - u) + uv = v$. By invoking the Jacobian theorem, the joint density of U, V is

$$f_{U,V}(u, v) = e^{-uv}e^{-v(1-u)}|v| = ve^{-v},$$

$0 < u < 1, v > 0$.

Thus, the joint density of U, V has factorized into a product form on a rectangle; the marginals are

$$f_U(u) = 1, 0 < u < 1; f_V(v) = ve^{-v}, v > 0,$$

and the rectangle being $(0,1) \times (0,\infty)$. Therefore, we have proved that if X, Y are independent standard exponentials, then $\frac{X}{X+Y}$ and $X + Y$ are independent,and they are respectively uniform and a Gamma.

Example 1.122. (A Relation Between Gamma and Beta). The previous example generalizes in a nice way. Let X, Y be independent variables, distributed respectively as $G(\alpha, 1), G(\beta, 1)$. Let again, $U = \frac{X}{X+Y}, V = X + Y$. Then, from our previous example, the Jacobian determinant is still $J = v$. Therefore, the joint density of U, V is

$$f_{U,V}(u,v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} e^{-v}(uv)^{\alpha-1}(v(1-u))^{\beta-1}v$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}e^{-v}v^{\alpha+\beta-1},$$

$0 < u < 1, v > 0$.

Once again, we have factorized the joint density of U and V as the product of the marginal densities, with (U, V) varying in the rectangle $(0,1) \times (0,\infty)$, the marginal densities being

$$f_V(v) = \frac{e^{-v}v^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)}, v > 0,$$

$$f_U(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, 0 < u < 1.$$

That is, if X, Y are independent Gamma variables, then $\frac{X}{X+Y}$ and $X+Y$ are independent, and they are respectively distributed as a Beta and a Gamma. Of course, we already knew from an mgf argument that $X + Y$ is a Gamma.

Example 1.123. (Product of n Uniforms). Let $X_1, X_2, \cdots, X_n$ be independent $U[0,1]$ variables, and suppose we want to find the density of the product $U = U_n = \prod_{i=1}^n X_i$. According to our general discussion, we have to choose $n-1$ other functions, and then apply the Jacobian theorem. Define

$$u_1 = x_1, u_2 = x_1x_2, u_3 = x_1x_2x_3, \cdots, u_n = x_1x_2\cdots x_n.$$

This is a one-to-one transformation, and the inverse functions are $x_i = \frac{u_i}{u_{i-1}}, 2 \le i \le n; x_1 = u_1$. Thus, the Jacobian matrix of the partial derivatives is lower trianglular, and therefore the Jacobian determinant equals the product of the diagonal elements

$$J = \prod_{i=1}^n \frac{\partial x_i}{\partial u_i} = \frac{1}{\prod_{i=1}^{n-1} u_i}.$$

Now applying the Jacobian density theorem, the joint density of $U_1, U_2, \cdots, U_n$ is

$$f_{U_1,U_2,\cdots,U_n}(u_1, u_2, \cdots, u_n) = \frac{1}{\prod_{i=1}^{n-1} u_i},$$

$0 < u_n < u_{n-1} < \cdots < u_1 < 1$.

On integrating out $u_1, u_2, \cdots, u_{n-1}$, we get the density of U_n:

$$f_{U_n}(u) = \int_u^1 \int_{u_{n-1}}^1 \cdots \int_{u_2}^1 \frac{1}{u_1 u_2 \cdots u_{n-1}} du_1 du_2 \cdots du_{n-2} du_{n-1}$$

$$= \frac{|(\log u)^{n-1}|}{(n-1)!},$$

$0 < u < 1$. This example illustrates that applying the Jacobian theorem needs careful manipulation with multiple integrals, and skills in using the Jacobian technique are very important in deriving distributions of functions of many variables.

Example 1.124. (Transformation to Polar Coordinates). We have already worked out an example where we transformed two variables to their polar coordinates, in order to calculate expectations of suitable functions, when the variables have a spherically symmetric density. We now use a transformation to polar coordinates to do distributional calculations. *In any spherically symmetric situation, transformation to polar coordinates is a technically useful device, and gets the answers out quickly for many problems.*

Let (X, Y) have a spherically symmetric joint density given by $f(x, y) = g(\sqrt{x^2 + y^2})$. Consider the polar transformation $r = \sqrt{X^2 + Y^2}, \theta = \arctan(\frac{Y}{X})$. This is a one-to-one transformation, with the inverse functions given by

$$x = r\cos\theta, y = r\sin\theta.$$

The partial derivatives of the inverse functions are

$$\frac{\partial x}{\partial r} = \cos\theta, \frac{\partial x}{\partial \theta} = -r\sin\theta, \frac{\partial y}{\partial r} = \sin\theta, \frac{\partial y}{\partial \theta} = r\cos\theta.$$

Therefore, the Jacobian determinant is $J = r\cos^2\theta + r\sin^2\theta = r$. By the Jacobian theorem, the density of $(r, \theta)$is

$$f_{r,\theta}(r, \theta) = rg(r),$$

with r, θ belonging to a suitable rectangle, which will depend on the exact set of values (x, y) on which the original joint density $f(x, y)$ is strictly positive. But, in any case, we have established that the joint density of (r, θ) factorizes into the product form on a rectangle, and so *in any spherically symmetric situation, the polar coordinates r and θ are independent, a very convenient fact.* Always, in a spherically symmetric case, r will have the density $crg(r)$ on some interval and for some suitable normalizing constant c, and θ

will have a uniform density on some interval.
Consider now the specific case of two independent standard normals. Indeed, in this case, the joint density is *spherically symmetric*, namely,

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

Thus, $g(r) = \frac{1}{2\pi} e^{-r^2/2}, r > 0$. Therefore, in this case r has the *Weibull density* $re^{-r^2/2}, r > 0$, and θ is uniform on $(-\pi, \pi)$.

1.22.2 n-Dimensional Polar Coordinates

We saw evidence of practical advantages of transforming to polar coordinates in two dimensions in the previous section. As a matter of fact, in any spherically symmetric situation in any number of dimensions, transformation to the n *dimensional polar coordinates* is a standard technical device. The transformation from rectangular to the polar coordinates often greatly simplifies the algebraic complexity of the calculations. We first present the n dimensional polar transformation in this section.

Definition 1.58. For $n \geq 3$, the n dimensional polar transformation is a one-to-one mapping from $\mathcal{R}^n \to [0,\infty) \times \prod_{i=1}^{n-2} \Theta_i \times \Theta_{n-1}$, where $\Theta_{n-1} = [0, 2\pi]$, and for $i \leq n-2, \Theta_i = [0, \pi]$, with the mapping defined by

$$x_1 = \rho \cos\theta_1,$$

$$x_2 = \rho \sin\theta_1 \cos\theta_2,$$

$$x_3 = \rho \sin\theta_1 \sin\theta_2 \cos\theta_3,$$

$$\vdots$$

$$x_{n-1} = \rho \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2} \cos\theta_{n-1},$$

$$x_n = \rho \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2} \sin\theta_{n-1}.$$

The transformation has the useful property that $x_1^2 + x_2^2 + \cdots + x_n^2 = \rho^2 \ \forall\, (x_1, x_2, \cdots, x_n) \in \mathcal{R}^n$, i.e., ρ is the length of the vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. The Jacobian determinant of the transformation equals

$$J = \rho^{n-1} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}.$$

Consequently, by the Jacobian density theorem, we have the following result.

Theorem 1.66. (Joint Density of Polar Coordinates) Let $X_1, X_2, \cdots, X_n$ be n continuous random variables with a joint density $f(x_1, x_2, \cdots, x_n)$. Then the joint density of $(\rho, \theta_1, \theta_2, \cdots, \theta_{n-1})$ is given by

$$p(\rho, \theta_1, \theta_2, \cdots, \theta_{n-1}) = \rho^{n-1} f(x_1, x_2, \cdots, x_n) |\sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}|,$$

where in the right side, one writes for $x_1, x_2, \cdots, x_n$, their defining expressions in terms of $\rho, \theta_1, \theta_2, \cdots, \theta_{n-1}$, as provided above.

In particular, if $X_1, X_2, \cdots, X_n$ have a spherically symmetric joint density

$$f(x_1, x_2, \cdots, x_n) = g(\sqrt{x_1^2 + +x_2^2 + \cdots + x_n^2})$$

for some function g, then the joint density of $(\rho, \theta_1, \theta_2, \cdots, \theta_{n-1})$ equals

$$p(\rho, \theta_1, \theta_2, \cdots, \theta_{n-1}) = \rho^{n-1} g(\rho) |\sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}|,$$

and ρ is distributed independently of the angles $(\theta_1, \theta_2, \cdots, \theta_{n-1})$.

Example 1.125. (Chi Square Distribution from Polar Coordinates). Suppose that $X_1, X_2, \cdots, X_n$ are independent standard normals, so that their joint density is $f(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}, -\infty < x_i < \infty, i = 1, 2, \cdots, n$. Thus, the joint density is spherically symmetric with

$$f(x_1, x_2, \cdots, x_n) = g(\sqrt{x_1^2 + +x_2^2 + \cdots + x_n^2}),$$

where $g(\rho) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\rho^2}{2}}$. Therefore, from our general theorem above, $\rho = \sqrt{\sum_{i=1}^n X_i^2}$ has the density $c\rho^{n-1}e^{-\frac{\rho^2}{2}}$ for some normalizing constant c. Making the transformation $W = \rho^2$, we get from the general formula for the density of a monotone transformation in one dimension that W has the density

$$f_W(w) = ke^{-w/2} w^{n/2\ -1}, w > 0.$$

It follows that $W = \rho^2 = \sum_{i=1}^n X_i^2$ has a χ_n^2 density, because the constant k must necessarily be $\frac{1}{2^{n/2}\Gamma(\frac{n}{2})}$, which makes $f_W(w)$ exactly equal to the χ_n^2 density. Note that this can be proved also by using mgfs; but now we have a polar transformation proof of it.

1.22.3 General Spherically Symmetric Facts

Transformation to polar coordinates also results in some striking formulas and properties for general spherically symmetric distributions. They are collected together in the following theorem.

Theorem 1.67. (General Spherically Symmetric Facts) Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ have a spherically symmetric joint density $f(x_1, x_2, \cdots, x_n) = g(\sqrt{x_1^2 + \cdots + x_n^2})$. Then,

(a) For any $m < n$, the distribution of $(X_1, X_2, \cdots, X_m)$ is also spherically symmetric.
(b) $\rho = ||X||$ has the density $c\rho^{n-1}g(\rho)$, where c is the normalizing constant $\frac{1}{\int_0^\infty \rho^{n-1}g(\rho)d\rho}$.
(c) Let $\mathbf{U} = \frac{\mathbf{X}}{||X||}$, and $\rho = ||X||$. Then U and ρ are independent, and $\mathbf{U}$ is *distributed uniformly on the boundary of the n-dimensional unit sphere.*

(d) For any unit vector $\mathbf{c}, c_1X_1 + \cdots + c_nX_n$ has the same distribution as X_1.
(e) If $\mathbf{X}$ is uniformly distributed on the boundary of the n-dimensional unit sphere, then a lower dimensional projection $(X_1, X_2, \cdots, X_k)(k < n)$ has the density

$$f_k(x_1, x_2, \cdots, x_k) = c(1 - \sum_{i=1}^{k} x_i^2)^{(n-k)/2-1}, \sum_{i=1}^{k} x_i^2 < 1,$$

where the normalizing constant c equals

$$c = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n-k}{2})};$$

in particular, if $n = 3$, then each $|X_i| \sim U[0, 1]$, and each $X_i \sim U[-1, 1]$.

1.23 The Dirichlet Distribution

The Jacobian density formula, when suitably applied to a set of independent Gamma random variables, results in a hugely useful and important density for random variables in a *simplex*. In the plane, the standard simplex is the triangle with vertices at $(0, 0), (0, 1)$ and $(1, 0)$. In the general n dimensions, the standard simplex is the set of all n-dimensional vectors $\mathbf{x} = (x_1, \cdots, x_n)$ such that each $x_i \geq 0$, and $\sum_{i=1}^n x_i \leq 1$. If we define an additional x_{n+1} as $x_{n+1} = 1 - \sum_{i=1}^n x_i$, then $(x_1, \cdots, x_{n+1})$ forms a vector of proportions adding to one. Thus, the Dirichlet distribution can be used in any situation where an entity has to necessarily fall into one of $n + 1$ mutually exclusive subclasses, and we want to the study the proportion of individuals belonging to the different subclasses. Indeed, when statisticians want to model an ensemble of fractional variables adding to one, they often first look at the Dirichlet distribution as their model. Dirichlet distributions are also immensely important in Bayesian statistics. Fundamental work on the use of Dirichlet distributions in Bayesian modelling and on calculations using the Dirichlet distribution has been done in Ferguson (1973), Blackwell (1973), and Basu and Tiwari (1982).

Let $X_1, X_2, \cdots, X_{n+1}$ be independent Gamma random variables, with $X_i \sim G(\alpha_i, 1)$. Define $p_i = \frac{X_i}{\sum_{j=1}^{n+1} X_j}, 1 \leq i \leq n$, and denote $p_{n+1} = 1 - \sum_{i=1}^n p_i$. Then, we have the following theorem.

Theorem 1.68. $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ has the joint density

$$f(p_1, p_2, \cdots, p_n) = \frac{\Gamma(\sum_{i=1}^{n+1} \alpha_i)}{\prod_{i=1}^{n+1} \Gamma(\alpha_i)} \prod_{i=1}^{n+1} p_i^{\alpha_i - 1}.$$

Definition 1.59. (Dirichlet Density) An n-dimensional vector $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ is said to have *the Dirichlet distribution with parameter vector* $\alpha = (\alpha_1, \cdots, \alpha_{n+1}), \alpha_i > 0$, if it has the joint density

$$f(p_1, p_2, \cdots, p_n) = \frac{\Gamma(\sum_{i=1}^{n+1} \alpha_i)}{\prod_{i=1}^{n+1} \Gamma(\alpha_i)} \prod_{i=1}^{n+1} p_i^{\alpha_i - 1},$$

$p_i \geq 0, \sum_{i=1}^{n} p_i \leq 1$.

We write $\mathbf{p} \sim \mathcal{D}_n(\alpha)$.

Remark: When $n = 1$, the Dirichlet density reduces to a Beta density with parameters α_1, α_2.

Simple integrations give the following moment formulas.

Proposition Let $\mathbf{p} \sim \mathcal{D}_n(\alpha)$. Then,

$$E(p_i) = \frac{\alpha_i}{t},\ \mathrm{Var}(p_i) = \frac{\alpha_i(t - \alpha_i)}{t^2(t+1)}, \mathrm{Cov}(p_i, p_j) = -\frac{\alpha_i \alpha_j}{t^2(t+1)}, i \neq j,$$

where $t = \sum_{i=1}^{n+1} \alpha_i$.

Thus, notice that the covariances (and hence. the correlations) are always negative.

A convenient fact about the Dirichlet density is that lower dimensional marginals are also Dirichlet distributions. So are the conditional distributions of suitably renormalized subvectors given the rest.

Theorem 1.69. (Marginal and Conditional Distributions)

(a) Let $\mathbf{p} \sim \mathcal{D}_n(\alpha)$. Fix $m < n$, and let $\mathbf{p}_m = (p_1, \cdots, p_m)$, and $\alpha_m = (\alpha_1, \cdots, \alpha_m, t - \sum_{i=1}^{m} \alpha_i)$. Then $\mathbf{p}_m \sim \mathcal{D}_m(\alpha_m)$. In particular, each $p_i \sim Be(\alpha_i, t - \alpha_i)$;

(b) Let $\mathbf{p} \sim \mathcal{D}(\alpha)$. Fix $m < n$, and let $q_i = \frac{p_i}{1 - \sum_{i=1}^{m} p_i}, i = m+1, \cdots, n$. Let $\beta_m = (\alpha_{m+1}, \cdots, \alpha_{n+1})$. Then,

$$(q_{m+1}, \cdots, q_n) | (p_1, \cdots, p_m) \sim \mathcal{D}_{n-m}(\beta_m).$$

1.24 Multivariate Normal and Related Distributions

The multivariate normal distribution is the natural extension of the bivariate normal to the case of several jointly distributed random variables. Dating back to the works of Galton, Karl Pearson, Edgeworth, and later Ronald Fisher, the multivariate normal distribution has occupied the central place in modelling jointly distributed continuous

random variables. There are several reasons for its special status. Its mathematical properties show a remarkable amount of intrinsic structure; the properties are extremely well studied; statistical methodologies in common use often have their best or optimal performance when the variables are distributed as multivariate normal; and, there is the *multidimensional central limit theorem* and its various consequences which imply that many kinds of functions of independent random variables are approximately normally distributed, in some suitable sense. We present some of the multivariate normal theory and facts with examples in this section. Principal references are Anderson (1984), Rao (1973), and Tong (1980).

1.24.1 Definition and Some Basic Properties

As in the bivariate case, a general multivariate normal distribution is defined as the distribution of a linear function of a standard normal vector. Here is the definition.

Definition 1.60. Let $n \geq 1$. A random vector $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_n)$ is said to have an n-dimensional standard normal distribution if the Z_i are independent univariate standard normal variables, in which case their joint density is $f(z_1, z_2, \cdots, z_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n z_i^2}{2}}, -\infty < z_i < \infty, i = 1, 2, \cdots, n$.

Definition 1.61. Let $n \geq 1$, and let B be an $n \times n$ real matrix of rank $k \leq n$. Suppose $\mathbf{Z}$ has an n-dimensional standard normal distribution. Let μ be any n-dimensional vector of real constants. Then $\mathbf{X} = \mu + B\mathbf{Z}$ is said to have a *multivariate normal distribution with parameters* μ *and* Σ, where $\Sigma = BB'$ is an $n \times n$ real symmetric nonnegative definite (nnd) matrix. If $k < n$, the distribution is called a *singular multivariate normal*. If $k = n$, then Σ is positive definite and the distribution of $\mathbf{X}$ is called a *nonsingular multivariate normal* or often, just a multivariate normal. We use the notation $\mathbf{X} \sim N_n(\mu, \Sigma)$, or sometimes $\mathbf{X} \sim MVN(\mu, \Sigma)$.
We treat only nonsingular multivariate normals in this section.

