STA256: Probability and Statistics I

Kris Aujla - kris.aujla@mail.utoronto.ca

Summer 2025

Extensive course notes covering all material. Includes chapters 1-5, practice problem solutions from each chapter at the end of each section (from the list), slides and tutorial problems organized by chapter. This is constantly being updated and tweaked, for any mistakes contact me.

Entirety of chapters 1-4 completed in first-sub session

Textbook: Introduction to Mathematical Statistics (8th Edition) by Rober V. Hogg, Joseph W. McKean and Allen T. Craig, 2019.

Contents

1 Probability and Distributions

1.1 Introduction, Motivation and Intuition

We begin in this section to discuss, informally, a probability model which will be formalized later. Imagine you are a curious investigator that wants to investigate various things. You decide that if we want any significant conclusions, we must have some sort of repeated experimentation, where each experiment happens under the exact same conditions. For example, let's say that you decide to investigate the effect of a new drug and have some way to ensure the conditions are identical for each repeated iteration. However you notice that the outcomes are erratic and unpredictable. Or let's say you want to test the effect that a chemical fertilizer has on the yield of a cereal grain. You notice that each experiment results with a *outcome*. The key idea to notice here is that the outcomes of these experiments could not have been predicted with certainty prior to the experiment.
Now suppose you have created some experiment in such a way that you can obtain the collection of all possible outcomes before the experiment happens and you can ensure the conditions stay the same. This type of experiment is called a **Random Experiment**, and the collection of possible outcomes are called the **sample space**.

Definition 1.1.1

A random experiment is an act or process of observation that leads to a single outcome which cannot be predicted with certainty.

Definition 1.1.2

The sample space of an experiment is the set of all possible outcomes and is denoted by S.

Example 1.1.3

In a toss of a coin, let us define T to be tails and H to be heads. If we assume that we can repeatedly toss the coins under the same conditions, then this is an example of a random experiment. Since we know the possible outcomes, the sample space is $\mathrm{S} = \{T, H\}$

Example 1.1.4

In the game of shooting dice, with one red dice and one white dice. If we assume that we can shoot the dice under the exact same conditions everytime, then we have a random experiment. Let the set of outcomes, or sampling space, be given by the cartesian product $\mathrm{S} = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} = \{(1,1), \ldots, (1,6), (2,1), \ldots (6,6)\}$ where each element is the ordered pair corresponding to the number on the red and white dice respectively. The sampling space has 36 ordered pairs.

We usually use lowercase letters $a, b, c \ldots$ to denote the elements in our sampling space. We are sometimes interested in subsets of the sampling space. We denote these subsets of the sampling space with uppercase letters $(A, B, C, \ldots)$ and are usually called events. If an outcome from an experiment is an element in event A, we say that event A occurred.

Definition 1.1.5

Let S be a sample space of a random experiment. We call the subset $A \subseteq S$ event A. This event A is called simple if it contains only one sample point, that is $|A| = 1$. The event is called compound otherwise

Note there are variations to the above definition such as when the sample space is discrete, that is when the sample space is countable. Next, suppose that we have a random experiment that we have repeated N times. We can count the total number of times, f, that event A occurred. We are interested in the chances that event A occurs. The ratio f/N is called the relative frequency of event A in N experiments. However for small N, we can expect that the relative frequency will be erratic and unstable. Let's consider what happens when N increases. If we associate some number p to event A to which at this point the relative frequency starts to stabilize. Then p can be thought of as the number that after future iterations of the experiment, the relative frequency of event A will be approximately equal to p. Thus, although we cannot predict the outcome of a random experiment, we can, for a large value of N, predict approximately the relative frequency with which the outcome will be in A. This is formalized in the next section.

Example 1.1.6

Going back Example 1.1.4, let S be the discrete sample space of the random experiment. Let B be the set of all ordered pairs where the sum of both entries equals to 7. Here B is an event that is compound. Thus $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$. Assume that the dice are shot 1000 times and let f be the frequency that the sum equals seven. Assume that when $N = 1000$, $f = 160$. Then the relative frequency is $\frac{160}{1000} = 0.16$. Thus using a relative frequency approach we might say that the approximate probability of event B is some number p that is close to 0.16.

Remark 1.1. In the **relative frequency approach**, probability is based on repeating an experiment many times under the same conditions. The probability of an event is calculated by looking at how often the event happens over the long run.

Example: If you flip a fair coin many times, about half the flips will be heads. So, the probability of heads is about 0.5.

In the **subjective approach**, probability is a measure of personal belief about how likely an event is to happen. It does not require repeating the experiment many times.

For example, if you believe there is a 40% chance it will rain tomorrow, you can say the probability is 0.4, based on your judgment.

If you are a gambling man then you believe the probability of an event A is $\frac{2}{5}$, then you should be willing to bet with odds that match that belief:

$$\frac{p}{1-p} = \frac{2/5}{3/5} = \frac{2}{3}$$

This means you would be comfortable betting:

- Win 3 units if A happens, and lose 2 units if it does not, or
- Win 2 units if A does not happen, and lose 3 units if it does.

The math of probability works the same way whether you use the relative frequency or subjective belief approach.

1.2 Sets

I'm assuming you are familiar with what sets are and all the operations we can do with them including some set functions. There are some practice for set functions like integrating, its all intuitive and you can check the textbook for extra practice. Moreover throughout the rest of these notes I will be using topics from set theory that is taught in first year proof based courses such as MAT102, MAT137 or MAT157. Such topics include cardinality and powersets. So in this section we will go over the more abstract or new things that you might have not done. We begin we DeMorgan's Laws.

Theorem 1.2.1: DeMorgan's Laws

Let A and B be two sets and U be the universe of discourse. Then we have that

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c$$

Proof. There are a few ways to prove this, I'll do it using double subset inclusion. Let $x \in (A \cup B)^c$. Then $x \in U$ and $x \notin (A \cup B)$. This means $x \notin A$ **and** $x \notin B$. Thus by definition of complement this means $x \in A^c$ and $x \in B^c$. By definition of intersection we have that $x \in A^c \cap B^c$. Thus $(A \cup B)^c \subseteq A^c \cap B^c$. Let $x \in A^c \cap B^c$. Then by definition of complement, $x \notin A$ and $x \notin B$. Thus $x \notin A \cup B$. By definition then $x \in (A \cup B)^c$. Thus $A^c \cup B^c \subseteq (A \cap B)^c$. Thus we have shown that $(A \cup B)^c = A^c \cap B^c$. The second equality is done similarly and left as an exercise. □

Proposition 1.2.2: Distributive Law's

Let A, B, C be three sets. Then we have that

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

and

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Definition 1.2.3: Nondecreasing and Nonincreasing Sets

A sequence of sets $\{C_n\}$ is an increasing sequence if $C_n \subseteq C_{n+1}$ for all n, in which we write

$$\lim_{n\to\infty} C_n = \bigcup_{n=1}^{\infty} C_n.$$

A sequence of sets $\{C_n\}$ is a decreasing sequence if $C_{n+1} \subseteq C_n$ for all n, in which we write

$$\lim_{n\to\infty} C_n = \bigcap_{n=1}^{\infty} C_n.$$

Example 1.2.4:

Show that the sequence of sets $C_k = \{x : \frac{1}{k} \leq x \leq 3 - \frac{1}{k}\}$ for $k = 1, 2, 3, \ldots$ is nondecreasing and find

$$\lim_{k\to\infty} C_k.$$

Solution Listing out the first few sets we see that

$$C_1 = \{x : 1 \leq x \leq 2\}$$
$$C_2 = \{x : \frac{1}{2} \leq x \leq 2.5\}$$
$$C_3 = \{x : \frac{1}{3} \leq x \leq 2.6\overline{6}\}$$

We can see that C_1 is a subset of C_i for all $i \in \{1, 2, 3\}$. To see that this is nondecreasing look at the figure below.

More generally we see that $C_n \subseteq C_{n+1}$ for all $n \in \mathbb{N}$. To show this, notice that

$$C_n = \{x : \frac{1}{n} \leq x \leq 3 - \frac{1}{n}\}$$

and

$$C_{n+1} = \{x : \frac{1}{n+1} \leq x \leq 3 - \frac{1}{n+1}\}$$

Let $x \in C_n$. Then since $n + 1 \geq n$ we have

$$\frac{1}{n+1} \leq \frac{1}{n}$$

Since

$$\frac{1}{n+1} \leq \frac{1}{n} \leq x$$

We have the first part of the inequality. Moreover from above we also know that

$$3 - \frac{1}{n+1} \geq 3 - \frac{1}{n} \geq x$$

Thus we have that

$$\frac{1}{n+1} \leq x \leq 3 - \frac{1}{n+1}$$

which means that $x \in C_{n+1}$. Thus we have shown that $C_n \subseteq C_{n+1}$. Since n was arbitrary this holds for all n. So C_k is a nondecreasing sequence of sets. We now find the limit. First notice that since the sets are getting bigger, the limit will just be the union. Since the C_k are intervals of real numbers and the bounds of C_k depend on k, we can easily find the lower and upper bound. For the lower bound we have that

$$\lim_{k\to\infty} \frac{1}{k} = 0$$

and the upper bound we have that

$$\lim_{k\to\infty} 3 - \frac{1}{k} = 3$$

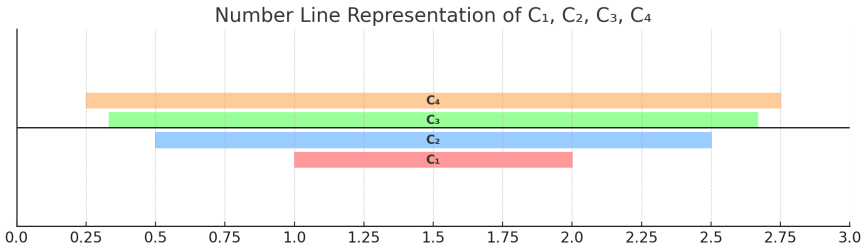


Figure 1: A number line showing the intervals C_1, C_2, C_3, C_4

Thus as we keep adding C_k to the union we get a bigger and bigger where the lower and upper bounds are approaching 0 and 3 respectively. Thus

$$\lim_{k\to\infty} C_k = \{x : 0 < x < 3\}$$

□

Note in the above example we do not include the bounds since none of the C_k ever attain 0 as a lower bound or 3 as a upper bound.

Example 1.2.5

Show that the sequence of sets $C_k = \{x : 2 < x \leq 2 + \frac{1}{k}\}$ for $k = 1, 2, 3, \ldots$ is nonincreasing and find

$$\lim_{k\to\infty} C_k.$$

Solution We begin with showing $C_{n+1} \subseteq C_n$ for all $n \in \mathbb{N}$. Let $x \in C_{n+1}$. Then we have that

$$2 < x \leq 2 + \frac{1}{n+1}$$

Since $\frac{1}{n+1} \leq \frac{1}{n}$ we have that $x \leq 2 + \frac{1}{n+1} \leq 2 + \frac{1}{n}$. Thus $2 < x \leq 2 + \frac{1}{n}$ and so $C_{n+1} \subseteq C_n$. To find the limit we see that the as n increases, the sets are getting smaller, so we need to find the lower and upper bound of that final limit set. Since Only the upper bound changes we simply evaluate the following limit for the upper bound

$$\lim_{k\to\infty} 2 + \frac{1}{k} = 2$$

Thus we have that

$$\lim_{k\to\infty} C_k = \{x : 2 < x \leq 2\} = \emptyset$$

□

1.3 The Probability Set Function

Before we go any further, we have to discuss the nagging issue of our sample space that a few of you may have noticed. More specifically the size of our sample space. If our sample space has only finitely many outcomes, then we can pretty easily assign probabilities to events (subsets) of our sample space. However what happens once our sample space is infinite? For example if our sample space has an uncountable number of possible outcomes, how would we possibly assign probabilities

to all subsets? So obviously we need our collections of events (subsets) to be nice enough so that we can do this. We have conditions that allow the collections of subsets to nice enough so that we can work with it. These well behaved collection of subsets are called $\sigma-$algebra. We only assign probabilities to a well-behaved collection of subsets of the sample space, denoted $\sigma-$algebra.

Definition 1.3.1

Let S be our sample space for a random experiment. A collection of events (subsets) $\mathcal{B} \subseteq \mathcal{P}(S)$ is well-behaved if and only if

1. It includes all the events we care about
2. is closed under complements (if an event is in the collection, then "not that event" is too)
3. Is closed under countable unions (if you take a countable number of events in the collection and unite them, the result is still in the collection),
4. And, using DeMorgan's Laws, it's also closed under countable intersections.

This collection is sometimes called $\sigma-$ algebra. Note that the use of this definition may not be consistently used in later definitions/theorems, however it should implicitly be known that this is always true for our collection of events we are investigating.

So now that we have our sample space S and our collection of well-behaved events $\mathcal{B}$, we can begin to assign probabilities to each event in $\mathcal{B}$ using a set function using the relative frequency approach we discussed earlier. However, before we do this, we need some axioms to have a solid foundation. These axioms are quite trivial but important. The first is that the probability of any event must be greater than or equal to zero for any N repetitions of a random experiment. The next is that the relative frequency of the entire sample space must be 1 since it includes all possible outcomes and something in our sample space must occur every time we do the experiment. The last is that if we have two events that are disjoint, the relative frequency of the union is equal to the sum of both of the two events. We now finally formalize the definition of probability using our relative frequency approach with the Probability Set Function.

Definition 1.3.2 : Probability Axioms

Let S be a sample space and $\mathcal{B}$ be the collection of well behaved events. Let P be a real valued function defined on $\mathcal{B}$. Then P is a probability set function if and only if

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If A_n is a sequence of events in $\mathcal{B}$ and is pairwise mutally exlusive ($A_m \cap A_n = \emptyset$ for all $n \neq m$), then we have that

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$$

If the disjoint union of these sets are equal to the sample space, then we say that the collection of events $\mathcal{B}$ is exhaustive, in which case,

$$\sum_{n=1}^{\infty} P(A_n) = 1$$

We also sometimes say that if a collection of events is mutually exclusive and exhaustive, it forms a partition of S.

Using this definition we can immediately come up with a few theorems. Note that only Theorem 1.3.3 and 1.3.7 are in the slides. For the first theorem, I will try the explain intuition. Suppose you are a weather forecaster and you believe that there is a probability of 0.25 that it will rain next week. Let A denote this outcome. In other words we have that $P(A) = 0.25$. Well then isn't it equivalent for you to tell your viewers that the probability that it will **not** rain is 0.75? In other words, since A denotes the outcome of raining, then A^c denotes the outcomes where it does not rain. So we have that $P(A^c) = 0.75$. Together we get that $P(A) = 1 - P(A^c)$.

Theorem 1.3.3

Let S be a sample space and let $\mathcal{B}$ denote the collection of well behaved events. Then for each event $A \in B$,

$$P(A) = 1 - P(A^c)$$

Proof. Notice that we have $S = A \cup A^c$. We know that $A \cap A^c = \emptyset$. So we have then from Probability Axioms part 3, $P(A \cup A^c) = P(A) + P(A^c)$. Moreover, from part 1 we know that $P(S) = 1$. Together we get that

$$1 = P(A) + P(A^c)$$

or

$$P(A) = 1 - P(A^c)$$

as needed. □

Next is the probability of the empty set is zero.

Theorem 1.3.4

The probability of the empty set is zero. That is $P(\emptyset) = 0$

Proof. From Theorem 1.3.3, if we let $A = \emptyset$ then we have

$$P(A) = 1 - P(\emptyset^c)$$

The complement of the empty set is just the entire sample space so $P(\emptyset^c) = P(A) = 1$, where the last equality we used Axiom 1 from Probability Axioms. Thus we get that $P(\emptyset) = 1 - 1 = 0$ as needed. □

For the next theorem, if we have two events A and B, and the event B contains more possible outcomes than A, then the probability for B must be greater than or equal to the probability of A. The proof of this theorem requires an identity with sets that you may or may not know, so it is not important for our needs. Nontheless we will prove it.

Theorem 1.3.5

Let S be a sample space and $\mathcal{B}$ a collection of well-behaved events. If $A \subset B$ then $P(A) \leq P(B)$.

Proof. Notice that $B = A \cup (A^c \cap B)$. That is since $A \subset B$, B is really just all of A plus all the stuff that is not in A but also in B. Thus clearly $A \cap (A^c \cap B) = \emptyset$ since using distributivity, $A \cap (A^c \cap B) = (A \cap A^c) \cup (A \cap B)$. We cannot have something that is in A and not in A. So from Probability Axioms part 3 we have that

$$P(B) = P(A) + P(A^c \cap B)$$

From part 1 of Probability Axioms we have that $P(A^c \cap B) \geq 0$. Thus we get that $P(B) \geq P(A)$ as needed. □

From this we get a simple lemma.

Lemma 1.3.6

Let S be a sample space and $\mathcal{B}$ a collection of well-behaved events. For each event $A \in \mathcal{B}$, $0 \leq P(A) \leq 1$.

Proof. Notice that $\emptyset \subset A \subset S$. From Theorem 1.3.5 and probability axioms we have that

$$P(\emptyset) \leq P(A) \leq P(S) \text{ or } 0 \leq P(A) \leq 1$$

□

Now for our final theorem we discuss the probability of the union of two events. We know that if A, B are two events that are mutually exclusive, then the probability of the union of these two events is just the sum of the respective probabilities. However what if we are considering two arbitrary events that may or may not be mutually exclusive. Is there some way to connect the probability of the union of these events with their respective probabilities? The answer is yes. If you are familiar with linear algebra, the formula looks similar to the dimension theorem for subspaces. Interesting.

I will also try to explain the intuition. Suppose we are trying to find the probability of picking something out of a hat. Suppose that A and B are two events that have some common elements between them. For the sake of the explanation, assume there is only common element between them x. Lets say we are interested in finding the probability of $A \cup B$. We know the probability of choosing something from event A or event B is going to be $P(A) + P(B)$, as we are basically making the desired pool elements we are choosing from the hat bigger. However there is some redundancies here. Since x is in both the pool of event A and event B, then we are basically adding the probability of choosing element x from the hat twice. So to combat this we simply subtract the redundancies. That is we subtract $P(A \cap B) = P(\{x\})$. Thus we get that the final probability of choosing something from either event A or event B is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Theorem 1.3.7 : The Addition Law of Probability

Let S be the sample space and $\mathcal{B}$ be a well behaved collection of events. Let $A, B \in \mathcal{B}$. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. To do this proof, using some identities from set theory will help us immensely. First notice that $A \cup B = A \cup (A^c \cap B)$ and $B = (A^c \cap B) \cup (A \cap B)$. Using the Probability Axiom part 3, since each part of the union of both of these sets are disjoint we have that

$$P(A \cup B) = P(A) + P(A^c \cap B) \text{ and } P(B) = P(A^c \cap B) + P(A \cap B)$$

Solving for $P(B)$ and substituting we get

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

as needed. □

Using these theorems of the Probability Set Function we now complete some examples.

Example 1.3.8

Let S be the sample space for our red/white die experiment in Example 1.1.4. The probability set functions assigns a probability of $\frac{1}{36}$ for each of the 36 points in S. If $A = \{(1,1),(3,3),(4,4)\}$ and $B = \{(1,2),(1,3),(1,4)\}$, then $P(A) = \frac{3}{36}, P(B) = \frac{3}{36}$. Then since $P(A \cap B) = 0$, we have that $P(A \cup B) = \frac{1}{6}$

Example 1.3.9

Let S be the sample space for our coin toss experiment from Example 1.1.3. This time two coins are to be tossed and the outcome is the ordered pair (face on the first coin, face on the second coin). Thus the sample space may be represented as $S = \{(H,H),(H,T),(T,H),(T,T)\}$. Let the probability set function assign a probability of $\frac{1}{4}$ to each element of S. Then Let $A = \{(H,H),(H,T)\}$, $B = \{(H,H),(T,H)\}$. Then $P(A) = P(B) = \frac{1}{2}$ and $P(A \cap B) = \frac{1}{4}$. So from Theorem 1.3.7 we have that $P(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$

Example 1.3.10

Gene expression profiling is a state-of-the-art method for determining the biology of cells. In Briefings in Functional Genomics and Proteomics (Dec. 2006), biologists at Pacific Northwest National Laboratory reviewed several gene expression profiling methods. The biologists applied two of the methods (A and B) to data collected on proteins in human mammary cells. The probability that the protein is cross-referenced (i.e., identified) by method A is 0.41, the probability that the protein is cross-referenced by method B is 0.42, and the probability that the protein is cross-referenced by both methods is 0.40.

1. Find the probability that the protein is cross-referenced by either method A or method B.
2. On the basis of your answer to part a, find the probability that the protein is not cross-referenced by either method.

Solution Let $S = \{\text{proteins identified by method } A, B, \text{ both, or neither}\}$. For 1, notice that we assign a probability of a protein being identified by method A of 0.41, so $P(A) = 0.41$. We do the same for method B and get $P(B) = 0.42$. The probability of the protein being identified by both method A and method B is 0.40 (Note here A and B refer to methods but we are using the probability set function on them. This means that A and B also refer to the set of proteins that have been cross referenced by method A and B). This means that the proteins $p \in S$ that the outcome of method A and method B is in the set $A \cap B$. So $P(A \cap B) = 0.40$. Thus the probability of a protein being identified by either method A or method B is $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.43$. For 2, we need to find the probability that the protein is not cross referenced by either method. This means event that we want to look for, contains no proteins that have been identified by method A or method B. Thus any protein that resides within $A \cup B$ has been identified by method A or B, so we take the complement of that set. So we need to find $P((A \cup B)^c)$. We know that probability of a protein being identified by method A and method B is 0.43 from part 1, so it is equivalent to say that the probability of **not** getting a protein identified by method A or method B is $1 - 0.43$ by Theorem 1.3.3. Thus $P((A \cup B)^c) = 1 - 0.43 = 0.57$ □

1.3.1 Sample Point Method

Now we are going to talk about assigning probabilities to a finite sample space. Since our sample space is finite, then we can determine and list out what the sample points are. Let $x_1, x_2, \ldots, x_m$ be the sample points. That is what the simple events are. These simple events will form our sample space $S = \{x_1, x_2, \ldots, x_m\}$. In order to assign probabilities to each of the simple events, we must make sure that our simple events are mutually exclusive (which they trivially are since they are simple), and that they are exhaustive (the union forms the set of all possible outcomes). Then for each simple event, we assign a probability. Again we have to be careful as in order for the probability set function to work, we need to abide to the Probability Axioms. So for each simple event, the probability must be greater than or equal to zero, and the summation of all these probabilities must equal 1. Formally, what we do is let $p_1, p_2, \ldots, p_m$ be numbers such that $0 \leq p_i \leq 1$ for all $i \in \{1, 2, \ldots, m\}$ and $\sum_{i=1}^{m} p_i = 1$. Then let A be any event from S. The probability of A is then $P(A) = \sum_{x_i \in A} p_i$. That is if $A = \{x_1, x_2\}$ then $P(A) = p_1 + p_2$. Then we have that $P(A) \geq 0$ and $P(S) = 1$. Moreover since S is finite, then for any two events A and B, $P(A \cup B) = P(A) + P(B)$, since $A \cap B = \emptyset$.

Example 1.3.11

Our experiment will be a six sided dice roll. Assume we have some way to roll the dice under the same conditions every time. Let E_i be that we observed a i for all $i \in \{1, 2, \ldots, 6\}$. This is a simple event. Thus the sample space is $S = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6$. We assign a probability $p_i = \frac{1}{6}$ for each i. Then notice that $P(S) = P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5) + P(E_6) = 1$

Definition 1.3.12

Let $C = \{x_1, x_2, \ldots, x_m\}$ be a finite sample space. Let $p_i = 1/m$ for all $i = 1, 2, \ldots, m$, and for all subsets A of C define

$$P(A) = \sum_{x_i \in A} \frac{1}{m} = \frac{|A|}{m},$$

where $|A|$ denotes the number of elements in A. Then P is a probability on C and it is referred to as the **equilikely case**.

There are a few pros and cons to this methods. The good thing is that its direct and straightforward as we can easily figure out the probability of any event. However since the sample space is finite, it can be prone to human error. This means not accounting for all possible outcomes in the experiment, or not defining the correct probability to each simple event. You can imagine why this method would impractical for large sample spaces such as a size of 1000 or even a 1,000,000. But the thing is that we are usually interested in finite sample spaces and assigning probabilities. For example, imagine we have a shuffled deck of 52 cards and we decide to pull 5 cards from it. Let's say you are interested in finding the probability of getting a pair (two of a kind) in a hand of five cards from our deck. To do this, we only need to know the total number of 5 card hands and the total number of two of a kind 5 card hands. That is we only need to know the number elements in our event to find the probability. To do this we need to learn about combinatorial analysis.

1.3.2 Counting Rules

We begin with the mn-rule or multiplication rule. We will explain this with an example. Suppose you are building an ice cream cone. You have three types of cones: waffle, sugar and plain. Here our $m = 3$. You also have 4 flavors of ice cream: vanilla, chocolate, strawberry, mint. Here our $n = 4$. You are wondering, how many cone/flavor combinations are there? Well what we can do is for each cone type, we list out all the possible cone/flavor combinations. For example f you choose waffle, you can pair it with: waffle + vanilla, waffle + chocolate , waffle + strawberry and waffle + mint. Counting all of them we get that the answer is 12. In other words what we did was, heuristically, 1 times 4 + 1 times 4 + 1 times 4 or 4×3. In general we see that if a task can be done in m ways, and for each of those ways a second task can be done in n ways, then the entire two-step process can be done in $m \times n$ ways. For those set theorist out there, this is exactly a cartesian product between sets.

Theorem 1.3.13

Given an experiment that consist of two steps, the first with m steps. Denote it $A = \{x_1, \ldots, x_m\}$ and another step that consist of n steps. Denote it $B = \{y_1, \ldots, y_n\}$, then the total possible outcomes form the set

$$A \times B = \{(x_1, y_1), \ldots, (x_m, x_n)\}$$

This is known as the mn-rule or multiplication rule.

For those you of wondering, this does not need to hold for two dimensions, as we can extend this to any number of sets we'd like.

For the next rule we go back to our ice cream example. Imagine you've upgrades your ice cream resources and now have 5 flavors, Vanilla, Chocolate, Strawberry, Mint, Mango. Suppose you are making a 3 scoop ice cream cone for your hungry friend. You want to scoop 3 different flavors, in a specific order (because the order of scoops matters — top, middle, bottom). How many different ordered 3-scoop cones can you make? Well using the mn-rule we know that there are $5 \times 5 \times 5 = 125$ different total combinations. To see this, we can for example start with the first scoop being vanilla and second scoop being vanilla. Then for our vanilla-vanilla cone, we simply list out all the ice cream cones we can make by adding the third scoop from our pool of flavors, so (vanilla - vanilla - vanilla), (vanilla - vanilla - chocolate), etc. In order words if we let $A = \{\text{Vanilla}, \text{Chocolate}, \text{Strawberry}, \text{Mint}, \text{Mango}\}$, then we are simply doing $|A \times A \times A| = 125$. However we are interested in ice cream cones where each scoop is a different flavor, no repeats. Well, for the first scoop, there 5 flavors we can choose from for the first scoop, 4 flavors for the second, and 3 flavors for the third. Thus what we have now are three different sets of flavors with different sizes. Using the mn-rule, we know that the total possible combinations are simple $5 \times 4 \times 3 = 60$. We call this a permutation.

Definition 1.3.14

Let A be a set with n elements and suppose we are interested in finding $k-$tuples where each entry is a distinct element from A. Then by $mn-$rule, there are $n(n-1)\cdots(n-k+1))$ such k tuples. We denote this P_k^n (Permutation). Moreover we can also express it as

$$P_k^n = \frac{n!}{(n-k)!}$$

Example 1.3.15 : Birthday Problem

Assume there are $n < 365$ unrelated people in a room. Find the probability of the event A that at least 2 people have the same birthday.