Theorem 1.70. (Density of Multivariate Normal) Suppose B is full rank. Then, the joint density of $X_1, X_2, \cdots, X_n$ is

$$f(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)},$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{R}^n$. It follows from the linearity of expectations, and the linearity of covariance, that $E(X_i) = \mu_i$, and $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, the (i, j)th element of Σ. The vector of expectations is usually called the *mean vector*, and the matrix of pairwise covariances is called the *covariance matrix*. Thus, we have the following facts about the physical meanings of μ, Σ:

Proposition The mean vector of $\mathbf{X}$ equals $E(\mathbf{X}) = \mu$; The covariance matrix of $\mathbf{X}$ equals Σ.

An important property of a multivariate normal distribution is the property of *closure under linear transformations*, i.e., any number of linear functions of $\mathbf{X}$ will also have a multivariate normal distribution. The precise closure property is as follows.

Theorem 1.71. (Density of Linear Transformations) Let $\mathbf{X} \sim N_n(\mu, \Sigma)$, and let $A_{k\times n}$ be a matrix of rank $k, k \leq n$. Then $\mathbf{Y} = A\mathbf{X} \sim N_k(A\mu, A\Sigma A')$. In particular, marginally, $X_i \sim N(\mu_i, \sigma_i^2)$, where $\sigma_i^2 = \sigma_{ii}, 1 \leq i \leq n$, and all lower dimensional marginals are also multivariate normal in the corresponding dimension. **Corollary (MGF of Multivariate Normal)** Let $\mathbf{X} \sim N_n(\mu, \Sigma)$. Then the mgf of $\mathbf{X}$ exists at all points in $\mathcal{R}^n$ and is given by

$$\psi_{\mu,\Sigma}(\mathbf{t}) = e^{t'\mu + \frac{t'\Sigma t}{2}}.$$

This follows on simply observing that the theorem above implies that $\mathbf{t}'\mathbf{X} \sim N(t'\mu, t'\Sigma t)$, and by then using the formula for the mgf of a univariate normal distribution. ♣

As we just observed, any linear combination $\mathbf{t}'\mathbf{X}$ of a multivariate normal vector is univariate normal. A remarkable fact is that the converse is also true.

Theorem 1.72. (Characterization of Multivariate Normal) Let $\mathbf{X}$ be an n-dimensional random vector, and suppose each linear combination $\mathbf{t}'\mathbf{X}$ has a univariate normal distribution. Then $\mathbf{X}$ has a multivariate normal distribution.

Example 1.126. Suppose $(X_1, X_2, X_3) \sim N_3(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

We want to find the joint distribution of $(X_1 - X_2, X_1 + X_2 + X_3)$. We recognize this to be a linear function $A\mathbf{X}$, where

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Therefore, by the theorem above,

$$(X_1 - X_2, X_1 + X_2 + X_3) \sim N_2(A\mu, A\Sigma A'),$$

and by direct matrix multiplication,

$$A\mu = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

and

$$A\Sigma A' = \begin{pmatrix} 3 & -2 \\ -2 & 12 \end{pmatrix}$$

In particular, $X_1 - X_2 \sim N(-1, 3), X_1 + X_2 + X_3 \sim N(3, 12)$, and $\text{Cov}(X_1 - X_2, X_1 + X_2 + X_3) = -2$. Therefore, the correlation between $X_1 - X_2$ and $X_1 + X_2 + X_3$ equals $\frac{-2}{\sqrt{3}\sqrt{12}} = -\frac{1}{3}$.

1.24.2 Multivariate Normal Conditional Distributions

As in the bivariate normal case, zero correlation between a particular pair of variables implies that the particular pair must be independent. And, as in the bivariate normal case, all lower dimensional conditional distributions are also multivariate normal.

Theorem 1.73. Suppose $\mathbf{X} \sim N_n(\mu, \Sigma)$. Then, X_i, X_j are independent if and only if $\sigma_{ij} = 0$. More generally, if $\mathbf{X}$ is partitioned as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ is k-dimensional, and $\mathbf{X_2}$ is $n - k$ dimensional, and if Σ is accordingly partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where Σ_{11} is $k \times k$, Σ_{22} is $(n - k) \times (n - k)$, then $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent if and only if Σ_{12} is the null matrix. An important result that follows is the following; it says that once you take out the effect of $\mathbf{X_2}$ on $\mathbf{X_1}$, $\mathbf{X_2}$ and the *residual* will actually become independent.

Corollary Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}) \sim N_n(\mu, \Sigma)$. Then, $\mathbf{X_1} - E(\mathbf{X_1} \,|\mathbf{X_2})$ and $\mathbf{X_2}$ are independent.

Example 1.127. Let $(X_1, X_2, X_3) \sim N(\mu, \Sigma)$, where $\mu_1 = \mu_2 = \mu_3 = 0, \sigma_{ii} = 1, \sigma_{12} = \frac{1}{2}, \sigma_{13} = \sigma_{23} = \frac{3}{4}$. It is easily verified that Σ is positive definite. We want to find the conditional distribution of (X_1, X_2) given $X_3 = x_3$.

By the above theorem, the conditional mean is

$$(0, 0) + (\sigma_{13}, \sigma_{23})\frac{1}{\sigma_{33}}(x_3 - 0) = (\frac{3}{4}, \frac{3}{4})x_3$$

$= (\frac{3x_3}{4}, \frac{3x_3}{4})$. And the conditional covariance matrix is found from the formula in the above theorem as

$$\begin{pmatrix} \frac{7}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{7}{16} \end{pmatrix}$$

In particular, given $X_3 = x_3, X_1 \sim N(\frac{3x_3}{4}, \frac{7}{16})$; the distribution of X_2 given $X_3 = x_3$ is the same normal distribution. Finally, given $X_3 = x_3$, the correlation between X_1 and X_2 is $-\frac{\frac{1}{16}}{\sqrt{\frac{7}{16}}\sqrt{\frac{7}{16}}} = -\frac{1}{7} < 0$, although the unconditional correlation between X_1 and X_2 is positive, because $\sigma_{12} = \frac{1}{2} > 0$. The correlation between X_1 and X_2 given $X_3 = x_3$ is called *the partial correlation between X_1 and X_2 given $X_3 = x_3$.*

Example 1.128. Suppose $(X_1, X_2, X_3) \sim N_3(\mu, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

Suppose we want to find all a, b such that $X_3 - aX_1 - bX_2$ is independent of (X_1, X_2). The answer to this question depends only on Σ; so, we leave the mean vector μ unspecified.

To answer this question, we have to first note that the three variables $X_3 - aX_1 - bX_2, X_1, X_2$ together have a multivariate normal distribution, by the general theorem on multivariate normality of linear transformations. Because of this fact, $X_3 - aX_1 - bX_2$ is independent of (X_1, X_2) if and only if each of $\text{Cov}(X_3 - aX_1 - bX_2, X_1), \text{Cov}(X_3 - aX_1 - bX_2, X_2)$ is zero. These are equivalent to

$$\sigma_{31} - a\sigma_{11} - b\sigma_{21} = 0; \text{ and } \sigma_{32} - a\sigma_{12} - b\sigma_{22} = 0$$

$$\Leftrightarrow 4a + b = 0; a + 2b = 1$$

$$\Rightarrow a = -\frac{1}{7}, b = \frac{4}{7}.$$

1.25 Order Statistics

The ordered values of a sample of observations are called the order statistics of the sample, and the smallest and the largest called the extremes. Order statistics and extremes are among the most important functions of a set of random variables that we study in probability and statistics. There is natural interest in studying the highs and lows of a sequence, and the other order statistics help in understanding concentration of probability in a distribution, or equivalently, the diversity in the population represented by the distribution. Order statistics are also useful in statistical inference, where estimates of parameters are often based on some suitable functions of the order statistics. In particular, the median is of very special importance.

Distribution theory for order statistics when the observations are from a discrete distribution is complex, both notationally and algebraically, because of the fact that there could be several observations which are actually equal. These ties among the sample values make the distribution theory cumbersome. We therefore concentrate on the continuous

case. Principal references for this section are the books by David (1980). Reiss (1989), Galambos (1987), Resnick (2007), and Leadbetter, Lindgren, and Rootzén (1983).

1.25.1 Basic Distribution Theory

Definition 1.62. Let $X_1.X_2, \cdots, X_n$ be any n real valued random variables. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the ordered values of $X_1.X_2, \cdots, X_n$. Then, $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are called the *order statistics* of $X_1.X_2, \cdots, X_n$.

Remark: Thus, the minimum among $X_1.X_2, \cdots, X_n$ is the first order statistic, and the maximum the nth order statistic. The middle value among $X_1.X_2, \cdots, X_n$ is called the median. But it needs to be defined precisely, because there is really no middle value when n is an even integer. Here is our definition.

Definition 1.63. Let $X_1.X_2, \cdots, X_n$ be any n real valued random variables. Then, the *median* of $X_1.X_2, \cdots, X_n$ is defined to be $M_n = X_{(m+1)}$ if $n = 2m + 1$ (an odd integer), and $M_n = X_{(m)}$ if $n = 2m$ (an even integer). hat is, in either case, the median is the order statistic $X_{(k)}$ where k is the smallest integer $\geq \frac{n}{2}$.

An important connection to understand is the connection order statistics have with the *empirical CDF*, a function of immense theoretical and methodological importance in both probability and statistics.

Definition 1.64. Let $X_1, X_2, \cdots, X_n$ be any n real valued random variables. The *empirical CDF* of $X_1.X_2, \cdots, X_n$, also called the empirical CDF of the sample, is the function

$$F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n},$$

i.e., $F_n(x)$ measures the proportion of sample values that are $\leq x$ for a given x.

Remark: Therefore, by its definition, $F_n(x) = 0$ whenever $x < X_{(1)}$, and $F_n(x) = 1$ whenever $x \geq X_{(n)}$. It is also a constant , namely, $\frac{k}{n}$, for all x-values in the interval $[X_{(k)}, X_{(k+1)})$. So F_n satisfies all the properties of being a valid CDF. Indeed, it is the CDF of a discrete distribution, which puts an equal probability of $\frac{1}{n}$ at the sample values $X_1, X_2, \cdots, X_n$.

Definition 1.65. Let $Q_n(p) = F_n^{-1}(p)$ be the quantile function corresponding to F_n. Then, $Q_n = F_n^{-1}$ is called the *quantile function* of $X_1, X_2, \cdots, X_n$, or the empirical quantile function.

We can now relate the median and the order statistics to the quantile function F_n^{-1}.

Proposition Let $X_1, X_2, \cdots, X_n$ be n random variables. Then,

$$(a) X_{(i)} = F_n^{-1}(\frac{i}{n});$$

$$(b) M_n = F_n^{-1}(\frac{1}{2}).$$

We now specialize to the case where $X_1, X_2, \cdots, X_n$ are independent random variables with a common density function $f(x)$ and CDF $F(x)$, and work out the fundamental distribution theory of the order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$.

Theorem 1.74. (Joint Density of All the Order Statistics) Let $X_1, X_2, \cdots, X_n$ be independent random variables with a common density function $f(x)$. Then, the joint density function of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is given by

$$f_{1,2,\cdots,n}(y_1, y_2, \cdots, y_n) = n! f(y_1) f(y_2) \cdots f(y_n) I_{\{y_1 < y_2 < \cdots < y_n\}}.$$

Proof: A verbal heuristic argument is easy to understand. If $X_{(1)} = y_1, X_{(2)} = y_2, \cdots, X_{(n)} = y_n$, then exactly one of the sample values $X_1, X_2, \cdots, X_n$ is y_1, exactly one is y_2, etc., but we can put any of the n observations at y_1, any of the other $n-1$ observations at y_2, etc., and so the density of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is $f(y_1) f(y_2) \cdots f(y_n) \times n(n-1) \cdots 1 = n! f(y_1) f(y_2) \cdots f(y_n)$, and obviously if the inequality $y_1 < y_2 < \cdots < y_n$ is not satisfied, then at such a point the joint density of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ must be zero.

Example 1.129. (Uniform Order Statistics). Let $U_1, U_2, \cdots, U_n$ be independent $U[0,1]$ variables, and $U_{(i)}, 1 \le i \le n$, their order statistics. Then, by our theorem above, the joint density of $U_{(1)}, U_{(2)}, \cdots, U_{(n)}$ is

$$f_{1,2,\cdots,n}(u_1, u_2, \cdots, u_n) = n! I_{0 < u_1 < u_2 < \cdots < u_n < 1}.$$

Once we know the joint density of all the order statistics, we can find the marginal density of any subset, by simply integrating out the rest of the coordiantes, but *being extremely careful in using the correct domain over which to integrate the rest of the coordinates.* For example, if we want the marginal density of just $U_{(1)}$, that is of the sample minimum, then we will want to integrate out $u_2, \cdots, u_n$, and the correct domain of integration would be, for a given u_1, a value of $U_{(1)}$, in (0,1),

$$u_1 < u_2 < u_3 < \cdots < u_n < 1.$$

So, we will integrate down in the order $u_n, u_{n-1}, \cdots, u_2$, to obtain

$$\begin{aligned} f_1(u_1) &= n! \int_{u_1}^1 \int_{u_2}^1 \cdots \int_{u_{n-1}}^1 du_n du_{n-1} \cdots du_3 du_2 \\ &= n(1-u_1)^{n-1}, 0 < u_1 < 1. \end{aligned}$$

Likewise, if we want the marginal density of just $U_{(n)}$, that is of the sample maximum, then we will want to integrate out $u_1, u_2, \cdots, u_{n-1}$, and now the answer will be

$$f_n(u_n) = n! \int_0^{u_n} \int_0^{u_{n-1}} \cdots \int_0^{u_2} du_1 du_2 \cdots du_{n-1}$$

$$= nu_n^{n-1}, 0 < u_n < 1.$$

However, it is useful to note that for the special case of the minimum and the maximum, we could have obtained the densities much more easily and directly. Here is why. First take the maximum. Consider its CDF; for $0 < u < 1$:

$$P(U_{(n)} \le u) = P(\cap_{i=1}^n \{X_i \le u\}) = \prod_{i=1}^n P(X_i \le u)$$

$$= u^n,$$

and hence, the density of $U_{(n)}$ is $f_n(u) = \frac{d}{du}[u^n] = nu^{n-1}, 0 < u < 1$.

Likewise, for the minimum, for $0 < u < 1$, the tail CDF is:

$$P(U_{(1)} > u) = P(\cap_{i=1}^n \{X_i > u\}) = (1-u)^n,$$

and so the density of $U_{(1)}$ is

$$f_1(u) = \frac{d}{du}[1-(1-u)^n] = n(1-u)^{n-1}, 0 < u < 1.$$

For a general $r, 1 \le r \le n$, the density of $U_{(r)}$ works out to a Beta density:

$$f_r(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{n-r}, 0 < u < 1,$$

which is the $Be(r, n-r+1)$ density.

As a rule, if the underlying CDF F is symmetric about its median, then the sample median will also have a density symmetric about the median of F. Additionally, the density of the sample maximum will generally be skewed to the right, and that of the sample minimum skewed to the left.

For general CDFs, the density of the order statistics usually will not have a simple formula in terms of elementary functions; but approximations for large n are often possible. This will be treated in a later section.

Example 1.130. (Density of One and Two Order Statistics). The joint density of any subset of the order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ can be worked out from their joint density, which we derived in the preceding section. The most important case in applications is the joint density of two specific order statistics, say $X_{(r)}$ and $X_{(s)}, 1 \le r < s \le n$, or the density of a specific one, say $X_{(r)}$. A verbal heuristic argument helps in understanding the formula for the joint density of $X_{(r)}$ and $X_{(s)}$, and also the density of a specific one $X_{(r)}$.

First consider the density of just $X_{(r)}$. Fix u. To have $X_{(r)} = u$, we must have exactly one observation at u, another $r-1$ below u, and $n-r$ above u. This will *suggest* that the

density of $X_{(r)}$ is

$$f_r(u) = nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(1-F(u))^{n-r}$$

$$= \frac{n!}{(r-1)!(n-r)!}(F(u))^{r-1}(1-F(u))^{n-r}f(u),$$

$-\infty < u < \infty$. This is in fact the correct formula for the density of $X_{(r)}$.

Next, consider the case of the joint density of two order statistics, $X_{(r)}$ and $X_{(s)}$. Fix $0 < u < v < 1$. Then, to have $X_{(r)} = u, X_{(s)} = v$, we must have exactly one observation at u, another $r-1$ below u, one at v, $n-s$ above v, and $s-r-1$ between u and v. This will *suggest* that the joint density of $X_{(r)}$ and $X_{(s)}$ is

$$f_{r,s}(u,v) = nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(n-r)f(v)\binom{n-r-1}{n-s}(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}$$

$$= \frac{n!}{(r-1)!(n-s)!(s-r-1)!}(F(u))^{r-1}(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}f(u)f(v),$$

$-\infty < u < v < \infty$.

Again, this is indeed the joint density of two specific order statistics $X_{(r)}$ and $X_{(s)}$.

The arguments used in this example lead to the following theorem.

Theorem 1.75. (Density of One and Two Order Statistics and Range) Let $X_1, X_2, \cdots, X_n$ be independent observations from a continuous CDF $F(x)$ with density function $f(x)$. Then,

$$(a)X_{(n)} \text{ has the density } f_n(u) = nF^{n-1}(u)f(u), -\infty < u < \infty;$$

$$(b)X_{(1)} \text{ has the density } f_1(u) = n(1-F(u))^{n-1}f(u), -\infty < u < \infty;$$

$$(c)\text{For a general } r, 1 \le r \le n, X_{(r)} \text{ has the density}$$

$$f_r(u) = \frac{n!}{(r-1)!(n-r)!}F^{r-1}(u)(1-F(u))^{n-r}f(u), -\infty < u < \infty;$$

$$(d)\text{For general } 1 \le r < s \le n, (X_{(r)}, X_{(s)}) \text{ have the joint density}$$

$$== \frac{n!}{(r-1)!(n-s)!(s-r-1)!}(F(u))^{r-1}(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}f(u)f(v),$$

$-\infty < u < v < \infty$;

$$(e)\text{The minimum and the maximum, } X_{(1)} \text{ and} X_{(n)} \text{ have the joint density}$$

$$f_{1,n}(u,v) = n(n-1)(F(v)-F(u))^{n-2}f(u)f(v), -\infty < u < v < \infty;$$

$$(f)\textbf{(CDF of Range)}\ W = W_n = X_{(n)} - X_{(1)} \text{ has the CDF}$$

$$F_W(w) = n\int_{-\infty}^{\infty}[F(x+w) - F(x)]^{n-1}f(x)dx;$$

(g)**(Density of Range)** $W = W_n = X_{(n)} - X_{(1)}$ has the density

$$f_W(w) = n(n-1)\int_{-\infty}^{\infty}[F(x+w) - F(x)]^{n-2}f(x)f(x+w)dx.$$

Example 1.131. (Moments of Uniform Order Statistics). The general formulas in the above theorem lead to the following moment formulas in the uniform case.

In the $U[0,1]$ case,

$$E(U_{(1)}) = \frac{1}{n+1}, E(U_{(n)}) = \frac{n}{n+1},$$

$$\text{var}(U_{(1)}) = \text{var}(U_{(n)}) = \frac{n}{(n+1)^2(n+2)}; 1-U_{(n)} \stackrel{\mathcal{L}}{=} U_{(1)}; \text{Cov}(U_{(1)}, (U_{(n)}) = \frac{1}{(n+1)^2(n+2)},$$

$$E(W_n) = \frac{n-1}{n+1}, \text{var}(W_n) = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

For a general order statistic, it follows from the fact that $U_{(r)} \sim Be(r, n-r+1)$, that

$$E(U_{(r)}) = \frac{r}{n+1}; \text{var}(U_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)}.$$

Furthermore, it follows from the formula for their joint density that

$$\text{Cov}(U_{(r)}, U_{(s)}) = \frac{r(n-s+1)}{(n+1)^2(n+2)}.$$

Example 1.132. (Exponential Order Statistics). A second distribution of importance in the theory of order statistics is the Exponential distribution. The mean λ essentially arises as just a multiplier in the calculations. So, we will treat only the standard Exponential case.

Let $X_1, X_2, \cdots, X_n$ be independent standard Exponential variables. Then, by the general theorem on the joint density of the order statistics, in this case the joint density of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is

$$f_{1,2,\cdots,n}(u_1, u_2, \cdots, u_n) = n!e^{-\sum_{i=1}^n u_i},$$

$0 < u_1 < u_2 < \cdots < u_n < \infty$. Also, in particular, the minimum $X_{(1)}$ has the density

$$f_1(u) = n(1-F(u))^{n-1}f(u) = ne^{-(n-1)u}e^{-u} = ne^{-nu},$$

$0 < u < \infty$. In other words, we have the quite remarkable result that the *minimum of n independent standard Exponentials is itself an Exponential with mean* $\frac{1}{n}$. Also, from the general formula, the maximum $X_{(n)}$ has the density

$$f_n(u) = n(1-e^{-u})^{n-1}e^{-u} = n\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}e^{-(i+1)u}, 0 < u < \infty.$$

As a result,

$$E(X_{(n)}) = n\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}\frac{1}{(i+1)^2} = \sum_{i=1}^{n}(-1)^{i-1}\binom{n}{i}\frac{1}{i},$$

which also equals $1 + \frac{1}{2} + \cdots + \frac{1}{n}$. We will later see in the section on spacings that by another argument, it will also follow that in the standard Exponential case, $E(X_{(n)}) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$.

Example 1.133. (Normal Order Statistics). Another clearly important case is that of the order statistics of a normal distribution. Because the general $N(\mu, \sigma^2)$ random variable is a location-scale transformation of a standard normal variable, we have the distributional equivalence that $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ have the same joint distribution as $(\mu + \sigma Z_{(1)}, \mu + \sigma Z_{(2)}, \cdots, \mu + \sigma Z_{(n)})$. So, we consider just the standard normal case.

Because of the symmetry of the standard normal distribution around zero, for any r, $Z_{(r)}$ has the same distribution as $-Z_{(n-r+1)}$. In particular, $Z_{(1)}$ has the same distribution as $-Z_{(n)}$. From our general formula, the density of $Z_{(n)}$ is

$$f_n(x) = n\Phi^{n-1}(x)\phi(x), -\infty < x < \infty.$$

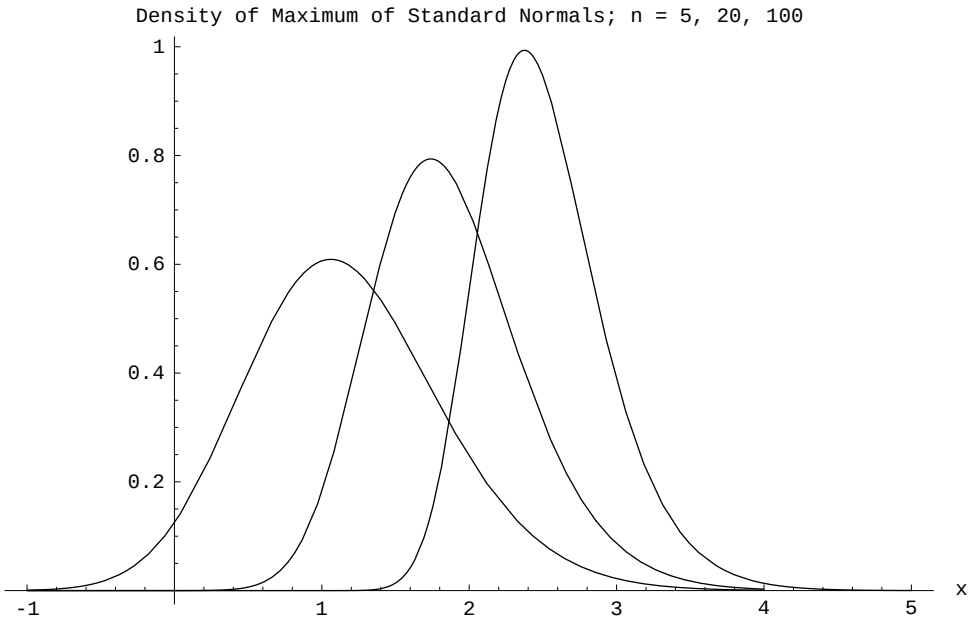
Again, this is a skewed density. It can be shown, either directly, or by making use of the general theorem on existence of moments of order statistics (see the next section) that every moment, and in particular the mean and the variance of $Z_{(n)}$ exist. Except for very small n, closed form formulas for the mean or variance are not possible. For small n, integration tricks do produce exact formulas. For example,

$$E(Z_{(n)}) = \frac{1}{\sqrt{\pi}}, \text{if } n = 2; E(Z_{(n)}) = \frac{3}{2\sqrt{\pi}}, \text{if } n = 3.$$

Such formulas are available for $n \le 5$; see David (1980).

We tabulate the expected value of the maximum for some values of n to illustrate the slow growth.

n	$E(Z_{(n)})$
2	.56
5	1.16
10	1.54
20	1.87
30	2.04
50	2.25
100	2.51
500	3.04
1000	3.24
10000	3.85



The density of $Z_{(n)}$ is plotted here for three values of n. We can see that the density is shifting to the right, and at the same time getting more peaked. Theoretical asymptotic (i.e., as $n \to \infty$) justifications for these visual findings are possible, and we will see some of them in a later section.

1.25.2 Quantile Transformation

Uniform order statistics play a very special role in the theory of order statistics, because many problems about order statistics of samples from a general density can be reduced, by a simple and common technique, to the case of uniform order statistics. It is thus especially important to understand and study uniform order statistics. The technique that makes helpful reductions of problems in the general continuous case to the case of a uniform distribution on [0,1] is one of making just the quantile transformation. We describe the exact correspondence below.

Theorem 1.76. (Quantile Transformation Theorem) Let $X_1, X_2, \cdots, X_n$ be independent observations from a continuous CDF $F(x)$ on the real line, and let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ denote their order statistics. Let $F^{-1}(p)$ denote the quantile function of F. Let $U_1, U_2, \cdots, U_n$ be independent observations from the $U[0,1]$ distribution, and let $U_{(1)}, U_{(2)}, \cdots, U_{(n)}$ denote their order statistics. Then, the following equalities in distributions hold:

$$(a) F(X_1) \sim U[0,1], \text{ that is, } F(X_1) \text{ and } U_1 \text{ have the same distribution;}$$

We write this equality in distribution as $F(X_1) \stackrel{\mathcal{L}}{=} U_1$;

$$(b) F^{-1}(U_1) \stackrel{\mathcal{L}}{=} X_1;$$

$$(c) F(X_{(i)}) \stackrel{\mathcal{L}}{=} U_{(i)};$$

$$(d) F^{-1}(U_{(i)}) \stackrel{\mathcal{L}}{=} X_{(i)};$$

$$(e)(F(X_{(1)}), F(X_{(2)}), \cdots, F(X_{(n)})) \stackrel{\mathcal{L}}{=} (U_{(1)}, U_{(2)}, \cdots, U_{(n)});$$

$$(f)(F^{-1}(U_{(1)}), F^{-1}(U_{(2)}), \cdots, F^{-1}(U_{(n)})) \stackrel{\mathcal{L}}{=} (X_{(1)}, X_{(2)}, \cdots, X_{(n)}).$$

Remark: This theorem says that any distributional uestion about the set of order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ of a sample from a general continuous distribution can be rephrased in terms of the set of order statistics from the $U[0,1]$ distribution. For this, all we need to do is to substitute $F^{-1}(U_{(i)})$ in place of $X_{(i)}$, where F^{-1} is the quantile function of F.

So, at least in principle, as long as we know how to work skillfully with the joint distribution of the uniform order statistics, we can answer questions about any set of order statistics from a general continuous distribution.This has proved to be a very useful technique in the theory of order statistics.

An important consequence of the above theorem is the following.

Proposition Let $X_1, X_2, \cdots, X_n$ be independent observations from a continuous CDF F. Then, for any $r, s, \text{Cov}(X_{(r)}, X_{(s)}) \geq 0$.

1.25.3 The Empirical CDF

The empirical CDF $F_n(x)$, defined in Section 3.25, is a tool of tremendous importance in statistics and probability. The reason for its effectiveness as a tool is that if sample observations arise from some CDF F, then the empirical CDF F_n will be very close to F for large n. So, we can get a very good idea of what the true F is by looking at F_n. Furthermore, since $F_n \approx F$, it can be expected that if $T(F)$ is a nice functional of F, then the empirical version $T(F_n)$ would be close to $T(F)$. Perhaps the simplest example of this is the mean $T(F) = E_F(X)$. The empirical version then is $T(F_n) = E_{F_n}(X) = \frac{\sum_{i=1}^n X_i}{n}$, because F_n assigns the equal probability $\frac{1}{n}$ to just the observation values $X_1, \cdots, X_n$. This means that the mean of the sample values should be close to the expected value under the true F. And, this is indeed true under simple conditions, and we have already seen some evidence for it in the form of the central limit theorem. We provide some basic properties and applications of the empirical CDF in this section.

Theorem 1.77. Let $X_1, X_2, \cdots$ be independent observations from a CDF F. Then,

(a) For any fixed $x, nF_n(x) \sim Bin(n, F(x))$;

(b)**(DKW Inequality)** Let $\Delta_n = sup_{x \in \mathcal{R}} |F_n(x) - F(x)|$. Then, for all $n, \epsilon > 0$, and all F,

$$P(\Delta_n > \epsilon) \leq 2e^{-2n\epsilon^2};$$

(c) Assume that F is continuous. For any given n, and $\alpha, 0 < \alpha < 1$, there exist positive constants D_n, independent of F, such that whatever be F,

$$P(\forall x \in \mathcal{R}, F_n(x) - D_n \le F(x) \le F_n(x) + D_n) \ge 1 - \alpha.$$

Remark: Part (b), the DKW inequality, was first proved in Dvoretzky, Kiefer, and Wolfowitz (1956), but in a weaker form. The inequality stated here is proved in Massart (1990). Furthermore, the constant 2 in the inequality is the best possible choice of the constant, i.e., the inequality is false with any other constant $C < 2$. The inequality says that *uniformly in* x, for large n, the empirical CDF is arbitrarily close to the true CDF with a very high probability, and the probability of the contrary is *subgaussian*. We will see more precise consequences of this in a later chapter. Part (c) is important for statisticians, as we show in our next example.

Example 1.134. (Confidence Band for a Continuous CDF). This example is another important application of the quantile transformation method. Imagine a hypothetical sequence of independent $U[0,1]$ variables, $U_1, U_2, \cdots$, and let G_n denote the empirical CDF of this sequence of uniform random variables; i.e.,

$$G_n(t) = \frac{\#\{i : U_i \le t\}}{n}.$$

By the quantile transformation,

$$\Delta_n = sup_{x \in \mathcal{R}} |F_n(x) - F(x)| \stackrel{\mathcal{L}}{=} sup_{x \in \mathcal{R}} |G_n(F(x)) - F(x)|$$

$$= sup_{0<t<1} |G_n(t) - t|,$$

which shows that as long as F is a continuous CDF, so that the quantile transformation can be applied, for any n, the distribution of Δ_n is the same for all F. This common distribution is just the distribution of $sup_{0<t<1}|G_n(t) - t|$. Consequently, if D_n is such that $P(sup_{0<t<1}|G_n(t) - t| > D_n) \le \alpha$, then D_n also satisfies (the apparently stronger statement)

$$P(\forall x \in \mathcal{R}, F_n(x) - D_n \le F(x) \le F_n(x) + D_n) \ge 1 - \alpha.$$

The probability statement above provides the assurance that with probability $1 - \alpha$ or more, the true CDF $F(x)$, as a function, is caught between the pair of functions $F_n(x) \pm D_n$. As a consequence, the band $F_n(x) - D_n \le F(x) \le F_n(x) + D_n, x \in \mathcal{R}$, is called a $100(1-\alpha\%$ *confidence band* for F. This is of great use in statistics, because statisticians often consider the true CDF F to be not known.

The constants D_n have been computed and tabulated for small and moderate n.

1.26 Fundamental Limit Theorems

Distributional calculations in probability are typically such that exact calculations are difficult or impossible. For example, one of the simplest possible functions of n variables is their sum, and yet in most cases, we cannot find the distribution of the sum for fixed n in an exact closed form. But the central limit theorem allows us to conclude that in some cases the sum will behave like a normally distributed random variable, when n is large. Similarly, the role of general asymptotic theory is to provide an approximate answer to exact solutions in many types of problems, and often very complicated problems. The nature of the theory is such that the approximations have remarkable unity of character, and indeed nearly unreasonable unity of character. Asymptotic theory is the single most unifying theme of probability and statistics. Particularly, in statistics, nearly every method or rule or tradition has its root in some result in asymptotic theory. No other branch of probability and statistics has such an incredibly rich body of literature, tools, and applications, in amazingly diverse areas and problems. Skills in asymptotics are nearly indispensable for a serious statistician or probabilist.
In this introductory section, we lay out the basic concepts of asymptotics with concrete applications. More specialized tools will be separately treated in many of the subsequent chapters. Principal references for this section are DasGupta (2008), Serfling (1980), van der Vaart (1998).

1.26.1 Some Basic Notation and Convergence Concepts

Some basic definitions, notation, and concepts are put together in this section.

Definition 1.66. Let a_n be a sequence of real numbers. We write $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$. We write $a_n = O(1)$ if $\exists\, K < \infty \ni |a_n| \le K\, \forall n \ge 1$.

More generally, if $a_n, b_n > 0$ are two sequences of real numbers, we write $a_n = o(b_n)$ if $\frac{a_n}{b_n} = o(1)$; we write $a_n = O(b_n)$ if $\frac{a_n}{b_n} = O(1)$.

Remark: Note that the definition allows the possibility that a sequence a_n which is $O(1)$ is also $o(1)$. The converse is always true, i.e., $a_n = o(1) \Rightarrow a_n = O(1)$.

Definition 1.67. Let a_n, b_n be two real sequences. We write $a_n \sim b_n$ if $\frac{a_n}{b_n} \to 1$, as $n \to \infty$. We write $a_n \asymp b_n$ if $0 < \liminf \frac{a_n}{b_n} \le \limsup \frac{a_n}{b_n} < \infty$, as $n \to \infty$.

Example 1.135. Let $a_n = \frac{n}{n+1}$. Then, $|a_n| \le 1 \forall n \ge 1$; so $a_n = O(1)$. But, $a_n \to 1$. as $n \to \infty$; so a_n is not $o(1)$.

However, suppose $a_n = \frac{1}{n}$. Then, again, $|a_n| \le 1 \forall n \ge 1$; so $a_n = O(1)$. But, this time $a_n \to 0$. as $n \to \infty$; so a_n is both $O(1)$ and $o(1)$. But $a_n = O(1)$ is a weaker statement in this case than saying $a_n = o(1)$.

Next, suppose $a_n = -n$. Then $|a_n| = n \to \infty$, as $n \to \infty$; so a_n is not $O(1)$, and therefore also cannot be $o(1)$.

Example 1.136. Let $c_n = \log n$, and $a_n = \frac{c_n}{c_{n+k}}$, where $k \geq 1$ is a fixed positive integer. Thus, $a_n = \frac{\log n}{\log(n+k)} = \frac{\log n}{\log n + \log(1+\frac{k}{n})} = \frac{1}{1+\frac{1}{\log n}\log(1+\frac{k}{n})} \to \frac{1}{1+0} = 1$. Therefore, $a_n = O(1), a_n \sim 1$, but a_n is not $o(1)$. The statement that $a_n \sim 1$ is stronger than saying $a_n = O(1)$.

To motivate the first probabilistic convergence concept, we give an illustrative example.

Example 1.137. For $n \geq 1$, consider the simple discrete random variables X_n with the pmf $P(X_n = \frac{1}{n}) = 1 - \frac{1}{n}, P(X_n = n) = \frac{1}{n}$. Then, for large n, X_n is close to zero with a large probability. Although for any given n, X_n is never equal to zero, the probability of it being far from zero is very small for large n. For example, $P(X_n > .001) \leq .001$ if $n \geq 1000$. More formally, for any given $\epsilon > 0$, $P(X_n > \epsilon) \leq P(X_n > \frac{1}{n}) = \frac{1}{n}$, if we take n to be so large that $\frac{1}{n} < \epsilon$. As a consequence, $P(|X_n| > \epsilon) = P(X_n > \epsilon) \to 0$, as $n \to \infty$. This example motivates the following definition.