Solution Assign a number 1 to n for each person in the room. Let $A = \{\text{dates in a year}\}$, so $|A| = 365$. We then use $n-$tuples to denote the birthdays of each person in order, so first entry is the birthday of person 1 etc. In order to find the probability that at least 2 people have the same birthday, we first need to know the total number of possible birthday combinations between n people. In other words we need to use the $mn-$rule. We get that there are 365^n possible $n-$tuple birthday combinations. What we did here was find the cardinality of the set $\prod^n A$. Thus this is the size of our sample space. That is $|S| = 365^n$. Now in order to use the probability set function

we need to assign probabilities. We can assume that the probability for a birthday or occur on a certain day is equilikely (Definition 1.3.12). Thus the probability of a single element in our sample space is 365^{-n}. If we let B denote the event that at least 2 people have the same birthday, then the complement of B is the set of birthdays that are distinct. As we know from sample-point method, all we we need now is total number of elements in B^c. Thus since we need each entry to be unique, we simply find the permutation. So by Theorem 1.3.3 we have that

$$P(B) = 1 - P(B^c) = 1 - \frac{P_n^{365}}{365^n}$$

□

We now end with the our final counting rule. Like before, we have a set A with n elements. We are interested in finding, without caring about the order, the total number of subsets with k elements taken from A. I'll derive the formula using an example. Let $A = \{1, 2, 3, 4, 5, 6\}$. We are interested in the total number of subsets with 3 elements taken from A. We will denote $\binom{6}{3}$ as this number. How would we go about this? Well one way would be to take an arbitrary set with 3 elements, B, and find the total number of permutations of that set. That is the total number of ways we can order this set, P_3^3. Once we find this number, we then know for each unique 3 element set from A, will have P_3^3 different ordered arrangements. These ordered arrangements will be different for each unique 3 element set. Now, if look at the total number of permutation of 3 elements from A, that is P_3^6 how can we relate this to $\binom{6}{3}$? Well each unique 3 element subset has 3! permutations. So, if we take all the permutations and group them by which subset they came from, each subset gives us 3! permutations. Thus we get that

$$P_3^6 = \binom{6}{3} \cdot P_3^3$$

Or

$$\binom{6}{3} = \frac{P_3^6}{3! \times 3!}$$

Definition 1.3.16

Given a set A with n elements, the total number of unordered subsets of size k without replacement is

$$C_k^n = \binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!}$$

We usually say there are C_k^n combinations of k things taken from a set of n things.

Like the $mn-$rule we can extend this so that we have a formula for the total number of k subsets from a set with n elements, where each subset has its own unique size and each element only appears once in a subset. The quickest way is this: Choose who goes in group 1: $\binom{n}{n_1}$ ways. Then choose group 2 from what's left: $\binom{n-n_1}{n_2}$. Continue until group $k-1$; the last group is forced.

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{k-1}}{n_k} = \frac{n!}{n_1!\,n_2!\cdots n_k!}$$

Definition 1.3.17

he number of ways of partitioning n distinct objects into k distinct groups containing $n_1, n_2, \ldots, n_k$ objects, respectively, where each object appears in exactly one group and $\sum_{i=1}^{k} n_i = n$, is

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

The terms $\binom{n}{n_1, n_2, \ldots, n_k}$ are called the **multinomial coefficients** because they occur in the expansion of the multinomial term $(y_1 + y_2 + \cdots + y_k)^n$.

Example 1.3.18 : Poker Hands

We go back to our 52 shuffled deck example that motivated our counting rules. Let a card be drawn at random from an ordinary deck of 52 playing cards that has been well shuffled. The sample space S consists of 52 elements, each element represents one and only one of the 52 cards. We are going to assume that the probability of drawing a card is $\frac{1}{52}$. If E_2 are the outcomes of drawing a spade, then $P(E_1) = \frac{|E_1|}{52} = \frac{13}{52}$ because there are 13 spades in the deck. If E_2 is the set of outcomes of drawing a king then $P(E_2) = \frac{|E_2|}{52} = \frac{4}{52}$, since there are 4 kings in the deck. These are quite trivially, so now suppose we draw 5 cards from the deck instead. We do not care about the order so really we are looking at the subset of 5 cards from S. We know there are $\binom{52}{5}$ different poker hands. We will assume the probability of drawing any 5 card hand is $1/\binom{52}{5}$. Going back to our original discussion, we now can compute the probability of drawing certain hands such as the two of a kind or a flush because we have the total number of poker hands and can get the total number of certain hands then easily calculate the probability using our equilikely formula. Let E_1 be the event of drawing a flush. In a flush, all the cards have the same suit. We need to find the total number of flushes in our shuffled deck. There are 4 suits to choose from in a flush, lets say spades. Each suit has 13 cards. We want to find the number of 5 card hands from 13 cards. That is $\binom{13}{5}$. So using the multiplication rule, we have 4 different suits and $\binom{13}{5}$ different hands, so there are $4 \times \binom{13}{5}$ possible flushes. Thus we get that

$$P(E_1) = \frac{4 \times \binom{13}{5}}{\binom{52}{5}} = \frac{4 \times 1287}{2598960} = 0.00198$$

We now find the probability of the event E_2 of drawing a 3 of a kind. That is 3/5 of the cards are the same rank (face value not necessarily same suit) and the other two are distinct. So there are 13 different ranks. Now given our 3 same rank cards, there are $\binom{4}{3}$ combinations of the suits of these 3 cards. And the other 2 distinct cards have $\binom{12}{2} = 66$ different combinations. Then there are 16 possible suits (4×4) for the two distinct cards. So the number of 3 of a kinds in our deck is $13 \times 4 \times 66 \times 16 = 54912$. Thus the probability of drawing a 3 of a kind is

$$P(E_2) = \frac{54912}{2598960} = 0.0211$$

We do more of these in the practice questions at the end of the section.

Example 1.3.19

We have $n = 3$ people (A, B, and C) and $r = 2$ job positions. How many ways (permutations) are there to assign the positions to the people?

Solution Using our formula we get that the answer is

$$P_2^3 = \frac{3!}{(3-2)!} = 3! = 6$$

□

Example 1.3.20

We have $n = 3$ people (A, B, and C)and pick $r = 2$ people to create a committee. How many combination of two types of three?

Solution We are looking for the total number of 2 element subsets. We can use our formula and get that it is

$$\binom{3}{2} = \frac{3!}{2!(3-1)!} = \frac{6}{2} = 3$$

□

Example 1.3.21

You need to assign 9 aircrafts to 3 sorties of size 4, 3, and 2 aircrafts. How many ways can the 9 aircrafts be assigned to the 3 groups?

Solution Let A, B, C be the three sorties with capacity 4,3 and 2 respectively. We start with the number of combinations of 4 aircrafts from 9, $\binom{9}{4}$, then 3 airfracts from 5, $\binom{5}{3}$ and 2 airfracts in a size of 2, $\binom{2}{2}$. Together we get

$$\binom{9}{4} \cdot \binom{5}{3} \cdot \binom{2}{2} = \frac{9!}{4!(5!)} \cdot \frac{5!}{3!(2!)} \cdot \frac{2!}{1(2-2)} = \frac{9!}{4!3!2!}$$

□

To summarize our counting rules we have the following path that could help.

	With replacement	**Without replacement**
Order important	n^r (using MN rule)	Permutations
Order not important	Not discussed	Combinations ($\binom{n}{r}$)

1.3.3 Additional Properties of Probability

We end this section with additional properties of probability. Recall definition 1.2.3 of a sequence of nondecreasing sets. Let $\{C_n\}$ be a nondecreasing sequences of events. Then $C_n \subset C_{n+1}$ for all n, and $\lim_{n\to\infty} C_n = \bigcup_{n=1}^{\infty} C_n$. Consider $\lim_{n\to\infty} P(C_n)$. The question here is that can we bring the limit inside? The answer is yes. We show this by proving the following theorem:

Theorem 1.3.22 : Continuity Theorem of Probability

Let $\{C_n\}$ be a nondecreasing sequence of events. Then

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n) = P(\bigcup_{n=1}^{\infty} C_n)$$

Let $\{C_n\}$ be a nonincreasing sequence of events. Then

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n) = P(\bigcap_{n=1}^{\infty} C_n)$$

Proof. My proof of this uses some topics from calculus 2 sequences. The first thing to notice is that by Theorem 1.3.5, for all n, $P(C_n) \leq P(C_{n+1})$ since $C_n \subset C_{n+1}$. So that means $P(C_n)$ is an increasing sequence. Then we also know that this sequence is bounded above by $\bigcup_{n=1}^{\infty} C_n$ since for all n, $C_n \subset \bigcup_{n=1}^{\infty} C_n$. So that is $P(C_n) \leq P(\bigcup_{n=1}^{\infty} C_n)$. Using the monotone convergence theorem we know this converges to the supremum. Thus

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n) = P(\bigcup_{n=1}^{\infty} C_n)$$

as needed. The second proof is analogous as we basically just replace increasing with decreasing and supremum with infimum. However this proof is not fully correct and we are skipping over some steps that need much more rigor and justification which I don't feel like doing so heres an alternate (boring-er) proof does not require calculus 2 stuff:

We know that $P(\bigcup_{n=1}^{\infty} C_n) = 1$. We also know that $0 \leq P(C_n) \leq 1$. Moreover we know that $P(C_n) = 1 - P(C_n^c)$. Thus we get that

$$0 \leq \lim_{n\to\infty} P(C_n) = \lim_{n\to\infty} 1 - P(C_n^c) \leq 1$$

We end up getting that $\lim_{n\to\infty} P(C_n^c) = 0$, so thus we get that

$$\lim_{n\to\infty} P(C_n) = \lim_{n\to\infty} 1 - P(C_n^c) = 1 - 0 = 1 = P(\bigcup_{n=1}^{\infty} C_n)$$

However this proof isn't always correct and only correct when $\bigcup_{n=1}^{\infty} C_n$ is the entire sample space. So the actual final proof is:

Set

$$\lim_{n\to\infty} C_n = C_\infty := \bigcup_{n=1}^{\infty} C_n.$$

Define

$$D_1 := C_1, \qquad D_n := C_n \setminus C_{n-1} \quad (n \geq 2).$$

Then $D_1, D_2, \ldots$ are pairwise disjoint and, for every m,

$$\bigcup_{k=1}^{m} D_k = C_m, \qquad \bigcup_{k=1}^{\infty} D_k = C_\infty.$$

Using finite additivity for C_m and countable additivity for C_∞ from Probability Axiom part 3,

$$P(C_m) = \sum_{k=1}^{m} P(D_k), \qquad P(C_\infty) = \sum_{k=1}^{\infty} P(D_k).$$

Because the partial sums on the right-hand side form a non-decreasing, bounded sequence,

$$\lim_{m\to\infty} P(C_m) = \lim_{m\to\infty} \sum_{k=1}^{m} P(D_k) = \sum_{k=1}^{\infty} P(D_k) = P(C_\infty).$$

Hence

$$\lim_{n\to\infty} P(C_n) = P\Big(\bigcup_{n=1}^{\infty} C_n\Big).$$

For the next part, assume $C_1 \supseteq C_2 \supseteq \cdots$ and set $C_\infty := \bigcap_{n=1}^{\infty} C_n$.
Consider the complements C_n^c. They form an increasing sequence:

$$C_1^c \subseteq C_2^c \subseteq \cdots, \qquad \bigcup_{n=1}^{\infty} C_n^c = C_\infty^c.$$

Apply part 1 to $\{C_n^c\}$:

$$\lim_{n\to\infty} P(C_n^c) = P\Big(\bigcup_{n=1}^{\infty} C_n^c\Big) = P(C_\infty^c).$$

Finally use $P(A^c) = 1 - P(A)$ from Theorem 1.3.3:

$$\lim_{n\to\infty} P(C_n) = 1 - \lim_{n\to\infty} P(C_n^c) = 1 - P(C_\infty^c) = P(C_\infty).$$

Thus

$$\lim_{n\to\infty} P(C_n) = P\Big(\bigcap_{n=1}^{\infty} C_n\Big).$$

□

To not drag on about this proof, i'll just say:

think of probability as mass spread over outcomes. If you keep adding events $C_1 \subseteq C_2 \subseteq \cdots$, the mass covered can only grow. It seems natural that the mass should level off exactly at the mass of the *ultimate* event $C_\infty = \bigcup_{n=1}^{\infty} C_n$. Continuity-from-below says "yes the limit of the partial masses is the total mass you ultimately cover." Why introduce the pieces D_n? The main difficulty is turning the informal idea above into an equation. A standard trick is to carve the growing sets into disjoint slices*, so that we can directly apply the axiom of countable additivity (Probability Axiom 3).

The slices are really just the new part added for each subsequent event. That is D_1 is just the first event C_1. D_2 is the new part that appears when you pass from C_1 to C_2: $D_2 = C_2 \setminus C_1$. D_3 is the new part added at stage 3, and so on.

By construction the D_n are mutally exclusive so adding up the first m slices rebuilds C_m,

$$\bigcup_{k=1}^{m} D_k = C_m,$$

adding all the slices gives the final union,

$$\bigcup_{k=1}^{\infty} D_k = C_\infty.$$

Another useful result for arbitrary unions is given by Boole's Inequality. The probability that at least one of the events happens is no more than the sum of the individual probabilities. This is especially useful when the events might overlap, so you can't simply add the probabilities exactly.

Theorem 1.3.23 : Boole's Inequality

Let $\{C_n\}$ be an arbitrary sequence of events that are well behaved. Then

$$P(\bigcup_{n=1}^{\infty} C_n) \leq \sum_{n=1}^{\infty} P(C_n)$$

Proof. If our union of events is a finite set then the result follows directly from Theorem 1.3.7. In order to do anything with our union of sets, we either need these sets to be disjoint or nondecreasing/nonincreasing. To achieve this, we define the partial union by

$$B_n = \bigcup_{i=1}^{n} C_i$$

The sequence of our partial unions are a nondecreasing sequence of events. That is $B_n \subset B_{n+1}$. more over we have that

$$\lim_{n\to\infty} B_n = \bigcup_{n=1}^{\infty} C_n$$

We now can use our previous theorems with nondecerasing events. That is using Theorem 1.3.22, we have that

$$P(\bigcup_{n=1}^{\infty} C_n) = \lim_{n\to\infty} P(B_n)$$

Thus from the finite case we get the inequality easily

$$\lim_{n\to\infty} P(B_n) \leq \lim_{n\to\infty} \sum_{i=1}^{n} P(C_i) = \sum_{i=1}^{\infty} P(C_i)$$

as needed. □

Theorem 1.3.7 gave us the formula for the probability of two arbitrary unions. However can extend this to any finite number of arbitrary unions. That is the addition law of The Addition Law of Probability can be extended to k events. A key thing to notice, which you probably have, is that $A_1 \cup A_2 \cup A_3 \cup \ldots = (A_1 \cup A_2) \cup A_3 \cup \ldots$. That is when we have a large string of unions, we evaluate the first two unions then unionize that result with the next set. We can see that this has a sort of inductive nature, so we will prove this theorem by induction.

Theorem 1.3.24 : Inclusion-Exclusion Formula

Let $A_1, A_2, \ldots, A_k$ be k events that are well-behaved. Then

$$P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{k-1} P(A_1 \cap A_2 \cap \cdots \cap A_k)$$

Proof. For our base case $k = 2$ we see that from Theorem 1.3.7,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

We also do $k = 3$. Let $B_1 = A_1 \cup A_2$. Then $P(B_1) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. We now then consider $P(A_1 \cup A_2 \cup A_3) = P(B_1 \cup B_2)$, where $A_3 = B_2$. Using the formula we get

$$P(B_1 \cup B_2) = P(B_1) + P(B_2) - P(B_1 \cap B_2) = (P(A_1) + P(A_2) - P(A_1 \cap A_2)) + P(A_3) - P((A_1 \cup A_2) \cap A_3)$$

Using distributive laws this simplifies to

$$P(B_1 \cup B_2) = (P(A_1) + P(A_2) + P(A_3)) - P(A_1 \cap A_2) - P((A_1 \cap A_3) \cup (A_3 \cap A_2))$$

Then using Theorem 1.3.7 again we get that

$$P((A_1 \cap A_3) \cup (A_3 \cap A_2)) = P(A_1 \cap A_3) + P(A_3 \cap A_2) - P((A_1 \cap A_3) \cap (A_3 \cap A_2))$$

Using the fact that $(A_1 \cap A_3) \cap (A_3 \cap A_2) = A_1 \cap A_2 \cap A_3$, we get that

$$\begin{aligned} P(B_1 \cup B_2) &= P(A_1 \cup A_2 \cup A_3) \\ &= (P(A_1) + P(A_2) + P(A_3)) - P(A_1 \cap A_2) - (P(A_1 \cap A_3) + P(A_3 \cap A_2) - P(A_1 \cap A_2 \cap A_3) \\ &= (P(A_1) + P(A_2) + P(A_3)) - (P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_3 \cap A_2)) + P(A_1 \cap A_2 \cap A_3) \\ &= \sum_{i=1}^{3} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + (-1)^2 P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

as needed. Next assume that $P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{k-1} P(A_1 \cap A_2 \cap \cdots \cap A_k)$ for some $2 \leq k \leq n$ and consider $n + 1$ events. Notice that

$$\bigcup_{i=1}^{n+1} A_i = (\bigcup_{i=1}^{n} A_i) \cup A_{n+1}$$

Using Theorem 1.3.7 we then get that

$$P((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) = P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1})$$

Using the induction hypothesis we get that

$$P((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) = (\sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_k)) + P(A_{n+1}) -$$

$$P((\bigcup_{i=1}^{n} A_i \cap A_{n+1}))$$

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

We simplify the last part using induction hypothesis again:

$$P((\bigcup_{i=1}^{n} A_i \cap A_{n+1})) = \sum_{i=1}^{n} P(A_i \cap A_{n+1}) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j \cap A_{n+1}) + \cdots + (-1)^{n-1} P\Big(\bigcap_{i=1}^{n} A_i \cap A_{n+1}\Big).$$

Plugging back in we get

$$P\Big(\bigcup_{i=1}^{n+1} A_i\Big) = \sum_{i=1}^{n+1} P(A_i) - \sum_{1\leq i<j\leq n+1} P(A_i \cap A_j) + \cdots + (-1)^{n} P\Big(\bigcap_{i=1}^{n+1} A_i\Big),$$

which is exactly the inclusion–exclusion formula for $n+1$ events.

Therefore, by induction, the formula holds for every $k \geq 2$. □

It looks like there is a lot going on in this proof because of the equalities and tedious notation, but it is rather quite simple and I encourage you to do it to fully understand. Using these inequalities we can come up with other results quite easily as well. For instance the Bonferroni's Inequality.

Theorem 1.3.25

Let C_1, C_2 be two events. Then

$$P(C_1 \cap C_2) \geq P(C_1) + P(C_2) - 1$$

Proof. From Theorem 1.3.7 and Theorem 1.3.23 we have that

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2) \leq 1$$

Re arranging we get our result. □

Now that we done this section, we can complete our talk about the probability space. We learn that given a random experiment, we can assign probabilities to certain outcomes. To do this we need a probability space. For our probability space to be valid and make sense, it consist of three parts. The first is the sample space, the set of all possible outcomes from our random experiment. The next is the σ-algebra. That is the well-behaved collection of events. We need this in order to be able to assign probabilities to events in a well-defined manner. Then finally, the way we assign probabilities is through the probability set function. We learn that this set function is a probability set function if it adheres to the Probability Axioms. If it does, we have many useful theorems and results. We now learn of a variant of the probability set function which is mainly based of intuition.

1.4 Conditional Probability and Independence

If you recall Theorem 1.3.22, one of my proofs failed because it would work when the limit of the non-decreasing sequence of events was the entire sample space. Sometimes in random experiments, if we are given some event A that's a subset of a larger sample space S, we are really only interested in the outcomes of our event A. So this means that our sample space switches from S to A. Now

if I take a new subset $B \subset S$, what would the probability of B relative to A be? To explain, I will use an example. Imagine I am holding a card I picked at random out of a standard deck of playing. What is the probability that it is the ace of hearts? Well we know that its just $1/52$. Now, if I tell you the card I am holding is an ace, what is the chance I am holding the ace of hearts? Well we know that there are 4 aces in the deck, so it's obviously $1/4$. This is conditional probability. The conditional probability of event B given event A, written $P(B|A)$ (where "|" reads as given), measures how likely B is once we know that A has happened. Using our example, $P(A) = 1/52$. The probability of card that is an ace is $P(B) = 1/4$. The probability of an ace given that the card is just some random card picked from the deck is $P(B|A)$. Well to find this we have to find the total number of cards that satisfy A and B, that is $P(A \cap B) = 4/52$. Then our final probability of B given A is $P(B|A) = \frac{P(A \cap B)}{P(A)} = 1/4$. The reason why we divide by $P(A)$ instead of 52 is because like we said, our sample space switched from S to A. That is we are now looking at the outcomes in A instead.

Before we jump the gun and define this new type of probability, notice that we are basically creating a probability set function on the event A relative to the probability set function on the sample space S. This means that we have to adhere to the Probability Axioms. First, since A is our new sample space, and we only care about the elements that are in B but also in A, then we have that $P(B|A) = P(A \cap B|A)$. Moreover, its trivial to see that the probability of A given A has happened is 1, that is $P(A|A) = 1$.

Definition 1.4.1 : Conditional Probability

Let A and B be two events with $P(B) > 0$. We define the conditional probability of B given A as

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

Moreover we have that

1. $P(B|A) = P(A \cap B|A) \geq 0$
2. $P(A|A) = 1$
3. $P(\bigcup_{n=1}^{\infty} B_n|A) = \sum_{n=1}^{\infty} P(B_n|A)$

For part 3 we are adhering to Probability Axiom part 3. To see this, assume that B_n are mutually exclusive. Moreover, it is also true that $(B_n \cap A) \cap (B_m \cap A) = \emptyset$. Using this and probability axiom part 3 we get that

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} B_n|A) &= \frac{P(\bigcup_{n=1}^{\infty} B_n \cap A)}{P(A)} \\ &= \sum_{n=1}^{\infty} \frac{P(B_n \cap A)}{P(A)} \\ &= \sum_{n=1}^{\infty} P(B_n|A) \end{aligned}$$

Thus our conditional probability set function adheres to the Probability Axioms. Thus we call this the conditional probability set function given A. This is only true if event A has a non zero probability.

Example 1.4.2

A hand of five cards is to be dealt at random without replacement from an ordinary deck of 52 playing cards. Find the conditional probability of an all-spade hand (B), relative to the hypothesis that there are at least four spades in the hand (A).

Solution Let S be the sample of the entire deck of cards. Then let A be the event where we draw a 5 card hand with at least 4 spades. Let B be the event that it is an all-spade hand. We are given that A has occurred. We need to find $P(B|A)$. To do this we use our formula and get that

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

We first find $P(A \cap B)$. This is the event where our 5 card hand as atleast 4 four spades and has an all spade hand. Thus $A \cap B = B$. So we need to find $P(B)$. To do this, we need to find the total number of all spade hands. There are 13 spades within a deck of 52 cards. Since we do not care about order, we are looking for $\binom{13}{5}$. We then divide by the the total number of 5 card hands that is $\binom{52}{5}$. So

$$P(B) = \frac{\binom{13}{5}}{\binom{52}{5}}$$

We now find $P(A)$. Then using multiplication rule, the total number of combinations with at least 4 spades and 1 non spade is $\binom{13}{4}\binom{39}{1}$. We use multiplication rule here. We also add the total combinations of 5 spade hands. Thus we get that

$$P(A) = \frac{\binom{13}{4}\binom{39}{1} + \binom{13}{5}}{\binom{52}{5}}$$

Together we get that

$$P(B|A) = \frac{\binom{13}{5}/\binom{52}{5}}{[\binom{13}{4}\binom{39}{1} + \binom{13}{5}]/\binom{52}{5}} = 0.0441$$

□

From Definition 1.4.1, we can re arrange and get that $P(A \cap B) = P(A)P(B|A)$ (this is called multiplication rule for probabilities). Of course on it's own this is useless because we have no way of finding out what $P(B|A)$ is. However depending on the nature of the random experiment, we can sometimes assign probabilities to $P(A)$ and $P(B|A)$ so that we can find what $P(A \cap B)$ is. To show this we will do some examples.

Example 1.4.3

A bowl contains eight chips. Three of the chips are red and the remaining five are blue. Two chips are to be drawn successively, at random and without replacement. We want to compute the probability that the first draw results in a red chip (A) and that the second draw results in a blue chip (B).

Solution We can immediately find $P(A) = 3/8$. After drawing a red chip, the bowl now has 7 chips total with 2 red chips and 5 blue chips. Then given that event A has occurred, the probability of B given A is $P(B|A) = 5/7$. Thus then we have that $P(A \cap B) = P(A)P(B|A) = (\frac{3}{8})(\frac{5}{7}) = \frac{15}{56}$ □

Example 1.4.4

From an ordinary deck of playing cards, cards are to be drawn successively, at random and without replacement. Find the probability that the third spade appears on the sixth draw.

Solution Let A be the event that 2 spades are drawn from the first 5 draws. Let B be the event that the sixth draw is a spade. We need to find $P(A \cap B)$. Since there are 13 spades in deck, then there are $\binom{13}{2}$ combinations of 2 spade cards. There is then $\binom{39}{3}$ combinations of the other 3 cards not being spades. Then in total using multiplication rule there are $\binom{13}{2}\binom{39}{2}$ combinations of the desired 5 cards we'd like from event A. Thus the probability of this is

$$P(A) = \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}}$$

Now there are only 11 spades left in the deck and we have drawn 5 so $P(B|A) = \frac{11}{47}$. Thus the probability that event A and event B have occured is

$$P(A \cap B) = \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}} \cdot \frac{11}{47} = 0.0642$$

□

The multiplication rule for probabilities can be extended to 3 or more events. We apply it on $P(A \cap B \cap C) = P((A \cap B) \cap C) = P(A \cap B)P(C|(A \cap B)) = P(A)P(B|A)P(C|A \cap B)$. This can be extended to any number of k events.

Theorem 1.4.5 : Multiplication Law of Probability

Let $A_1, A_2, \ldots, A_k$ be events from a sample space S with probability greater than zero. Then

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

Proof. We will prove this by induction on k. The base case $k = 2, 3$ have been shown above. Assume that the result holds for some k events where $3 \leq k \leq n$ and consider $n+1$ events. Using the induction hypothesis we get that

$$\begin{aligned} P(A_1 \cap A_2 \cap \ldots \cap A_n \cap A_{n+1}) &= P((A_1 \cap A_2 \cap \ldots \cap A_n) \cap A_{n+1}) \\ &= P(A_1 \cap A_2 \cap \ldots \cap A_n)P(A_{n+1}|A_1 \cap A_2 \cap \ldots \cap A_n) \\ &= P(A_1)P(A_2|A_1)...P(A_n|A_1 \cap A_2 \ldots \cap A_{n-1})P(A_{n+1}|A_1 \cap A_2 \cap \ldots \cap A_n) \end{aligned}$$

as needed. Thus by induction the result holds for all $k \in \mathbb{N}$. □

Example 1.4.6

Four cards are to be dealt successively, at random and without replacement, from an ordinary deck of playing cards. Find the probability of receiving a spade, a heart, a diamond, and a club, in that order.

Solution Let A, B, C, D be the respective events. Then using Theorem 1.4.5 we see that

$$P(A \cap B \cap C \cap D) = P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C) - \frac{13}{52}\frac{13}{51}\frac{13}{50}\frac{13}{49}$$

□

Example 1.4.7

A university consists of 60% female and 40% male students. If a student is female, 70% and 30% are the chances of being local and international respectively. Further, half of of the female local students are undergraduate, while 30% and 20% are master and PhD, respectively. If a student is randomly chosen, find the probability of choosing a female local undergraduate student.

Solution Let S be our sample space, the set of all university students. Let A be the event where the student is female. Let B be the event where the student is local. Let C be the event where the student is an undergrad student. Then it is clear that $P(A) = 0.6$. We are told that given a student is female, the probability that the student is local is 0.7. That is $P(B|A) = 0.7$. Then we are told that given a female local student the probability that they are an undergrad student is 0.5. That is $P(C|A \cap B)$. Using the multiplication rule of probability we see that

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = 0.21$$

□

Recall from Definition 1.3.2, if we have k events $A_1, A_2, \ldots, A_k$ that form a partition on C with $P(A_i) > 0$ for all $i \in \{1, 2, \ldots, k\}$ then they are mutually exclusive and exhaustive. The probabilities of each of these events do not need to be equilikely. Now consider another event B. Since our k events are exhaustive, then that means $B = B \cap (A_1 \cup A_2 \cup \ldots \cup A_k)$. Using distributive laws we get thet $B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_k)$. Note that $(B \cap A_i)$ for all $i \in \{1, 2, \ldots, k\}$ are mutually exclusive. Then see that $P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k)$. Using Multiplication Law of probability,

$$\begin{aligned} P(B) &= P(A_1)(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k) \\ &= \sum_{i=1}^{k} P(A_i)P(B|A_i) \end{aligned}$$

Theorem 1.4.8 : Law of Total Probability

Let S be a sample space and $\mathcal{B}$ be a well-behaved collection of events. Let $A_1, A_2, \ldots, A_k \in \mathcal{B}$ form a partition on C such that $P(A_i) > 0$ for all $i \in \{1, 2, \ldots, k\}$. Then for any event $B \in \mathcal{B}$,

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

Proof. Done above □

Example 1.4.9

Of the voters in a city, 40% are Republican and 60% are Democrats. Among the Republicans, 70% are in favor of a bond issue, while 80% of the Democrats favor this issue. If a voter is selected at random, what is the probability that he/she will favor the bond issue?

Solution Let S be the set of voters in a city. Let A_1 be the set of voters that are Democrats, and let A_2 be the set of voters that are Republican. Then $A_1 \cup A_2 = S$. We also know that $P(A_1) = 0.6$ and $P(A_2) = 0.4$. Let C be the event where the voters are in favor of the bond issue. Then we have that given a voter is republican, the probability that voter is in favor of a bond issue is 0.7. That is $P(C|A_2) = 0.7$. If the voter is democratic, then $P(C|A_1) = 0.8$. Using the Law of Total Probability, we have

$$\begin{aligned} P(C) &= \sum_{i=1}^{2} P(A_i)P(C|A_i) \\ &= P(A_1)P(C|A_1) + P(A_2)P(C|A_2) \\ &= 0.6 \times 0.8 + 0.4 \times 0.7 \\ &= 0.76 \end{aligned}$$

□

Theorem 1.4.8 leads to an important theorem.

Theorem 1.4.10 : Bayes' Theorem

Let S be a sample space and $\mathcal{B}$ be a well-behaved collection of events. Let $A_1, A_2, \ldots, A_k \in \mathcal{B}$ form a partition on S such that $P(A_i) > 0$ for all $i \in \{1, 2, \ldots, k\}$. Let $B \in \mathcal{B}$ be any event. Then

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_1)P(B|A_i)}$$

Proof. By the definition of conditional property we have

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

Using Multiplication Law of Probability, $P(A_j \cap B) = P(A_j)P(B|A_j)$ and the Law of Total Probability we get

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_1)P(B|A_i)}$$

as required. □

This theorem allows us to calculate the probability of A_j given B has occurred from the probabilities of $A_1, A_2, \ldots, A_k$ and the probability of B given A_i has occurred for $i = 1, 2, \ldots, k$.. What are the applications of *Bayes' Theorem*?

1. Invert conditional probabilities: You know the various conditional probabilities $P(B|A_1).P(B|A_2), ..., P(B|A_k)$ but we want instead to find $P(A_j|B)$.

2. You know/obtain $P(A_1), P(A_2), .., P(A_k)$ and then subsequently observe/collect $P(B|A_1).P(B|A_2), ..., P(B|A_k)$.

3. Update your probabilities conditioned on what you have observed by calculating $P(A_1|B), P(A_2|B), .., P(A_k|B)$

We now do some examples using this theorem.

Example 1.4.11

Suppose that bowl A_1 contains 3 red and 7 blue chips where as bowl A_2 contains 8 red and 2 blue chips. All chips are identical in shape and size. A die is cast and bowl A_1 is selected if five or six spots show on the side that is up; otherwise, bowl A_2 is selected. For this situation, it seems reasonable to assign $P(A_1) = \frac{2}{6}$ and $P(A_2) = \frac{4}{6}$. The selected bowl is handed to another person and one chip is taken at random. Say that this chip is red, an event which we denote by B. By considering the contents of the bowl, we can assign probabilities of choosing a red chip from bowl A_1 or A_2. That is $P(B|A_1) = \frac{3}{10}$ and $P(B|A_2) = \frac{8}{10}$. Thus we can find the probability that the bowl selected was A_1 or A_2 given the the chip was red. That is

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{3}{19}$$

In this example. $P(A_1) = \frac{2}{6}$ and $P(A_2) = \frac{4}{6}$ are called **Prior Probabilities** of A_1, A_2 respectively. This is because these were known before performing the random act of choosing the bowls. After the chip is taken and is observed to be red, the conditional probabilities $P(A_1|B) = \frac{3}{19}$ and $P(A_2|B) = \frac{16}{19}$ are called **posterior probabilities.** It makes sense that $P(A_2|B)$ has a higher probability than $P(A_1|B)$ because the chances of having bowl A_2 are better once that a red chip is observed than before a chip is taken. Bayes' theorem provides a method of determining exactly what those probabilities are.

Example 1.4.12

Three plants, A_1, A_2, and A_3, produce respectively, 10%, 50%, and 40% of a company's output. Although plant A_1 is a small plant, its manager believes in high quality and only 1% of its products are defective. The other two, A_2 and A_3, are worse and produce items that are 3% and 4% defective, respectively. All products are sent to a central warehouse. One item is selected at random and observed to be defective, say event B. Find the conditional probability that it comes from plant A_1.

Solution We are given that $P(A_1) = 0.1, P(A_2) = 0.5, P(A_3) = 0.4$. These are the prior probabilities of A_i. Thus clearly $S = A_1 \cup A_2 \cup A_3$. That is they form a partition on our sample space. We are told that the probability that a product is defective given that it is from plant A_i is : $P(B|A_1) = 0.01, P(B|A_2) = 0.03, P(B|A_3) = 0.04$. Now we need to update our probability that a product is from plant A_1 given that the product is defective. That is we need to find $P(A_1|B)$. Using Bayes's theorem we see that

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ &= \frac{(0.1)(0.01)}{(0.1)(0.01) + (0.5)(0.03) + (0.4)(0.04)} \\ &= \frac{1}{32} \end{aligned}$$

This is much smaller than the prior probability of A_1 . This is as it should be because the fact that the item is defective decreases the chances that it comes from the high-quality plant A_1. □

Example 1.4.13

Suppose we want to investigate the percentage of abused children in a certain population. The events of interest are: a child is abused (A) and its complement a child is not abused ($N = A^c$). For the purposes of this example, we assume that $P(A) = 0.01$ and, hence, $P(N) = 0.99$. The classification as to whether a child is abused or not is based upon a doctor's examination. Because doctors are not perfect, they sometimes classify an abused child (A) as one that is not abused (N_D, where N_D means classified as not abused by a doctor). On the other hand, doctors sometimes classify a non-abused child (N) as abused (A_D). Suppose these error rates of misclassifications are

$$P(N_D \mid A) = 0.04 \quad \text{and} \quad P(A_D \mid N) = 0.05;$$

thus the probabilities of correct decisions are

$$P(A_D \mid A) = 0.96 \quad \text{and} \quad P(N_D \mid N) = 0.95.$$

Compute the probability that a child taken at random is classified as abused by a doctor

Solution What we need to find is the probability that a child is classified as abused by a doctor A_D given that the child can be either abused A or not abused N. That is we need to find $P(A_D|A \cup N)$. That is we get that

$$\begin{aligned} P(A_D|A \cup B) = P(A_D) &= \frac{P(A_D \cap (A \cup N))}{P(A \cup B)} \\ &= \frac{P(A_D \cap A) + P(A_D \cap N)}{P(A) + P(N)} \\ &= \frac{P(A)P(A_D|A) + P(N)P(A_D|N)}{1} \\ &= (0.01)(0.96) + (0.99)(0.05) \\ &= 0.0591 \end{aligned}$$

Notice this is quite higher than the probability that a child is actually abused (0.01). Further, we see that the probability that the child is abused given that the doctor classified them as abused is

$$\begin{aligned} P(A|A_D) = \frac{P(A \cap A_D)}{P(A_D)} \\ &= \frac{(0.96)(0.01)}{0.0591} \\ &= 0.1624 \end{aligned}$$

which is quite low. We can similarly find that the probability that a child is not abused when the doctor classified them as abused is 0.8376. The reason that these probabilities are so poor at recording the true situation is that the doctors' error rates are so high relative to the fraction 0.01 of the population that is abused. An investigation such as this would, hopefully, lead to better training of doctors for classifying abused children. □

Example 1.4.14 : Drug Testing

Based on an investigation, it is believed that 1% of the sailors use illegal drugs. A test detects drug use as follows:

1. If someone has used illegal drugs, there is a 99% chance it will detect it (true positive).
2. If someone does not use illegal drugs, there is a 0.5% chance the test will indicate it (false positive).

What is the chance that a positive drug test signals a drug user?

Solution Let U_+ denote the event where the sailor uses illegal drugs and U_- denote the event where the the sailor does not use illegal drugs. Then $P(U_+) = 0.01, P(U_-) = 0.99$. Let T_+ denote the event where the drug test detects the if someone has used illegal drugs. We are told the probability that the test detects if someone used illegal drugs given that the person has used illegal drugs is $P(T_+|U_+) = 0.99$. We are told the probability that the test detects if someone used illegal drugs given that they did not (false positive) is $P(T_+|U_-) = 0.005$. We are asked what is the probability that the if an illegal drug user uses the test, it comes back as positive. That is $P(U_+|T_+)$. Using Bayes's Theorem we get that

$$\begin{aligned} P(U_+|T_+) &= \frac{P(U_+)P(T_+|U_+)}{P(U_+)P(T_+|U_+) + P(U_+)P(T_+|U_-)} \\ &= \frac{(0.01)(0.99)}{(0.01)(0.99) + (0.99)(0.005)} \\ &= \frac{2}{3} \end{aligned}$$

□

Example 1.4.15 : Maize Seeds

The genetic origin and properties of maize (modern-day corn) were investigated in Economic Botany. Seeds from maize ears carry either single spikelets or paired spikelets, but not both. Progeny tests on approximately 600 maize ears revealed the following information: 40% percent of all seeds carry single spikelets, while 60% carry paired spikelets. A seed with single spikelets will produce maize ears with single spikelets 29% of the time and paired spikelets 71% of the time. A seed with paired spikelets will produce maize ears with single spikelets 26% of the time and paired spikelets 74% of the time. Find the probability that a randomly selected maize ear seed carries a single spikelet and produces ears with single spikelets. Find the probability that a randomly selected maize ear seed produces ears with paired spikelets.

Solution Let A_1 be the event where the maize seed ears carry single spikelets. Let A_2 be the event where the maize seed ears carry double spikelets. We are given that $P(A_1) = 0.4, P(A_2) = 0.6$ Let B be the event where the seed produces single spikelets. Let D be the event where the seed produces double spikelets.

We are told that a seed will produce maize ears with single spikelets 29% of the time given that the seed is carries single spikelets. That is $P(B|A_1) = 0.29$. Thus $P(D|A_1) = 0.71$. We are told that a seed will produce maize seed ears with single spikelets 26% of the time given that the seed

carries double spikelets. That is $P(D|A_2) = 0.26$. Thus $P(B|A_2) = 0.74$.

We need to find $P(A_1 \cap B)$. We get

$$P(A_1 \cap B) = P(A_1)P(B|A_1) = (0.4)(0.29) = 0.116$$

We need to find the probability that a seed produces double spikelets given that the seed carries either single or double spikelets. That is we need to find $P(D|A_1 \cup A_2) = P(D)$. Using conditional probability definition and Bayes's Theorem we get that

$$\begin{aligned} P(D|A_1 \cup A_2) = P(D) &= \frac{P(D \cap (A_1 \cup A_2))}{P(A_1 \cup A_2)} \\ &= \frac{P(D \cap A_1) + P(D \cap A_2)}{1} \\ &= P(A_1)P(D|A_1) + P(A_2)P(D|A_2) \\ &= (0.4)(0.71) + (0.6)(0.74) \\ &= 0.728 \end{aligned}$$

□

1.4.1 Independence

Sometimes given an event A has occurred, the probability of another event B does not change. That is when $P(A) > 0$, $P(B|A) = P(B)$. When this happens, we say that A and B are independent events. Another way to think about it is that knowledge about event A does not affect your probabilistic assessment of B. This is the intuitive notion that A and B are independent. We also see that then $P(A \cap B) = P(A)P(B|A) = P(A)P(B)$ This then implies that

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

This leads us to our definition of independence.

Definition 1.4.16

Let A and B be two events. We say that A and B are independent if one of the following conditions hold:

- $P(B|A) = P(B)$
- $P(A|B) = P(A)$
- $P(A \cap B) = P(A)P(B)$

Otherwise the events are dependent.

We also have the following result knowing that A and B are independent.

Proposition 1.4.17

Let A and B be two independent events. Then the three pairs of events are independent.

1. A^c and B
2. A and B^c
3. A^c and B^c

Proof. Let A and B be independent. Then $P(A\cap B) = P(A)P(B)$. For 1: Using the disjoint union $B = (A^c \cap B) \cup (A \cap B)$ we get

$$P(B) = P(A^c \cap B) + P(A \cap B)$$

Re arranging we get

$$P(A^c \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)(1 - P(A))$$

Using the result $P(A^c) = 1 - P(A)$ we get

$$P(A^c \cap B) = P(A^c)P(B)$$

For 2 the proof is analogous to part 1. For 3: We use DeMorgan's Law $A^c \cap B^c = (A \cup B)^c$. Then we get that

$$\begin{aligned} P(A^c \cap B^c) = P((A \cup B^c)) &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= (1 - P(A))(1 - P(B)) \\ &= P(A^c)P(B^c) \end{aligned}$$

We used Inclusion-Exclusion formula for the third equality and Theorem 1.3.3 in the fourth. □

Example 1.4.18

It is known that a patient with a disease will respond to treatment with probability 0.9. If three patients with the disease are treated and they respond independently, find the probability that at least one will respond.

Solution Let A_i denote the event where the patient i with a disease will respond to treatment for $i = 1, 2, 3$. Then $P(A_i) = 0.9$ and $P(A_i^c) = 0.1$. We are asked to find the probability that at least one of the three patients respond independently. That is we need to find $P(A_1 \cup A_2 \cup A_3)$. Notice that using DeMorgan's Law we have that $A_1 \cup A_2 \cup A_3 = (A_1^c \cap A_2^c \cap A_3^c)^c$. Thus using independence of the events A_i and Theorem 1.3.3, we get that

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = P((A_1^c \cap A_2^c \cap A_3^c)^c) &= 1 - P(A_1^c \cap A_2^c \cap A_3^c) \\ &= 1 - P(A_1^c)P(A_2^c)P(A_2^c) \\ &= 1 - (0.1)^3 \\ &= 0.999 \end{aligned}$$

□

Example 1.4.19

Consider the following events in the toss of a single balanced six-sided die:

- A= Observe an odd number
- B = Observe an even number
- C = Observe a 1 or a 2.

Question 1: Are A and B independent events.
Question 2: Are A and C independent events.

Solution Question 1: For A and B to independent events, then the occurrence of event A or B should not affect B or A respectively. First assume that event A has occurred. Then what is the probability of event B given that event A has occurred? It is zero since we observed an odd number and it cannot be even. That is $P(A \cap B) = 0 \neq P(A)P(B) = 1/4$. Thus A and B are not independent events.

Question 2: For A and C to independent events, then the occurrence of event A or C should not affect B or A respectively. First assume that event A has occurred. Then we have observed an odd number. Then what is the probability of event C given that event A has occurred? It is $1/6$. That is $P(A \cap C) = P(A)P(C) = \frac{2}{6}\frac{3}{6} = 1/6$. Thus A and C are independent events. □

As expected, we can extend the idea of independence to any number of sets. For example take the three events A_1, A_2, A_3. We say that they are mutually independent if and only if they are pairwise independent: $P(A_1 \cap A_2) = P(A_1)P(A_2), P(A_1 \cap A_3) = P(A_1)P(A_3), P(A_2 \cap A_3) = P(A_2)P(A_3)$ and $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$. More generally we have the following definition:

Definition 1.4.20

n events $A_1, A_2, .., A_n$ are mutually independent if and only if for every collection of k of these events, $2 \leq k \leq n$ and for every permutation $d_1, d_2, .., d_k$ of $1, 2, .., k$,

$$P(A_{d_1} \cap A_{d_2} \cap \cdots \cap A_{d_k}) = P(A_{d_1})P(A_{d_2}) \cdots P(A_{d_k}).$$

In particular, if $A_1, A_2, .., A_n$ are mutually independent then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n)$$

Its key to note that pairwise independence does not imply mutual independence. We illustrate this with an example.

Example 1.4.21

Asan example, suppose we twice spin a fair spinner with the numbers 1, 2, 3, and 4. Let A_1 be the event that the sum of the numbers spun is 5, let A_2 be the event that the first number spun is a 1, and let A_3 be the event that the second number spun is a 4. Then $P(A_i) = 1/4$ for $i = 1, 2, 3, 4$ and for $i \neq j$, $P(A_i \cap A_j) = 1/16$. So they are pairwise independent. But the event $A_1 \cap A_2 \cap A_3$ is the event where the pair of numbers (1,4) are spun. That has a probability of 1/16 however $P(A_1 \cap A_2 \cap A_3) = 1/16 \neq P(A_1)P(A_2)P(A_3) = 1/64$. Hence they are not mutually independent.

Sometimes we repeat random experiments like flipping a coin or rolling a die more than once, and the outcome of one doesn't affect the outcome of the others.

When that happens, we say the experiments are independent. So when we say things like "independent coin flips" or "independent die rolls," we mean that the outcome of one flip or roll has no influence on the others. For example, If you flip a coin 5 times, each flip is independent — knowing the result of one flip doesn't tell you anything about the others. If you roll a die, put it back, and roll again each roll is independent. "Independent outcomes" just means: knowing one result doesn't change the probabilities of the others.

Example 1.4.22

A coin is flipped independently several times. Let the event A_i represent a head (H) on the ith toss; thus A_i^c represents a tail (T). Assume that A_i and A_i^c are equally likely; that is, $P(A_i) = P(A_i^c) = \frac{1}{2}$. Thus the probability of an ordered sequence like HHTH is, from independence,

$$P(A_1 \cap A_2 \cap A_3^c \cap A_4) = P(A_1)P(A_2)P(A_3^c)P(A_4) = \left(\tfrac{1}{2}\right)^4 = \tfrac{1}{16}.$$

Similarly, the probability of observing the first head on the third flip is

$$P(A_1^c \cap A_2^c \cap A_3) = P(A_1^c)P(A_2^c)P(A_3) = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8}.$$

Also, the probability of getting at least one head on four flips is

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup A_4) &= 1 - P\left[(A_1 \cup A_2 \cup A_3 \cup A_4)^c\right] \\ &= 1 - P(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c) \\ &= 1 - \left(\tfrac{1}{2}\right)^4 = \tfrac{15}{16}. \end{aligned}$$

1.5 Random Variables

To discuss Random Variables, we will look at the factors and reasoning that motivate our definition. Let S be a sample space. As you've seen, sometimes our sample space isn't always numbers such as in Exercise 1.4.22 or a coin toss. To address this, we can create rule(s) for each element $x \in S$ so that they can be represented by numbers. For example, let the random experiment be a coin toss. Then we see that $S = \{H, T\}$. Then we can define a real-valued function $X(T) = 0, X(H) = 1$. This function takes our sample space and translates it to the space of real numbers $\mathcal{D} = \{0, 1\}$.

Definition 1.5.1

Consider a random experiment with a sample space S. A function X, which assigns each element $c \in S$ one and only one number $X(c) = x$, is called a random variable. The **space** or **range** of X is the set of real numbers $\mathcal{D} = \{x : x = X(c), c \in S\}$.

Example 1.5.2

A fair die is thrown twice. The sample points are $(1,1), (1,2), \ldots, (6,6)$. There are 36 sample points. Let us assign the same probability 1/36 for each of these points. Suppose we are interested in the sum of the numbers of each outcome. Then we can define a random variable $x = i + j$ associated with the outcome (i, j). List the possible values of X and find the probability of each value.

Solution We follow the rule of the function X and assign each element from our sample space a number. We get the following :

Value of X	**Probability**
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

□

Usually $\mathcal{D}$ is either a countable set or an interval of real numbers. The first is called Discrete Random Variables and the second is called Continuous Random Variables. We first discuss Discrete Random Variables then move on to Continuous Random Variables. But first we discuss Probability Distributions. Given a random variable X, besides giving us the real numbered sample space $\mathcal{D}$, we are given a different type of probability (some call it an induced probability).

We begin with a Discrete Random Variable like in the example above. That is $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$. The events we are interested in are the subsets of $\mathcal{D}$. This new probability is also quite easy to see. We define a function

$$p_X(d_i) = P(\{c : X(c) = d_i\}) \text{ for } i = 1, 2, \ldots, m$$

To understand this we simply follow the route we took to get here. We first begin with a random experiment which gave us a discrete sample space. That is our sample space is countable. Then we define a discrete random variable X. With this we got the range of X which was $\mathcal{D}$ which was our sample space now in real numbers. Then this new function X induced a new probability. We define the function p_X as above. This function takes in an element from our range $d_i \in \mathcal{D}$ and out puts the probability of some event which we denote $A \subseteq S$. All the elements in A map to this

number d_i. That is for all $c \in A$, $X(c) = d_i$. Thus we can say the induced probability of a discrete random variable X is

$$P(D) = \sum_{d_i \in D} p_X(d_i) \ , \ D \in \mathcal{D}$$

We formalize this in the next section. Note the following:

- Capital letters denote random variables, typically X, Y , and Z and small letters denote the value of a random variable, typically x, y, and z. They are the outcomes of the experiment.
- We write $P(X = x) = p_X(x)$ which means the probability that random variable X takes on x.
- Every random variable is associated with a probability distribution that specifies the possible random variable values and the probability each value will occur

1.5.1 Discrete Random Variables

As we've seen above, if the sample space is countable, then the domain for the function X can only take in and output a countable amount of real numbers. That is the range of X, $\mathcal{D}$, is countable.

Definition 1.5.4

A random variable is a Discrete Random Variable if it's sample space is countable.

We've already seen examples of this such as Example 1.5. Next, as we said above, a random variable induces a new probability, called the probability distribution.

Definition 1.5.5

The probability distribution for a discrete random variable X can be represented by a formula, table, or a graph that provides $P(X = x) = p_X(x)$ for all $x \in D$. This is often called the Probability Mass Function (PMF).

Basically, the probability distribution of the random variable X tells us how the total probability of 1 is "distributed" among all the possible values that X can have. To illustrate this we do an example.

Example 1.5.6

Two fair 6-sided dice are rolled, and the sum of the up-faces is recorded as a random variable X. Let the sample space be the set of all ordered pairs (i, j) where $1 \leq i, j \leq 6$. Since the dice are fair, each pair has equal probability, i.e., $P(i, j) = \frac{1}{36}$. The possible values of X range from 2 to 12. We can list them out can get

Range value x	2	3	4	5	6	7	8	9	10	11	12
Probability $p_X(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

1. Let $B_1 - \{x : x = 7 \text{ or } 11\}$. What is $p_X(B_1)$
2. Let $B_2 - \{x : x = 2, 3, 12\}$. What is $p_X(B_2)$

Solution Before we answer these two questions I want to show how we got our probability distribution above. Lets look at $x = 4$. If $x = 4$, then $p_X = P[\{c : X(c) = 4\}]$. Let $A = \{c : X(c) = 4\}$ denote this event. What is A here? It is the ordered pairs where the sum of the up-faces equals 4. That is through enumeration (we can do this since our sample space is countable), $A = \{(2,2),(3,1),(1,3)\}$. Then since the probability are equal, we see that $P(A) = 3/36$.

For (1): We are given $B_1 = \{x : x = 7 \text{ or } 11\}$. We want to compute:

$$P_X(B_1) = \sum_{x \in B_1} p_X(x) = p_X(7) + p_X(11)$$

From the table, we know: $p_X(7) = \frac{6}{36}$ and $p_X(11) = \frac{2}{36}$.Thus

$$P_X(B_1) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36}$$

Alternatively, we could define the event:

$$A_1 = \{c \in C : X(c) \in B_1\} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1),(5,6),(6,5)\}$$

Since each outcome in the sample space has probability $1/36$, we get:

$$P(A_1) = \frac{8}{36}$$

which confirms our earlier calculation. For (2): Let $B_2 = \{x : x = 2, 3, 12\}$. Again, we compute:

$$P_X(B_2) = \sum_{x \in B_2} p_X(x) = p_X(2) + p_X(3) + p_X(12)$$

From the table above: $p_X(2) = \frac{1}{36}$, $p_X(3) = \frac{2}{36}$, $p_X(12) = \frac{1}{36}$ So:

$$P_X(B_2) = \frac{1}{36} + \frac{2}{36} + \frac{1}{36} = \frac{4}{36}$$

□

For discrete random variables, the probability distribution is called the probability mass function. The reason why is because you can think of it like a Riemann sum. You can guess what the Continuous Probability Distribution would be like then.

We then have the following theorem which solidifies a few things.

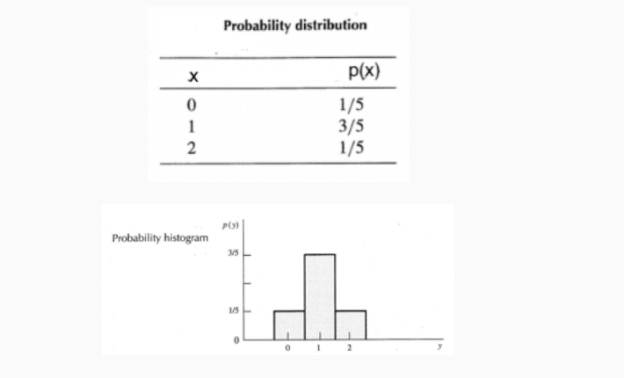
Theorem 1.5.7

Let X be a discrete random variable and $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ be its range. Then following must be true:

1. $0 \leq p_X(x) \leq 1$
2. $\sum_{i=1}^{m} p_X(d_i) = 1$

Example 1.5.8

A recruiting manager has six internship applications, three men and three women. He has to choose two for a special project. So as not to play favorites, he decides to select the two at random. Let X denote the number of women selected. Find the probability distribution for X.



Probability distribution

x	p(x)
0	1/5
1	3/5
2	1/5

Figure 2: Discrete Probability Distribution of X women selected. We see that the probability is low at the extremes (two and zero) and high for the average (one).

Solution Let S be the sample space of three women and three men. Let X be the total number of women selected. Then we see that $X = \{0, 1, 2\}$.We need to find $p_X(x)$ for all $x \in X$. We will find the general formula by considering an arbitrary $x \in X$. First, notice there are $\binom{6}{2}$ ways we can select two people from the three men and three women for the special projects. Then if there are x women selected, then there are $\binom{3}{x}$ ways we can select x women. Then there are $\binom{3}{2-x}$ ways we can choose the men. Thus the formula of X is

$$p_X = \frac{\binom{3}{x}\binom{3}{2-x}}{\binom{6}{2}}$$

To build the probability distribution we calculate $P(X = x)$. We get that $P(X = 0) = 0.2$, $P(X = 1) = 0.6$, $P(X = 0) = 0.2$. Note that what we are finding here is how the probability that when choosing two people from our sample space S is distributed, where our random variable is the number of women. So we can think of $P(X)$ as the distribution of the probability that x women are selected. We can see that the probability is low for two and zero women but is high for one women.

□

We next talk about the Cumulative Distribution Function (CDF). The CDF gives you the cumulative probability up to a certain value. It's the probability that X is "less than or equal to" that value. To understand this better, I'll explain this with an example and application.

Let's go back to our earlier example: a recruiting manager selects 2 applicants at random from a pool of 6 people — 3 men and 3 women. Let X be the number of women selected. From earlier, we found the probability distribution:

$$P(X = 0) = 0.2, \quad P(X = 1) = 0.6, \quad P(X = 2) = 0.2$$

Using this, we constructed the CDF:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.2 & \text{if } 0 \leq x < 1 \\ 0.8 & \text{if } 1 \leq x < 2 \\ 1.0 & \text{if } x \geq 2 \end{cases}$$

This means, for example, that there's a 20% chance that at most 0 women are selected, and an 80% chance that at most 1 woman is selected. Now let's use the CDF to understand percentiles,

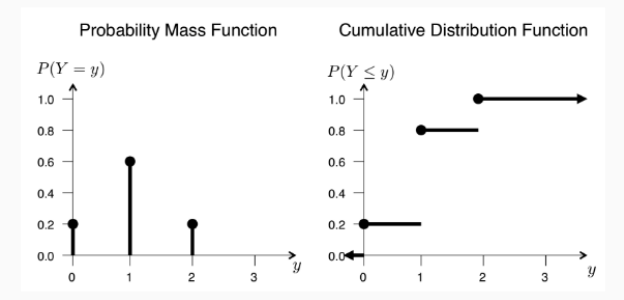


Figure 3: The graph of the Probability Mass Function of X compared to the Cumulative Distribution Function of X.

which are specific cutoffs based on probability. The p-th percentile is the smallest value of x such that $F(x) \geq p$. Let's find the 5th percentile in this example:

We're looking for the smallest x such that $F(x) \geq 0.25$. From the CDF:

$$F(0) = 0.2, F(1) = 0.8$$

So the smallest x where $F(x) \geq 0.25$ is $x = 1$.

Therefore, the 25th percentile is 1, meaning there is a 25% chance or more that at most 1 woman is selected.

Similarly, if we asked for the 90th percentile, we would find the smallest x such that $F(x) \geq 0.9$. Since:

$$F(1) = 0.8 < 0.9, F(2) = 1.0 \geq 0.9$$

The 90th percentile is 2, meaning there is at least a 90% chance that 2 or fewer women are selected.