Definition 1.68. Let $X_n, n \geq 1$, be an infinite sequence of random variables defined on a common sample space Ω. We say that X_n *converges in probability to* c, a specific real constant, if for any given $\epsilon > 0, P(|X_n - c| > \epsilon) \to 0$, as $n \to \infty$. Equivalently, X_n converges in probability to c if given any $\epsilon > 0, \delta > 0$, there exists an $n_0 = n_0(\epsilon, \delta)$ such that $P(|X_n - c| > \epsilon) < \delta \, \forall n \geq n_0(\epsilon, \delta)$.

If X_n converges in probability to c, we write $X_n \stackrel{\mathcal{P}}{\Rightarrow} c$, or sometimes also, $X_n \stackrel{P}{\Rightarrow} c$.

Sometimes the sequence X_n may get close to some random variable, rather than a constant. Here is an example of such a situation.

Example 1.138. Let X, Y be two independent standard normal variables. Define a sequence of random variables X_n as $X_n = X + \frac{Y}{n}$. Then, intuitively, we feel that for large n, the $\frac{Y}{n}$ part is small, and X_n is very close to the fixed random variable X. Formally, $P(|X_n - X| > \epsilon) = P(|\frac{Y}{n}| > \epsilon) = P(|Y| > n\epsilon) = 2[1 - \Phi(n\epsilon)] \to 0$, as $n \to \infty$. This motivates a generalization of the previous definition.

Definition 1.69. Let $X_n, X, n \geq 1$ be random variables defined on a common sample space Ω. We say that X_n *converges in probability to* X if given any $\epsilon > 0, P(|X_n - X| > \epsilon) \to 0$, as $n \to \infty$. We denote it as $X_n \stackrel{\mathcal{P}}{\Rightarrow} X$, or $X_n \stackrel{P}{\Rightarrow} X$.

Definition 1.70. A sequence of random variables X_n is said to be *bounded in probability* or *tight* if, given $\epsilon > 0$, one can find a constant k such that $P(|X_n| > k) \leq \epsilon$ for all $n \geq 1$.

Important Notation If $X_n \stackrel{\mathcal{P}}{\Rightarrow} 0$, then we write $X_n = o_p(1)$. More generally, if $a_n X_n \stackrel{\mathcal{P}}{\Rightarrow} 0$ for some positive sequence a_n, then we write $X_n = o_p(\frac{1}{a_n})$.

If X_n is bounded in probability, then we write $X_n = O_p(1)$. If $a_n X_n = O_p(1)$, we write $X_n = O_p(\frac{1}{a_n})$.

Proposition Suppose $X_n = o_p(1)$. Then, $X_n = O_p(1)$. The converse is, in general, not true.

Definition 1.71. Let $\{X_n, X\}$ be defined on the same sample space. We say that X_n *converges almost surely to* X (or X_n *converges to* X *with probability 1*) if $P(\omega : X_n(\omega) \to X(\omega)) = 1$. We write $X_n \stackrel{a.s}{\to} X$ or $X_n \stackrel{a.s}{\Rightarrow} X$.

Remark: If the limit X is a finite constant c, we write $X_n \stackrel{a.s}{\Rightarrow} c$.Almost sure convergence is a stronger mode of convergence than convergence in probability.
Next, we introduce the concept of convergence in mean. It often turns out to be a convenient method for establishing convergence in probability.

Definition 1.72. Let $X_n, X, n \geq 1$ be defined on a common sample space Ω. Let $p \geq 1$, and suppose $E(|X_n|^p), E(|X|^p) < \infty$. We say that X_n *converges in pth mean to* X or X_n *converges in* L_p *to* X if $E(|X_n - X|^p) \to 0$, as $n \to \infty$. If $p = 2$, we also say that X_n *converges to* X *in quadratic mean.* If X_n converges in L_p to X, we write $X_n \stackrel{L_p}{\Rightarrow} X$.

The next result says that convergence in L_p is a useful method for establishing convergence in probability.

Proposition Let $X_n, X, n \geq 1$ be defined on a common sample space Ω. Suppose X_n converges to X in L_p for some $p \geq 1$. Then $X_n \stackrel{\mathcal{P}}{\Rightarrow} X$.
Proof: Simply observe that, by using Markov's inequality,

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^p > \epsilon^p) \leq \frac{E(|X_n - X|^p)}{\epsilon^p}$$

$\to 0$, by hypothesis.

Example 1.139. (Some Counterexamples). Let X_n be the sequence of two point random variables with pmf $P(X_n = 0) = 1 - \frac{1}{n}, P(X_n = n) = \frac{1}{n}$. Then, X_n converges in probability to zero. But, $E(|X_n|) = 1 \ \forall n$, and hence X_n does not converge in L_1 to zero. In fact, it does not converge to zero in L_p for any $p \geq 1$.

Now take the same sequence X_n as above, and assume moreover that they are independent. Take an $\epsilon > 0$, and positive integers m, k. Then,

$$P(|X_n| < \epsilon \ \forall m \leq n \leq m + k)$$

$$= P(X_n = 0 \ \forall m \leq n \leq m + k) = \prod_{n=m}^{m+k} (1 - \frac{1}{n})$$

$$= \frac{m - 1}{m + k}.$$

For any m, this converges to zero as $k \to \infty$. Therefore, $\lim_{m\to\infty} P(|X_n| < \epsilon \ \forall n \geq m)$ cannot be one, and so, X_n does not converge almost surely to zero.

Next, let X_n have the pmf $P(X_n = 0) = 1 - \frac{1}{n}, P(X_n = \sqrt{n}) = \frac{1}{n}$. Then, X_n again converges in probability to zero. Furthermore, $E(|X_n|) = \frac{1}{\sqrt{n}} \to 0$, and so X_n converges in L_1 to zero. But, $E(X_n^2) = 1 \ \forall n$, and hence X_n does not converge in L_2 to zero.

1.26.2 Laws of Large Numbers

The definitions and the treatment in the previous section are for general sequences of random variables. Averages and sums are sequences of special importance in applications. The classic *laws of large numbers*, which characterize the long run behavior of averages, are given in this section.

A very useful tool for establishing almost sure convergence is stated first.

Theorem 1.78. (Borel-Cantelli Lemma)

Let $\{A_n\}$ be a sequence of events on a sample space Ω. If

$$\sum_{n=1}^{\infty} P(A_n) < \infty,$$

then P(infinitely many A_n occur) = 0.
If $\{A_n\}$ are pairwise independent, and

$$\sum_{n=1}^{\infty} P(A_n) = \infty,$$

then P(infinitely many A_n occur) = 1.
Proof: We prove the first statement. In order that infinitely many among the events $A_n, n \geq 1$, occur, it is necessary and sufficient, that given any m, there is at least one event among $A_m, A_{m+1}, \cdots$ that occurs. In other words,

$$\text{Infinitely many } A_n \text{ occur } = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n.$$

On the other hand, the events $B_m = \cup_{n=m}^{\infty} A_n$ are decreasing in m, i.e., $B_1 \supseteq B_2 \supseteq \cdots$. Therefore,

$$P(\text{infinitely many } A_n \text{ occur}) = P(\cap_{m=1}^{\infty} B_m) = \lim_{m\to\infty} P(B_m)$$

$$= \lim_{m\to\infty} P(\cup_{n=m}^{\infty} A_n) \leq \limsup_{m\to\infty} \sum_{n=m}^{\infty} P(A_n)$$

$$= 0,$$

since, by assumption, $\sum_{n=1}^{\infty} P(A_n) < \infty$.

Remark: Although pairwise independence suffices for the conclusion of the second part of the Borel-Cantelli lemma, common applications involve cases where the A_n are mutually independent.
Here is a pretty application of the Borel-Cantelli lemma.

Example 1.140. (Almost Sure Convergence of Binomial Proportion). Let $X_1, X_2, \cdots$ be an infinite sequence of independent $Ber(p)$ random variables, where $0 < p < 1$. Let $\bar{X_n} = \frac{S_n}{n} = \frac{X_1+X_2+\cdots+X_n}{n}$. Then, from our previous formula in Chapter 6 for Binomial distributions, $E(S_n - np)^4 = np(1-p)[1+3(n-2)p(1-p)]$. Thus, by Markov's inequality,

$$P(|\bar{X_n} - p| > \epsilon) = P(|S_n - np| > n\epsilon)$$

$$= P((S_n - np)^4 > (n\epsilon)^4) \leq \frac{E(S_n - np)^4}{(n\epsilon)^4}$$

$$= \frac{np(1-p)[1+3(n-2)p(1-p)]}{(n\epsilon)^4} \leq \frac{3n^2(p(1-p))^2 + np(1-p)}{n^4\epsilon^4}$$

$$\leq \frac{C}{n^2} + \frac{D}{n^3},$$

for finite constants C, D. Therefore,

$$\sum_{n=1}^{\infty} P(|\bar{X_n} - p| > \epsilon) \leq C \sum_{n=1}^{\infty} \frac{1}{n^2} + D \sum_{n=1}^{\infty} \frac{1}{n^3}$$

$$< \infty.$$

It follows from the Borel-Cantelli lemma that the Binomial sample proportion $\bar{X_n}$ converges almost surely to p. In fact, the convergence of the sample mean $\bar{X_n}$ to $E(X_1)$ (i.e., the common mean of the X_i) holds in general. The general results, due to Khintchine and Kolmogorov, are known as the *laws of large numbers*, stated below.

Theorem 1.79. (Weak Law of Large Numbers) Suppose $X_1, X_2, \ldots$ are independent and identically distributed (iid) random variables (defined on a common sample space Ω), such that $E(|X_1|) < \infty$, and $E(X_1) = \mu$. Let $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\bar{X_n} \stackrel{\mathcal{P}}{\Rightarrow} \mu$.

Theorem 1.80. (Strong Law of Large Numbers) Suppose $X_1, X_2, \ldots$ are independent and identically distributed (iid) random variables (defined on a common sample space Ω). Then, $\bar{X_n}$ has an a.s. (almost sure) limit iff $E(|X_1|) < \infty$, in which case $\bar{X_n} \stackrel{a.s.}{\Rightarrow} \mu = E(X_1)$.

Remark: It is not very simple to prove either of the two laws of large numbers in the generality stated above. If the X_i have a finite variance, then Markov's inequality easily leads to the weak law; we demonstrated this in Section 3.5.
We close this section with an important result on the uniform closeness of the empirical CDF to the underlying CDF in the iid case.

Theorem 1.81. (Glivenko-Cantelli Theorem) Let F be any CDF on the real line, and $X_1, X_2, \cdots$ iid with common CDF F. Let $F_n(x) = \frac{\#\{i \leq n : X_i \leq x\}}{n}$ be the sequence of empirical CDFs. Then, $\Delta_n = sup_{x \in \mathcal{R}} |F_n(x) - F(x)| \stackrel{a.s.}{\Rightarrow} 0$.

The Glivenko-Cantelli theorem justifiies the common use of the empirical CDF F_n to estimate an unknown population CDF F. In a sense, the Glivenko-Cantelli theorem justifiies the very concept of using a sample to make inferences about a population. You should consider the Glivenko-Cantelli theorem as fundamental.

1.26.3 Rules of Convergence Preservation

In statistics, we often want to study transformations of some random variable, or some sequence of random variables. This section addresses the question of when convergence properties are preserved if we suitably transform a sequence of random variables.

The next important theorem gives some frequently useful results, that are analogous to corresponding results in basic calculus.

Theorem 1.82.

$$(a) X_n \stackrel{\mathcal{P}}{\Rightarrow} X, Y_n \stackrel{\mathcal{P}}{\Rightarrow} Y \Rightarrow X_n \pm Y_n \stackrel{\mathcal{P}}{\Rightarrow} X \pm Y;$$

$$(b) X_n \stackrel{\mathcal{P}}{\Rightarrow} X, Y_n \stackrel{\mathcal{P}}{\Rightarrow} Y \Rightarrow X_n Y_n \stackrel{\mathcal{P}}{\Rightarrow} XY;$$

$$X_n \stackrel{\mathcal{P}}{\Rightarrow} X, Y_n \stackrel{\mathcal{P}}{\Rightarrow} Y, P(Y \neq 0) = 1 \Rightarrow \frac{X_n}{Y_n} \stackrel{\mathcal{P}}{\Rightarrow} \frac{X}{Y};$$

$$(c) X_n \stackrel{a.s.}{\Rightarrow} X, Y_n \stackrel{a.s.}{\Rightarrow} Y \Rightarrow X_n \pm Y_n \stackrel{a.s.}{\Rightarrow} X \pm Y;$$

$$(d) X_n \stackrel{a.s.}{\Rightarrow} X, Y_n \stackrel{a.s.}{\Rightarrow} Y \Rightarrow X_n Y_n \stackrel{a.s.}{\Rightarrow} XY;$$

$$X_n \stackrel{a.s.}{\Rightarrow} X, Y_n \stackrel{a.s.}{\Rightarrow} Y, P(Y \neq 0) = 1 \Rightarrow \frac{X_n}{Y_n} \stackrel{a.s.}{\Rightarrow} \frac{X}{Y};$$

$$(e) X_n \stackrel{L_1}{\Rightarrow} X, Y_n \stackrel{L_1}{\Rightarrow} Y \Rightarrow X_n + Y_n \stackrel{L_1}{\Rightarrow} X + Y;$$

$$(f) X_n \stackrel{L_2}{\Rightarrow} X, Y_n \stackrel{L_2}{\Rightarrow} Y \Rightarrow X_n + Y_n \stackrel{L_2}{\Rightarrow} X + Y.$$

Example 1.141. Suppose $X_1, X_2, \cdots$ are independent $N(\mu_1, \sigma_1^2)$ variables, and $Y_1, Y_2, \cdots$ are independent $N(\mu_2, \sigma_2^2)$ variables. For $n, m \geq 1$, let $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \bar{Y_m} = \frac{1}{m}\sum_{j=1}^{m} Y_j$. By the strong law of large numbers (SLLN), as $n, m \to \infty, \bar{X_n} \stackrel{a.s.}{\Rightarrow} \mu_1$, and $\bar{Y_m} \stackrel{a.s.}{\Rightarrow} \mu_2$. Then, by the theorem above, $\bar{X_n} - \bar{Y_m} \stackrel{a.s.}{\Rightarrow} \mu_1 - \mu_2$.

Also, by the same theorem, $\bar{X_n}\bar{Y_m} \stackrel{a.s.}{\Rightarrow} \mu_1\mu_2$.

Definition 1.73. (The Multidimensional Case) Let $\mathbf{X_n}, n \geq 1, \mathbf{X}$ be d-dimensional random vectors, for some $1 \leq d < \infty$. We say that $\mathbf{X_n} \stackrel{\mathcal{P}}{\Rightarrow} \mathbf{X}$ if $X_{n,i} \stackrel{\mathcal{P}}{\Rightarrow} X_i$, for each $i = 1, 2, \cdots, d$. That is, each coordinate of $\mathbf{X_n}$ converges in probability to the corresponding coordinate of $\mathbf{X}$. Similarly, $\mathbf{X_n} \stackrel{a.s.}{\Rightarrow} \mathbf{X}$ if $X_{n,i} \stackrel{a.s.}{\Rightarrow} X_i$, for each $i = 1, 2, \cdots, d$.

The next result is one of the most useful results on almost sure convergence and convergence in probability. It says that convergence properties are preserved if we make smooth transformations.

Theorem 1.83. (Continuous Mapping) Let $\mathbf{X_n}, \mathbf{X}$ be d-dimensional random vectors, and $f : S \subseteq \mathcal{R}^d \to \mathcal{R}^p$ a continuous function. Then,

(a)

$$\mathbf{X_n} \stackrel{\mathcal{P}}{\Rightarrow} \mathbf{X} \Rightarrow f(\mathbf{X_n}) \stackrel{\mathcal{P}}{\Rightarrow} f(\mathbf{X});$$

(b)

$$\mathbf{X_n} \stackrel{a.s.}{\Rightarrow} \mathbf{X} \Rightarrow f(\mathbf{X_n}) \stackrel{a.s.}{\Rightarrow} f(\mathbf{X}).$$

Here are some important applications of this theorem.

Example 1.142. (Convergence of Sample Variance). Let $X_1, X_2, \cdots, X_n$ be independent observations from a common distribution F, and suppose that F has finite mean μ and finite variance σ^2. The sample variance, of immense importance in statistics, is defined as $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. The purpose of this example is to show that $s^2 \stackrel{a.s.}{\Rightarrow} \sigma^2$, as $n \to \infty$.

First note that if we can prove that $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \stackrel{a.s.}{\Rightarrow} \sigma^2$, then it follows that s^2 also converges almost surely to σ^2, because $\frac{n}{n-1} \to 1$ as $n \to \infty$. Now,

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2$$

(an algebraic identity). Since F has a finite variance, it also possesses a finite second moment, namely, $E_F(X^2) = \sigma^2 + \mu^2 < \infty$. By applying the SLLN (strong law of large numbers) to the sequence $X_1^2, X_2^2, \cdots$, we get $\frac{1}{n}\sum_{i=1}^n X_i^2 \stackrel{a.s.}{\Rightarrow} E_F(X^2) = \sigma^2 + \mu^2$. By applying the SLLN to the sequence $X_1, X_2, \cdots$, we get $\bar{X} \stackrel{a.s.}{\Rightarrow} \mu$, and therefore by the continuous mapping theorem, $(\bar{X})^2 \stackrel{a.s.}{\Rightarrow} \mu^2$. Now, by the theorem on preservation of convergence, we get that $\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2 \stackrel{a.s.}{\Rightarrow} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$, which finishes the proof.

Example 1.143. (Convergence of Sample Correlation). Suppose $F(x, y)$ is a joint CDF in $\mathcal{R}^2$, and suppose that $E(X^2), E(Y^2)$ are both finite. Let $\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ be the correlation between X and Y. Suppose $(X_i, Y_i), 1 \le i \le n$ are n independent observations from the joint CDF F. The sample correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

The purpose of this example is to show that r converges almost surely to ρ.