As we seen, the CDF is crucial for finding percentiles because it accumulates probability in a way that allows you to ask questions like: "What value of X is reached with 90% certainty?" or "What's the smallest number such that 25% of the time, X is below it?" This is widely used in statistical reporting, standardized testing, and real-world decision making.

Definition 1.5.9

Let X denote a random variable. The Cumulative Distribution Function (cdf) of X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x) = P_X((-\infty, x]) = P(\{c \in S : X(c) \leq x\})$$

See Figure 3 for how F_X looks graphically and try to understand why it looks that way.

We now prove a theorem related to the CDF. Before we do I'll try to explain the intuition behind each one. There are three parts. For the first one, let $F_X(x)$ be a CDF for X. We then know that $F_X(x) = P(\{c \in S : X(c) \leq x\})$. What happens when x becomes smaller. More specifically, what is the behavior of F_X as x approaches $-\infty$. Well looking to the graph above and through just guessing, we are expecting that it should get smaller. Why? Well we define $F_X(x) = P(\{c \in S : X(c) \leq x\})$. So then

$$\lim_{x \to -\infty} F_X(x) = \lim_{x \to -\infty} P(\{c \in S : X(c) \leq x)\})$$

Notice that $0 \leq P(\{c \in S : X(c) \leq x)\}) \leq 1$. If $x_1 < x_2$, then $\{c \in S : X(c) \leq x_1\} \subset \{c \in S : X(c) \leq x_2\}$. With this it should be clear what happens to the probability at both $-\infty$ and ∞.

Theorem 1.5.9

Let X be a random variable. Let F_X be a Cumulative Distribution Function (CDF. Then

1. $\lim_{x \to -\infty} F_X(x) = 0$
2. $\lim_{x \to \infty} F_X(x) = 1$
3. $\lim_{x \to x_o} F_X(x) = F(x_0)$

Proof. For 1: As we seen above, if $x_2 \leq x_1$ then $\{c \in S : X(c) \leq x_2\} \subseteq \{c \in S : X(c) \leq x_1\}$. Thus

$$0 \leq P(\{c \in S : X(c) \leq x_2\}) \leq P(\{c \in S : X(c) \leq x_1\}) \leq 1$$

What we can do then is define a non-increasing sequence of events. Choose any sequence $x_1 > x_2 > x_3$.. with $x_n \to -\infty$ and set $\{C_n\}$ by $C_n = (-\infty, x_n]$ where $x_n < x_{n-1}$. Then we see clearly that

$$\bigcap_{n=1}^{\infty} C_n = \emptyset.$$

Thus we get by using Continuity Theorem

$$\lim_{n \to \infty} P(C_n) = P(\bigcap_{n=1}^{\infty} C_n) = P(\emptyset) = 0$$

But we see that $P(C_n) = F_x(x_n)$ so

$$\lim_{n \to \infty} F_X(x_n) = 0$$

Moreover since F_X is monotone we get that

$$\lim_{x \to -\infty} F_X(x) = 0$$

as needed. For 2: We approach this with the same method as we did in part 1. We see that when $x_1 \leq x_2$, $\{c \in S : X(c) \leq x_1\} \subseteq \{c \in S : X(c) \leq x_2\}$. We define $\{D_n\}$ by $D_n = (-\infty, x_n]$ where $x_{n-1} < x_n$ and $x_n \to \infty$. Thus this is a non decreasing sequence of events and we get that

$$\lim_{n \to \infty} D_n = \bigcup_{n=1}^{\infty} D_n = \mathbb{R}$$

Thus we get by using Continuity Theorem

$$\lim_{n \to \infty} P(D_n) = P(\bigcup_{n=1}^{\infty} C_n) = P(\mathbb{R}) = 1$$

Like above we see that $P(D_n) = F_x(x_n)$ so

$$\lim_{n \to \infty} F_X(x_n) = 1$$

Since F_X is monotone we get that

$$\lim_{x\to\infty} F_X(x) = 1$$

For 3: Let $\{x_n\}$ be any sequence of real numbers such that $x_n \to x_o$. Then define the non-increasing sequence of events $\{B_n\}$ by $B_n = (-\infty, x_n]$. Then we see that

$$\lim_{n\to\infty} B_n = \bigcap_{n=1}^{\infty} B_n = (-\infty, x_0]$$

Thus using the Continuity Theorem we get that

$$\lim_{n\to\infty} P(B_n) = P(\bigcap_{n=1}^{\infty} B_n) = P((-\infty, x_0]) = F_X(x_o)$$

as needed. □

Note the following: This theorem holds for both Continuous and Discrete Random Variables as you see in the proof we do not specify what type of random variable X is. But there is a difference between of the cdf of discrete and continuous random variables. The cdf for a discrete random variables is a step function. They jump at each x value at which there is positive probability. Otherwise, in between, the function is flat. This occurs because the cdf only increases at finite or countable number of points with positive probability. Moreover we see that F_X is a non-decreasing function always. The reason why this is true is because as said above, if $x_1 < x_2$, then $\{c \in S : X(c) \leq x_1\} \subset \{c \in S : X(c) \leq x_2\}$ which implies

$$F_X(x_1) = P(\{c \in S : X(c) \leq x_1\}) \leq P(\{c \in S : X(c) \leq x_2\}) = F_X(x_2)$$

The next theorem is helpful for computing CDF.

Theorem 1.5.10

Let X be a random variable and F_X be a CDF. Then for $a < b$, $P([a \leq X \leq b]) = F_X(b) - F_X(a)$.

Proof. Notice that $\{-\infty \leq X \leq b\} = \{-\infty \leq X \leq a\} \cup \{a \leq X \leq b\}$. Since the right hand side is a disjoint union the result follows trivially. □

1.5.2 Transformations

A problem often encountered in statistics is the following. We have a random variable X and we know its distribution. We are interested, though, in a random variable Y that is some transformation of X, say $Y = g(X)$. In particular, we want to determine the distribution of Y.

Assume X is discrete with support S_X. Then the support of Y is

$$S_Y = \{\, g(x) : x \in S_X \,\}.$$

If g is one-to-one, the pmf of Y is obtained as

$$p_Y(y) = P(Y = y) = P\big[g(X) = y\big] = P\big[X = g^{-1}(y)\big] = p_X\big(g^{-1}(y)\big).$$

Example 1.5.11

Let X have the probability mass function $p_X(x) = 1/3$ for $x = 1, 2, 3$ and zero elsewhere. Find the probability mass function of $Y = 2X + 1$

Solution Since $g(x) = 2x + 1$ is injective then for $y = 3, 5, 7$ we have that

$$p_Y(y) = P(Y = y) = P(2X + 1 = y) = P(X = \frac{y-1}{2}) = p_X(g^{-1}(y)) = p_X(\frac{y-1}{2}) = \frac{1}{3}$$

□

For the next example consider the following: Consider a sequence of independent flips of a coin, each resulting in a head (H) or a tail (T). Moreover, on each flip, we assume that H and T are equally likely; that is, $P(H) = P(T) = 1/2$. The sample space S consist of sequences like $TTHTHHT\cdots$. Let X be the random variable of the number of flips needed to obtain the first head. so $X(TTHTHHT) = 3$. So the space of x is $\mathcal{D} = \{1, 2, 3, \dots\}$. We see that when $x = 1$ the sequence starts with a heads and so $P(X = 1) = 1/2$. Likewise when $x = 2$ the second flip is heads so we get that $P(X = 2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. More generally if $X = x$ then there are $x - 1$ tails followed by a head. That is $TTT\cdots TH$. Thus we have a geometric sequence of probabilities. That is

$$P(X = x) = \left(\frac{1}{2}\right)^{x-1}\frac{1}{2} = \left(\frac{1}{2}\right)^{x}$$

Example 1.5.12

Let X be the random variable above. That is the flip number where the first head appears. Let Y be the number of flips before we get the first head.

Solution Like above we see that then $Y = X - 1$ flips before the first head. Since this transformation is one-to-one we can solve for the pmf easily:

$$p_Y(y) = P(Y = y) = P(X - 1 = y) = P(X = y + 1) = p_X(y + 1) = \left(\frac{1}{2}\right)^{y+1}$$

□

The case when $g(x)$ is not one-to-one,instead of developing an overall rule, for most applications with discrete random variables, the pmf of Y can be obtained in a straightforward manner.

Example 1.5.13

Let X have have the pmf $p_X(x) = 1/3$ for $x = -1, 0, 1$ and zero elsewhere. Find the pmf of $Y = X^2$

Solution We see that the support of Y is $S_Y = \{0, 1\}$. However notice that $g(x)$ here is not one-to-one. However we see that

$$p_Y(1) = P(Y = 1) = P(X^2 = 1) = P(X = 1) + P(X = -1) = \frac{2}{3}$$

and we see that

$$p_Y(0) = P(Y = 0) = P(X^2 = 0) = P(X = 0) = p_X(0) = \frac{1}{3}$$

□

Example 1.5.14

Consider a geometric random variable like in Example 1.5.12. Suppose we are playing a game against the "house" (say, a gambling casino). If the first head appears on an odd number of flips, we pay the house one dollar, while if it appears on an even number of flips, we win one dollar from the house Let Y denote our net gain. Find the pmf of Y.

Solution The space of Y is $\{-1, 1\}$. We need to find $P(X \in \{1, 3, 5, 7, \dots\})$. We get that

$$P(X \in \{1, 3, 5, 7, \dots\}) = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^{2x-1} = \frac{1}{2}\sum_{x=1}^{\infty} \left(\frac{1}{4}\right)^{x-1} = \frac{1/2}{1-(1/4)} = \frac{2}{3}$$

Thus $p_Y(-1) = 2/3$ and $p_Y(1) = 1/3$. For the last equality we just do the same above for even numbers instead. □

Example 1.5.15

Let X be a geometric random variable like in Example 1.5.12. Let $Z = (X-2)^2$. Find the pmf of Z.

Solution The space of Z is $S_Z = \{0, 1, 4, 9, 16, \dots\}$. Notice that when $Z = 0$ we have that $X = 2$ so $p_Z(2) = p_X(2) = 1/4$. Similiary we find that $p_Z(1) = p_X(1) + p_X(3) = 5/8$. For all other values of X we have a one-to-one correspondence given by $x = \sqrt{z} + 2$ for $z \in \{4, 9, 16, \dots\}$. Hence the pmf of z is

$$p_Z(z) = \begin{cases} p_X(2) = 1/4 & \text{for } z = 0 \\ p_X(1) + p_X(2) = \frac{5}{8} & \text{for } z = 1 \\ \sqrt{z} + 2 & \text{for } z \in \{4, 9, 16, \dots\} \end{cases}$$

□

1.5.3 Continuous Random Variable

The case where our random variable is continuous, our space $\mathcal{D}$ is uncountable. That is it is a interval of real numbers. Usually continuous random variables are in the form of measurements. For example the weight or height of an adult. We might be interested in the probability that someone is 5'7 feet tall but also the probability that someone is over 6'0 feet tall. Like for discrete, we can find the induced probability distribution of X. Do this we can usually find a non-negative function $f_X(x)$. We get that for any interval of real numbers,

$$P_X[(a, b)] = P(\{c : a \leq X(c) \leq b\}) = \int_b^a f_X(x)dx$$

Note that it is not always true that if the Space is uncountable it implies that X is continuous as mixture random variable exists. Instead the formal definition is the following:

Definition 1.5.16

A random variable X with CDF $F_X(x)$ is said to be continuous if $F_X(x)$ is continuous for $-\infty < x < \infty$.

Definition 1.5.17

Let $F_X(x)$ be a be the distribution function for a continuous random variable X. Then, the probability density function (pdf) $f_X(x)$ is given by

$$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x)$$

It follows that

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

This is only true when f_X is continuous. Because an interval of real numbers has no measure, we have that $P(X = x) = 0$ for all x. Thus we have that then $P(a \leq X \leq b) = P(a < X < b)$. However for the most part, the intuition with discrete random variables carries over as we will see in the following examples. Of course we have that the PDF must be greater than or equal to zero. As well as the integral from negative infity to positive infinity must equal to 1.

Theorem 1.5.19

$f_X(x)$ is a pdf for a continuous random variable X if and only if

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

Proof. Will be completed once I'm doing MAT257 alongside Ehsaan or Duncan. □

Example 1.5.20

Given $f_X(x) = cx^2$, $0 \leq x \leq 2$, find c so that $f_X(x)$ is a valid pdf. Also find $P(1 \leq X \leq 2)$.

Solution We know that $0 \leq \int_0^2 f_X(x)dx = 1$. We get that

$$\begin{aligned}\int_0^2 f_X(x)dx &= \int_0^2 cx^2 \\ &= c\left[\frac{x^3}{3}\right]_0^2 \\ &= \frac{8c}{3} = 1\end{aligned}$$

Thus we get that $c = \frac{3}{8}$ which means $f_X(x) = \frac{3x^2}{8}$. Then we evaluate the $P(1 \leq X \leq 2)$ easily to get

$$P(1 \leq X \leq 2) = \int_1^2 \frac{3x^2}{8} = \frac{7}{8}$$

□

Example 1.5.21

Suppose we select a point at random in the interior of a circle of radius 1. Let X be the distance of the selected point from the origin. Find the CDF and PDF of X and $P(\frac{1}{4} \leq X \leq \frac{1}{2})$.

Solution The sample space is $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$. Because points are chosen at random the events with equal area are equilikely. Thus given some event $A \subset S$ the probability that a point is chosen within that event is proportional to the area inside the circle. That is

$$P(A) = \frac{\text{area of A}}{\pi}$$

To find the CDF, we need a function $F_X(x) = P(\{c \in S : X(c) \leq x\})$. Notice that $X[c = (x, y)] = \sqrt{x^2 + y^2}$ for all $c \in S$. So we see that when $x < 0$, $F_X(x) = 0$. If $0 \leq x < 1$ we that

$$F_X(x) = P(\{(x, y) \in S : x^2 + y^2 < x\})$$

That it, it is the probability that a point is chosen in a circle with radius x. Using the above defintion of $P(A)$ we see that

$$F_X(x) = \frac{\pi x^2}{\pi} = x^2$$

For $x > 1$, all events in S are contained within the circle with radius 1, so the probability is 1. Thus we get that

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

Then taking the dervative of $F_X(x)$ we get that

$$f_X(x) \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Using this we see that

$$P(\frac{1}{4} < X < \frac{1}{2}) = \int_{\frac{1}{4}}^{\frac{1}{2}} f_X(x)dx = \int_{\frac{1}{4}}^{\frac{1}{2}} 2xdx = [x^2]_{\frac{1}{4}}^{\frac{1}{2}} = \frac{3}{16}$$

□

Example 1.5.22

Let the random variable be the time in seconds between incoming telephone calls at a busy switchboard. Suppose that a reasonable probability model for X is given by the pdf

$$f_X(x) \begin{cases} \frac{1}{4}e^{-x/4} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Show that it is a valid pdf and find the probability that the time between successive phone calls exceeds 4 seconds.

Solution It is clear to see that $f_X \geq 0$ for all x. Next we see that

$$\int_0^\infty \frac{1}{4}e^{-x/4} = \lim_{b\to\infty}[-e^{-x/4}]_0^b = \lim_{b\to\infty}\frac{-1}{e^{b/4}} + 1 = 1$$

Now what we need to find is $P(X > 4)$. Since $F_X(x) = \int_{-\infty}^x f_X(t)dt$, we simply flip the bounds that get that

$$P(X > 4) = \int_4^\infty \frac{1}{4}e^{-x/4}dx = e^{-1} = 0.3679$$

The distribution (f_X) graphically has a long right tail and no left tail. We say that this distribution is skewed right or positively skewed. □

1.5.4 Quantile

we now formalize quantiles (Percentiles) that we talked about earlier.

Definition 1.5.23

Let $0 < p < 1$. The quantile of order p of the distribution of a random variable X is a value ξ_p such that $P(X \leq \xi_p) \leq p$ and $P(X \leq \xi_p) \geq p$. It is also known as the $(100p)$the percentile of X.

ξ_p is really just the smallest value of x that satisfies $F_X(x) \geq p$. Some examples include the median (or sometimes called the second quartile) $\xi_{0.5}$. That is $P(X \leq \xi_{0.5}) \leq 0.5$. It is a point in the domain of X that divides the mass of the pdf into its lower and upper halves. The first and third quartiles divide each of these halves into quarters. They are, respectively $\xi_{0.25}$ and $\xi_{0.75}$. We label these quartiles as q_1, q_2 and q_3, respectively. The difference $IQR = q_3 - q_1$ is called the interquartile range of X which is used as a measure of spread or dispersion of the distribution of X. Quantiles need not be unique even for continuous random variables with pdfs. However if ξ_p is in the support of a continuous random variable X with cdf $F_X(x)$, then ξ_p is the unique solution to the equation $\xi_p = F_X^{-1}(p)$.

Example 1.5.24

Find the 0.20 quantile (20th percentile) of the distribution that has pdf $f(x) = 4x^3$, for $0 < x < 1$, and zero elsewhere

Solution We find that

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f(t)dt \\ &= \lim_{b\to-\infty}\int_b^x 4t^3 \\ &= \lim_{b\to-\infty} x^4 - x^b \\ &= x^4 \end{aligned}$$

Thus since our CDF is absolutely continuous and is invertible for positive values, we see that the 20th percentile $\xi_{0.20}$ is the unique solution to the equation

$$\xi_{0.20} = F_X^{-1}(0.20) = \sqrt[4]{0.20} = 0.67$$

□

Example 1.5.25

Find the median of random variable X that has the following cdf.

$$F_X(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{8}, & 2 \leq x < 2.5 \\ \frac{3}{16}, & 2.5 \leq x < 4 \\ \frac{1}{2}, & 4 \leq x < 5.5 \\ \frac{5}{8}, & 5.5 \leq x < 6 \\ \frac{11}{16}, & 6 \leq x < 7 \\ 1, & x \geq 7 \end{cases}$$

Solution We can immediately find that this random variable is not continuous because the CDF has jump discontinuities (step function). The median value of this cdf is the smallest value $\xi_{0.5}$ such that $P(X \leq \xi_{0.5}) \leq 0.5$. That is we need to find the smallest value such that $F_X(\xi_{0.5}) \leq 0.5$. We see that when $\xi_{0.5} = 4$, $F_X(\xi_{0.5}) = 0.5$ and for all values of $x < \xi_{0.5}$, we have that $F_X(x) < 0.5$ and for all values of $x > \xi_{0.5}$ we have that $F_X(x) > 0.5$. Thus the median of the random variable X is $\xi_{0.5} = 4$. □

1.5.5 Transformations: The cdf Technique

Let X be a continuous random variable with a known pdf f_X . As in the discrete case, we are often interested in the distribution of a random variable Y which is some transformation of X, say, $Y = g(X)$. Often we can obtain the pdf of Y by first obtaining its cdf. We illustrate this with two examples.

Example 1.5.26

Let X be a random variable with pdf $f_X(x) = \frac{1}{2}$ for $-1 < x < 1$. Find the pdf of $Y = X^2$.

Solution The cdf of Y is obatined by

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

Since $f_Y(y) = F'_Y(x)$, we get that using chain rule

$$f_Y(y) = F'_X(\sqrt{y}) - F'_X(-\sqrt{y}) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}) = \frac{1}{2} \cdot \frac{1}{2\sqrt{y}} + \frac{1}{2} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}$$

for $0 < y < 1$. □

Example 1.5.27

Suppose we select a point at random in the interior of a circle of radius 1. Let X be the distance of the selected point from the origin. Let $Y = X^2$ be a random variable. Find the CDF and PDF of Y.

Solution We saw from Example 1.5.21 that the support of X was $S_X = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\} = (0,1)$. The support for Y is the same. We also saw that

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

and

$$f_X(x) \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Thus by using F_X and the fact that the support is only positive real numbers we get that

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y}^2 = y$$

Thus we easily find that the pdf is $f_Y(y) = 1$ for $0 < y < 1$ and zero elsewhere. □

These examples illustrate the cumulative distribution function technique. As we seen in Example 1.5.27, then the transformation is injective over the support, then we can find the formula for the pdf of Y in terms of the pdf of X. We formalize this in the following theorem.

Theorem 1.5.28 : The Jacobian Method

Let X be a continuous random variable with pdf $f_X(x)$ and support S_X. Let $Y = g(X)$ where $g(x)$ is an injective differentiable function on S_X. Denote the inverse of g by $x = g^{-1}(y)$ and let $dx/dy = dg^{-1}(y)/dx$. Then the pdf of Y is given by

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

for $y \in S_Y$ where $S_Y = \{y = g(x) : x \in S_X\}$ is the support of Y.

Proof. Since g is injective and continuous then it is strictly monotone. Assume that it is increasing. The CDF is given by

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

Thus taking the derivative with respect to y to both sides we get that

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

where dx/dy is the derivative of the function $x = g^{-1}(y)$. Assume that it is decreasing. Then we have that $Y \leq y$ and $g(X) \leq y$. Thus the inequality flips since it is decreasing and we get that $X \geq g^{-1}(y)$. Thus $F_Y(y) = 1 - F_X(g^{-1}(y))$. Thus the pdf of Y is $f_Y = f_X(g^{-1}(y))\frac{-dx}{dy}$. Since g is decreasing then $dx/dy < 0$ thus $-dx/dy = |dx/dy|$. In both cases the equation holds. □

We refer to $dx/dy = dg^{-1}(y)/dx$ as the Jacobian (denoted by J) of the transformation. Without going too in depth here as this will be taught in your Calc 3 class, the Jacobian determinant measures how space is stretched or compressed by a transformation of variables. In one dimension we see that it is just the derivative but we can generalize this to n dimensions. When transforming random variables, the Jacobian accounts for how the probability density changes due to this

stretching or compression. Because area/volume/space can change under transformations. The Jacobian determinant tells you how much. If the transformation stretches space, density must decrease (to conserve probability). If it shrinks space, density must increase. The Jacobian scales the PDF accordingly so that total probability stays 1.

We can summarize the Jacobian Method into an algoritm. Assume that X is a continous random variable and $Y = g(X)$ is injective an differentiable. The following steps lead to the pdf of Y.

1. Find the support of Y

2. Solve for the inverse of the transformation. That is find $x = g^{-1}(y)$.

3. Obtain $\frac{dx}{dy}$.

4. The pdf Y is $f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}|$

Example 1.5.29

Let X be a random variable with pdf $f_X(x) = 2x$ for $0 < x < 1$. Find the pdf of $Y = X^2$.

Solution Since $g(x) = x^2$ is injective and continuous on $(0, 1)$ we can use the Jacobian method. Following the algorithm above, we first find the support of Y which is $S_Y = \{y = g(x), x \in S_X\} = (0, 1)$. Then the inverse is $g^{-1}(y) = \sqrt{y}$. We have then

$$\frac{dx}{dy} = \frac{dg^{-1}(y)}{dy} = \frac{1}{2\sqrt{y}}$$

Thus the pdf of Y is

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = 2\sqrt{y} \cdot \frac{1}{2\sqrt{y}} = 1$$

□

Example 1.5.30

Let X be a random variable with pdf $f_X(x) = 2(1 - x)$ for $0 < x < 1$. Find the pdf of $Y = -\ln(1 - X)$.

Solution Let $g(x) = -\ln(1 - x)$. We see that g is differentiable for $0 < x < 1$. We can also see that this is strictly increasing (this is easy to show). Thus we can use the Jacobian method. The

support of Y is $S_Y = \{y = g(x) : x \in S_X\} = (0, \infty)$. We now find the inverse of g.

$$\begin{aligned} y &= -\ln(1-x) \\ e^y &= e^{-\ln(1-x)} \\ e^y &= e^{\ln(\frac{1}{1-x})} \\ e^y &= \frac{1}{1-x} \\ 1-x &= \frac{1}{e^y} \\ x &= 1 - \frac{1}{e^y} = g^{-1}(y) \end{aligned}$$

Then we see easily that

$$\frac{dx}{dy} = \frac{dg^{-1}(y)}{dy} = e^{-y}$$

Thus the pdf is

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = -2e^{-y} \cdot e^{-y} = -2e^{-2y}$$

for $y > 0$ and 0 elsewhere. □

1.6 Expectation of a Random Variable

In this section we introduce the expectation operator. Before we go into the definition I will explain the intuition and reasoning for this operator. If we think of a random variable X as a game that produces a numerical outcome every time you play. *The expectation* $E(X)$ is the long-run average outcome you would see if you repeated the game over and over and kept taking the mean of the results.

For example, imagine rolling a fair die $N = 600$ times. Think of one roll of the die as a lottery ticket: If the face shows 1, you "earn" 1 point. If the face shows 2, you earn 2 points, ... , If the face shows 6, you earn 6 points. The expected value is what you would average per ticket if you bought an enormous stack of identical tickets (rolled the die many, many times). Pretend we roll the die $N = 600$ times. Because the die is fair, the expected number of times each face appears is $N/6 = 100$. If we add up the total points we get:

Face x	**Expected rolls**	**Points each time**	**Points contributed to the grand total**
1	100	1	$1 \times 100 = 100$
2	100	2	$2 \times 100 = 200$
3	100	3	$3 \times 100 = 300$
4	100	4	$4 \times 100 = 400$
5	100	5	$5 \times 100 = 500$
6	100	6	$6 \times 100 = 600$

Total points $= 100 + 200 + 300 + 400 + 500 + 600 = 2100$. Divide by the number of rolls to get the average:

$$\frac{2100}{600} = 3.5$$

If we were to do this with $N = 100,000$ we would get a value close to 3.5 as well. In general, imagine rolling the die N times and taking the usual average

$$\frac{1}{N}\sum_{k=1}^{N} X_k.$$

As N grows, the **relative frequency** of each face x approaches its probability $p(x) = 1/6$. So, in the long run, about $Np(x)$ of the N rolls will show x. The running average therefore looks like

$$\frac{1}{N}\Big[1 \cdot Np(1) + 2 \cdot Np(2) + \cdots + 6 \cdot Np(6)\Big] = \sum_x x\, p(x),$$

exactly the formula for our discrete random variable expectation. Those products $x\,p(x)$ are literally the amount each face contributes to the grand average.

Definition 1.6.1

Let X be a random variable. If X is continuous random with pdf $f_X(x)$ and

$$\int_{-\infty}^{\infty} |f_X(x)|dx < \infty$$

then the expectation of X is

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

If X is a discrete random variable with pmf $p_X(x)$ and

$$\sum_x |x| p_X(x) < \infty$$

then the expectation of X is

$$E(X) = \sum_x x p_X(x)$$

We often use μ as shorthand to denote the expected value. i.e, $\mu = E(X)$. Sometime, the expectation $E(x)$ is called the expected value of X, or the mean of X.

Example 1.6.2

Back to the Example of choosing 2 out of 6 applications which consists of 3 men and 3 women. Here X is the number of female applications in the sample. The probability mass function is as follows:

$$p_X(0) = 1/5$$
$$p_X(1) = 3/5$$
$$p_X(2) = 1/5$$

Find the expectation of X.

Solution The support of X is $S_X = \{0, 1, 2\}$. Using the formula above we get that

$$E(X) = \sum_{x=0}^{2} x p_X(x) = 0 \cdot \frac{1}{5} + 1 \cdot \frac{3}{5} + 2 \cdot \frac{1}{5} = 1$$

□

Example 1.6.7

Let X be a random variable with pdf $f_X(x) = \frac{x+1}{18}$ for $-2 < x < 4$ and zero elsewhere. Find $E(X)$.

Solution The expectation of a continuous random variable is defined by

$$E(X) = \int_{-2}^{4} x f_X(x) dx = \int_{-2}^{4} x \cdot \frac{x+1}{18} dx.$$

We get that

$$\begin{aligned} E(X) &= \int_{-2}^{4} x \cdot \frac{x+1}{18} dx \\ &= \frac{1}{18} \left[\left[\frac{x^3}{3} \right]_{-2}^{4} + \left[x^2 \right]_{-2}^{4} \right] = 2 \end{aligned}$$

□

Let X be a random variable. Let $Y = g(X)$ be a transformation of X. To find the expectation of Y we can find the distribution of Y. However we can find it using the expectation of X.

Theorem 1.6.8

Let X be a random variable and $Y = g(X)$ for some function g.

1. Suppose X is a discrete random variable with pmf $p_X(x)$ and support denote by S_X. If $\sum_{x \in S_X} |g(x)| p_X(x) < \infty$, then the expectation of Y exist and is given by

$$E(Y) = \sum_{x \in S_X} g(x) p_X(x).$$

2. Suppose X is a continuous random variable with pdf $f_X(x)$. If $\int_{-\infty}^{\infty} |g(x)| f_X(x) dx v < \infty$, then the expectation of the Y exist and

$$E(Y) = \int_{-\infty}^{\infty} g(x) p_X(x) dx$$

Proof. Proof requires some calc so will be skipping this for now □

Example 1.6.9

Let X be a discrete random variable with pmf $p_X(x) = \left(\frac{1}{2}\right)^{x+1}$, for $x = 1, 2, \ldots$ and $Y = e^{-X}$. Find $E(Y)$

Solution Let $g(x) = e^{-x}$. Then we see that

$$\sum_{x \in S_X} \frac{\left(\frac{1}{2}\right)^{x+1}}{e^x} = \frac{1}{2} \sum_{x \in S_X} \left(\frac{1}{2e}\right)^x$$

which converges. Thus we see that the expectation of Y is the geometric series

$$E(Y) = \frac{1}{2} \sum_{x \in S_X}^{\infty} \left(\frac{1}{2e}\right)^x = \frac{e}{2e-1}.$$

□

Example 1.6.10

Let X be a continuous random variable with pdf $f_X(x) = 2x$ for $0 < x < 1$ and $Y = \frac{1}{1+X}$. Find $E(X)$.

Solution We see that

$$E(Y) = \int_0^1 \frac{2x}{1+x} dx = \int_0^1 \frac{2x+2-2}{1+x} dx = 2\int_0^1 1 dx - 2\int_0^1 \frac{1}{1+x}$$

Evaluating this we get that

$$E(Y) = 2 - 2\ln(2) = 2(1 - \ln(2)).$$

□

We now show that the expectation operator on random variable $E(X)$ is a linear operator.

Theorem 1.6.11

Let $g_1(X)$ and $g_2(X)$ be function on a random variable X. Suppose the expectation of g_1, g_2 exist. Then for any constants k_1 and k_2, the expectation $k_1 g_1(X) + k_2 g_2(X)$ exist and is given by

$$E(k_1 g_1(X) + k_2 g_2(X)) = k_1 E[g_1(X)] + k_2 E[g_2(X)]$$

Proof. We will prove this for the discrete case as the continuous case is similar. We see that from the defintion of expectation,

$$E(k_1 g_1(X) + k_2 g_2(X)) = \sum_{x \in S_X} (k_1 g_1(x) + k_2 g_2(x)) p_X(x) = \sum_{x \in S_X} k_1 g_1(x) p_X(x) + \sum_{x \in S_X} k_2 g_2(x) p_X(x)$$

The result follows. We were able to split up the sum since by assumption g_1 and g_2 expectations exist which means their series are absolutely convergent which allows us to split the sum. □

Example 1.6.12

Let X have the pmf $p_x(x) = \frac{x}{6}$ for $x = 1, 2, 3$ and zero elsewhere. Find $E(6X^3 + X)$.

Solution Using the linearity of expectation we get that

$$E(6X^3 + X) = 6E(X^3) + E(X) = 6\sum_{x=1}^{3} \frac{x^4}{6} + \sum_{x=1}^{3} \frac{x^2}{6} = \frac{301}{3}$$

□

1.7 Some Special Expectations

There are many different kinds of expectations that we end up using in many different fields. First let X be a discrete random variable with pmf $p_X(x)$. Let the support of X be $S_X\{a_1, a_2, \dots\}$. It follows that

$$E(X) = \sum_{x \in S_X} x p_X(x) = a_1 p_X(a_1) + a_2 p_X(a_2) + a_3 p_X(a_3) + \cdots$$

This sum can be seen as the weighted average of the values a_i with the weights being $p_X(a_i)$. Thus we call this the mean value of X.

Definition 1.7.1

Let X be a random variable whose expectations exist. The mean value μ of X is defined by $\mu = E(X)$.

We now look at how much the values of X vary or spread out around their mean μ. The variance of X is defined as

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

This means:

- Take how far each value is from the mean: $(x - \mu)$
- Square that deviation: $(x - \mu)^2$
- Multiply it by the probability that x occurs : $(x - \mu)^2 p_X(x)$
- Add all of those up (i.e., take the expected value)

So the formula is

$$E[(X - \mu)^2] = \sum_x (x - \mu)^2 p_X(x) = (a_1 - \mu)^2 p_X(a_1) + (a_2 - \mu)^2 p_X(a_2) + \cdots$$

Intuitively, you're computing a weighted average of the squared deviations from the mean. The "weights" are the probabilities of each value. It tells you how much the values of X typically deviate/varies from the mean. A small variance means values are tightly clustered around the mean. A large variance means values are spread out more. This is used heavily in Machine learning. Specifically in the Regression learning model cost function.

Definition 1.7.2

For a random variable X with mean $E(X) - \mu$, the variance of X is defined as the expected value of $(X - \mu)^2$. That is

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

It is easy to see that

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

Using the linearity of E we get that

$$\begin{aligned}\sigma^2 &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2\end{aligned}$$

which gives us an easier way to compute the variance of X.

The standard deviation σ is just the positive square root of the variance σ^2. That is $\sigma = \sqrt{E[(X-\mu)^2]}$. While variance gives us a measure of how spread out the values are, it's measured in squared units, which can be hard to interpret. Standard deviation brings the units back to the same as the original data. So:

- If X is in dollars, μ is in dollars.
- Variance σ^2 is in dollars^2.
- Standard Deviation σ is in dollars.

This makes standard deviation easier to interpret and compare.

Example 1.7.3

Consider two users. One receives either 8 or 12 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 20 emails, also with a 50-50% chance. Find the expectation, variance, and standard deviation of these two random variables.

Solution Let X be the random variable of the number of emails that user one receives. Then

$$E(X) = 8 \cdot \frac{1}{2} + 12 \cdot \frac{1}{2} = 10$$

Then

$$E(X^2) = 8^2 \cdot \frac{1}{2} + 12^2 \cdot \frac{1}{2} = 104$$

Hence the variance of X is $\sigma_X^2 = E(X^2) - \mu^2 = 104 - 100 = 4$. The standard deviation is then $\sigma_X = \sqrt{4} = 2$.

Similarly let Ybe the random variable of the number of emails that user two receives. We get that $E(Y) = 10$ and $E(Y^2) = 200$. Then $\sigma_Y^2 = 200 - 100 = 100$. Thus the standard deviation is $\sigma_Y = 10$. □

Even though variance is not a linear operator we have the following result.

Theorem 1.7.4

Let X be a random variable with a finite mean μ and variance σ^2. Then for all constants a and b,

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

Proof. Using the linearity of E we get that $E(aX + b) = a\mu + b$. Thus using the definition of variance we get that

$$\text{Var}(aX + b) = E\left[\left[(aX + b) - (a\mu + b)^2\right]\right] = E\left[[a(X - \mu)]^2\right] = E[a^2(X - \mu)^2] = a^2\text{Var}(X).$$

□

Based on this theorem, for standard deviations, $\sigma_{aX+b} = |a|\sigma$. We show this with an example

Example 1.7.5

Let X be a random variable with pdf $f_X(x) = 1/(2a)$ for $-a < x < a$ and zero elsewhere. Find the mean and variance of X. Let $Y = 2X$. Find the standard deviation of Y.

Solution We see that

$$\mu = E(X) = \int_{-a}^{a} x\frac{1}{2a}dx = \frac{1}{2a}\left[\frac{x^2}{2}\right]_{-a}^{a} = 0$$

and then

$$\sigma^2 = \int_{-a}^{a}(x-\mu)^2\frac{1}{2a}dx = \frac{1}{2a}\int_{-a}^{a}x^2dx = \frac{1}{2a}\left[\frac{x^3}{3}\right]_{-a}^{a} = \frac{a^2}{3}$$

Thus the standard deviation of X is $\sigma = a/\sqrt{3}$. Next, to find the standard deviation of Y, we need to find expectation and variance of Y which means we need to find the pdf of Y. Since $g(x) = 2x$ is an injective function, we can use the Jacobian Method. The inverse is $g^{-1}(y) = \frac{1}{2}y$ and $dx/dy = 1/2$. Thus we see that

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = \frac{1}{2a}\cdot\frac{1}{2} = \frac{1}{4a}$$

for $-2a < y < 2a$ and zero elsewhere. We see that $\sigma_Y = (2a)/\sqrt{3}$. Hence, the standard deviation of Y is twice that of X, reflecting the fact that the probability for Y is spread out twice as much (relative to the mean zero) as the probability for X. □

For our third special expectation, we find a way to basically encode information about our random variable into a single function.

Definition 1.7.6

Let X be a random variable such that for some $h > 0$, $E(e^{tX})$ exist for $-h < t < h$. The moment generating function (mgf) of X is defined to be the function $M(t) = E(e^{tX})$ for $-h < t < h$.

The MGF doesn't necessarily exist for all values of t, just some small open interval around 0 is enough for the MGF to be valid. It's called "moment generating" because it generates the moments (like the mean, variance, etc.) of X by taking derivatives. You can think of the MGF as a way to encode the entire distribution of a random variable into a single function. From that function, you can: Extract the mean, variance, and all higher moments. Sometimes uniquely identify the distribution.

Example 1.7.7

Let X be a random variable with pmf

$$p_X(x) = \frac{1}{3}\left(\frac{2}{3}\right)^{x-1}$$

for $x = 1, 2, 3, \ldots$. Find the MGF of X.

Solution We see that

$$M(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} \frac{1}{3} \left(\frac{2}{3}\right)^{x-1} = \frac{1}{3} e^t \sum_{x=1}^{\infty} \left(e^t \cdot \frac{2}{3}\right)^{x-1} = \frac{1}{3} e^t \left(1 - e^t \frac{2}{3}\right)$$

given that $e^t(2/3) < 1$ or $t < \ln(2/3)$. This last interval is an open interval of 0; hence, the mgf of X exists. □

Going back to our discussion above, the exponential function e^{tX} is very "smooth" and has a power series expansion:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

Now take the expected value:

$$M_X(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots$$

Each term in the expansion contains the moments $E[X], E[X^2], E[X^3], \ldots$ — that's why it's called the moment generating function. To get the mean $\mu = E[X]$, take the first derivative of $M(t)$ and evaluate at $t = 0$:

$$M'(0) = E[X]$$

To get the variance, you use:

$$\text{Var}(X) = E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2$$

What we found were the 1st and 2nd moments of the distribution or in general, the kth moment of X. We formalize this:

Theorem 1.7.8

If $M(t)$ exist, then for any positive integer k

$$M^{(k)}(0) = E(X^k)$$

Proof. The general idea of the proof was shown above. □

Example 1.7.9

Let $M(t) = \frac{1}{6}e^t + \frac{2}{6}e^{2t} + \frac{3}{6}e^{3t}$. Find $E(X)$, $\text{Var}(X)$ and the distribution of X.

Solution We see that

$$E(X) = M^{(1)}(0) = \frac{1}{6}e^0 + 2 \cdot \frac{2}{6}e^{2 \cdot 0} + 3 \cdot \frac{3}{6}e^{3 \cdot 0} = \frac{14}{6}$$

For the variance we first find $E(X^2)$.

$$E(X^2) = M^{(2)}(0) = \frac{1}{6}e^0 + 2 \cdot \frac{4}{6}e^{2 \cdot 0} + 3 \cdot \frac{9}{6}e^{3 \cdot 0} = 6$$

Thus we get that $\text{Var}(X) = E(X^2) - (E(X))^2 = 6 - (14/6)^2 = 0.556$. For the distribution, we see that the mgf has the form $M(t) = E(e^{tX}) = \sum_x p(x)e^{tx}$. We see that easily $P(X = 1) = 1/6, P(X = 2) - 2/6, P(X = 3) = 3/6$ □

One of our applications of the mgf was that it could uniquely identify distributions. Let X and Y be two random variables with mfg. If X and Y have the same distribution, that is $F_X(z) = F_Y(z)$ for all z, then we see clearly that $M_X(t) = M_Y(t)$ in a neighborhood around zero. But like we said above, the converse of this statement is true. The proof of the converse is out of the scope of probably every course you'll take in your undergrad. It is because of this theorem that allows us to extract information (moments) from the mgf for our distribution.

Theorem 1.7.10 : Uniqueness of MGF

If there exist a $\delta > 0$ such that $M_X(t) = M_Y(t) < \infty$ for all $t \in (-\delta, \delta)$, then $F_X(t) = F_Y(t)$ for all $t \in \mathbb{R}$.

Proof. I'll attempt to prove this in third year. This proof is so advanced that it's omitted in almost all undergrad texts. □

1.8 Important inequalities

In this section we talk about important inequalities about expectations. The first is the Markov inequality. It gives an upper bound on the probability that a random variable exceeds a threshold. In plain terms: The probability that a nonnegative quantity $u(X)$ is large (greater than or equal to c) is at most the average of that quantity divided by c. So, if the mean is small, the chance of large values is small.

Theorem 1.8.1 : Markov's Inequality

Let $u(X)$ be a non-negative function on the random variable X. If $E[u(X)]$ exist then for every positive constant c,

$$P(u(X) \geq c) \leq \frac{E[u(X)]}{c}$$

Proof. The proof will be done when X is a continuous random variable as in the discrete case we just switch the integrals with sums. The idea of the proof will be to bound $E[u(X)]$ from below by $P(u(X) \geq c) = \int_A f_X(x)dx$ by splitting up the integral. We begin by defining $A = \{x : u(x) \geq c\}$. Then we see that we can split up the integral :

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f_X(x)dx = \int_A u(x)f_X(x)dx + \int_{A^c} u(x)f_X(x)dx$$

The key thing to notice here is that the two integrals are non-negative. This is because u and f are both non-negative functions. Thus we get that

$$E[u(X)] \geq \int_A u(x)f_x(x)dx$$

However since $u(x) \geq c$ then we can simply apply the integral of $f(x)$ over A to both sides to get

$$E[u(X)] \geq \int_A u(x)f_x(x)dx \geq c\int_A f_x(x)dx$$

Notice that

$$\int_A f_x(x)dx = P(x \in A) = P(u(X) \geq c).$$

thus we get our result

$$E[u(X)] \geq cP(u(X) \geq c).$$

□

The preceding theorem is a generalization of an inequality that is often called Chebyshev's Inequality. Before we do this, we first prove an interesting result.

Lemma 1.8.2

Let X be a random variable and let m be a positive integer. Suppose $E[X^m]$ exist. If k is a positive integer such that $k \leq m$ then $E[X^k]$ exist.

Proof. We prove this for the continous case as the discrete case is similar. Let $f_X(x)$ be the pdf. Then we see that

$$\begin{aligned}
\int_{-\infty}^{\infty} |x^k| f(x)dx &= \int_{|x|\leq 1} |x^k| f(x)dx + \int_{|x|>1} |x^k| f(x)dx \\
&\leq \int_{|x|\leq 1} f(x)dx + \int_{|x|>1} |x^m| f(x)dx \\
&\leq \int_{-\infty}^{\infty} f(x)dx + \int_{|x|>1} |x^m| f(x)dx \\
&= 1 + E[X^m] < \infty.
\end{aligned}$$

In the second line we simply used the fact that $|x|^k \leq 1$ where $|x| \leq 1$ thus its compressing f. In the third line we simply just extend the integral to $\mathbb{R}$. Overall we get our desired result easily. □

Theorem 1.8.3 : Chebyshev's Inequality

Let X be a random variable with a finite variance σ^2 (by Lemma 1.8.2 this implies $\mu - E(X)$ exist). Then for every $k \geq 0$,

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

Proof. From the Markov's inequality, we simply let $u(X) = (X - \mu)^2$ and $c = k^2\sigma^2$. Then we have

$$P(u(X) \geq c) = P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{\text{Var}(X)}{k^2\sigma^2} = \frac{1}{k^2}$$

□

Hence we get that $1/k^2$ is an upper bound for the probability $P((X - \mu)^2 \geq k\sigma)$.

Example 1.8.4

Let X have the pdf $f_X(x) = \frac{1}{2\sqrt{3}}$ for $-\sqrt{3} < x < \sqrt{3}$. Find $P(|X - \mu| \geq k\sigma)$ for $k = 3/2$ and compare it by the upper bound provided by Chebyshev's inequality.

Solution We first need to find the mean μ in order to find the standard deviation. We see that

$$E(X) = \mu = \int_{-\sqrt{3}}^{\sqrt{3}} x \frac{1}{2\sqrt{3}} dx = 0$$

We then find $E(X^2)$:

$$E(X^2) = \int_{-\sqrt{3}}^{\sqrt{3}} x^2 \frac{1}{2\sqrt{3}} dx = 1$$

Thus we see that

$$\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2 = 1 - 0 = 1.$$

So we get $\sigma = 1$. Thus we need to find $P(|X| \geq 3/2) = 1 - P(|X| \leq 3/2)$. We get that

$$P(|X| \geq 3/2) = 1 - P(|X| \leq 3/2) = 1 - \int_{-3/2}^{3/2} \frac{1}{2\sqrt{3}} dx \approx 0.134.$$

Chebyshev's Inequality gives us the upper bound $1/k^2 = 4/9 \approx 0.444$. □

There is talk about convex functions and Jensen's inequality in the textbook but this is extra material not needed, so I will not go over it.

1.9 Practice Problems

These are the practice problem solutions from the list (questions are from the textbook) and tutorial problems organized by each section. Note some questions are left unfinished.

1.9.1 Section 1.2 Answers

These questions are trivial and I am not doing them.

1.9.2 Section 1.3 Answers

1.3.3

A coin is to be tossed as many times as necessary to turn up one head. Thus the elements c of the sample space C are $H, TH, TTH, TTTH$, and so forth. Let the probability set function P assign to these elements the respective probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and so forth. Show that $P(C) = 1$. Let $C_1 = \{c : c \text{ is } H, TH, TTH, TTTH, \text{ or } TTTTH\}$. Compute $P(C_1)$. Next, suppose that $C_2 = \{c : c \text{ is } TTTTTH \text{ or } TTTTTTH\}$. Compute $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

Solution Suppose that the coin is tossed n times so that the last coin tossed is heads. Then the sequence of heads/tails are $T_1T_2...T_{n-1}H_n$. Thus by definition of C, we have that $T_1T_2...T_{n-1}H_n \in C$. Thus for any arbitrary iteration of this random experiment we have that the outcome always lives in C. Thus $P(C) = 1$. Next, notice that C_1 is composed of simple events being $A_1 = \{H\}, A_2 = \{TH\}, \ldots$. Thus since these sets are disjoint then by Probability Axiom part 3 we have that

$$P(C_1) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} = \frac{31}{32}$$

Similarly we find that

$$P(C_2) = \frac{1}{64} + \frac{1}{128} = \frac{3}{128}$$

Also notice that $C_1 \cap C_2 = \emptyset$. Thus by probability axiom part 2 we have that $P(C_1 \cap C_2) = 0$. Then since C_1 and C_2 are disjoint then we can use probability axiom part 3 again to get

$$P(C_1 \cup C_2) = P(C_2) + P(C_2) = \frac{31}{32} + \frac{3}{128} = \frac{127}{128}$$

□

1.3.9

Let $A_1, A_2, \ldots, A_k$ be k events that are well-behaved. Then

$$P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{k-1} P(A_1 \cap A_2 \cap \cdots \cap A_k)$$

Solution Done in Theorem 1.3.24 □

1.3.10

A bowl contains 16 chips, of which 6 are red, 7 are white, and 3 are blue. If four chips are taken at random and without replacement, find the probability that: (a) each of the four chips is red; (b) none of the four chips is red; (c) there is at least one chip of each color.

Solution For a: There are $\binom{16}{4}$ combinations we can have by choosing 4 random chips without replacement. There are $\binom{6}{4}$ combinations we can choose 4 red chips from the 16. Thus we get that

$$P(A) = \frac{\binom{6}{4}}{\binom{16}{4}} = 0.0082.$$

For B: We need the probability that none of the chips are red. Then we want to choose 0 red chips from the 6 red chips and 4 from the 10 chips of the other colors. We get that

$$P(B) = \frac{\binom{6}{0}\binom{10}{4}}{\binom{16}{4}} = 0.1154.$$

For c: We need the probability that there is at least one chip of each color. We can have 2 red, 1 white and 1 blue or 2 white, 1 red, 1 blue or 2 blue 1 white 1 red. We get that

$$P(C) = \frac{\binom{6}{2}\binom{7}{1}\binom{3}{1} + \binom{6}{1}\binom{7}{2}\binom{3}{1} + \binom{6}{1}\binom{7}{1}\binom{3}{2}}{\binom{16}{4}} = 0.45.$$

□

1.3.11

A person has purchased 10 of 1000 tickets sold in a certain raffle. To determine the five prize winners, five tickets are to be drawn at random and without replacement. Compute the probability that this person wins at least one prize. Hint: First compute the probability that the person does not win a prize.

Solution Let A denote the event where the person wins at least one prize. Following the hint we will find $P(A^c)$. That is we find the probability that the person wins no prize. There are $\binom{1000}{5}$

combinations of tickets drawn at random and without replacement. If the person wins no prizes then that means there 10 tickets were chosen from the 990 tickets that were not the winning tickets. Thus we get that $P(A^c) = \frac{\binom{990}{5}}{\binom{1000}{5}}$. Then using the theorem we have that

$$P(A) = 1 - P(A^c) = 1 - \frac{\binom{990}{5}}{\binom{1000}{5}} = 0.0491$$

□

1.3.13

Three distinct integers are chosen at random from the first 20 positive integers. Compute the probability that: (a) their sum is even; (b) their product is even.

Solution For a: Since we are choosing three distinct numbers and need to look at their sum, order does not matter since addition is commutative. Moreover we are choosing these numbers without replacement. Thus there are $\binom{20}{3}$ combinations for us to choose these three distinct integers. In order for this sum to be even, we need the three distinct integers to satisfy, not in this order, (even,odd,odd) or (even,even,even). There are 10 odd and 10 even numbers in the first 20 positive integers. There are $\binom{10}{3}$ ways to choose three distinct even numbers from the 20 positive integers. Then there are $\binom{10}{2}\binom{10}{1}$ ways to choose two odd and one even number. Thus we get that

$$P(A) = \frac{\text{nCr}\,(10,3) \;+\; \text{nCr}\,(10,1)\,\text{nCr}\,(10,1)\,\text{nCr}\,(9,1))}{\text{nCr}\,(20,3)} = 0.5$$

For b: For the product of three distinct positive integers to be even, then only one of the three integers need to be even. However instead of adding all possible combinations of one even two odd, two even one odd, three even zero odd, we can instead find the probability that the product is odd. Then all factors need to be odd. Thus there are $\binom{10}{3}$ ways we can choose these distinct integers so that they are all odd. We get that then

$$P(B) = 1 - P(B^c) = 1 - \frac{\binom{10}{3}}{\binom{20}{3}} = 0.895$$

□

1.3.15

In a lot of 50 light bulbs, there are 2 bad bulbs. An inspector examines five bulbs, which are selected at random and without replacement.

(a) Find the probability of at least one defective bulb among the five.

(b) How many bulbs should be examined so that the probability of finding at least one bad bulb exceeds $\frac{1}{2}$?

Solution **a)** Suppose that the probability of selecting a bulb from the 50 light bulbs are equilikely. Let A denote the event where at least one of the bulb is defective. We need to find $P(A)$. Instead we can find the probability that all 5 bulbs are not defective. That is we can find $P(A^c)$ and then use Theorem 1.3.3 to get $P(A) = 1 - P(A^c)$. We then get that

$$P(A^c) = \frac{48}{50}\frac{47}{49}\frac{46}{48}\frac{45}{47}\frac{44}{46} = 0.809$$

Then $P(A) = 1 - 0.809 = 0.191$

b) Let A be the event where at least one bulb is defective from the n bulbs selected by the investigator. Then we find the probability $P(A)$ as we do in part A. To find $P(A^c)$, we find the total number of ways to choose n good light bulbs divided by the total number of ways to choose n light bulbs. That is

$$P(A^c) = \frac{\binom{48}{n}}{\binom{50}{n}}$$

We then need the smallest n such that

$$1 - P(A^c) > \frac{1}{2}$$

or

$$\frac{\binom{48}{n}}{\binom{50}{n}} < \frac{1}{2}$$

We see that $n = 13$ works. □

1.9.3 Section 1.4 Answers

1.4.1

If $P(A_1) > 0$ and if $A_2, A_3, A_4, \ldots$ are mutually disjoint sets, show that

$$P(A_2 \cup A_3 \cup \cdots \mid A_1) = P(A_2 \mid A_1) + P(A_3 \mid A_1) + \cdots .$$

Solution We begin with the defintion of condtional probability

$$\begin{aligned} P(A_2 \cup A_3 \cup \cdots \mid A_1) &= \frac{P[A_1 \cap \bigcup_{n=2}^{\infty} A_n]}{P(A_1)} \\ &= \frac{P(\bigcup_{n=2}^{\infty} A_1 \cap A_n)}{P(A_1)}. \end{aligned}$$

Since $A_1 \cap A_n$ are disjoint we see that

$$\begin{aligned} P(A_2 \cup A_3 \cup \cdots \mid A_1) &= \frac{P(\bigcup_{n=2}^{\infty} A_1 \cap A_n)}{P(A_1)} \\ &= \frac{\sum_{n=2}^{\infty} P(A_1 \cap A_n)}{P(A_1)} \\ &= \frac{\sum_{n=2}^{\infty} P(A_1) P(A_n|A_1)}{P(A_1)} \\ &= \sum_{n=2}^{\infty} P(A_n|A_1) \end{aligned}$$

□

1.4.3

Suppose we are playing draw poker. We are dealt (from a well-shuffled deck) five cards, which contain four spades and another card of a different suit. We decide to discard the card of a different suit and draw one card from the remaining cards to complete a flush in spades (all five cards spades). Determine the probability of completing the flush.

Solution Let A denote the event where we draw four spades and another card of a different suit. Let B denote the event where after we discard the extra card we draw a spade. We need to find $P(B|A)$. Using the definition of conditional probability we get that

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

However notice that since A and B are independent then we have that $P(A \cap B) = P(A)P(B)$. So $P(B|A) = P(B)$. Since there are 13 spades in a deck and we have 4 of them. There were 52 cards but we removed 4 and added 1 back so we have 48 cards left. Thus $P(B|A) = P(B) = 9/48$. □

1.4.4

From a well-shuffled deck of ordinary playing cards, four cards are turned over one at a time without replacement. What is the probability that the spades and red cards alternate?

Solution There are two combinations of the four cards where the spades and red cards alternate. It can either be spade, red, spade, red or red first instead. There are 13 spades and 26 red cards. Thus the total number of sequences are $P_4^{52} = 52 \cdot 51 \cdot 50 \cdot 49$. We then choose 2 spades from the 13 spades giving us $\binom{13}{2}$ ways. Similar for 2 red there are $\binom{26}{2}$ combinations. There are 2 valid alternating patterns: Pattern 1: Spade, Red, Spade, Red or Pattern 2: Red, Spade, Red, Spade. So we multiply by 2. Within each pattern: There are 2 spade spots which means we can permute the 2 spades: 2!. There are 2 red spots which means we can permute the 2 reds: 2. Thus we get that

$$P(A) = \frac{2 \cdot \binom{13}{1}\binom{26}{1}2! \cdot 2!}{52 \cdot 51 \cdot 50 \cdot 49} = 0.31.$$

□

1.4.5

A hand of 13 cards is to be dealt at random and without replacement from an ordinary deck of playing cards. Find the conditional probability that there are at least three kings in the hand given that the hand contains at least two kings.

Solution Let A denote the event such that there are at least two kings in a hand of 13 cards. Note that there are $\binom{52}{13}$ ways to to choose 13 cards from a deck of 52 cards. Note that there are 4 kings in a deck. Then there are $\binom{4}{2}$ ways to choose two kings from the 4 kings. Then there are $\binom{48}{11}$ other ways to choose the other 11 cards that are not kings. Using mn-rule we get that there are $N_2 = \binom{4}{2}\binom{48}{11}$ total ways to have two kings in a hand of 13 cards. Similarly we get that there are $N_3 = \binom{4}{3}\binom{48}{10}$ to choose three kings in a hand of 13 cards. And finally there are $N_4 = \binom{4}{4}\binom{48}{9}$ ways to have 4 kings in a hand of 13 cards. Thus

$$P(A) = N_2 + N_3 + N_4$$

Let B denote the event that there are at least three kings in the hand of 13 cards. We need to find the probability that our hand of 13 cards has three kings given that our hand already has at least 2 kings. That is $P(B|A)$. Notice that $A \cap B = B$ since at least three kings implies there are

at least two kings, that is $A \subseteq B$. We get that

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(B)}{N_2 + N_3 + N_4} \end{aligned}$$

Solving for $P(B)$ we get that

$$P(B) = N_3 + N_4$$

Thus

$$P(B|A) = \frac{N_3 + N_4}{N_2 + N_3 + N_4} = \frac{\binom{4}{3}\binom{48}{10} + \binom{4}{4}\binom{48}{9}}{\binom{4}{2}\binom{48}{11} + \binom{4}{3}\binom{48}{10} + \binom{4}{4}\binom{48}{9}}$$

□

1.4.6

A drawer contains eight different pairs of socks. If six socks are taken at random and without replacement, compute the probability that there is at least one matching pair among these six socks. Hint: Compute the probability that there is not a matching pair.