It is convenient to rewrite r in the equivalent form

$$r = \frac{\frac{1}{n}\sum_{i=1}^n X_iY_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

By the SLLN, $\frac{1}{n}\sum_{i=1}^n X_iY_i$ converges almost surely to $E(XY)$, and $\bar{X}, \bar{Y}$ converge almost surely to $E(X), E(Y)$. By the previous example, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ converges almost surely to $\text{Var}(X)$, and $\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ converges almost surely to $\text{Var}(Y)$. Now consider the function $f(s,t,u,v,w) = \frac{s-tu}{\sqrt{v}\sqrt{w}}, -\infty < s,t,u < \infty, v,w > 0$. This function is continuous on the set $S = \{(s,t,u,v,w) : -\infty < s,t,u < \infty, v,w > 0, (s-tu)^2 \le vw\}$. The joint distribution of $\frac{1}{n}\sum_{i=1}^n X_iY_i, \bar{X}, \bar{Y}, \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2, \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ assigns probability one to the set S. By the continuous mapping theorem, it follows that $r \stackrel{a.s.}{\Rightarrow} \rho$.

1.27 Convergence in Distribution and CLT

Studying distributions of random variables is of paramount importance in both probability and statistics. The relevant random variable may be a member of some sequence X_n. Its exact distribution may be cumbersome. But it may be possible to approximate its distribution by a simpler distribution. We can then approximate probabilities for the true distribution of the random variable by probabilities in the simpler distribution. The type of convergence concept that justifies this sort of approximation is called *convergence in distribution* or *convergence in law*. Of all the convergence concepts we are discussing, convergence in distribution is among the most useful in answering practical questions. For example, statisticians are usually much more interested in constructing confidence intervals than just point estimators, and a central limit theorem of some kind is necessary to produce a confidence interval.

We start with an illustrative example.

Example 1.144. Suppose $X_n \sim U[\frac{1}{2} - \frac{1}{n+1}, \frac{1}{2} + \frac{1}{n+1}], n \ge 1$. Since the interval $[\frac{1}{2} - \frac{1}{n+1}, \frac{1}{2} + \frac{1}{n+1}]$ is shrinking to the single point $\frac{1}{2}$, intuitively we feel that the distribution of X_n is approaching a distribution concentrated at $\frac{1}{2}$, i.e., a *one point distribution*. The CDF of the distribution concentrated at $\frac{1}{2}$ equals the function $F(x) = 0$ for $x < \frac{1}{2}$, and $F(x) = 1$ for $x \ge \frac{1}{2}$. Consider now the CDF of X_n; call it $F_n(x)$. Fix $x < \frac{1}{2}$. Then, for all large $n, F_n(x) = 0$, and so $\lim_n F_n(x)$ is also zero. Next fix $x > \frac{1}{2}$. Then, for all large $n, F_n(x) = 1$, and so $\lim_n F_n(x)$ is also one. Therefore, if $x < \frac{1}{2}$, or if $x > \frac{1}{2}$, $\lim_n F_n(x) = F(x)$. If x is exactly equal to $\frac{1}{2}$, then $F_n(x) = \frac{1}{2}$. But $F(\frac{1}{2}) = 1$. So $x = \frac{1}{2}$ is a problematic point, and the only problematic point, in that $F_n(\frac{1}{2}) \not\to F(\frac{1}{2})$. Interestingly, $x = \frac{1}{2}$ is also exactly the only point at which F is *not* continuous. However, we do not want this one problematic point to ruin our intuitive feeling that X_n is approaching the

one point distribution concentrated at $\frac{1}{2}$. That is, we do not take into account any points where the limiting CDF is not continuous.

Definition 1.74. Let $X_n, X, n \geq 1$, be real valued random variables defined on a common sample space Ω. We say that X_n *converges in distribution (in law)* to X if $P(X_n \leq x) \to P(X \leq x)$ as $n \to \infty$, at every point x which is a continuity point of the CDF F of the random variable X.

We denote convergence in distribution by $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$.

If $\mathbf{X_n}, \mathbf{X}$ are d-dimensional random vectors, then the same definition applies by using the joint CDFs of $\mathbf{X_n}, \mathbf{X}$, i.e., $\mathbf{X_n}$ converges in distribution to $\mathbf{X}$ if $P(X_{n1} \leq x_1, \cdots, X_{nd} \leq x_d) \to P(X_1 \leq x_1, \cdots, X_d \leq x_d)$ at each point $(x_1, \cdots, x_d)$ which is a continuity point of the joint CDF $F(x_1, \cdots, x_d)$ of the random vector $\mathbf{X}$.

An important point of caution is the following:

Caution In order to prove that d-dimensional vectors $\mathbf{X_n}$ converge in distribution to $\mathbf{X}$, it is not, in general, enough to prove that each coordinate of $\mathbf{X_n}$ converges in distribution to the corresponding coordinate of $\mathbf{X}$.
By definition of convergence in distribution, if $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$, and if X has a continuous CDF F (continuous everywhere), then $F_n(x) \to F(x)\ \forall x$ where $F_n(x)$ is the CDF of X_n. The following theorem says that much more is true, namely that the convergence is actually uniform.

Theorem 1.84. (Pólya's Theorem) Let $X_n, n \geq 1$ have CDF F_n, and let X have CDF F. If F is everywhere continuous, and if $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$, then

$$sup_{x \in \mathcal{R}} |F_n(x) - F(x)| \to 0,$$

as $n \to \infty$.

A large number of equivalent characterizations of convergence in distribution are known. A few basic ones are the folloowing.

Theorem 1.85. (Portmanteau Theorem). Let $\{X_n, X\}$ be random variables taking values in a finite dimensional Euclidean space. The following are characterizations of $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$:

(a) $E(g(X_n)) \to E(g(X))$ for all bounded continuous functions g;

(b) $E(g(X_n)) \to E(g(X))$ for all bounded uniformly continuous functions g.

Here is a simple application of this theorem.

Example 1.145. Consider $X_n \sim Uniform\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}$. Then, it can be shown easily that the sequence X_n converges in law to the $U[0,1]$ distribution. Consider now the function $g(x) = x^{10}, 0 \le x \le 1$. Note that g is continuous and bounded. Therefore, by part (a) of the *Portmanteau Theorem*, $E(g(X_n)) = \sum_{k=1}^{n} \frac{k^{10}}{n^{11}} \to E(g(X)) = \int_0^1 x^{10} dx = \frac{1}{11}$.

This can be proved by using convergence of Riemann sums to a Riemann integral. But it is interesting to see the link to convergence in distribution.

Example 1.146. (Convergence in Distribution is Weaker than Convergence of Density). Suppose X_n is a sequence of random variables on $[0,1]$ wit h density $f_n(x) = 1 + \cos(2\pi n x)$. Then, $X_n \stackrel{\mathcal{L}}{\Rightarrow} U[0,1]$ by a direct verification of the definition using CDFs. Indeed, $F_n(x) = x + \frac{\sin(2n\pi x)}{2n\pi} \to x \, \forall x \in (0,1)$. However, note that the densitities f_n *do not* converge to the uniform density 1 as $n \to \infty$.

Convergence of densities is useful to have when true, because it ensures a much stronger form of convergence than convergence in distribution. Suppose X_n have CDF F_n and density f_n, and X has CDF F and density f. If $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$, then we can only assert that $F_n(x) = P(X_n \le x) \to F(x) = P(X \le x) \, \forall x$. However, if we have convergence of the densities, then we can make the much stronger assertion that for *any* event $A, P(X_n \in A) \to P(X \in A)$, not just for events A of the form $A = (-\infty, x]$.

The most important result on convergence in distribution is the *central limit theorem*, which we have already met. It is restated here, together with the multidimensional version, which is important.

Theorem 1.86. (CLT)
Let $X_i, i \ge 1$ be iid (independent and identically distributed) with $E(X_i) = \mu$ and $\text{Var}X_i) = \sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \stackrel{\mathcal{L}}{\Rightarrow} Z \sim N(0,1).$$

We also write

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1).$$

The multidimensional central limit theorem is stated next.

Theorem 1.87. (**Multivariate CLT**)
Let $\mathbf{X_i}, i \ge 1$, be iid d-dimensional random vectors with $E(\mathbf{X_1}) = \mu$, and covariance matrix $\text{Cov}(\mathbf{X_1}) = \Sigma$. Then,

$$\sqrt{n}(\bar{\mathbf{X}} - \mu) \stackrel{\mathcal{L}}{\Rightarrow} N_d(0, \Sigma).$$

1.27.1 Slutsky's Theorem and Applications

Akin to the results on preservation of convergence in probability and almost sure convergence under various operations, there are similar other extremely useful results on

preservation of convergence in distribution. The first theorem is of particular importance in statistics.

Theorem 1.88. (Slutsky's Theorem) Let $\mathbf{X_n}, \mathbf{Y_n}$ be d and p -dimensional random vectors for some $d, p \geq 1$. Suppose $\mathbf{X_n} \stackrel{\mathcal{L}}{\Rightarrow} \mathbf{X}$, and $\mathbf{Y_n} \stackrel{\mathcal{P}}{\Rightarrow} \mathbf{c}$. Let $h(x, y)$ be a scalar or a vector valued jointly continuous function in $(x, y) \in \mathcal{R}^d \times \mathcal{R}^p$. Then $h(\mathbf{X_n}, \mathbf{Y_n}) \stackrel{\mathcal{L}}{\Rightarrow} h(\mathbf{X}, \mathbf{c})$. The following are some particularly important consequences of Slutsky's theorem.

Corollary (a) Suppose $\mathbf{X_n} \stackrel{\mathcal{L}}{\Rightarrow} \mathbf{X}, \mathbf{Y_n} \stackrel{\mathcal{P}}{\Rightarrow} \mathbf{c}$, where $\mathbf{X_n}, \mathbf{Y_n}$ are of the same order. Then, $\mathbf{X_n} + \mathbf{Y_n} \stackrel{\mathcal{L}}{\Rightarrow} \mathbf{X} + \mathbf{c}$;

(b) Suppose $\mathbf{X_n} \stackrel{\mathcal{L}}{\Rightarrow} \mathbf{X}, Y_n \stackrel{\mathcal{P}}{\Rightarrow} c$, where Y_n are scalar random variables. Then $Y_n \mathbf{X_n} \stackrel{\mathcal{L}}{\Rightarrow} c\mathbf{X}$;

(c) Suppose $\mathbf{X_n} \stackrel{\mathcal{L}}{\Rightarrow} \mathbf{X}, Y_n \stackrel{\mathcal{P}}{\Rightarrow} c \neq 0$, where Y_n are scalar random variables. Then $\frac{\mathbf{X_n}}{Y_n} \stackrel{\mathcal{L}}{\Rightarrow} \frac{\mathbf{X}}{c}$.

Example 1.147. (Convergence of the t to Normal). Let $X_i, i \geq 1$, be iid $N(\mu, \sigma^2), \sigma > 0$, and let $T_n = \frac{\sqrt{n}(\bar{X}-\mu)}{s}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, namely the sample variance. We saw in Section 3.21 that T_n has the central $t(n-1)$ distribution. Write

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{s/\sigma}.$$

We have seen that $s^2 \stackrel{a.s.}{\Rightarrow} \sigma^2$. Therefore, by the *continuous mapping theorem*, $s \stackrel{a.s.}{\Rightarrow} \sigma$, and so $\frac{s}{\sigma} \stackrel{a.s.}{\Rightarrow} 1$. On the other hand, by the central limit theorem, $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$. Therefore, now by *Slutsky's theorem*, $T_n \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$.
This argument shows that whatever be the common distribution of the X_i, *if* $\sigma^2 = Var(X_1) < \infty$ *and* > 0, *then* $T_n \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$, *although the exact distribution of* T_n *is no longer the central* t *distribution,* unless *the common distribution of the* X_i *is normal.*

Example 1.148. (An Important Statistics Example). Let $X = X_n \sim Bin(n, p), n \geq 1, 0 < p < 1$. In statistics, p is generally treated as an unknown parameter, and the usual estimate of p is $\hat{p} = \frac{X}{n}$. Define $T_n = |\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}}|$. The goal of this example is to find the limiting distribution of T_n. First, by the central limit theorem,

$$\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} = \frac{X - np}{\sqrt{np(1-p)}} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1).$$

Next, by the WLLN, $\hat{p} \stackrel{\mathcal{P}}{\Rightarrow} p$, and hence by the continuous mapping theorem for convergence in probability, $\sqrt{\hat{p}(1-\hat{p})} \stackrel{\mathcal{P}}{\Rightarrow} \sqrt{p(1-p)}$. This gives, by Slutsky's theorem, $\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$. Finally, because the absolute value function is continuous, it follows that

$$T_n = |\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}}| \stackrel{\mathcal{L}}{\Rightarrow} |Z|,$$

the absolute value of a standard normal.

1.28 Limiting Distributions of Order Statistics

We discussed the importance of order statistics and sample percentiles in detail in this chapter. The exact distribution theory of one or several order statistics was presented there. Although closed form in principle, the expressions are complicated, except in some special cases, such as the uniform and the Exponential. However, once again it turns out that just like sample means, order statistics and sample percentiles also have a very structured limiting distribution theory. We present a selection of the fundamental results on this limit theory for order statistics and sample percentiles. Principal references for this section are Galambos (1977), Serfling (1980), Reiss (1989), de Haan (2006) and DasGupta (2008).

First, we fix some notation. Suppose $X_i, i \geq 1$, are iid random random variables with CDF F. We denote the order statistics of $X_1, X_2, \cdots, X_n$ by $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$, or sometimes, simply as $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The empirical CDF is denoted as $F_n(x) = \frac{\#\{i:X_i \leq x\}}{n}$ and $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$ will denote the empirical quantile function. The population quantile function is $F^{-1}(p) = \inf\{x : F(x) \geq p\}$. $F^{-1}(p)$ will also be denoted as ξ_p.

Consider the kth order statistic $X_{k:n}$ where $k = k_n$. Three distinguishing cases are:

(a) $\sqrt{n}(\frac{k_n}{n} - p) \to 0$, for some $0 < p < 1$. This is called *the central case*.

(b) $n - k_n \to \infty$ and $\frac{k_n}{n} \to 1$. This is called *the intermediate case*.

(c) $n - k_n = O(1)$. This is called *the extreme case*.

Different asymptotics apply to the three cases; the case of central order statistics is considered here.

We start with an illustrative example.

Example 1.149. (Uniform Case). As a simple motivational example, consider the case of iid $U[0,1]$ observations $U_1, U_2, \cdots$. The goal of this example is to find the limiting distribution of the kth order statistic, $U_{(k)}$, where $U_{(k)}$ is a central order statistic. For simplicity, we take $k = np$, for some $p, 0 < p < 1$, and assume that np is an integer. The final goal in this example is to show that in this central case, $U_{(k)}$ is approximately normally distributed for large n. More precisely, if $q = 1 - p$, and $\sigma^2 = pq$, then we want to show that $\frac{\sqrt{n}(U_{(k)} - p)}{\sigma}$ converges in distribution to the standard normal.

This can be established in a number of ways. The following heuristic argument is illuminating. We let $F_n(u)$ denote the empirical CDF of $U_1, \cdots, U_n$, and recall that $nF_n(u) \sim Bin(n, F(u))$, which is $Bin(n, u)$ in the uniform case.
$i = 1, 2, \ldots, r$.

An important consequence of the joint asymptotic normality of a finite number of order statistics is that linear combinations of a finite number of order statistics will be asymptotically univariate normal. A precise statement is as follows.

Corollary Let $c_1, c_2, \cdots, c_r$ be r fixed real numbers, and let $\mathbf{c}' = (c_1, \cdots, c_r)$. Under the assumptions of the above theorem,

$$\sqrt{n}(\sum_{i=1}^{r} c_i X_{k_i:n} - \sum_{i=1}^{r} c_i \xi_{q_i}) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \mathbf{c}'\Sigma\mathbf{c}).$$

Here is an important statistical application.

Example 1.150. (The Interquartile Range). The 25th and the 75th percentile of a set of sample observations are called the first and the third quartile of the sample. The difference between them gives information about the spread in the distribution from which the sample values are coming. The difference is called the *interquartile range*, and we denote it as IQR. In statistics, suitable multiples of the IQR are often used as measures of spread, and are then compared to traditional measures of spread,such as the sample standard deviation s.

Let $k_1 = \lfloor \frac{n}{4} \rfloor, k_2 = \lfloor \frac{3n}{4} \rfloor$. Then $IQR = X_{k_2:n} - X_{k_1:n}$. It follows on some calculation from the above Corollary that

$$\sqrt{n}(IQR - (\xi_{\frac{3}{4}} - \xi_{\frac{1}{4}})) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{16}[\frac{3}{f^2(\xi_{\frac{3}{4}})} + \frac{3}{f^2(\xi_{\frac{1}{4}})} - \frac{2}{f(\xi_{\frac{1}{4}})f(\xi_{\frac{3}{4}})}]).$$

Here, the notation f means the density of F. Specializing to the case when F is the CDF of $N(\mu, \sigma^2)$, on some algebra and computation, for normally distributed iid observations,

$$\sqrt{n}(IQR - 1.35\sigma) \stackrel{\mathcal{L}}{\Rightarrow} N(0, 2.48\sigma^2)$$

$$\Rightarrow \sqrt{n}(\frac{IQR}{1.35} - \sigma) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \frac{2.48}{1.35^2}\sigma^2) = N(0, 1.36\sigma^2).$$

1.29 At Instructor's Discretion

1.29.1 Twenty Useful Inequalities

Inequalities play an extremely useful role whenever a mathematical quantity cannot be calculated exactly. For example, something as simple as Chebyshev's inequality is of some use in bounding probabilities in regions far from the mean of the distribution. Probability inequalities form an incredibly rich subject. Here, we only mention a few fundamental inequalities; a more comprehensive collection may be seen in DasGupta (2008).