Solution We choose 6 pairs of socks from the 8 and get $\binom{8}{6}$ combinations. To have no matching pairs, every sock you choose must come from a different pair. So there are 2^6 ways we can choose a sock from the six socks. We get that

$$P(A) = 1 - \frac{\binom{8}{6} \cdot 2^6}{\binom{16}{6}} = 0.7763.$$

□

1.4.7

A pair of dice is cast until either the sum of seven or eight appears.

(a) Show that the probability of a seven before an eight is 6/11.

(b) Next, this pair of dice is cast until a seven appears twice or until each of a six and eight has appeared at least once. Show that the probability of the six and eight occurring before two sevens is 0.546.

Solution For (a): Let X denote the random variable of the sum of the two faces of the dice. Then $P(X = 7) = 6/36 = 1/6$. Moreover $P(X = 8) = 8/36$. Since we only care about the rolls that are seven or eight we are asked to find

$$P(7|7 \text{ or } 8) = \frac{6/36}{(6+5)/36} = \frac{6}{11}$$

. For (b): Similarly we need to find $P(6 \text{ or } 8|\text{two } 7)$ □

1.4.8

In a certain factory, machines I, II, and III are all producing springs of the same length. Machines I, II, and III produce 1%, 4%, and 2% defective springs, respectively. Of the total production of springs in the factory, Machine I produces 30%, Machine II produces 25%, and Machine III produces 45%.

(a) If one spring is selected at random from the total springs produced in a given day, determine the probability that it is defective.

(b) Given that the selected spring is defective, find the conditional probability that it was produced by Machine II.

Solution Let S be our sample space of all springs produced by the factories. Then let A_i be the event where the spring is produced by factory i. So A_1 corresponds to Machine I. Since the total production is made up by these factories, then A_i are mutually exclusive and exhaustive. Thus they form a partition on S. That is $S = A_1 \cap A_2 \cap A_3$. We are told that $P(A_1) = 0.3$, $P(A_2) = 0.25$, $P(A_3) = 0.45$.

Let B denote the event where a product selected from the total products produced by all three factories is defective.. Let B_i be the event where the product produced in factory i is defective. Then $P(B_1|A_1) = 0.01$, $P(B_2|A_2) = 0.04$, $P(B_3|A_3) = 0.02$.

For (a), we need to find the probability that a product selected from the total products produced by all three factories is defective. That is we need to find $P(B|A_1 \cup A_2 \cup A_3) = P(B|S) = P(B)$. We get that using Law of Total Probability

$$\begin{aligned} P(B|A_1 \cup A_2 \cup A_3) &= \frac{P(B \cap (A_1 \cap A_2 \cap A_3))}{P(A_1 \cup A_2 \cup A_2)} \\ &= \frac{P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_2)}{P(A_1) + P(A_2) + P(A_2)} \\ &= \frac{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)}{1} \\ &= (0.3)(0.01) + (0.25)(0.04) + (0.45)(0.02) \\ &= 0.022 \end{aligned}$$

For (b), we need to find the probability that a spring was produced by Machine II given that it is defective. That is we need to find $P(A_2|B)$. We use Bayes's Theorem. That is

$$\begin{aligned} P(A_2|B) &= \frac{P(A_2 \cap B)}{P(B)} \\ &= \frac{P(A_2)P(B|A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ &= \frac{(0.25)(0.04)}{(0.3)(0.01) + (0.25)(0.04) + (0.45)(0.02)} \\ &\approx 0.4545 \end{aligned}$$

□

1.4.9

Bowl I contains six red chips and four blue chips. Five of these 10 chips are selected at random and without replacement and put in bowl II, which was originally empty. One chip is then drawn at random from bowl II. Given that this chip is blue, find the conditional probability that two red chips and three blue chips are transferred from bowl I to bowl II.

Solution Let A denote the event where the □

1.4.12

Let C_1 and C_2 be independent events with $P(C_1) = 0.6$ and $P(C_2) = 0.3$. Compute (a) $P(C_1 \cap C_2)$, (b) $P(C_1 \cup C_2)$, and (c) $P(C_1 \cup C_2^c)$.

Solution For (a): Since C_1 and C_2 are independent events, then $P(C_1 \cap C_2) = P(C_1)P(C_2|C_1) - P(C_2)P(C_2) = 0.6 \times 0.3 = 0.18$.

For (b): We have that through Inclusion/Exclusion formula, $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2) = 0.6 + 0.3 - 0.18 = 0.72$.

For (c): $P(C_1 \cup C_2^c) = P(C_1) + (1 - P(C_2)) - P(C_1 \cap C_2^c)$. See that from Proposition 1.4.17 $P(C_1 \cap C_2^c) = P(C_1)P(C_2^c|C_1) = P(C_1)P(C_2^c) = (0.6)(1 - 0.3)$. So we have $P(C_1 \cup C_2^c) = 0.6 + 0.7 - (0.6)(0.7) = 0.88$. □

1.4.25

Let the three mutually independent events C_1, C_2, and C_3 be such that $P(C_1) = P(C_2) = P(C_3) = \frac{1}{4}$. Find $P\left[(C_1^c \cap C_2^c) \cup C_3\right]$.

Solution Using the Inclusion/Exclusion formula we have that

$$P((C_1^c \cap C_2^c) \cup C_3) = P(C_1^c \cap C_2^c) + P(C_3) - P((C_1^c \cap C_2^c) \cap C_3)$$

We know that C_1^c and C_2^c are independent so we have that $P(C_1^c \cap C_2^c) = P(C_1^c)P(C_2^c) = (1 - \frac{1}{4})^2 = 0.5625$.

We also know that $P((C_1^c \cap C_2^c) \cap C_3) = P(C_1^c)P(C_2^c)P(C_3) = (0.5625)(\frac{1}{4}) = 0.140625$. So together we have that

$$P((C_1^c \cap C_2^c) \cup C_3) = 0.5625 + \frac{1}{4} - 0.140625 = 0.671875$$

□

1.4.26

Each bag in a large box contains 25 tulip bulbs. It is known that 60% of the bags contain bulbs for 5 red and 20 yellow tulips, while the remaining 40% of the bags contain bulbs for 15 red and 10 yellow tulips. A bag is selected at random and a bulb taken at random from this bag is planted.

(a) What is the probability that it will be a yellow tulip?

(b) Given that it is yellow, what is the conditional probability it comes from a bag that contained 5 red and 20 yellow bulbs?

Solution Let S be our sample space containing bags with 25 tulip bulbs. We are told that the sample space is split between bags with 5 red and 20 yellow tulips and 15 red and 10 yellow tulips. Let A_1 be the event where the bag contains 5 red and 20 yellow tulips. Let A_2 be the event where the bag contains 15 red and 10 yellow tulips. Then notice that $A_1 \cup A_2$ forms a partition on S.

For (a): We need to the find the probability that after we select a random bag and a random bulb the bulb color is yellow. Denote B to be the event where a randomly selected tulip from a bag is yellow. Suppose that the probability of randomly selecting a tulip from a bag is equilikely. We need to find $P(B)$. Using the Law of Total Probability, we see that

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) = (0.6) \cdot \frac{20}{25} + (0.4) \cdot \frac{10}{25} = 0.64$$

For (b): We now need to find the probability that the tulip comes from a bag with 5 red and 20 yellow bulbs given that the tulip is yellow. That is we need to find $P(A_1|B)$. Using Bayes's Theorem we have that

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1 \cap B)}{P(B)} \\ &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \\ &= \frac{(0.6)(0.8)}{0.64} \\ &= 0.75 \end{aligned}$$

□

1.9.4 Section 1.5 Answers

Note this covers all answers from section 1.5, 1.6, and 1.7 in the textbook from the list of practice problems.

1.5.2

For each of the following, find the constant c so that $p(x)$ satisfies the condition of being a pmf of one random variable X.

(a) $p(x) = c(\frac{2}{3})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

(b) $p(x) = cx$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere.

Solution The conditions for a function to be a valid pmf for a random variable is that $0 \leq p(x) \leq 1$ for all x and $\sum_{i=1}^{m} p(d_i) = 1$ where $d_i \in S_X = \{d_1, d_2, \ldots, d_m\}$ is the countable support.

For (a): The first condition is trivially satisfied. What we need now is

$$\sum_{n=1}^{\infty} c\left(\frac{2}{3}\right)^n = 1$$

$$\sum_{n=1}^{\infty} \left(\frac{2}{3}\right)^n = \frac{1}{c}$$

Using the geometric series formula we get that

$$\frac{2/3}{1-(2/3)} = 2 = \frac{1}{c} \text{ or } c = \frac{1}{2}$$

For (b): Since we need $0 \leq cx \leq 1$ for $x = 1, 2, 3, 4, 5, 6$, we know that $c \leq 1/6$. Then we get that

$$\sum_{n=1}^{6} cn = c\sum_{n=1}^{6} n = c \cdot \frac{6 \cdot 7}{2} = 21c = 1 \text{ or } c = \frac{1}{21}$$

□

1.5.4

Let $p_X(x)$ be the pmf of a random variable X. Find the cdf $F(x)$ of X and sketch its graph along with that of $p_X(x)$ if:

(a) $p_X(x) = 1$, $x = 0$, zero elsewhere.

(b) $p_X(x) = \frac{1}{3}$, $x = -1, 0, 1$, zero elsewhere.

(c) $p_X(x) = \frac{x}{15}$, $x = 1, 2, 3, 4, 5$, zero elsewhere.

Solution The cdf of a random variable X is defined by $F_X(x) = P(X \leq x) = P(\{c \in S : X(c) \leq x\})$. The graphs of each CDF will be given as a desmos link.

For (a): Since the pmf is only 1 when $x = 0$, then we can easily see that $F_X(x) = 1$ when $x \geq 0$ and 0 otherwise. Graph.

For (b): We can easily find the CDF since the pmf is constant. For $x < -1$, $F_X(x) = 0$. For $-1 \leq x < 0$, $F_X(x) = 1/3$. For $0 \leq x < 1$, $F_X(x) = 2/3$. For $1 \leq x$ we have $F_X(x) = 1$. Together we get that

$$F_X(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{3}, & -1 \leq x < 0 \\ \frac{2}{3}, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

Graph.

For (c): We know that when $x < 1$, $F_X(x) = 0$. When $1 \leq x < 2$ we have that $F_X(x) = 1/15$. Then when $2 \leq x < 3$ we have that $F_X(x) = 1/15 + 2/15 = 3/15 = 1/5$. When $3 \leq x < 4$ we have that $F_X(x) = 3/15 + 3/15 = 6/15$. Then when $4 \leq x < 5$ we have that $F_X(x) = 6/15 + 4/15 = 10/15$. Then when $5 \leq x < 6$ we have that $F_X(x) = 10/15 + 5/15 = 1$. Putting it together we get that

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{15}, & 1 \leq x < 2 \\ \frac{3}{15}, & 2 \leq x < 3 \\ \frac{6}{15}, & 3 \leq x < 4 \\ \frac{10}{15}, & 4 \leq x < 5 \\ 1, & x \geq 5 \end{cases}$$

Graph. □

1.5.6

Let the probability set function of the random variable X be

$$P_X(D) = \int_D f(x)\,dx,$$

where $f(x) = \frac{2x}{9}$, for $x \in \mathcal{D} = \{x : 0 < x < 3\}$. Define the events

$$D_1 = \{x : 0 < x < 1\} \quad \text{and} \quad D_2 = \{x : 2 < x < 3\}.$$

Compute $P_X(D_1)$, $P_X(D_2)$, and $P_X(D_1 \cup D_2)$.

Solution

$$P_X(D_1) = \frac{2}{9}\int_0^1 x dx = \frac{1}{9}$$

$$P_X(D_2) = \frac{2}{9}\int_2^3 x dx = 1 - \frac{4}{9} = \frac{5}{9}$$

Since $D_1 \cap D_2 = \emptyset$, then

$$P_X(D_1 \cup D_2) = P_X(D_1) + P_X(D_2) = \frac{1}{9} + \frac{5}{9} = \frac{2}{3}$$

□

1.6.1

Let X equal the number of heads in four independent flips of a coin. Using certain assumptions, determine the pmf of X and compute the probability that X is equal to an odd number.

Solution Our sample space S consist of 4 length sequences such as $HTHT, TTTT, HHHT, \ldots,$. Then $X(HTHT) = 2$, $X(HHHT) = 3$, $H(TTTT) = 0$. More specifically,

$$S = \{x_1x_2x_3x_4 : x_i = H \text{ or T}, i = 1, 2, 3, 4\}$$

Given four independent flips of a coin $c \in S$, there are $X(c) = n$ heads in the sequence. The probability that n heads occur when flipping four coins is

$$P(c \in S : X(c) = n) = \frac{1}{2^n}$$

Thus the probability mass function of X is going to be $p_X(x) = P(X = x) = \frac{1}{2^x}$. We are now interested int he probability that X is equal to an odd number. That is we are interested in the probability that the number of heads are either 1 or 3. That is

$$p_X(\{1, 3\}) = P(\{c \in S : X(c) = 1, 3\}).$$

Since the events are mutually exclusive we find the probabilities that the number of heads equals one and three separately. Let A denote the event where the number of heads equals one. There are $2^4 = 16$ possible combinations of four independent coin flips we can have. Since they are independent we are assuming that each sequence is equilikely. There are $\binom{4}{1}$ ways we can place the

head in our sequences such as $TTTH, HTTT, TTHT$,etc. Thus we have that $P(A) = \binom{4}{1} \cdot \frac{1}{16} = 0.25$ probability. Similarly, let B denote the event where the sequence has three heads. We get that $P(B) = \binom{4}{3}\frac{1}{16} = 0.25$. Thus

$$p_X(\{1,3\}) = 0.5$$

□

1.6.8

Let X have the pmf $p_X(x) = (\frac{1}{2})^x$ for $x = 1, 2, 3, \ldots$, zero elsewhere. Find the pmf of $Y = X^3$.

Solution The support of Y is $S_Y = \{1, 8, 27, \ldots\}$.

$$p_Y(y) = P(Y = y) = P(X^3 = y) = P(X = \sqrt[3]{y}) = \left(\frac{1}{2}\right)^{\sqrt[3]{y}}$$

for $y = 1, 8, 27, \ldots$, and zero elsewhere. □

1.6.10

Let X have the pmf

$$p_X(x) = \left(\frac{1}{2}\right)^{|x|}, x = -1, -2, -3, \ldots$$

Find the pmf of $Y = X^4$.

Solution The support of Y is $S_Y = \{1, 16, 81, \ldots\}$. Then we see that

$$p_Y(y) = P(Y = y) = P(X^4 = y) = P(X = \sqrt[4]{y}) = \left(\frac{1}{2}\right)^{|\sqrt[4]{y}|}$$

for $y = 1, 16, 81, \ldots$, and zero elsewhere. □

1.7.3

Let the subsets

$$C_1 = \left\{\frac{1}{4} < x < \frac{1}{2}\right\} \quad \text{and} \quad C_2 = \left\{\frac{1}{2} \leq x < 1\right\}$$

of the space

$$\mathcal{C} = \{x : 0 < x < 1\}$$

of the random variable X be such that

$$P_X(C_1) = \frac{1}{8} \quad \text{and} \quad P_X(C_2) = \frac{1}{2}.$$

Find $P_X(C_1 \cup C_2)$, $P_X(C_1^c)$, and $P_X(C_1^c \cap C_2^c)$.

Solution Since C_1 and C_2 are mutually exclusive, we can simply add their probabilities for the union of these events

$$P_X(C_1 \cup C_2) = P_X(C_1) + P_X(C_2) = \frac{5}{8}.$$

Then using Theorem 1.3.3 we get that

$$PX(C_1^c) = 1 - \frac{1}{8} = \frac{7}{8}.$$

Next notice that $C_1^c \cap C_2^c = (C_1 \cup C_2)^c$ using DeMorgan's Laws. So we have that

$$P(C_1^c \cap C_2^c) = P((C_1 \cup C_2)^c) = 1 - P(C_1 \cup C_2) = \frac{3}{8}.$$

□

1.7.6

For each of the following pdfs of X, find $P(|X| < 1)$ and $P(X^2 < 9)$.

(a) $f(x) = \frac{x^2}{18}, \quad -3 < x < 3$, zero elsewhere.

(b) $f(x) = \frac{x+2}{18}, \quad -2 < x < 4$, zero elsewhere.

Solution Notice that $P(|X| < 1) = P(-1 < X < 1)$ and $P(X^2 < 9) = P(-3 < X < 3)$.

For (a):

$$P(|X| < 1) = \frac{1}{18}\int_{-1}^{1} x^2 dx = \frac{1}{54} + \frac{1}{54} = \frac{1}{27}.$$

$$P(-3 < X < 3) = \frac{1}{18}\int_{-3}^{3} x^2 = 1$$

For (b):

$$P(|X| < 1) = \frac{1}{18}\int_{-1}^{1} (x+2) dx = \frac{1}{4} - \frac{1}{36} = \frac{8}{36}$$

$$P(-3 < X < 3) = \frac{1}{18}\int_{-2}^{3} (x+2) dx = \frac{1}{18} \cdot \frac{25}{2} = \frac{25}{36}$$

□

1.7.9

The median and quantiles, in general, are discussed in Section 1.7.1. Find the median of each of the following distributions:

(a) $p(x) = \frac{4!}{x!(4-x)!}\left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{4-x}, \quad x = 0, 1, 2, 3, 4$, zero elsewhere.

(b) $f(x) = 3x^2, \quad 0 < x < 1$, zero elsewhere.

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

Solution The median is defined as the smallest value $\xi_{0.5}$ such that $P(X \leq \xi_{0.5}) = 0.5$. Thus we have to find the CDF and find the smallest value such that $F_X(x) \geq 0.5$.

For (a): Since this is discrete we can just plug in the values $x = 0, 1, 2, 3, 4$ into $p(x)$ and find the median. We see that $P(X \leq 0) = 0.316$ and $P(X \leq 1) = 0.316 + 0.422 = 0.738$. Thus the median is 1.

For (b): We see that the CDF is

$$F_X(x) = \int_0^x 3t^2 dt = x^3.$$

Thus since this is continous we can just solve for the value $\xi_{0.5} = \sqrt[3]{0.5} \approx 0.7937$.

For (c): We find the CDF:

$$F_X(x) = \int_{-\infty}^x \frac{1}{\pi(1+t^2)} dx = \lim_{b\to-\infty} \int_b^x \frac{1}{\pi(1+t^2)} dx = \lim_{b\to-\infty} \frac{\arctan(x)}{\pi} - \frac{\arctan(b)}{\pi} = \frac{\arctan(x)}{\pi} + \frac{1}{2}$$

Since this function is continuous we find median by solving for the inverse. That is

$$\xi_{0.5} = \tan(\pi\,(0.5 - 0.5)) = 0.$$

□

1.7.10

Let $0 < p < 1$. Find the 0.20 quantile (20th percentile) of the distribution that has pdf $f_X(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

Solution To find the 0.20 quantile, we need to find the smallest value p such that $F_X(p) \leq 0.20$, where F_X is the cdf of X. We find the cdf by integrating the pdf. That is

$$F_X(x) = \int 4x^3 dx = x^4$$

Then we need to find some p such that p^4. We get

$$p^4 \leq 0.20 \text{ or } p \leq \sqrt[4]{0.20}$$

Since our cdf is continuous we see that the 20th percentile is $p = \sqrt[4]{0.20} = 0.669$. □

1.7.22

Let X have the pdf $f_X(x) = x^2/9$, $0 < x < 3$ and zero elsewhere. Find the pdf of $Y = X^3$

Solution Since $g(x) = x^3$ is injective and differentiable over the support of X we can use the Jacobian method. First the support of Y is now $S_Y = (0, 27)$. Next we find the inverse $g^{-1}(y) = \sqrt[3]{y}$. Then we find the Jacobian $dx/dy = dg^{-1}(y)/dy = \frac{1}{3}y^{-2/3}$. Together we get that

$$p_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = \frac{\frac{y^{2/3}}{9}}{\frac{1}{3}y^{-2/3}} = \frac{1}{27}$$

for $y \in (0, 27)$ and zero elsewhere. □

1.7.23

If the pdf of X is $f_X(x) = 2xe^{-x^2}$ for $0 < x < \infty$ and zero elsewhere. Find the pdf of $Y = X^2$.

Solution The support of Y is $(0,\infty)$. Since $g(x) = x^2$ is injective for positve reals we can use the jacobian method. We see that the inverse is the positive square root. That is $g^{-1}(y) = \sqrt{y}$. Then the jacobian is $dx/dy = \frac{1}{2\sqrt{y}}$. We then see that

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = \frac{2\sqrt{y}e^{-y}}{2\sqrt{y}} = e^{-y}$$

□

1.7.25

Let X have the pdf $f_X(x) = 4x^3$, $0 < x < 1$ and zero elsewhere. Find the pdf of $Y = -\ln X^4$

Solution The support of Y is $S_Y = (0,\infty)$. This is because the singularity occurs are zero and the function blows up. We see that $g(x) = -\ln x^4$ is decreasing and differentiable on the support of Y. We can use the Jacobian method here. The inverse of g is

$$\begin{aligned} e^y &= e^{-\ln x^4} \\ e^y &= \frac{1}{x^4} \\ x^4 &= \frac{1}{e^y} \\ x &= \frac{1}{\sqrt[4]{e^y}} = g^{-1}(x). \end{aligned}$$

Next we find the Jacobian $dx/dy = dg^1(x)/dy = \frac{-1}{4}e^{\frac{-y}{4}}$. Together we get that

$$p_Y(y) = f_X(g^{-1}(x))|\frac{dx}{dy}| = 4e^{\frac{-3y}{4}} \cdot \frac{1}{4}e^{\frac{-y}{4}} = e^{-y}$$

for $y > 0$ and zero elsewhere. □

1.9.5 Section 1.6 Answers

1.8.4

Suppose that $p(x) = \frac{1}{5}$, $x = 1,2,3,4,5$, zero elsewhere, is the pmf of the discrete-type random variable X. Compute $E(X)$ and $E(X^2)$. Use these two results to find $E[(X+2)^2]$ by writing

$$(X+2)^2 = X^2 + 4X + 4.$$

Solution We being by finding $E(X)$. We see that

$$E(X) = \sum_{x=1}^{5} xp(x) = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 5 \cdot \frac{1}{5} = 3$$

Then similarly we find that

$$E(X^2) = \sum_{x=1}^{5} x^2 p(x) = 1^2 \cdot \frac{1}{5} + 2^2 \cdot \frac{1}{5} + 3^2 \cdot \frac{1}{5} + 4^2 \cdot \frac{1}{5} + 5^2 \cdot \frac{1}{5} = 11.$$

We now need to find $E[(X+2)^2] = E[X^2+4X+4]$. Using the linearity of expectation we get that

$$E[X^2+4X+4] = E(X^2)+4E(X)+4 = 11+4\cdot 3+4 = 27.$$

□

1.8.7

Let X have the pdf $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere. Consider a random rectangle whose sides are X and $(1-X)$. Determine the expected value of the area of the rectangle.

Solution The area of the rectangle will be $X(1-X) = -X^2+X$. The expected value of the area then is

$$E[X(1-X)] = E[-X^2+X] = -E(X^2)+E(X).$$

We first find $E(X)$:

$$E(X) = \int_0^1 3x^3 dx = \frac{3}{4}.$$

Then we find $E(X^2)$:

$$E(X^2) = \int_0^1 3x^4 dx = \frac{3}{5}.$$

Together we get that the expected value of the area of the rectangle is

$$E(X(1-X)) = -E(X^2)+E(X) = \frac{-3}{5}+\frac{3}{4} = 0.15.$$

□

1.8.8

A bowl contains 10 chips, of which 8 are marked \$2 each and 2 are marked \$5 each. Let a person choose, at random and without replacement, three chips from this bowl. If the person is to receive the sum of the resulting amounts, find his expectation.

Solution Let X be the random variable which outputs the sum of the amounts from the three selected chips. In order to find the expectation, we need to first find the probability distribution of X. Notice that X can either be the sum equaling \$6, \$9 or \$12. Thus $S_X = \{6, 9, 12\}$. Thus we find the pmf of X. We find that

$$p_X(6) = P(X=6) = P(\{\text{all three chips are \$2 }\}) = \frac{\binom{8}{3}}{\binom{10}{3}} \approx 0.467$$

$$p_X(9) = P(X=9) = P(\{\text{two chips are \$2 and one is \$5 }\}) = \frac{\binom{8}{2}\cdot\binom{2}{1}}{\binom{10}{3}} \approx 0.467$$

$$p_X(12) = P(X=12) = P(\{\text{two chips are \$5 and one is \$2 }\}) = \frac{\binom{8}{1}\cdot\binom{2}{2}}{\binom{10}{3}} \approx 0.067$$

Thus we can now find the expected value of X:

$$E(X) = \sum_{x\in S_X} x p_X(x) = 6\cdot 0.467 + 9\cdot 0.467 + 12\cdot 0.067 = 7.809.$$

□

1.8.12

Let X have the pdf $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere.

(a) Compute $E(X^3)$.

(b) Show that $Y = X^3$ has a uniform$(0, 1)$ distribution.

(c) Compute $E(Y)$ and compare this result with the answer obtained in part (a).

Solution For (a): We see that

$$E(X^3) = \int_0^1 3x^5 dx = 0.5.$$

For (b): Recall that any random variable whose pdf/pmf is constant on the support is considered uniform. Since $g(x) = x^3$ is injective and continuous we can use the Jacobian method. we being by finding the inverse $g^{-1}(y) = \sqrt[3]{y}$. Then $dx/dy = \frac{1}{3y^{2/3}}$. Thus we see that

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}| = \frac{3y^{2/3}}{3y^{2/3}} = 1$$

for $0 < y < 1$ and zero elsewhere. Thus we Y is uniform.

For (c): We find the

$$E(Y) = \int_0^1 x \cdot 1 dx = 0.5 = E(X^3)$$

□

1.9.6 Section 1.7 Answers

1.9.1

Find the mean and variance, if they exist, of each of the following distributions.

(a) $p(x) = \dfrac{3!}{x!(3-x)!}\left(\dfrac{1}{2}\right)^3$, $x = 0, 1, 2, 3$, zero elsewhere.

(b) $f(x) = 6x(1-x)$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = \dfrac{2}{x^3}$, $1 < x < \infty$, zero elsewhere.

Solution Recall that the mean $\mu = E(X)$ and $\text{Var}(X) = E(X^2) - E(X)$. So for each part we will find $E(X)$ then $E(X^2)$.

For (a):

$$E(X) = \sum_{x=0}^{3} xp(x) = 1 \cdot \frac{3!}{1!(3-1)!}\left(\frac{1}{2}\right)^3 + 2 \cdot \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^3 + 3 \cdot \frac{3!}{3!(3-3)!}\left(\frac{1}{2}\right)^3 = \frac{3}{2}.$$

$$E(X^2) = \sum_{x=0}^{3} x^2 p(x) = 1^2 \cdot \frac{3!}{1!(3-1)!}\left(\frac{1}{2}\right)^3 + 2^2 \cdot \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^3 + 3^2 \cdot \frac{3!}{3!(3-3)!}\left(\frac{1}{2}\right)^3 = 3$$

Thus

$$\sigma^2 = E(X^2) - \mu^2 = 3 - \frac{9}{4} = \frac{3}{4}$$

For (b):

$$E(X) = \int_0^1 6x^2(1-x)dx = 0.5$$

$$E(X^2) = \int_0^1 6x^3(1-x)dx = 0.3$$

Thus

$$\sigma^2 = E(X^2) - \mu^2 = 0.3 - 0.5^2 = \frac{1}{20}.$$

For (c) :

$$E(X) = \int_1^\infty x \cdot \frac{2}{x^3}dx = \lim_{b\to\infty}\int_1^b \frac{2}{x^2}dx = 2(\lim_{b\to\infty}\frac{-1}{b} + 1) = 2$$

$$E(X^2) = \int_1^\infty x^2 \cdot \frac{2}{x^3}dx = \lim_{b\to\infty}\int_1^b \frac{2}{x}dx = DNE.$$

Thus the variance does not exist. □

1.9.2

Let $p(x) = \left(\frac{1}{2}\right)^x$ for $x = 1, 2, 3, \ldots$, and zero elsewhere, be the pmf of a random variable X. Find the mfg, the mean and the variance.