Improved Bonferroni Inequality Given $n \geq 3$, events $A_1, \cdots, A_n, S_{1,n} = \sum P(A_i), S_{2,n} = \sum_{1 \leq i < j \leq n} P(A_i \cap A_j), S_{3,n} = \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k)$,

$$S_{1,n} - S_{2,n} + \frac{2}{n-1} S_{3,n} \leq P(\cup A_i) \leq S_{1,n} - \frac{2}{n} S_{2,n}.$$

Galambos, J. and Simonelli, I. (1996), Bonferroni-Type Inequalities with Applications, Springer-Verlag.

Cantelli's Inequality If $E(X) = \mu, Var(X) = \sigma^2 < \infty$, then

$$P(X - \mu \geq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}, \lambda > 0,$$

$$P(X - \mu \leq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}, \lambda < 0.$$

Rao, C. R. (1973), Linear Statistical Inference and Applications, Wiley.

Alon-Spencer Inequality Given a nonnegative integer valued random variable X with a finite variance,

$$P(X = 0) \leq \frac{\text{Var}(X)}{E(X^2)}.$$

Alon, N. and Spencer, J. (2000). The Probabilistic Method, Wiley.

Anderson Inequality Given $X \sim N(0, \Sigma_1), Y \sim N(0, \Sigma_2), \Sigma_2 - \Sigma_1$ nnd, and C any symmetric convex set,

$$P(X \in C) \geq P(Y \in C).$$

Tong, Y. (1980). Probability Inequalities in Multivariate Distributions, Academic Press.

Bernstein Inequality Given a nonnegative random variable X, with a finite mgf $\psi(a) = E(e^{aX})$,

$$P(X \geq t) \leq \inf_{a>0} e^{-at}\psi(a).$$

Hoeffding Inequality Given constants $a_1, \cdots, a_n$, and iid *Rademacher variables* $\epsilon_1, \cdots, \epsilon_n, P(\epsilon_i = \pm 1) = \frac{1}{2}$, for any $t > 0$,

$$P(\sum a_i \epsilon_i \geq t) \leq e^{-\frac{t^2}{2\sum a_i^2}}.$$

Hoeffding, W. (1963), JASA, 58, 13-30.

Hotelling-Solomon Inequality Given a random variable X, with $E(X) = \mu, \text{Var}(X) = \sigma^2 < \infty$, and median m,

$$|\mu - m| \leq \sigma.$$

Hotelling, H. and Solomons, L. (1932), Ann. Math. Statist., 3, 141-142.

3σ Rule For Unimodal Variables If X is unimodal with an absolutely continuous distribution, $\mu = E(X), \sigma^2 = Var(X) < \infty$, then

$$P(|X - \mu| \geq 3\sigma) \leq \frac{4}{81} < .05.$$

Dharmadhikari, S. and Joag-Dev, K. (1988). Unimodality, Convexity and Applications, Academic Press.

Vysochanskiĭ-Petunin Inequality For Unimodal Variables If X is unimodal with an absolutely continuous distribution, $\alpha \in \mathcal{R}, \tau^2 = E(X-\alpha)^2$, , then

$$P(|X-\alpha| \geq k) \leq \frac{4\tau^2}{9k^2}, k \geq \frac{\sqrt{8}\tau}{\sqrt{3}},$$

$$P(|X-\alpha| \geq k) \leq \frac{4\tau^2}{3k^2} - \frac{1}{3}, k \leq \frac{\sqrt{8}\tau}{\sqrt{3}}.$$

Pukelsheim, F. (1994), Amer. Statist., 48, 88-91.

Johnson-Rogers Inequality for Unimodal Variables If X is unimodal around M with an absolutely continuous distribution, $E(X) = \mu, Var(X) = \sigma^2$, then

$$|\mu - M| \leq \sigma\sqrt{3}.$$

Dharmadhikari, S. and Joag-Dev, K. (1988). Unimodality, Convexity and Applications, Academic Press.

Inequality for Normal Distributions Given $X \sim N(\mu, \sigma^2)$, for any $k > 0$,

$$P(|X-\mu| > k\sigma) < \frac{1}{3k^2}.$$

DasGupta, A. (2000), Metrika, 51, 185-200.

Berge Inequality for Bivariate Distributions If $X = (X_1, X_2)$ is a two dimensional variable with coordinate means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ, then,

$$P(\max(\frac{|X_1-\mu_1|}{\sigma_1}, \frac{|X_2-\mu_2|}{\sigma_2}) \geq k) \leq \frac{1+\sqrt{1-\rho^2}}{k^2}.$$

Dharmadhikari, S. and Joag-Dev, K. (1988). Unimodality, Convexity and Applications, Academic Press.

Multivariate Chebyshev Inequality of Olkin and Pratt I

Given $X = (X_1, \cdots, X_n), E(X_i) = \mu_i, Var(X_i) = \sigma_i^2, corr(X_i, X_j) = 0$ for $i \neq j$,

$$P(\cup_{i=1}^n \frac{|X_i-\mu_i|}{\sigma_i} \geq k_i) \leq \sum_{i=1}^n k_i^{-2}.$$

Olkin, I. and Pratt, J. (1958), Ann. Math. Statist., 29, 226-234.

Positive Dependence Inequality for Multivariate Normal Given $X_{n\times 1} \sim MVN(\mu, \Sigma), \Sigma$ p.d., and such that $\sigma^{ij} \leq 0$ for all $i, j, i \neq j$,

$$P(X_i \geq c_i, i = 1, \cdots, n) \geq \prod_{i=1}^n P(X_i \geq c_i),$$

for any constants $c_1, \cdots, c_n$.

Tong, Y. (1980). Probability Inequalities in Multivariate Distributions, Academic Press.

Chernoff Inequality Given $X \sim N(\mu, \sigma^2)$, and an absolutely continuous function $g(z)$ such that $Var(g(X)), E[|g'(X)|] < \infty$,

$$\sigma^2(Eg'(X))^2 \leq Var(g(X)) \leq \sigma^2 E[(g'(X))^2].$$

Chernoff, H. (1981), Ann. Prob., 9, 533-535.

von Bahr-Esseen Inequality Given independent random variabls $X_1, \cdots, X_n$, symmetric about zero, $1 \leq p \leq 2$,

$$E|\sum X_i|^p \leq \sum E|X_k|^p.$$

von Bahr, B. and Esseen, C. (1965), Ann. Math. Statist., 36, 299-303.

Inequality for Mills Ratio With $\Phi(x), \phi(x)$ as the $N(0,1)$ CDF and PDF, $R(x) = \frac{1-\Phi(x)}{\phi(x)}$ the *Mills ratio*,

$$\frac{x}{x^2+1} \leq R(x) \leq \frac{1}{x}.$$

Patel, J. and Read, C. (1996). Handbook of the Normal Distribution, CRC Press.

Inequality for Poisson Distributions Given $X \sim Poisson(\lambda)$,

$$P(X \geq k) \geq 1 - \Phi((k-\lambda)/\sqrt{\lambda}).$$

Bohman, H. (1963), Skand. Actuarietidskr., 46, 47-52.

Inequality for Binomial Distributions Given $X \sim Bin(n,p)$,

$$P(X \geq k) \leq \binom{n}{k} p^k \leq (\frac{enp}{k})^k.$$

Shorack, G. and Wellner, J. (1986). Empirical Processes with Applications to Statistics, Wiley.

LeCam Inequality Given $X_i \overset{indep.}{\sim} Ber(p_i), 1 \leq i \leq n, \lambda = \sum p_i, X = \sum X_i, Y \sim Poisson(\lambda)$,

$$\sum_{k \geq 0} |P(X=k) - P(Y=k)| \leq 2 \sum p_i^2.$$

LeCam, L. (1965), in Bernoulli, Bayes, Laplace Anniversary, 179-202, Univ. Calif. Press.

1.29.2 Multivariate Geometry Formulas

$$(a)\text{Volume of n Dimensional Unit Sphere } = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}.$$

$$(b)\text{Surface Area of n Dimensional Unit Sphere } = \frac{n\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}.$$

$$(c)\text{Volume of n Dimensional Simplex } = \int_{x_i \geq 0, x_1+\cdots+x_n \leq 1} dx_1 \cdots dx_n = \frac{1}{n!}.$$

$$(d) \int_{x_i \geq 0, \sum_{i=1}^n (\frac{x_i}{c_i})^{\alpha_i} \leq 1} f(\sum_{i=1}^n (\frac{x_i}{c_i})^{\alpha_i}) \prod_{i=1}^n x_i^{p_i-1} dx_1 \cdots dx_n$$

$$= \frac{\prod_{i=1}^n c_i^{p_i}}{\prod_{i=1}^n \alpha_i} \frac{\prod_{i=1}^n \Gamma(\frac{p_i}{\alpha_i})}{\Gamma(\sum_{i=1}^n \frac{p_i}{\alpha_i})} \int_0^1 f(t) t^{\sum_{i=1}^n \frac{p_i}{\alpha_i}-1} dt,$$

$(c_i, \alpha_i, p_i > 0)$.

$$(e) \int_{\sum_{i=1}^n x_i^2 \leq 1} (\sum_{i=1}^n p_i x_i)^{2m} dx_1 \cdots dx_n = \frac{(2m-1)!}{2^{2m-1}(m-1)!} \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+m+1)} \times (\sum_{i=1}^n p_i^2)^m,$$

$\forall p_1, \cdots, p_n$, $\forall m \geq 1$.

1.29.3 Exponential and Uniform Spacings

Another set of statistics, closely linked to the order statistics, and helpful in both statistics and probability are the *spacings*. They are the gaps between successive order statistics. For some particular underlying distributions, their mathematical properties are extraordinarily structured, and those in turn lead to results for other distributions. Two instances are the spacings of uniform and Exponential order statistics. Some basic facts about spacings are discussed below.

Definition 1.75. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the order statistics of a sample of n observations $X_1, X_2, \cdots, X_n$. Then, $W_i = X_{(i+1)} - X_{(i)}, 1 \leq i \leq n-1$ are called the *spacings* of the sample, or the spacings of the order statistics.

First we discuss Exponential spacings. The spacings of an Exponential sample have the characteristic property that the spacings are all independent Exponentials as well. Here is the precise result.

Theorem 1.89. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the order statistics from an $Exp(\lambda)$ distribution. Then $W_0 = X_{(1)}, W_1, \cdots, W_{n-1}$ are independent, with $W_i \sim Exp(\frac{\lambda}{n-i}), i =$

$0, 1, \cdots, n-1$.**Corollary (Réyni)** The joint distribution of the order statistics of an Exponential distribution with mean λ have the representation

$$(X_{(r)})|_{r=1}^{n} \stackrel{\mathcal{L}}{=} (\sum_{i=1}^{r} \frac{X_i}{n-i+1})|_{r=1}^{n},$$

where $X_1, \cdots, X_n$ are themselves independent Exponentials with mean λ. ♣

Remark: Verbally, the order statistics of an Exponential distribution are linear combinations of independent Exponentials, with a very special sequence of coefficients. In an obvious way, the representation can be extended to the order statistics of a general continuous CDF by simply using the quantile transformation.

Example 1.151. (Moments and Correlations of Exponential Order Statistics). From the representation in the above Corollary, we immediately have

$$E(X_{(r)}) = \lambda \sum_{i=1}^{r} \frac{1}{n-i+1};\ \mathrm{Var}(X_{(r)}) = \lambda^2 \sum_{i=1}^{r} \frac{1}{(n-i+1)^2}.$$

Furthermore, by using the same representation, for $1 \le r < s \le n$, $\mathrm{Cov}(X_{(r)}, X_{(s)}) = \lambda^2 \sum_{i=1}^{r} \frac{1}{(n-i+1)^2}$, and therefore the correlation $\rho_{X_{(r)},X_{(s)}} = \sqrt{\frac{\sum_{i=1}^{r} \frac{1}{(n-i+1)^2}}{\sum_{i=1}^{s} \frac{1}{(n-i+1)^2}}}$. In particular,

$$\rho_{X_{(1)},X_{(n)}} = \frac{\frac{1}{n}}{\sqrt{\sum_{i=1}^{n} \frac{1}{i^2}}} \approx \frac{\sqrt{6}}{n\pi},$$

for large n. In particular, $\rho_{X_{(1)},X_{(n)}} \to 0$, as $n \to \infty$. *In fact, in large samples the minimum and the maximum are in general approximately independent.*

Next, we come to uniform spacings. Actually, the results above for Exponential spacings lead to some highly useful and neat representations for the spacings and the order statistics of a uniform distribution. The next result describes the most important properties of uniform spacings and order statistics.

Theorem 1.90. Let $U_1, U_2, \cdots, U_n$ be independent $U[0,1]$ variables, and $U_{(1)}, U_{(2)}, \cdots, U_{(n)}$ the order statistics. Let $W_0 = U_{(1)}, W_i = U_{(i+1)} - U_{(i)}, 1 \le i \le n-1$, and $V_i = \frac{U_{(i)}}{U_{(i+1)}}, 1 \le i \le n-1, V_n = U_{(n)}$. Let also $X_1, X_2, \cdots, X_{n+1}$ be $(n+1)$ independent standard Exponentials, independent of the $U_i, i = 1, 2, \cdots, n$. Then,

$(a) V_1, V_2^2, \cdots, V_{n-1}^{n-1}, V_n^n$ are independent $U[0,1]$ variables, and $(V_1, V_2, \cdots, V_{n-1})$ are independent of V_n;

$(b) (W_0, W_1, \cdots, W_{n-1}) \sim \mathcal{D}(\alpha)$, a Dirichlet distribution with parameter vector $\alpha_{n+1\times 1} = (1, 1, \cdots, 1)$. That is, $(W_0, W_1, \cdots, W_{n-1})$ is uniformly distributed in the n-dimensional

simplex;

$(c)(W_0, W_1, \cdots, W_{n-1}) \stackrel{\mathcal{L}}{=} (\frac{X_1}{\sum_{j=1}^{n+1} X_j}, \frac{X_2}{\sum_{j=1}^{n+1} X_j}, \cdots, \frac{X_n}{\sum_{j=1}^{n+1} X_j})$;

$(d)(U_{(1)}, U_{(2)}, \cdots, U_{(n)}) \stackrel{\mathcal{L}}{=} (\frac{X_1}{\sum_{j=1}^{n+1} X_j}, \frac{X_1+X_2}{\sum_{j=1}^{n+1} X_j}, \cdots, \frac{X_1+X_2+\cdots+X_n}{\sum_{j=1}^{n+1} X_j})$.

Remark: *Part (d) of this theorem, representing uniform order statistics in terms of independent exponentials is one of the most useful results in the theory of order statistics.*

1.29.4 Wishart Distributions

Much as in the case of one dimension, structured results are available for functions of a set of n-dimensional independent random vectors, each distributed as a multivariate normal. The applications of most of these results are in statistics. Perhaps the foremost concepts in the multivariate case are those of the *mean vector* and the *sample covariance matrix*, analogous to $\bar{X}$ and s^2 in the univariate case. A classic reference is Anderson (1984).
First, we need some notation. Given N independent p-dimensional random vectors, $\mathbf{X_i}, 1 \leq i \leq N$, each $X_i \sim N_n(\mu, \Sigma)$, we define the *sample mean vector* and the *sample covariance matrix* as

$$\bar{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X_i},$$

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})',$$

where $\sum$ in the definitions above is defined as vector addition, and for a vector $\mathbf{u}, \mathbf{uu}'$ means a matrix product. In plain words, S is the matrix whose diagonal elements are the sample variances on the individual coordinates of the vector random variable, and the off-diagonal elements are the pairwise sample covariances;

$$s_{jj} = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \bar{X}_j)^2;$$

$$s_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k).$$

The primary use of S is in estimating Σ, or interesting functions of Σ, for example the ordered eigenvalues of Σ, Note that S is always a symmetric matrix. So, by the distribution of S, we mean the joint distribution of $s_{11}, s_{12}, \cdots, s_{1p}, s_{22}, s_{23}, \cdots, s_{pp}$. In one dimension, i.e., when $n = 1$, $\bar{X}$ is also distributed as a normal, and $\bar{X}, s^2$ are independently distributed. Moreover, in one dimension, $(N-1)s^2/\sigma^2$ has a χ^2_{N-1} distribution. Analogs of all of these

results exist in this general multivariate case. This part of the multivariate normal theory is very classic. Here is the definition of the Wishart distribution.

Definition 1.76. (Wishart Distribution) Let $W_{p\times p}$ be a symmetric positive definite random matrix with elements $w_{ij}, 1 \leq i, j \leq p$. W is said to have a *Wishart distribution* with k degrees of freedom ($k \geq p$) and scale matrix Σ, if the joint density of its elements,$w_{ij}, 1 \leq i \leq j \leq p$ is given by

$$f(W) = c|W|^{(k-p-1)/2} e^{-\frac{1}{2}tr(\Sigma^{-1}W)},$$

where the normalizing constant c equals

$$c = \frac{1}{2^{kp/2}|\Sigma|^{k/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma(\frac{k-i+1}{2})}.$$

We write $W \sim W_p(k, \Sigma)$.

The most classic distributional result on S and $\bar{\mathbf{X}}$ in the multivariate normal case is the following theorem.

Theorem 1.91. Let $\mathbf{X_i}$ be independent $N_p(\mu, \Sigma)$ random vectors, $1 \leq i \leq N$. Then,

$$(a)\, \bar{\mathbf{X}} \sim N_p(\mu, \frac{\Sigma}{N});$$

$$(b)\, \text{For } N > p, S \text{ is positive definite with probability one;}$$

$$(c)\, \text{For } N > p, (N-1)S \sim W_p(N-1, \Sigma);$$

$$(d)\, \bar{\mathbf{X}} \text{ and } S \text{ are independently distributed.}$$

Part (a) of this theorem follows easily from the representation of a multivariate normal vector in terms of a multivariate standard normal vector. For part (b), see Eaton and Perlman (1973) and Dasgupta(1971). Part (c) is classic. Numerous proofs of part (c) are available. Specifically, see Mahalanobis, Bose, and Roy (1937).
The following moment formulas are also fundamental.