Solution Once we find the mgf we can find the mean and variance by finding the first and second moment. We see that $M(t) = E(e^{tX})$ for some $t \in (-h, h)$ in some open interval around zero. We then see that

$$E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx}\left(\frac{1}{2}\right)^x = \sum_{x=1}^{\infty} = \left(e^t \cdot \frac{1}{2}\right)^x$$

This converges only when $t \leq \ln 2$. Thus this is a valid mfg. We then see that this converges to the geometric series:

$$M(t) = E(e^{tX}) = \frac{1}{1 - \frac{e^t}{2}} = \frac{2}{2 - e^t}.$$

We then find the first moment by finding $M'(0)$:

$$M'(0) = E(X) = \mu = 2$$

and similarly

$$M''(0) = E(X^2) = E(X^2) = 6$$

the variance is then

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

□

1.9.7

Show that the moment generating function of the random variable X having the pdf

$$f(x) = \frac{1}{3}, \quad -1 < x < 2, \text{ zero elsewhere,}$$

is

$$M(t) = \begin{cases} \dfrac{e^{2t} - e^{-t}}{3t} & t \neq 0 \\ 1 & t = 0. \end{cases}$$

Solution The mgf of a random variable is defined as $M(t) = E(e^{tX})$ for all $t \in (-a, b)$ where $(-a, b)$ is an open interval around zero. We see that using Theorem 1.6.8 we get that

$$E(e^{tX}) = \int_{-1}^{2} e^{tx} f(x)dx = \frac{1}{3}\int_{-1}^{2} e^{tx}dx = \frac{1}{3}\left[\frac{e^{tx}}{t}\right]_{-1}^{2} = \frac{e^{2t}}{3t} - \frac{e^{-t}}{3t} = \frac{e^{2t} - e^{-t}}{3t}.$$

for $t \neq 0$. We also see that when $t = 0$ that

$$M(0) = \int_{-1}^{2} e^{0} f(x)dx = \int_{-1}^{2} f(x)dx = 1.$$

□

1.9.8

Let X be a random variable such that $E[(X-b)^2]$ exists for all real b. Show that $E[(X-b)^2]$ is a minimum when $b = E(X)$.

Solution To show that $E[(X-b)^2]$ is at a minimum when $b = E(X) = \mu$, we can define a function $h(b) = E[(X-b)^2]$ for all real b, and show that $h'(b) = 0$ if and only if $b = E(X)$. First we see that $E[(X-b)^2] = E(X^2) - bE(X) + b^2$ by linearity of the expectation. Thus we have

$$h(b) = b^2 - 2bE(X) + E(X^2).$$

Then we see that

$$h'(b) = 2b - 2E(X)$$

Solving for when $h'(b) = 0$ we get that

$$2b = 2E(X) \text{ or } b = E(X)$$

as needed. □

1.9.7 Section 1.8 Answers

1.10.2

Let X be a random variable such that $P(X \leq 0) = 0$ and let $\mu = E(X)$ exist. Show that $P(X \geq 2\mu) \leq \frac{1}{2}$

Solution From Markov's ineqaulity, we know that $P(u(X) \geq c) \leq \frac{E(X)}{c}$. Let $u(X) = X$ and $c = 2\mu$. Then we see that

$$P(X \geq 2\mu) \leq \frac{\mu}{2\mu} = \frac{1}{2}.$$

□

1.10.3

If X is a random variable such that $E(X) = 3$ and $E(X^2) = 13$, use Chebyshev's inequality to determine a lower bound for the probability $P(-2 < X < 8)$.

Solution We first find the variance $\sigma^2 = E(X^2) - E(X) = 13 - 9 = 4$. Thus the standard deviation is $\sigma = \sqrt{4} = 2$. We then see that $P(-2 < X < 8) = P(-5 < X - 3 < 5) = P(|X - 3| < 5)$. Then using Chebyshev's inequality we get that

$$P(|X - 3| > 5) = P(|X - 3| > \frac{5}{2}\sigma) \leq \frac{1}{\left(\frac{5}{2}\right)^2} = 0.16.$$

We have that $P(|X - 3| < 5) = 1 - P(|X - 3| > 5)$ which implies

$$P(|X - 3| < 5) \geq 0.84 \geq 0.16 \geq P(|X - 3| > 5).$$

□

1.10.5

Let X be a random variable with mgf $M(t)$, $-h < t < h$. Prove that

$$P(X \geq a) \leq e^{-at}M(t), \quad 0 < t < h,$$

and that

$$P(X \leq a) \leq e^{-at}M(t), \quad -h < t < 0.$$

Solution We begin recall that $M(t) = E(e^{tX})$. First, assume $0 < t < h$. Define $u(X) = e^{tX}$. Thus using Markov's inequality we see that

$$P(e^{tX} \geq e^{ta}) \leq e^{-at}M(t)$$

Since e^{tx} is an increasing function when $t > 0$ we have that then (through inverses)

$$P(X \geq a) = P(e^{tX} \geq e^{at}) \leq e^{-at}M(t).$$

Now assume $-h < t < 0$. The idea is similar above. We let $u(X) = e^{tX}$. Notice that e^{tx} is a decreasing function. We then get that

$$P(X \leq a) = P(e^{tX} \geq a) \leq e^{-at}M(t)$$

as needed □

2 Multivariate Distribution

2.1 Distributions of Two Random Variables

This section extends what we did in the previous section to account for multiple random variables. We begin with the following example. A coin is flipped three times. This time the result comes in an ordered pair: number of heads on the first two tosses and the number of heads on all three tosses. Let our sample space be $S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$. Let X_1 be the random variable denoting the number of heads on the first two tosses and X_2 denote the number of heads on all three tosses. We can represent this by a pair of random variables (X_1, X_2). For example $(X_1(HTH), X_2(HTH))$ gives us the ordered pair $(1, 2)$. Thus we see that our pair of random variables take us from the sample space to space of ordered number pairs.

$$\mathcal{D} = \{(0,0), (0,1), (1,1), (1,2), (2,2), (2,3)\}.$$

So we see that X_1 and X_2 are two random variables defined on the sample space S and the space is a subset of two-dimensional euclidean space $\mathbb{R}$. Thus this is a vector valued function.

Definition 2.1.1

Given a random experiment with sample space S, consider two random variables X_1 and X_2, which assign to each element c of S one and only one ordered pair of numbers $X_1(c) = x_1$, $X_2(c) = x_2$. Then we say that X_1, X_2 is a random vector. The space of (X_2, X_2) is the set $\mathcal{D} = \{(x_1, x_2) : x_1 = X_1(c), x_2 = X_2(c)\}$.

As with single random variables, we are interested in two types of random vectors, discrete and continuous. We begin with the discrete. A joint or bi-variate probability distribution is a probability distribution on two random variables. That is, it gives the probability on the simultaneous outcome of the random variables.

Definition 2.1.2

A random vector (X_1, X_2) is a discrete random vector if its space $\mathcal{D}$ is countable. The joint probability mass function (pmf) for (X_1, X_2) is

$$p_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

for all $(x_1, x_2) \in \mathcal{D}$.

Like with random variables the pmf satisfies the Probability Axioms

1. $0 \leq p_{X_1,X_2}(x_1, x_2) \leq 1$ for all $(x_!, x_2) \in \mathcal{D}$.
2. $\sum_{x_1} \sum_{x_2} p_{X_1,X_2}(x_1, x_2) = 1$ for all non zero (x_1, x_2).

Then for any event $B \subseteq \mathcal{D}$ we have that

$$P((X_1, X_2) \in B) = \sum_B \sum_B p_{X_1,X_2}(x_1, x_2)$$

Example 2.1.3

Find the pmf of the random vector defined at the start of the section.

Solution We are given a random vector (X_1, X_2) where X_1 is the random variable representing the total number of heads after the first two tosses and X_2 is the random variable of the total number of heads after the three tosses. We can can assign probabilities easily and create a table. For example $p_{X_1,X_2}(1,2)$ refers to the probability that we see only one head after the first two tosses then another head for the third toss, giving us the sequence HTH or THH. Since we have 8 elements in our sample space S and each sequence is equilikely then we have that $p_{X_1,X_2}(1,2) = 2/8$. Following this logic we get the following table:

	Support of X_2			
Support of X_1	0	1	2	3
0	1/8	1/8	0	0
1	0	2/8	2/8	0
2	0	0	1/8	1/8

□

Just like with random variables, we can define a Cumulative Distribution Function (cdf) for random vectors. The cdf is given by

$$F_{X_1.X_2}(x_1, x_2) = P[\{X_1 \leq x_1\} \cap \{X \leq x_2\}],$$

for all $(x_1, x_2) \in \mathbb{R}^2$. We often denote it as $P[X_1 \leq x_1, X_2 \leq x_2]$.

Definition 2.1.4

For random variables X_1 and X_2, the joint or bivariate cumulative distribution function (cdf) is

$$F_{X_1.X_2}(x_1, x_2) = P[X_1 \leq x_1, X_2 \leq x_2] = P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}]$$

For the discrete case we then can express the cdf as

$$F_{X_1,X_2}(x_1, x_2) = \sum_{z_1 \leq x_1} \sum_{z_2 \leq x_2} p(z_1, z_2).$$

We use an example to illustrate this.

Example 2.1.5

Using the same random vector in Example 2.1.3, find $P[X_1 \geq 2, X_2 \geq 2]$.

Solution Since the random vector is discrete we see that the CDF will be discontinuous. We can find the CDF by expressing it as a sum:

$$F_{X_1,X_2}(2,2) = \sum_{z_1 \geq 2} \sum_{z_2 \geq 2} p_{X_1,X_2}(z_1, z_2) = p_{X_1,X_2}(2,2) + p_{X_1,X_2}(2,3) = 2/8.$$

□

Just like with random variables we have the following theorem for joint cdf:

Theorem 2.1.6

If (X_1, X_2) is a random vector with joint cdf $F(x_1, x_2)$ then the following are true

1. $F(-\infty, -\infty) = F(-\infty, x_2) = F(x_2, -\infty) = 0$
2. $F(\infty, \infty) = 1$
3. If $a \leq b$ and $c \leq d$ then

$$P(a \leq X_1 \leq b, c \leq X_2 \leq d) = F(b, d) - F(b, c^-) - F(a^-, d) + F(a^-, c^-)$$

Proof. For 1: We begin by looking at

$$\lim_{a \to -\infty} \lim_{b \to -\infty} F(a, b).$$

We see that $F(a, b) = P[\{X_1 \leq a\} \cap \{X_2 \leq b\}]$. If we show that the cdf tends to zero when both x_1 and x_2 go to negative infinity, it follows that the single variable limit will equal zero since empty set intersect with any set is empty. To do this we start by choosing any sequence $x_1 > x_2 > x_3 \ldots$ with $x_n \to -\infty$ and set $\{C_n\} = (-\infty, x_n]$ where $x_n < x_{n-1}$. We see clearly that

$$\bigcap_{n=1}^{\infty} C_n = \emptyset.$$

However notice that using the Continuity theorem we get that

$$\lim_{n \to \infty} P(C_n \cap \{X_2 \leq b\}) = P(\emptyset \cap \{X_2 \leq b\}) = P(\emptyset) = F(x_n, b) = 0.$$

Moreover, just like with random variables the joint cdf is monotone so we get that

$$\lim_{a \to -\infty} F(a, b) = 0.$$

We can repeat this for b and get the same result. As you can see the proof of this is identical to the random variable version. This is because if we hold one direction constant/static, our cdf behaves like a random variable.

For 2: This proof is similar to the random variable version as we simply just define to non-decreasing sequence of events. We define $\{D_n\}$ by $D_n = (-\infty, x_n]$ where $x_{n-1} < x_n$ and $x_n \to \infty$. And define a similar sequence $\{S_n\}$. Then we see that

$$\lim_{n \to \infty} D_n = \bigcup_{n=1}^{\infty} D_n = \mathbb{R}.$$

Same with S_n. Using the continuity theorem we get that

$$\lim_{a \to \infty} \lim_{b \to \infty} F(a, b) = \lim_{n \to \infty} F(D_n \cap S_n) = P(\mathbb{R}) = 1.$$

For 3: Define $R = (a, b] \times (c, d]$. □

We now discuss joint contentious random variables. We say that a random vector is continuous if its cdf is continuous.

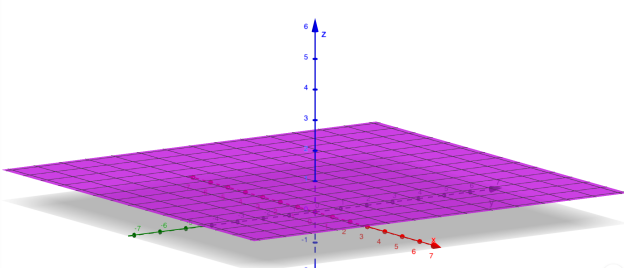


Figure 4: Uniform distribution for a joint continuous random variable

Definition 2.1.7

Let (X_1, X_2) be a joint continuous random variables with the joint cdf $F_{X_1.X_2}(x_1, x_2)$. f there exists a non-negative function $f_{X_1,X_2}(x_1, x_2)$ such that

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{X_1,X_2}(w_1, w_2)dw_1dw_2.$$

for all $(x_1, x_2) \in \mathbb{R}^2$, then (X_1, X_2) is said be joint continuous random variables. The function f_{X_1,X_2} is said to be the joint probability density function (pdf). Then

$$\frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} = f(x_1, x_2).$$

We then have the following properties of the pdf :

1. $f(x_!, x_2) \geq 0$ for all $(x_1, x_2) \in \mathbb{R}^2$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_!, x_2)dx_1dx_2 = 1$

As you can guess, the volume under the joint pdf surface correspond to probabilities. That is for any event $A \subseteq \mathcal{D}$ we have that

$$P[(X_1, X_2) \in A] = \iint_A f(x_1, x_2)dx_1dx_2$$

Example 2.1.8

Suppose a radioactive particle is located in a square with sides of unit length. Let X_1 and X_2 denote the particle's location and assume it is uniformly distributed in square; $f(x_1, x_2) = 1$ for $0 \leq x_1, x_2 \leq 1$ and zero elsewhere. Find $F(0.2, 0.4)$. Find $P(0.1 \leq X_1 \leq 0.3, 0 \leq X_2, 0.5)$.

Solution We see that

$$F(0.2, 0.4) = \int_0^{0.4}\int_0^{0.2} f(x_1, x_2)dx_1dx_2 = \int_0^{0.4}\int_0^{0.2} 1dx_1dx_2.$$

This is equivalent to finding the volume of a 2-dimensional sheet as seen in Figure 4. thus we get that

$$F(0.2, 0.4) = 0.4 \times 0.2 = 0.08.$$

Next we see that

$$P(0.1 \leq X_1 \leq 0.3, 0 \leq X_2, 0.5) = \int_0^{0.5} \int_{0.1}^{0.3} f(x_1, x_2) dx_1 dx_2 = \int_0^{0.5} \int_{0.1}^{0.3} 1 dx_1 dx_2 = 0.2 \times 0.5 = 0.10.$$

□

Example 2.1.9

Gasoline is stored on a FOB in a bulk tank. Let X_1 denote the proportion of the tank available at the beginning of the week after restocking. Let X_2 denote the proportion of the tank that is dispensed over the week. Note that X_1 and X_2 must be between 0 and 1 and $x_2 \leq x_1$. Let the joint pdf $f(x_1, x_2) = 3x_1$ for $0 \leq x_2 \leq x_1 \leq 1$ and zero otherwise. Find $P(0 \leq X_1 \leq 0.5, 0.25 \leq X_2)$.

Solution We see that we need to find the probability that there is between 0 and 0.5 fuel available at the start of the week and then there is more than 0.25 fuel left after the end of the week (note that this value is bounded above by x_1). That is we need to find

$$P(0 \leq X_1 \leq 0.5, 0.25 \leq X_2) = \int_0^{0.5} \int_{0.25}^{x_1} 3x_1 dx_2 dx_1.$$

Computing the inner integral we get that

$$\int_{0.25}^{x_1} 3x_1 dx_2 = 3x_1 \int_{0.25}^{x_1} 1 dx_2 = 3x_1(x_1 - 0.25).$$

Thus the inner integral becomes

$$\int_0^{0.5} 3x_1(x_1 - 0.25) dx_1 = \frac{5}{128}.$$

□

Example 2.1.10

Suppose an electrical component has two batteries. Let X and Y denote the lifetimes in standard units of the respective batteries. Assume that the pdf of (X, Y) is

$$f(x, y) = \begin{cases} 4xye^{-(x^2+y^2)} & x > 0,\ y > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Find

$$P\left(X > \frac{\sqrt{2}}{2}, Y > \frac{\sqrt{2}}{2}\right)$$

Define

$$A = \{(x, y) : |x-(1/2)| < 0.3,\ |y-(1/2)| < 0.3\} \quad \text{and} \quad B = \{(x, y) : |x-2| < 0.3,\ |y-2| < 0.3\}.$$

Find $P[(X_1, X_2) \in A]$ and $P[(X_1, X_2) \in B]$.

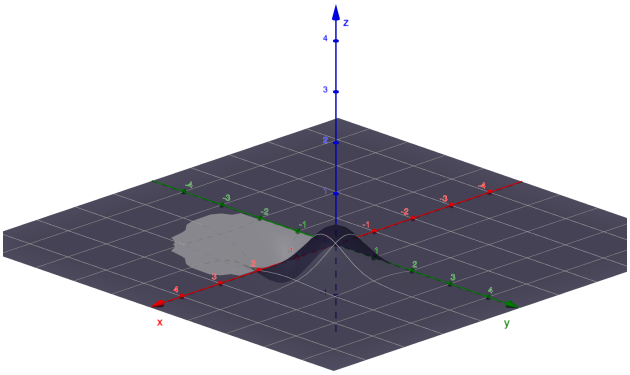


Figure 5: Image of pdf from Example 2.1.0. This can be found here

Solution We see that

$$P\left(X > \frac{\sqrt{2}}{2}, Y > \frac{\sqrt{2}}{2}\right) = \int_{\sqrt{2}/2}^{\infty}\int_{\sqrt{2}/2}^{\infty} 4xye^{-(x^2+y^2)}dxdy = \int_{\sqrt{2}/2}^{\infty}\int_{\sqrt{2}/2}^{\infty} 4xye^{-x^2}e^{-y^2}dxdy.$$

We start with the inner integral:

$$\int_{\sqrt{2}/2}^{\infty} 4xye^{-x^2}e^{-y^2}dx = 4ye^{-y^2}\lim_{b\to\infty}\int_{\sqrt{2}/2}^{b} xe^{-x^2}dx.$$

Letting $z = x^2$ then $dz = 2xdy$. We get that then

$$4ye^{-y^2}\lim_{b\to\infty}\int_{\sqrt{2}/2}^{b} xe^{-x^2}dx = 4ye^{-y^2}\lim_{b\to\infty}\frac{1}{2}\int_{0.5}^{b} e^{-z}dz = 4ye^{-y^2}\lim_{b\to\infty}\left[\frac{-1}{2e^b} + \frac{1}{2\sqrt{e}}\right] = \frac{4ye^{-y^2}}{2\sqrt{e}}.$$

Then the outer integral becomes easy:

$$\int_{\sqrt{2}/2}^{\infty}\frac{4ye^{-y^2}}{2\sqrt{e}}dy = \frac{4}{2\sqrt{e}}\int_{\sqrt{2}/2}^{\infty} ye^{-y^2}dy = \frac{4}{2\sqrt{e}}\lim_{b\to\infty}\int_{\sqrt{2}/2}^{b} ye^{-y^2}dy. \tag{1}$$

Doing the same substitution $w = y^2$ and $dw - 2ydy$ we get that

$$\frac{4}{2\sqrt{e}}\lim_{b\to\infty}\frac{1}{2}\int_{0.5}^{b} e^{-w}dw = \frac{4}{2\sqrt{e}}\lim_{b\to\infty}\left[\frac{-1}{2e^b} + \frac{1}{2\sqrt{e}}\right] = \frac{1}{e} = 0.3679.$$

Next, we simply compute the volume under the surface of events A and B to compute the probability.

$$P[(X,Y) \in A] = \int_{-0.2}^{0.8}\int_{-0.2}^{0.8} 4xye^{-x^2}e^{-y^2}dxdy = 0.1879$$

and

$$P[(X,Y) \in B] = \int_{-1.7}^{2.3}\int_{-1.7}^{2.3} 4xye^{-x^2}e^{-y^2}dxdy = 0.0026.$$

□

2.1.1 Marginal Distributions

Marginal distributions connect the concept of joint distributions to uni-variate distributions. When you have a joint distribution of two or more random variables, the marginal distribution of one of them is what you get when you "ignore" or "sum out" the others. Let (X_1, X_2) be a random vector. Then we can obtain the distributions for X_1 and X_2 in terms of the join distribution as follows:

$$\{X_1 \leq x\} = \{X_1 \leq x\} \cap \{-\infty < X_2 < \infty\} = \{X_1 \leq x, -\infty < X_2 < \infty\}$$

We get that $F_{X_1}(x_1) = P[X_1 \leq x, -\infty < X_2 < \infty]$.

Definition 2.1.11

Let (X_1, X_2) be a random vector with support $\mathcal{D}_{X_1}$ and $\mathcal{D}_{X_2}$ for X_1 and X_2 respectively. If (X_1, X_2) is of the discrete type then the marginal probability mass functions of X_1 and X_2 are respectively

$$p_1(x_1) = \sum_{x_2} p_1(x_1, x_2) \text{ for all } x_1 \in \mathcal{D}_{X_1}$$

and

$$p_2(x_2) = \sum_{x_1} p_1(x_1, x_2) \text{ for all } x_2 \in \mathcal{D}_{X_2}.$$

If (X_1, X_2) is of the continuous type then the marginal probability density functions of X_1 and X_2 are respectively

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \text{ for all } x_1 \in \mathcal{D}_{X_1}$$

and

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 \text{ for all } x_2 \in \mathcal{D}_{X_2}$$

To better understand this please see the following examples and exercises at the end of the section.

Example 2.1.12

Find the marginal distribution for X_1 and X_2 from Example 2.1.3 (tossing coin three times).

Solution We see that to find the marginal distribution we simply sum across the rows for X_1 and sum across the columns for X_2. Because these distributions are recorded in the margins of the table, this is why we refer to them as marginal pmfs.

		Support of X_2				
		0	1	2	3	$p_{X_1}(x_1)$
Support of X_1	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0	$\frac{2}{8}$
	1	0	$\frac{2}{8}$	$\frac{2}{8}$	0	$\frac{4}{8}$
	2	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$
	$p_{X_2}(x_2)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

□

Example 2.1.13

Find the marginal distribution for X_1 and X_2 from Example 2.1.9 (Gasoline).

Solution We were given the pdf $f_{X_1,X_2}(x_1,x_2) = 3x_1$. Thus we see that

$$f_{X_1}(x_1) = \int_0^{x_1} f_{X_1,X_2} dx_2 = \int_0^{x_1} 3x_1 dx_2 = 3x_1^2$$

for $0 < x_1 < 1$ and zero elsewhere. We also get

$$f_{X_2}(x_2) = \int_{x_2}^1 f_{X_1,X_2} dx_1 = \int_{x_2}^1 3x_1 dx_1 = \left[\frac{3x_1^2}{2}\right]_{x_2}^1 = \frac{3}{2} - \frac{3x_2^2}{2}$$

for $0 < x_2 < 1$ and zero elsewhere. □

Example 2.1.14

Find the marginal pdf of X from Example 2.1.10 (Batteries) and find the median life θ of the batteries.

Solution Just like in Example 2.1.10 we use the change of variables $w = y^2$. We get that

$$f_X(x_1) = \int_0^\infty 4xye^{-x^2}e^{-y^2}dy = 2xe^{-x^2}\int_0^\infty e^{-w}dw = 2xe^{-x^2}\lim_{b\to\infty}\left(\frac{-1}{b}+1\right) = 2xe^{-x^2}$$

for $x > 0$. It directly follows from the symmetry of the pdf that the pdf of Y is the same. To find the median lifetime we have to solve $F(\theta) = 0.5$. That is

$$\frac{1}{2} = \int_0^\theta 2xe^{-x^2}dx = 1 - e^{-\theta^2}.$$

Solving this we get that $\theta = \sqrt{-\ln\left(\frac{1}{2}\right)} = 0.83$. So % 50 percent of the batteries have lifetimes exceeding 0.83 units. □

2.1.2 Expectation

Exception with random vectors is similar to how it was with random variables. Suppose (X_1, X_2) are random vectors. Let $Y = g(X_1, X_2)$ for some real valued function $g : \mathbb{R}^2 \to \mathbb{R}$. Then Y is a random variable and we can determine its expectation by finding its distribution. However we can extend Theorem 1.6.8 to random vectors as well. If (X_1, X_2) are discrete joint random variables, then $E(Y)$ exist if

$$\sum_{x_1}\sum_{x_2} |g(x_1,x_2)| p_{X_1,X_2}(x_1,x_2) < \infty$$

and $E(Y) = \sum_{x_1}\sum_{x_2} |g(x_1,x_2)| p_{X_1,X_2}(x_1,x_2)$.

Similarly if (X_1, X_2) are joint continuous random variables then $E(Y)$ exist if

$$\int_{-\infty}^\infty \int_{-\infty}^\infty |g(x_2,x_2)| f_{X_2,X_2}(x_2,x_2) dx_1 dx_2 < \infty$$

and $E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g(x_2,x_2)|f_{X_2,X_2}(x_2,x_2)dx_1dx_2$.

Furthermore, we can show that Expectation is a linear operator.

Theorem 2.1.15

Suppose (X_1, X_2) are joint random vectors. Let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ be random variables whose expectations exist. Then for all real numbers k_1 and k_2,

$$E(k_1Y_1 + k_2Y_2) = k_1E(Y_1) + k_2E(Y_2).$$

Proof. I will prove this for the continuous case as the discrete case is similar. We need to show that $E(k_1Y_1 + k_2Y_2)$ exist. To do this we use the linearity of the integral and the triangle inequality.

$$\begin{aligned}
E(k_1Y_1 + k_2Y_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|k_1g_1(x_1,x_2) + k_2g_2(x_1,x_2)|f_{X_1,X_2}(x_1,x_2)dx_1x_2 \\
&\leq |k_1|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g_1(x_1,x_2)|f_{X_1,X_2}(x_1,x_2)dx_1x_2 + |k_2|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g_2(x_1,x_2)|f_{X_1,X_2}(x_1,x_2)dx_1x_2 \\
&< \infty.
\end{aligned}$$

Thus $E(k_1Y_1 + k_2Y_2)$ exist. Then we use the linearity of the integral to show that expectation is linear:

$$\begin{aligned}
E(k_1Y_1 + k_2Y_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[k_1g_1(x_1,x_2) + k_2g_2(x_1,x_2)]f_{X_1,X_2}(x_1,x_2)dx_1x_2 \\
&= k_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g_1(x_1,x_2)f_{X_1,X_2}(x_1,x_2)dx_1x_2 + k_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g_2(x_1,x_2)f_{X_1,X_2}(x_1,x_2)dx_1x_2 \\
&= k_1E(Y_1) + k_2E(Y_2).
\end{aligned}$$

□

If we are given some random vector and some function $g(X_2)$, we can find the expectation of X_2 in two ways. First is just to integrate over both variables using the joint density $f(x_1, x_2)$ even though g is only a function of x_2. This works because you're summing over all possible values of both X_1 and X_2, and isolating how $g(x_2)$ contributes. The second way is to just get the marginal distribution of X_2 by integrating x_1 from the joint and just compute the exepctation in the normal random variable way.

$$E[g(X_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x_2)f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = \int_{-\infty}^{\infty}g(x_2)f_{X_2}(x_2)dx_2.$$

Example 2.1.16

Let X_1 and X_2 be random variables with the joint pdf

$$f(x_1,x_2) = 8x_1x_2, \quad \text{for } 0 < x_1 < x_2 < 1 \text{ and zero elsewhere.}$$

Find $E(X_1X_2^2)$, $E(X_2)$, and $E(7X_1X_2^2 + 5X_2)$.

Solution We can simply compute these expectations.

$$\begin{aligned}
E(X_1X_2^2) &= \int_0^1 \int_0^{x_2} x_1x_2^2 f(x_1,x_2)dx_1dx_2 \\
&= \int_0^1 \int_0^{x_2} 8x_1^2x_2^3 dx_1dx_2 \\
&= \int_0^1 8x_2^3 \left[\frac{x_1^3}{3}\right]_0^{x_2} dx_2 \\
&= \frac{8}{3}\int_0^1 x_2^6 dx_2 \\
&= \frac{8}{21}.
\end{aligned}$$

Next we find $E(X_2)$ by integrating x_1 first

$$\begin{aligned}
E(X_2) &= \int_0^1 \int_0^{x_2} x_2 f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&= \int_0^1 8x_2^2 \int_0^{x_2} x_1 dx_1 dx_2 \\
&= \int_0^1 8x_2^2 \left[\frac{x^2}{2}\right]_0^{x_2} dx_2 \\
&= \int_0^1 4x_2^4 dx_2 \\
&= \frac{4}{5}.
\end{aligned}$$

Thus using the linerity of expectation we get that

$$E(7X_1X_2^2 + 5X_2) = 7E(X_1X_2^2) + 5E(X_2) = 7 \cdot \frac{8}{21} + 5 \cdot \frac{4}{5} = \frac{20}{3}.$$

□

2.1.17

Continuing from Example 2.1.16, let $Y = X_1/X_2$. Find $E(Y)$.