Theorem 1.92. Let $S \sim W_p(k, \Sigma)$. Then,

$$E(S) = k\Sigma; \quad E(S^{-1}) = \frac{1}{k-p-1}\Sigma^{-1};$$

$$E[\text{tr}S] = k(\text{tr}\Sigma); \quad E[\text{tr}S^{-1}] = \frac{1}{k-p-1}(\text{tr}\Sigma^{-1}).$$

$$E[c'Sc] = c'\Sigma c, \text{ for any } p\text{-vector c;}$$

$$E(|S|) = \frac{\prod_{i=1}^{p}(N-i)}{(N-1)^p}|\Sigma|.$$

1.29.5 Noncentral Distributions

The so called noncentral distributions arise in testing of hypothesis problems. You will see that in evaluating *the power* of standard normal theory tests, these noncentral distributions become objects of fundamental importance. Also, from time to time, you may need the formula for their density functions in some calculation.

The Noncentral t Distribution Suppose $X_1, X_2, \cdots, X_n$ are independent univariate $N(\mu, \sigma^2)$ variables. The *noncentral t statistic* is defined as

$$t(a) = \frac{\sqrt{n}(\bar{X} - a)}{s},$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance, and a is a general real number. If we take $a = \mu$, then $t(a)$ has the ordinary t distribution with $n-1$ degrees of freedom. But, if we take $a \neq \mu$, then the distribution of $t(a)$ becomes *noncentral*, and it then depends on μ, σ. The distribution of $t(a)$ is given in the following result.

Theorem 1.93. The statistic $t(a)$ has the *noncentral t distribution* with $n-1$ degrees of freedom and noncentrality parameter $\delta = \frac{\sqrt{n}(\mu - a)}{\sigma}$, with the density function

$$f_{t(a)}(x) = ce^{-(n-1)\delta^2/(2(x^2+n-1))}(x^2 + n - 1)^{-n/2}$$
$$\times \int_0^\infty t^{n-1} e^{-(t - \frac{\delta x}{\sqrt{x^2+n-1}})^2/2} dt, -\infty < x < \infty,$$

where the normalizing constant c equals

$$c = \frac{(n-1)^{(n-1)/2}}{\sqrt{\pi}\Gamma(\frac{n-1}{2})2^{n/2-1}}.$$

Furthermore, for $n > 2$,

$$E(t(a)) = \delta\sqrt{n-1}\frac{\Gamma(\frac{n}{2} - 1)}{\Gamma(\frac{n-1}{2})}.$$

Moving on to the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, recall that for the normal case, $\frac{(n-1)s^2}{\sigma^2}$ has an ordinary chi-square distribution with $n-1$ degrees of freedom. That is, because we take deviations of the data values X_i from $\bar{X}$, μ does not appear in the distribution of s^2. It does, if we take deviations of the data values from a general real number a.

Suppose $X_1, X_2, \cdots, X_n$ are independent $N(\mu, \sigma^2)$ variables. Then the distribution of $S_a^2 = \frac{\sum_{i=1}^{n}(X_i - a)^2}{\sigma^2}$ is given in the next theorem.

Theorem 1.94. The statistic S_a^2 has the *noncentral chi square distribution* with n degrees of freedom and noncentrality parameter $\lambda = n\frac{(\mu - a)^2}{\sigma^2}$, with the density function given by a *Poisson mixture of ordinary chi squares*

$$f_{S_a^2}(x) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} g_{n+2k}(x),$$

where $g_j(x)$ stands for the density of an ordinary chi square random variable with j degrees of freedom. Furthermore,

$$E(S_a^2) = n + \lambda; \text{var}(S_a^2) = 2(n + 2\lambda).$$

It is important to understand that a here is a nonrandom general real number. If a is a random variiable, the distribution of S_a^2 is not a noncentral chi square, and usually you would not be able to say what the distribution is.

1.29.6 Berry-Esseen Theorem

The one dimensional CLT (central limit theorem says that if $X_1, X_2, \cdots$ are iid with mean μ and a finite variance σ^2, then the distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ converges to $N(0,1)$. Such a convergence result is usually used to approximate the true value of $P(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x)$ by $\Phi(x)$ for some given fixed n. However, the CLT by itself says absolutely nothing about the accuracy of such an approximation for a given n. So, we need some idea of the error of the approximation, namely $|P(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x) - \Phi(x)|$. The first result for the iid case in this direction is the classic Berry-Esseen theorem (Berry (1941), Esseen (1945)). It gives an upper bound on the error of the CLT approximation for any given n. Recall that by *Polya's theorem*, the uniform error

$$\Delta_n = \sup_{-\infty<x<\infty} |P(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x) - \Phi(x)|$$

$\to 0$, as $n \to \infty$. Bounds on Δ_n for any given n are called *uniform bounds*. Here is the classic Berry-Esseen uniform bound.

Theorem 1.95. (Berry-Esseen) Let $X_1, X_2, \cdots$ be iid with mean μ and a finite variance σ^2. Assume, in addition, that $E(|X_i|^3) < \infty$ and let $\rho = E(|X_1 - \mu|^3)$. Then there exists a universal constant C, not depending on n or the distribution of the X_i, such that

$$\Delta_n = \sup_{-\infty<x<\infty} |P(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x) - \Phi(x)| \le \frac{C\rho}{\sigma^3\sqrt{n}}.$$

The constant C may be taken to be 0.8. The constant C cannot be taken to be smaller than $\frac{3+\sqrt{10}}{6\sqrt{2\pi}} \doteq 0.41$. Better values of the constant C can be found for specific types of the underlying CDF, e.g., if it is known that the samples are iid from an Exponential distribution. Numerous refinements of this basic Berry-Esseen theorem are known, but we will not mention them here. For proofs and the refinements, see Petrov (1975). Also see Feller (1966), Serfling (1980), and Bhattacharya and Rao (1986).

1.29.7 Edgeworth Expansions

One valid criticism of the CLT approximation is that any normal distribution is symmetric about its mean, and so, by employing a normal approximation we necessarily ignore any

skewness that may be present in the true distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. For instance, if the individual X_i have an exponential density, then the true density of the sum $S_n = \sum_{i=1}^{n} X_i$ is a Gamma density, which always has a skewness. This means that the true density of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ also has a skewness. But a normal approximation ignores that skewness, and as a result, the quality of the approximation can be poor, unless n is quite large. Refined approximations which address this criticism are available. These refined approximations were formally introduced in Edgeworth (1904) and Charlier (1931). As such, they are usually called *Edgeworth densities* and the *Gram-Charlier series*. Although they are basically the same thing, there is a formal difference between the formulas in the Edgeworth density and the Gram-Charlier series. Modern treatments of these refined approximations are carefully presented in Bhattacharya and Rao (1986), Hall (1992), and Lahiri (2003).
We present here refined approximations that adjust the normal approximation for skewness, and another one which also adjusts for kurtosis. Suppose $X_1, X_2, \cdots, X_n$ are continuous random variables with a density $f(x)$. Suppose each individual X_i has four finite moments. Let $\mu, \sigma^2, \beta, \gamma$ denote the mean, variance, coefficient of skewness, and coefficient of kurtosis of the common distribution of the X_i. Let $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Define the following three successively more refined density approximations for the density of Z_n:

$$\hat{f}_{n,0}(x) = \phi(x);$$

$$\hat{f}_{n,1}(x) = (1 + \frac{\beta(x^3 - 3x)}{6\sqrt{n}})\phi(x);$$

$$\hat{f}_{n,2}(x) = (1 + \frac{\beta(x^3 - 3x)}{6\sqrt{n}} + [\gamma\frac{x^4 - 6x^2 + 3}{24} + \beta^2\frac{x^6 - 15x^4 + 45x^2 - 15}{72}]\frac{1}{n})\phi(x).$$

The functions $\hat{f}_{n,0}(x), \hat{f}_{n,1}(x)$, and $\hat{f}_{n,2}(x)$ are called the *CLT approximation, the first order Edgeworth expansion, and the second order Edgeworth expansion for the density of the mean.*

Remark: Of the three approximations, *only* $\hat{f}_{n,0}(x)$ is truly a density function. The other two functions $\hat{f}_{n,1}(x)$ and $\hat{f}_{n,2}(x)$ become negative for some values of x for a given n. As a result, if they are integrated to obtain approximations for the probability $P(Z_n \le x)$, then the approximations are not monotone nondecreasing functions of x, and can even become negative (or larger than one). For any given n, the refined approximations give inaccurate and even nonsensical answers for values of x far from zero. However, at any given x, the approximations become more accurate as n increases.
It is important to note that the approximations are of the form $\phi(x) + \frac{P_1(x)}{\sqrt{n}}\phi(x) + \frac{P_2(x)}{n}\phi(x) + \cdots$, for suitable polynomials $P_1(x), P_2(x)$, etc. The relevant polynomials $P_1(x), P_2(x)$ are related to some very special polynomials, known as *Hermite polynomials*. Hermite polynomials are obtained from successive differentiation of the standard normal

density $\phi(x)$. Precisely, the jth Hermite polynomial $H_j(x)$ is defined by the relation

$$\frac{d^j}{dx^j}\phi(x) = (-1)^j H_j(x)\phi(x).$$

In particular,

$$H_1(x) = x; H_2(x) = x^2 - 1; H_3(x) = x^3 - 3x; H_4(x) = x^4 - 6x^2 + 3;$$

$$H_5(x) = x^5 - 10x^3 + 15x; H_6(x) = x^6 - 15x^4 + 45x^2 - 15.$$

By comparing the formulas for the refined density approximations to the formulas for the Hermite polynomials, the connection becomes obvious. They arise in the density approximation formulas as a matter of fact; there is no intuition for it.

There arre also corresponding higher order Edgeworth approximations for the CDF of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Let us introduce a little notation for convenience. Let $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ and let $F_n(x)$ denote the true CDF of Z_n. The CLT says that $F_n(x) \to \Phi(x)$ for every x; the Berry-Esseen theorem says $|F_n(x) - \Phi(x)| = O(\frac{1}{\sqrt{n}})$, uniformly in x if X has three moments. If we change the approximation $\Phi(x)$ to $\Phi(x) + \frac{C_1(F)p_1(x)\phi(x)}{\sqrt{n}}$ for some suitable constant $C_1(F)$, and a suitable polynomial $p_1(x)$, we can assert that

$$|F_n(x) - \Phi(x) - \frac{C_1(F)p_1(x)\phi(x)}{\sqrt{n}}| \;=\; O\left(\frac{1}{n}\right),$$

uniformly in x. Expansions of the form

$$F_n(x) = \Phi(x) + \sum_{s=1}^{k} \frac{q_s(x)}{\sqrt{n}^s} + o(n^{-k/2}) \text{ uniformly in } x,$$

are known as Edgeworth Expansions for the CDF of Z_n. One needs some conditions on F, and enough moments of X to carry the expansion to k terms for a given k. But usually an expansion to two terms after the leading term is of the most practical importance.

We first state a major assumption which distinguishes the case of the so called *lattice distributions* from nonlattice ones.

Cramér's Condition A CDF F on the real line is said to satisfy Cramér's Condition if $\limsup_{t\to\infty} |E_F(e^{itX})| < 1$.

Remark: All distributions with a density satisfy Cramér's Condition. The expansion to be stated below does not hold for the usual discrete distributions, such as Poisson, binomial, geometric, etc. There are Edgeworth expansions for Z_n in these discrete cases. But they include extra terms not needed in the continuous case, and these extra terms need additional messy notation.

Theorem 1.96. (Two term Edgeworth Expansion) Suppose F satisfies the Cramér condition, and $E_F(X^4) < \infty$. Then

$$F_n(x) = \Phi(x) + \frac{C_1(F)p_1(x)\phi(x)}{\sqrt{n}} + \frac{C_2(F)p_2(x) + C_3(F)p_3(x)}{n}\phi(x) + O(n^{-\frac{3}{2}}),$$

uniformly in x, where

$$\begin{aligned} C_1(F) &= \frac{E(X-\mu)^3}{6\sigma^3} \text{ (skewness correction)}, \\ C_2(F) &= \frac{\frac{E(X-\mu)^4}{\sigma^4} - 3}{24} \text{ (kurtosis correction)}, \\ C_3(F) &= \frac{C_1^2(F)}{72}, \\ p_1(x) &= (1-x^2), \\ p_2(x) &= 3x - x^3 \text{ and} \\ p_3(x) &= 10x^3 - 15x - x^5. \end{aligned}$$

To summarize, the Berry-Esseen accuracy order of $n^{-1/2}$ can be improved by an Edgeworth expansion to the rate n^{-1} if we correct for skewness, which involves the first three moments, and it can be improved to the rate $n^{-3/2}$ if we also correct for kurtosis, which involves the first four moments.

1.29.8 Convergence of Moments

If some sequence of random variables X_n converges in distribution to a random variable X, then sometimes we are interested in knowing whether moments of X_n converge to moments of X. More generally, we may want to find approximations for moments of X_n. Convergence in distribution just by itself cannot ensure convergence of any moment. An extra condition that ensures convergence of appropriate moments is *uniform integrability*. Uniform integrability is a somewhat abstract concept to the beginner. However, simple sufficient conditions can be given for uniform integrability, and these sufficient conditions are often adequate in practice. Here is one such simple result on convergence of moments.

Theorem 1.97. Suppose $X, X_n, n \geq 1$ are such that $E(|X|^p) < \infty$ for some given p and that $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$. Suppose for some $\delta > 0$, $\sup_n E(|X_n|^{p+\delta}) < \infty$. Then, Then $E(X_n^k) \to E(X^k) \, \forall k \leq p$.

Remark: As an example, to conclude that $E(X_n) \to E(X)$ if $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$, it is enough to show that $E(X_n^2)$ is uniformly bounded in n.

Convergence of moments can be useful to establish convergence in distribution. Clearly, however, if we only know that $E(X_n^k)$ converges to $E(X^k)$ for each k, from that alone we cannot conclude that the distributions of X_n converge to the distribution of X. This is because, there could, in general be, another random variable Y, with a distribution distinct from that of X, but with all moments equal to the moments of X. However, if we rule out that possibility, then convergence in distribution actually does follow.

Theorem 1.98. Suppose for some sequence $\{X_n\}$ and a random variable X, $E(X_n^k) \to E(X^k) \, \forall k \geq 1$. If the distribution of X is determined by its moments, then $X_n \stackrel{\mathcal{L}}{\Rightarrow} X$.

When is a distribution determined by its sequence of moments? This is a hard analytical problem, and is commonly known as *the moment problem*. There is a huge and sophisticated literature on the moment problem; it is beautifully put together in Dette annd Studden (1997). Two easily understood conditions for determinacy by moments are given in the next result.

Theorem 1.99. (a) If a random variable X is uniformly bounded, then it is determined by its moments.

(b) If the mgf of a random variable X exists in a nonempty interval containing zero, then it is determined by its moments.

Nearly all standard distributions with a name, such as normal, Poisson, exponential, gamma, double exponential, any uniform, any Beta, are determined by their moments. The only common distribution not determined by its moments is the lognormal distribution.

1.29.9 Limiting Distributions of Extremes

The limiting distributions of sample extremes for iid observations is completely different from that of central order statistics. For one thing, the limiting distributions of extremes are never normal. General references for this section are Galambos (1977), Reiss (1989), Sen and Singer (1993), and DasGupta (2011). We start with a familiar easy example that illustrates the different kind of asymptotics that extremes have, compared to central order statistics.

Example 1.152. Let $U_1, \ldots, U_n \overset{iid}{\sim} U[0,1]$. Then

$$P(n(1-U_{n:n}) > t) = P(1 - U_{n:n} > \frac{t}{n}) = P(U_{n:n} < 1 - \frac{t}{n}) = (1 - \frac{t}{n})^n, \qquad \text{if } 0 < t < n.$$

So $P(n(1-U_{n:n}) > t) \to e^{-t}$ for all real t, which implies that $n(1-U_{n:n}) \overset{\mathcal{L}}{\Rightarrow} Exp(1)$.

Notice two key things; the limit is nonnormal and the norming constant is n, not $\sqrt{n}$. The norming constant in general depends on the tail of the underlying CDF F.

It turns out that if $X_1, X_2, \ldots$ are iid from some F, then the limit distributions of $X_{n:n}$, if it at all exists, can be only one of three types. Characterizations are available and were obtained rigorously by Gnedenko(1943), although some of his results were previously known to Frechet, Fisher, Tippett, and von Mises.

The usual characterization result, called *the convergence of types theorem*, is somewhat awkward to state and can be difficult to verify. Therefore, we only present more easily verifiable sufficient conditions here. See DasGupta (2011) for a more complete treatment.

We make the assumption that F is continuous.

A few necessary definitions are given first.

Definition 1.77. A CDF F on an interval with $\xi(F) = \sup\{x : F(x) < 1\} < \infty$ is said to have terminal contact of order m if $F^{(j)}(\xi(F)-) = 0$ for $j = 1, \ldots, m$ and $F^{(m+1)}(\xi(F)-) \neq 0$.

Example 1.153. Consider the *Beta* density $f(x) = (m+1)(1-x)^m, 0 < x < 1$. Then the CDF $F(x) = 1 - (1-x)^{m+1}$. For this distribution, $\xi(F) = 1$, and $F^{(j)}(1) = 0$ for $j = 1, \ldots, m$, while $F^{(m+1)}(1) = (m+1)!$. Thus, F has terminal contact of order m.

Definition 1.78. A CDF F with $\xi(F) = \infty$ is said to be of an exponential type if F is absolutely continuous and infinitely differentiable, and if for each fixed $j \geq 2$, $\frac{F^{(j)}(x)}{F^{(j-1)}(x)} \sim (-1)^{j-1}\frac{f(x)}{1-F(x)}$, as $x \to \infty$, where $\sim$ means that the ratio converges to 1.

Example 1.154. Suppose $F(x) = \Phi(x)$, the $N(0,1)$ CDF. Then

$$F^{(j)}(x) = (-1)^{j-1}H_{j-1}(x)\phi(x),$$

where $H_j(x)$ is the j-th Hermite polynomial and is of degree j (see Chapter 12). Therefore, for every j, $\frac{F^{(j)}(x)}{F^{(j-1)}(x)} \sim (-1)^{j-1}x$. Thus, $F = \Phi$ is a CDF of the exponential type.

Definition 1.79. A CDF F with $\xi(F) = \infty$ is said to be of a polynomial type of order k if $x^k(1 - F(x)) \to c$ for some $0 < c < \infty$, as $x \to \infty$.

Example 1.155. All t-distributions, including therefore the Cauchy, are of polynomial type. Consider the t distribution with α degrees of freedom and with median zero. Then, it is easily seen that $x^\alpha(1 - F(x))$ has a finite nonzero limit. Hence, a t_α distribution is of the polynomial type of order α.

We now present our sufficient conditions for convergence in distribution of the maximum to three different types of limit distributions. The first three theorems below are proved in de Haan (2006, pp 15-19); also see Sen and Singer (1993). The first result handles cases such as the uniform on a bounded interval.

Theorem 1.100. Suppose $X_1, X_2, \ldots$ are *iid* from a CDF with m^{th} order terminal contact at $\xi(F) < \infty$. Then for suitable a_n, b_n,

$$\frac{X_{n:n} - a_n}{b_n} \overset{\mathcal{L}}{\Rightarrow} G,$$

where $G(t) = \begin{cases} e^{-(-t)^{m+1}} & t \leq 0 \\ 1 & t > 0. \end{cases}$ Moreover, a_n can be chosen to be $\xi(F)$ and one can choose $b_n = \left\{\frac{(-1)^m(m+1)!}{nF^{(m+1)}(\xi(F)-)}\right\}^{\frac{1}{m+1}}$.

The second result handles cases such as the t distribution.

Theorem 1.101. Suppose $X_1, X_2, \ldots$ are *iid* from a CDF F of a polynomial type of order k. Then for suitable a_n, b_n,

$$\frac{X_{n:n} - a_n}{b_n} \stackrel{\mathcal{L}}{\Rightarrow} G,$$

where $G(t) = \begin{cases} e^{-t^{-k}} & t \geq 0 \\ 0 & t < 0. \end{cases}$ Moreover, a_n can be chosen to be 0 and one can choose $b_n = F^{-1}(1 - \frac{1}{n})$.

The last result handles in particular the important normal case.

Theorem 1.102. Suppose $X_1, X_2, \ldots$ are iid from a CDF F of an exponential type. Then for suitable a_n, b_n,

$$\frac{X_{n:n} - a_n}{b_n} \stackrel{\mathcal{L}}{\Rightarrow} G,$$

where $G(t) = e^{-e^{-t}},\ -\infty < t < \infty$.

Suppose $X_1, X_2, \ldots$ are iid $N(0,1)$. Then

$$\sqrt{2\log n}\Big(X_{n:n} - \sqrt{2\log n} + \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}\Big) \stackrel{\mathcal{L}}{\Rightarrow} G,$$

where $G(t) = e^{-e^{-t}}, -\infty < t < \infty$.

See de Haan (2006, pp 11-12) or Galambos (1977) for a proof. The distribution with the CDF $e^{-e^{-t}}$ is the *Gumbel distribution.*

Example 1.156. (Sample Maximum in Normal Case). The density of the Gumbel distribution is $g(t) = e^{-t}e^{-e^{-t}}, -\infty < t < \infty$. This distribution has mean $m = C$ (the Euler constant), and variance $v^2 = \frac{\pi^2}{6}$. The asymptotic distribution gives us a formal approximation for the density of $X_{n:n}$;

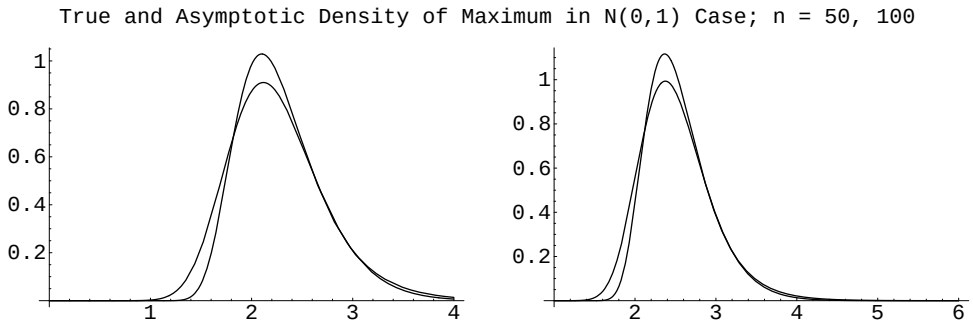
$$\hat{f}_n(x) = \sqrt{2\log n}\ g(\sqrt{2\log n}(x - \sqrt{2\log n} + \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}})).$$

Of course, the true density of $X_{n:n}$ is $n\phi(x)\Phi^{n-1}(x)$. The asymptotic and the true density are plotted here for $n = 50, 100$. The asymptotic density is more peaked at the center, and although it is fairly accurate at the upper tail, it is badly inaccurate at the center and the lower tail. Its lower tail dies too quickly. Hall (1979) shows that the rate of convergence of the asymptotic distribution is extremely slow, in a uniform sense.

Formal approximations to the mean and variance of $X_{n:n}$ are also obtained from the asymptotic distribution. For example,

$$E(X_{n:n}) \approx \sqrt{2\log n} - \frac{\log\log n}{2\sqrt{2\log n}} + \frac{C - \frac{1}{2}\log 4\pi}{\sqrt{2\log n}},$$

$$\mathrm{Var}(X_{n:n}) \approx \frac{\pi^2}{12\log n}.$$



1.29.10 Simulation

Simulation is a computer based exploratory exercise that aids in understanding how the behavior of a random or even a deterministic process changes in response to changes in inputs or the environment. It is essentially the only option left when exact mathematical calculations are impossible, or require an amount of effort that the user is not willing to invest. Even when the mathematical calculations are quite doable, a preliminary simulation can be very helpful in guiding the researcher to theorems that were not apriori obvious or conjectured, and also to identify the more productive corners of a particular problem. We give a short introduction to simulation in this section; specifically, we explain the technique of classic *Monte Carlo* and illustrate some textbook simulation techniques. For For conventional Monte Carlo and techniques of simulation, a few excellent references are Fishman (1995), Ripley (1987) and Robert and Casella (2004).

We start with the ordinary Monte Carlo. The ordinary Monte Carlo is a simulation technique for approximating the expectation of a function $\phi(X)$ for a general random variable X, when the exact expectation cannot be found analytically, or by other numerical means, such as *quadrature*. The basis for the ordinary Monte Carlo is Kolmogorov's SLLN (see Chapter 7), which says that if $X_1, X_2, \cdots$ are iid copies of X, the basic underlying random variable, then $\hat{\mu}_\phi = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$ converges almost surely to $\mu_\phi = E[\phi(X)]$, as long as we know that the expectation exists. This is because, if $X_1, X_2, \cdots$ are iid copies of X, then $Z_i = \phi(X_i), i = 1, 2, \cdots$ are iid copies of $Z = \phi(X)$ and therefore, by the canonical SLLN, $\frac{1}{n}\sum_{i=1}^{n} Z_i$ converges almost surely to $E[Z]$. Therefore, *provided that we can actually do the simulation in practice*, we can simply simulate a large number of iid copies of X and approximate the true value of $E[\phi(X)]$ by the sample mean $\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$. Of course, there will be a *Monte Carlo* error in this approximation, and if we run the simulation again, we will get a different approximated value for $E[\phi(X)]$.

We can also compute *confidence intervals* for the true value of μ_ϕ. By the central limit theorem, for a large Monte Carlo size $n, \hat{\mu} \approx N(\mu, \frac{\sigma^2}{n})$, where $\sigma^2 = \text{Var}(\phi(X))$. This CLT result can be used to produce a confidence interval at a given confidence level; see DasGupta (2011) for the details.

In the special case that we want to find a Monte Carlo estimate of a probability $P(X \in A)$, the function of interest becomes an indicator function $\phi(X) = I_{X \in A}$, and $Z_i = I_{X_i \in A}, i = 1, 2, \cdots, n$. Then, the Monte Carlo estimate of $p = P(X \in A)$ is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} Z_i = \frac{\#\{i : X_i \in A\}}{n}.$$

Example 1.157. (A Cauchy Distribution Expectation). Suppose we want to evaluate $\mu = E(\log |X|)$, where $X \sim C(0, 1)$, the standard Cauchy distribution. Actually, the value of μ can be analytically calculated, and $\mu = 0$ (this is a chapter exercise). We use the Monte Carlo simulation method to approximate μ, and then we investigate its accuracy. For this, we will calculate the Monte Carlo estimate for μ itself, and then a 95% t confidence interval for μ. We use four different values of the Monte Carlo sample size $n, n = 100, 250, 500, 1000$, to inspect the increase in accuracy obtainable with an increase in the Monte Carlo sample size.

n	Monte Carlo Estimate of $\mu = 0$	95% confidence interval
100	.0714	$.0714 \pm .2921$
250	-.0232	$-.0232 \pm .2185$
500	.0712	$.0712 \pm .1435$
1000	.0116	$.0116 \pm .1005$
hline		

The Monte Carlo estimate itself oscillates. But the confidence interval gets tighter as the Monte Carlo sample size n increases. Only for $n = 1000$, the results of the Monte Carlo simulation approach even barely acceptable accuracy. *It is common practice to use n in several thousands when applying ordinary Monte Carlo for estimating one single μ_ϕ. If we have to estimate μ_ϕ for several different choices of ϕ, and if the functions ϕ have awkward behavior at regions of low density of X, then the Monte Carlo sample size has to be increased.*

Example 1.158. (Monte Carlo Evaluation of π). In numerous probability examples, the number π arises in the formula for the probability of some suitable event A in some random experiment, say $p = P(A) = h(\pi)$. Then, we can form a Monte Carlo estimate for p, say $\hat{p}$, and estimate π as $h^{-1}(\hat{p})$, assuming that the function h is one-to-one. This is not a very effective method, as it turns out. But the idea has an inherent appeal that we describe such a method in this example.

Suppose we fix a positive integer N, and let X, Y be independent discrete uniforms on $\{1, 2, \cdots, N\}$. Then, it is known that

$$\lim_{N \to \infty} P(X, Y \text{ are coprime}) = \frac{6}{\pi^2}.$$

Here, coprime means that X, Y do not have any common factors > 1. So, in principle, we may choose a large N, choose n pairs (X_i, Y_i) independently at random from the discrete set $\{1, 2, \cdots, N\}$, and find a Monte Carlo estimate $\hat{p}$ for $p = P(X, Y \text{ are coprime})$, and invert it to form an estimate for the value of π as

$$\hat{\pi} = \sqrt{\frac{6}{\hat{p}}}.$$

The table below reports the results of such a Monte Carlo experiment.

N	n	Monte Carlo Estimate of $\pi = 3.14159$
500	100	3.0252
1000	100	3.0817
1000	250	3.1308
5000	250	3.2225
5000	500	3.2233
10000	1000	3.1629
10000	5000	3.1395

This is an interesting example of the application of Monte Carlo where two indices N, n have to simultaneously take large values. Only when $N = 10,000$ and $n = 5,000$, we come close to matching the second digit after the decimal.

The entire Monte Carlo method is based on the assumption that we can in fact simulate the Monte Carlo sample observations $X_1, \cdots, X_n$ from whatever distribution is the relevant one for the given problem. We give a basic description of a few widely used textbook simulation methods,

We start with the already familiar method of using the quantile transformation. Suppose F is a continuous CDF on the real line with the quantile function F^{-1}. Suppose $X \sim F$. Then $U = F(X) \sim U[0, 1]$. Therefore, to simulate a value of $X \sim F$, we can simulate a value of $U \sim U[0, 1]$ and use $X = F^{-1}(U)$. As long as the quantile function F^{-1} has a formula, this method will work for any one dimensional random variable X *with a continuous CDF*.

Example 1.159. (Exponential and Beta). Suppose we want to simulate a value for $X \sim Exp(1)$. The quantile function of the standard Exponential distribution is $F^{-1}(p) = -\log(1-p), 0 < p < 1$. Therefore, to simulate X, we can generate $U \sim U[0, 1]$, and use $X = -\log(1-U)$ ($-\log U$ will work too).

As another example, suppose we want to simulate $X \sim Be(\frac{1}{2}, \frac{1}{2})$, the Beta distribution with parameters $\frac{1}{2}$ each. The density of the $Be(\frac{1}{2}, \frac{1}{2})$ distribution is $\frac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1$. By direct integration, we get that the CDF is $F(x) = \frac{2}{\pi} \arcsin(\sqrt{x})$. Therefore, the quantile

function is $F^{-1}(p) = \sin^2(\frac{\pi}{2}p)$, and so, to simulate $X \sim Be(\frac{1}{2}, \frac{1}{2})$, we generate $U \sim U[0,1]$, and use $X = \sin^2(\frac{\pi}{2}U)$.

A hugely popular simulation technique is the *accept-reject* method. This method is useful when it is difficult to directly simulate from a target density $f(x)$ on the real line, but we can construct another density $g(x)$ such that $\frac{f(x)}{g(x)}$ is uniformly bounded, and it is much easier to simulate from g. Then we do simulate X from g, and retain it or toss it according to some specific rule. The set of X values that are retained may be treated as independent simulations from the original target density f. Since an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept-reject method. The density g is called the *envelope density*.

The method proceeds as follows:

(a) Find a density function g and a finite constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x.

(b) Generate $X \sim g$.

(c) Generate $U \sim U[0,1]$, independently of X.

(d) Retain this generated X value if $U \leq \frac{f(X)}{cg(X)}$.

(e) Repeat the steps until the required number of n values of X have been obtained.

The following result justifies this indirect algorithm for generating values from the actual target density f.

Theorem 1.103. Let $X \sim g$, and U, independent of X, be distributed as $U[0,1]$. Then the conditional density of X given that $U \leq \frac{f(X)}{cg(X)}$ is f.

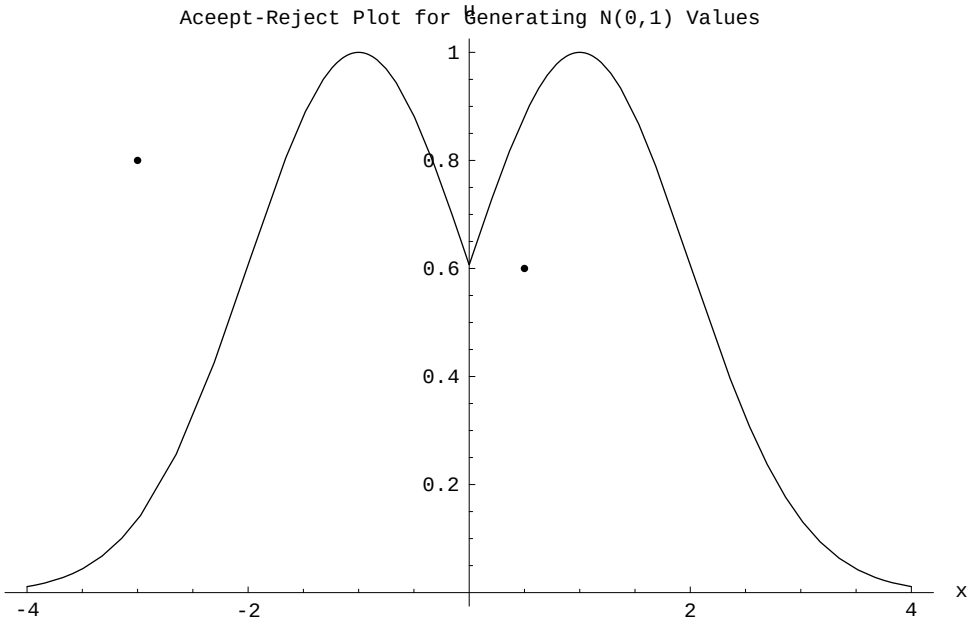
Proof Denote the CDF of f by F. Then,

$$P(X \leq x \,|\, U \leq \frac{f(X)}{cg(X)})$$

$$= \frac{P(X \leq x, U \leq \frac{f(X)}{cg(X)})}{P(U \leq \frac{f(X)}{cg(X)})} = \frac{\int_{-\infty}^{x} \int_{0}^{\frac{f(t)}{cg(t)}} g(t) du dt}{\int_{-\infty}^{\infty} \int_{0}^{\frac{f(t)}{cg(t)}} g(t) du dt}$$

$$= \frac{\int_{-\infty}^{x} \frac{f(t)}{cg(t)} g(t) dt}{\int_{-\infty}^{\infty} \frac{f(t)}{cg(t)} g(t) dt} = \frac{\int_{-\infty}^{x} f(t) dt}{\int_{-\infty}^{\infty} f(t) dt} = \frac{F(x)}{1} = F(x).$$

Example 1.160. (Generating from Normal via Accept-Reject). Suppose we want to generate $X \sim N(0,1)$. Thus our target density f is just the standard normal density. Since there is no formula for the quantile function of the standard normal, the quantile transform method is usually not used to generate from a standard normal distribution. We can, however, use the accept-reject method to generate standard normal values. For



this, we need an envelope density g such that $\frac{f(x)}{g(x)}$ is uniformly bounded, and furthermore, it should be easier to sample from this g.

One possibility is to use the standard double exponential density $g(x) = \frac{1}{2}e^{-|x|}$. Then,

$$\frac{f(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}{\frac{1}{2}e^{-|x|}}$$

$$= \sqrt{\frac{2}{\pi}}e^{|x|-x^2/2} \leq \sqrt{\frac{2e}{\pi}}$$

for all real x, by elementary differential calculus.

We take $c = \sqrt{\frac{2e}{\pi}}$ in the accept-reject scheme, and of course, g as the standard double exponential density. The scheme works out to the following: generate $U \sim U[0,1]$, and a double exponential value X, and retain X if $U \leq e^{|X|-X^2/2-\frac{1}{2}}$. Note that a double exponential value can be generated easily by several means:

(a) generate a standard exponential value Y and assign it a random sign (+ or − with an equal probability);

(b) generate two independent standard exponential values Y_1, Y_2 and set $X = Y_1 - Y_2$;

(c) use the quantile transform method, as there is a closed form formula for the quantile function of the standard double exponential.

It is helpful to understand this example by means of a plot. The generated X value is retained if and only if the pair (X, U) is *below* the graph of the function in the plot above, namely the function $u = e^{|x|-x^2/2-\frac{1}{2}}$. We can see that one of the two generated values will be accepted, and the other rejected.