Solution We can find $E(Y)$ by finding the distribution Y. To do that we could find its CDF then differentiate it to find its pdf. The second way is to use our expression we explained above and integrate it normally.

$$E(Y) = \int_0^1 \left[\int_0^{x_2} \left(\frac{x_1}{x_2}\right) 8x_1x_2 dx_1\right] dx_2 = \frac{8}{3}\int_0^1 x_2^3 dx_2 = \frac{2}{3}.$$

□

2.2 Transformations: Bivariate Random Variables

Let (X_1, X_2) be a random vector. Suppose we know the joint distribution of (X_1, X_2) and we seek the distribution of a transformation of (X_1, X_2), say $Y = g(X_1, X_2)$. We may be able to obtain the

cdf of Y. Another way is to use a transformation as we did for univariate random variables in the previous chapter. We begin with the discrete case.

Let $p(x_1, x_2)$ be the joint pmf of two discrete random variables X_1 and X_2 with $\mathcal{D}$ the set of points at which $p(x_1, x_2) > 0$. Let $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define an injective transformation that maps $\mathcal{D}$ onto $\mathcal{T}$.

The joint pmf of the two new random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$p_{Y_1,Y_2}(y_1, y_2) = \begin{cases} p_{X_1,X_2}\left[w_1(y_1, y_2), w_2(y_1, y_2)\right] & (y_1, y_2) \in \mathcal{T} \\ 0 & \text{elsewhere,} \end{cases}$$

where $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$ is the single-valued inverse of $y_1 = u_1(x_1, x_2)$, $y_2 = u_2(x_1, x_2)$. This is a change of variables for discrete random variables. To understand this we can look at it like this:

You have two original discrete random variables, say X_1 = number of heads and X_2 = number of tails and you define new variables as functions of these, like: $Y_1 = X_1 + X_2$ = total number of coin tosses and $Y_2 = X_2$ = number of tails

Then you ask: What is the joint probability distribution of Y_1 and Y_2 instead? Rather than working with $p_{X_1,X_2}(x_1, x_2)$, you're trying to write $p_{Y_1,Y_2}(y_1, y_2)$. To do this, you do a coordinate change, like switching from (x_1, x_2) to (y_1, y_2), as long as the transformation is one-to-one (you can go back and forth) and maps the region where $p(x_1, x_2) > 0$ to a new valid domain. Let: $y_1 = u_1(x_1, x_2), y_2 = u_2(x_1, x_2)$ This defines a transformation to new variables (Y_1, Y_2). Then you can compute their joint PMF by rewriting the old PMF in terms of the new variables:

$$p_{Y_1,Y_2}(y_1, y_2) = p_{X_1,X_2}(x_1, x_2) \quad \text{where } x_1 = w_1(y_1, y_2),\ x_2 = w_2(y_1, y_2)$$

and with our example we would get

$$p_{Y_1,Y_2}(y_1, y_2) = p_{X_1,X_2}(y_1 - y_2, y_2)$$

An example should make this more clear:

Example 2.2.1

In a large metropolitan area during flu season, suppose that two strains of flu, A and B, are occurring. For a given week, let X_1 and X_2 be the respective number of reported cases of strains A and B with the joint pmf

$$p_{X_1,X_2}(x_1, x_2) = \frac{\mu_1^{x_1}}{x_1!} \cdot \frac{\mu_2^{x_2}}{x_2!} \cdot e^{-(\mu_1+\mu_2)}, \quad x_1 = 0, 1, 2, \ldots, \quad x_2 = 0, 1, 2, \ldots$$

Find the pmf of the random variable of interest, $Y_1 = X_1 + X_2$, which is the total number of reported cases of A and B flu during a week.

Solution We let $Y_1 = X_1 + X_2$ and $Y_2 = X + 2$. This gives us an injective transformation $y_1 = x_1 + x_2$ and $y_2 = x + 2$ that maps $\mathcal{D}$ onto $\mathcal{T} = \{(y_1, y_2) : y_2 = 0, 1, 2, \ldots, y_1, \text{ and } y_2 = 0, 1, \ldots\}$. Solving for the inverses we get $x_1 = y_1 - y_2$ and $x_2 = y_2$. Thus we get that

$$p_{Y_1,Y_2}(y_1, y_2) = \frac{\mu_1^{y_1-y_2}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{(y_1 - y_2)!y_2!}$$

for all $(y_1, y_2) \in \mathcal{T}$ and zero elsewhere. Then to find the marginal pmf of Y_1 we simply sum up the pmf of Y_1, Y_2 while keeping each value of the support of Y_1 constant

$$P_{Y_1}(y_1) = \sum_{y_2=0}^{y_1} p_{Y_1,Y_2}(y_1, y_2) = \frac{(\mu_1 + \mu_2)^{y_1} e^{-\mu_1-\mu_2}}{y_1!}$$

Where the last equality is just the binomial expansion. □

For the continuous case we use an example to illustrate the cdf technique. This technique is finding the cdf of the transformation then differentiating it to find the pdf. Consider an experiment in which a person chooses at random a point (X_1, X_2) from the unit square. That is, the joint pdf is

$$f(x_1, x_2) = 1 \quad \text{for } 0 < x_1 < 1,\ 0 < x_2 < 1.$$

Suppose that our interest is not in X_1 or in X_2, but in $Z = X_1 + X_2$. The cdf of Z is denoted by $F_Z(z) = P(X_1 + X_2 \leq z)$. For $z < 0$ we have $F_Z(z) = 0$ since $X_1 + X_2$ is at least greater than zero. When $0 \leq z < 1$ we integrate over the region $0 < x_1 < z$ and $0 < x_2 < z - x_1$. We get

$$F_Z(z) = \int_0^z \int_0^{z-x_1} 1 dx_2 d_1 = \frac{z^2}{2}.$$

When $1 < z < 2$ we are integrating over the full square where $x_1 + x_2 \leq z$, but the boundary hits the upper edges. So, we rewrite it as:

$$F_Z(z) = 1 - P(X_1 + X_2 > z)$$

To compute that, consider the triangular region where $x_1 + x_2 > z$, which lies inside the unit square only when $z < 2$. The region is a triangle bounded below by the line $x_1 + x_2 = z$, and its area is:

$$\int_{z-1}^1 \int_{z-x_1}^1 1 \, dx_2 dx_1$$

Let's compute $F_Z(z)$ directly instead:

$$F_Z(z) = \int_0^1 \int_0^1 \mathbf{1}(x_1 + x_2 \leq z) \, dx_2 dx_1$$

But for $1 < z < 2$, this region is:

$$x_1 \in [z - 1, 1], \quad x_2 \in [0, z - x_1]$$

So:

$$F_Z(z) = \int_{z-1}^1 \int_0^{z-x_1} dx_2 dx_1 = \int_{z-1}^1 (z - x_1) \, dx_1 = \left[z x_1 - \frac{x_1^2}{2} \right]_{z-1}^1$$

Now plug in:

$$= z(1) - \frac{1}{2} - \left[z(z-1) - \frac{(z-1)^2}{2} \right] = z - \frac{1}{2} - \left[z^2 - z - \frac{(z^2 - 2z + 1)}{2} \right] = 1 - \frac{(2-z)^2}{2}$$

So

$$F_Z(z) = 1 - \frac{(2-z)^2}{2} \quad \text{for } 1 < z < 2$$

and $F_Z(x) = 1$ for $z \geq 2$ since sum is less than two. We get that

$$F_Z(z) = \begin{cases} 0 & \text{if } z < 0 \\ \displaystyle\int_0^z \int_0^{z-x_1} dx_2 dx_1 = \frac{z^2}{2} & \text{if } 0 \leq z < 1 \\ \displaystyle 1 - \int_{z-1}^1 \int_{z-x_1}^1 dx_2 dx_1 = 1 - \frac{(2-z)^2}{2} & \text{if } 1 \leq z < 2 \\ 1 & \text{if } z \geq 2 \end{cases}$$

We then differentiate to get the pdf

$$f_Z(z) = \begin{cases} z & \text{if } 0 < z < 1 \\ 2 - z & \text{if } 1 < z < 2 \\ 0 & \text{otherwise} \end{cases}$$

Recall in section 1.5.5 and Theorem 1.5.28 we described a way to find the pdf of a transformation when that transformation was injective. We talked about how that was the change-of-variables technique for single variable integration and how the dx/dy Jacobian term was the one-dimensional case of the Jacobian. So it makes sense that we can extend this to joint random variables and even several random variables.

Suppose (X_1, X_2) are joint continuous random variables with the joint pdf $f_{X_1,X_2}(x_1, x_2)$ and support set $\mathcal{D}$. Consider the transformed random vector $(Y_1, Y_2) = T(X_1, X_2)$ where T is a one-to-one continuous transformation. Let $\mathcal{T} = T(\mathcal{D})$ denote the support of (Y_1, Y_2).

Rewrite the transformation in terms of its components as

$$(Y_1, Y_2) = T(X_1, X_2) = \big(u_1(X_1, X_2), u_2(X_1, X_2)\big)$$

where the functions $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define T. Since the transformation is one-to-one, the inverse transformation T^{-1} exists. We write it as $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$.

Finally, we need the Jacobian of the transformation which is the *determinant* of order 2 given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

Note that J plays the role of dx/dy in the univariate case. We assume that these first-order partial derivatives are continuous and that the Jacobian J is not identically equal to zero in $\mathcal{T}$. Let B be any region in $\mathcal{T}$ and $A = T^{-1}(B)$. Because the transformation is one-to-one, then based on the change-in-variable technique, we have

$$\begin{aligned} P[(X_1, X_2) \in A] &= P[T(X_1, X_2) \in T(A)] = P[(Y_1, Y_2) \in B] \\ &= \iint\limits_A f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 = \iint\limits_{T(A)} f_{X_1,X_2}(T^{-1}(y_1, y_2))\, |J|\, dy_1 dy_2 \\ &= \iint\limits_B f_{X_1,X_2}\big(w_1(y_1, y_2), w_2(y_1, y_2)\big)\, |J|\, dy_1 dy_2 \end{aligned}$$

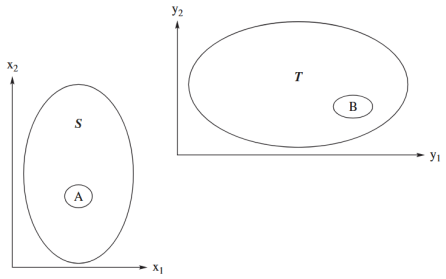


Figure 6: A general sketch of the supports of $(X1, X2)$, (S), and $(Y1, Y2)$, (T).

Since B is arbitrary, the last integrand must be the joint pdf of (Y_1, Y_2). That is, the joint pdf of (Y_1, Y_2) is

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}[w_1(y_1, y_2), w_2(y_1, y_2)] \cdot |J|, & (y_1, y_2) \in \mathcal{T} \\ 0, & \text{elsewhere.} \end{cases}$$

Example 2.2.2

Let X_1 and X_2 have the joint pdf

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{4} \exp\left(-\frac{x_1 + x_2}{2}\right), & 0 < x_1 < \infty,\ 0 < x_2 < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $Y_1 = \frac{1}{2}(X_1 - X_2)$.

Solution What we will do is define a transformation of (X_1, X_2) that is one-to-one and then find the marginal distribution of Y_1 from that. To do this we let $Y_1 = \frac{1}{2}(X_1 - X_2)$ and then for Y_2 we can choose any injective transformation and the choice could be the identity transformation or something similar. It's best to choose something linear or accordingly to make further calculations simpler. We let $Y_2 = \frac{1}{2}(X_1 + X_2)$. Then we get the inverses

$$X_1 = Y_1 + Y_2, X_2 = Y_1 - Y_2.$$

This is a one-to-one transformation of the support $\mathcal{S} = \{(x_1, x_2) : 0 < x_1 < \infty, 0 < x_2 < \infty\}$ to $\mathcal{T} = \{(y_1, y_2) : |y_1| < y_2\}$. We got this by noticing $x_2 = y_1 + y_2 > 0$ and $x_1 = y_1 - y_2 > 0$ implies $|y_1| < y_2$. The Jacobian matrix of the transformation is:

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

and then

$$|\det J| = |(1 \cdot 1) - (-1 \cdot 1)| = |2| = 2.$$

We get get our pdf using our formula

$$f_{Y_1,Y_2} = f_{X_1,X_2}(x_1, x_2)|\det J| = 2 \cdot f_{X_1,X_2}(y_1 + y_2, y_1 - y_2)|\det J|.$$

Therefore, the joint PDF is:

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{4}\exp\left(-\frac{(y_1+y_2)+(y_2-y_1)}{2}\right)\cdot 2, & y_2 > |y_1| \\ 0, & \text{otherwise} \end{cases}$$

Simplify the exponent:

$$(y_1+y_2)+(y_2-y_1) = 2y_2, \quad \Rightarrow \quad \exp\left(-\frac{2y_2}{2}\right) = e^{-y_2}$$

So

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{2}e^{-y_2}, & y_2 > |y_1| \\ 0, & \text{otherwise} \end{cases}$$

To get $f_{Y_1}(y_1)$, integrate out y_2:

$$f_{Y_1}(y_1) = \int_{|y_1|}^{\infty} \frac{1}{2}e^{-y_2}dy_2 = \frac{1}{2}\int_{|y_1|}^{\infty} e^{-y_2}dy_2 = \frac{1}{2}e^{-|y_1|}$$

as required. □

Example 2.2.3

Let X_1 and X_2 have the joint pdf

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 10x_1x_2^2, & 0 < x_1 < x_2 < 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $Y_1 = X_1/X_2$ and $Y_2 = X_2$.

Solution Since the transformations are one-to-one, we can apply the Jacobian method. We find the inverses $x_1 = y_1 \cdot y_2$ and $x_2 = y_2$. Thus we find that the support of our transformation $\mathcal{T}$ is $0 < y_1 \cdot y_2 < y_2$ and $0 < y_2 < 1$. Or simply $0 < y_1 < 1$ and $0 < y_2 < 1$.

Then we find the Jacobian of the transformation

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2.$$

Then we find that the pdf if the join continous random variable transformation is

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(y_1\cdot y_2, y_2)\cdot|y_2| = 10y_1y_2y_2^2|y_2| = 10y_1y_2^4$$

for $(y_1,y_2) \in \mathcal{T}$. We then find the marginal pdf Y_1 :

$$f_{Y_1}(y_1) = \int_0^1 10y_1y_2^4dy_2 = 2y_1$$

for $0 < y_1 < 1$ and the marginal pdf of Y_2 is

$$f_{Y_2}(y_2) = \int_0^1 10y_1y_2^4dy_1 = 5y_2^4$$

for $0 < y_2 < 1$. □

2.3 Conditional Distributions and Expectations

In this section, we discuss conditional distributions for random variables, similar to conditional probability in the previous chapter. The conditional distribution of a random variable given another one is the distribution of one of the random variables when the other has assumed a specific value. Most of this intuitively follows from Chapter 1 and section 2.1. We begin with the discrete case. From the definition of conditional probability we find that for all $x_1 \in S_{X_1}$

$$P(X_1 = x_1 | X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = X_2)} = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}.$$

Definition 2.3.1

Suppose X_1 and X_2 are discrete random variables with a joint probability mass function $p_{X_1,X_2}(x_1, x_2)$ and marginal probability mass functions $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$. Then the conditional pmf of X_1 given $X_2 = x_2$ is defined by

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}.$$

for all $x_1 \in S_{X_1}$ and any fixed x_2 with $p_{X_2}(x_2) > 0$.

This function satisfies the conditions of being a pmf since $p_{X_1|X_2}(x_1|x_2)$ is non-negative and

$$\begin{aligned}\sum_{x_1} p_{X_1|X_2}(x_1|x_2) &= \sum_{x_1} \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)} \\ &= \frac{1}{p_{X_2}(x_2)} \sum_{x_1} p_{X_1,X_2}(x_1, x_2) \\ &= \frac{p_{X_2}(x_2)}{p_{X_2}(x_2)} = 1.\end{aligned}$$

Next, let X_1 and X_2 be continuous random variables with marginal probability density functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. We can use the result from the discrete case to easily find that for any fixed x_2 with $f_{X_2}(x_2) > 0$ we have

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2(x_1,x_2)}}{f_{X_2}(x_2)}.$$

Definition 2.3.2

Suppose X_1 and X_2 are continuous random variables with a joint probability density function $f_{X_1,X_2}(x_1, x_2)$ and marginal probability density functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. Then the conditional pdf of X_1 given $X_2 = x_2$ is defined by

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

for all $x_1 \in S_{X_1}$.

Like above we see that $f_{X_1|X_2}(x_1|x_2)$ is non-negative and that

$$\begin{aligned}\int_{-\infty}^{\infty} f_{X_1|X_2}(x_1|x_2)dx_1 &= \int_{-\infty}^{\infty} \frac{f_{X_1,X_2(x_1,x_2)}}{f_{X_2}(x_2)}dx_1 \\ &= \frac{1}{f_{X_2}(x_2)}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1 \\ &= \frac{f_{X_2}(x_2)}{f_{X_2}(x_2)} = 1.\end{aligned}$$

If random variables are of the continuous type then the probability

$$P(a < X_1 < b|X_2 = x_2) = \int_a^b f_{X_1|X_2}(x_1|x_2)dx_1$$

is called the conditional probability that $a < X_1 < b$ given that $X_2 = x_2$.

Next we can naturally define the conditional distribution function by simply integrating our conditional pdf just like the CDF with random variables.

Definition 2.3.3

Suppose X_1 and X_2 are continuous random variables with a joint probability density function $f_{X_1,X_2}(x_1, x_2)$ and marginal probability density functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. Then the conditional distribution function of X_1 given X_2 is

$$F(x_1|x_2) = \int_{-\infty}^{x_1} \frac{f_{X_1,X_2}(u,x_2)}{f_{X_2}(x_2)}du$$

for all $x_1 \in \mathbb{R}$.

Example 2.3.4

A soft drink machine has a random amount X_2 (in gallons) in supply at the beginning of the day and dispenses a random amount X_1 during the day. It is not resupplied during the day, so $X_1 \leq X_2$, and the joint pdf is

$$f(x_1,x_2) = \begin{cases}\frac{1}{2}, & 0 \leq x_1 \leq x_2 \leq 2 \\ 0, & \text{elsewhere.}\end{cases}$$

What is the probability that less than half a gallon will be sold given that the machine contains 1.5 gallons at the start of the day?

Solution Just like with random variable conditional probability problems in 1.4, we first express what we need to find. We need to find $P(X_1 \leq 1/2|X_2 = 1.5)$. That is we need to find

$$f_{X_1|X_2}(x_1 \leq 1/2|1.5) = \int_0^{\frac{1}{2}} \frac{f_{X_1,X_2}(x_1,1.5)}{f_{X_2}(1.5)}dx_1.$$

So we first need to find the marginal pdf of X_2. We get that

$$\begin{aligned} f_{X_2}(x_2) &= \int_0^{x_2} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\ &= \int_0^{x_2} \frac{1}{2}dx_1dx_2 \\ &= \frac{x_2}{2} \end{aligned}$$

Then $f_{X_2}(1.5) = \frac{1.5}{2} = 0.75$. Then we get that the conditional probability is

$$\int_0^{\frac{1}{2}} \frac{f_{X_1,X_2}(x_1,1.5)}{f_{X_2}(1.5)}dx_1 = \int_0^{\frac{1}{2}} \frac{0.5}{0.75}dx_1 = \frac{1}{3}.$$

□

Next, since conditional probability works over one variable, we can easily define expectation just like how we did in section 1.6.

Definition 2.3.5

Suppose X_1 and X_2 are discrete random variables with joint pmf $p_{X_1,X_2}(x_2,x_2)$. Let $g(X_1)$ be a function of X_1. Then the conditional expectation of $g(X_1)$ given that $X_2 = x_2$, where x_2 is fixed, is given by

$$E(g(X_1)|X_2 = x_2) = \sum_{x_1} g(x_1)p_{X_1|X_2}(x_1|x_2).$$

Suppose X_1 and X_2 are continuous random variables with joint pmf $f_{X_1,X_2}(x_2,x_2)$. Let $g(X_1)$ be a function of X_1. Then the conditional expectation of $g(X_1)$ given that $X_2 = x_2$, where x_2 is fixed, is given by

$$E(g(X_1)|X_2 = x_2) = \int_{-\infty}^{\infty} g(x_1)f_{X_1|X_2}(x_1|x_2)dx_1.$$

Note that $E(g(X_1)|X_2 = x_2)$ is a function of x_2.

Example 2.3.6

For random variables X_1 and X_1 with the joint pdf $f(x_1,x_2) = 1/2$ for $0 \le x_2 \le x_1 \le 2$ and 0 otherwise. Find the conditional expectation of X_1 given that $X_2 = 1.5$.

Solution We need to find

$$E(X_1|X_2 = 1.5) = \int_0^{x_2} x_1)f_{X_1|X_2}(x_1|x_2)dx_1 = \int_0^{x_2} x_1 \cdot \frac{f(x_1,1.5)}{f_{X_2}(1.5)}.$$

So we need to find the marginal pdf of X_2. We find that

$$f_{X_2}(x_2) = \int_0^{x_2} f(x_1,x_2)dx_1 = \frac{x_2}{2}$$

for $0 \le x_2 \le 2$. Then $f_{X_2}(1.5) = 0.75$. Together we get

$$E(X_1|X_2 = 1.5) = \int_0^{1.5} x_1\frac{1/2}{0.75}dx_1 = 0.75.$$

□

We can define variance easily as well. Recall that $E(X_2|x_1)$ is the mean of X_2 given $X_1 = x_1$. Then the conditional variance of X_2 is defined by

$$\text{Var}(X_2|X_1 = x_1) = E\left[(X_2 - E(X_2|x_1))^2 |x_1\right].$$

We then see that through expansion and linearity

$$\text{Var}(X_2|X_1 = x_1) = E(X_2^2|x_1) - [E(X_2|x_1)]^2.$$

Example 2.3.7

Let X_1 and X_2 have the joint pdf $f(x_1, x_2) = 2$ for $0 < x_1 < x_2 < 1$ and zero elsewhere. Find $E(X_1|x_2)$ and $\text{Var}(X_1|x_2)$.

Solution We see that

$$E(X_1|x_2) = \int_0^{x_2} x_1 f_{X_1|X_2}(x_1|x_2)dx_1.$$

We find the marginal pdf of X_2

$$f_{X_2} = \int_0^{x_2} 2dx_1 = 2x_2$$

for $0 < x_2 < 1$. Thus we get that

$$E(X_1|x_2) = \int_0^{x_2} x_1 \cdot \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}dx_1 = \int_0^{x_2} x_1 \cdot \frac{1}{x_2}dx_1 = \frac{x_2}{2}.$$

Then we can find the condtional variance

$$\text{Var}(X_1|x_2) = \int_0^{x_2} \left(x_1 - \frac{x_2}{2}\right)^2 \cdot \left(\frac{1}{x_2}\right) dx_1 = \frac{x_2^2}{12},$$

for $0 < x_2 < 1$. □

Example 2.3.8

Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = \begin{cases} 6x_2, & 0 < x_2 < x_1 < 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Find $E(X_2|x_1)$. Since $E(X_2|x_1)$ is a function x_1 it is a random variable so let $Y = E(X_2|X_1)$. Find the cdf, distribution, means and variance of Y.

Solution To find $E(X_2|x_1)$, we first find the marginal pdf of X_1. We see that

$$f_{X_1}(x_1) = \int_0^{x_1} 6x_2 dx_2 = 3x_1^2$$

for $0 < x_1 < 1$ and zero elsewhere. We see that then

$$E(X_2|x_1) = \int_0^{x_1} x_2 \cdot \frac{f_{X_1,X_2}(x_1.x_2)}{f_{X_1}(x_1)}dx_2 = \int_0^{x_1} x_2 \cdot \frac{6x_2}{3x_1^2}dx_2 = \frac{2}{3}x_1$$

for $0 < x_1 < 1$. Next, let $Y = E(X_2|X_1) = 2X_1/3$. The cdf of Y is (cdf technique)

$$G(y) = P(Y \leq y) = P\left(X_1 \leq \frac{3y}{2}\right)$$

for $0 < y < 2/3$. Then since we have the marginal pdf of X_1 we can integrate and substitute $3y/2$ to find the cdf of Y.

$$G(y) = \int_0^{3y/2} 3x_1^2 dx_1 = \frac{27y^3}{8}$$

and zero elsewhere. Then the pdf of Y is

$$g(y) = \frac{d}{dy}\left[\frac{27y^3}{8}\right] = \frac{81y^2}{8}.$$

Next the mean is

$$E(Y) = \int_0^{2/3} y \cdot \frac{81y^2}{8} dy = \frac{1}{2}.$$

The variance is

$$\text{Var}(Y) = \int_0^{2/3} \left(y - \frac{1}{2}\right)^2 \cdot \frac{81y^2}{8} dy = \frac{1}{60}.$$

Since the marginal pdf of X_2 is

$$f_2(x_2) = \int_{x_2}^1 6x_2\, dx_1 = 6x_2(1 - x_2), \quad 0 < x_2 < 1,$$

zero elsewhere, it is easy to show that $E(X_2) = \frac{1}{2}$ and $\text{Var}(X_2) = \frac{1}{20}$. That is, here

$$E(Y) = E[E(X_2 \mid X_1)] = E(X_2)$$

and

$$\text{Var}(Y) = \text{Var}[E(X_2 \mid X_1)] \leq \text{Var}(X_2).$$

□

The last two observations are true in general which leads us to the following theorem.

Theorem 2.3.9

Let (X_1, X_2) be a random vector such that $\text{Var}(X_2)$ is finite. Then

(a) $E[E(X_2|X_1)] = E(X_2)$.

(b) $\text{Var}(E(X_2|X_1)) \leq \text{Var}(X_2)$.

Proof. We will prove this in the continuous case as the discrete case is analogous. We being with part (a). Since $E(X_2|X_1)$ is a function of x_1 we have that

$$E[E(X_2|X_1)] = \int_{-\infty}^{\infty} E(X_2|x_1) f_{X_1}(x_1) dx_1.$$

We then get that

$$\begin{aligned} E[E(X_2|X_1)] &= \int_{-\infty}^{\infty} E(X_2|x_1) f_{X_1}(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_2 \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} dx_2 \right] f_{X_1}(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f_{X_1,X_2}(x_1,x_2) dx_2 dx_1 \\ &= E(X_2). \end{aligned}$$

For (b), we will show that

$$\mathrm{Var}(X_2) = E[\mathrm{Var}(X_2 \mid X_1)] + \mathrm{Var}(E[X_2 \mid X_1]),$$

from which the inequality follows immediately. We begin by observing that

$$\begin{aligned} \mathrm{Var}(X_2) &= E[(X_2 - \mu_2)^2] \\ &= E\left\{[X_2 - E(X_2 \mid X_1) + E(X_2 \mid X_1) - \mu_2]^2\right\} \\ &= E\left\{[X_2 - E(X_2 \mid X_1)]^2\right\} + E\left\{[E(X_2 \mid X_1) - \mu_2]^2\right\} \\ &\quad + 2E\left\{[X_2 - E(X_2 \mid X_1)][E(X_2 \mid X_1) - \mu_2]\right\}. \end{aligned}$$

We show that the last term on the right-hand side is zero. It is equal to

$$2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_2 - E(X_2 \mid x_1)][E(X_2 \mid x_1) - \mu_2] f(x_1, x_2)\, dx_2\, dx_1,$$

$$= 2 \int_{-\infty}^{\infty} [E(X_2 \mid x_1) - \mu_2] \left\{ \int_{-\infty}^{\infty} [x_2 - E(X_2 \mid x_1)] \frac{f(x_1,x_2)}{f_1(x_1)} dx_2 \right\} f_1(x_1)\, dx_1.$$

But $E(X_2 \mid x_1)$ is the conditional mean of X_2 given $X_1 = x_1$. Since the inner integral is

$$E(X_2 \mid x_1) - E(X_2 \mid x_1) = 0,$$

the whole expression is zero. Therefore,

$$\mathrm{Var}(X_2) = E\left\{[X_2 - E(X_2 \mid X_1)]^2\right\} + E\left\{[E(X_2 \mid X_1) - \mu_2]^2\right\}.$$

The first term is nonnegative because it is the expected value of the nonnegative function $[X_2 - E(X_2 \mid X_1)]^2$. Since $E[E(X_2 \mid X_1)] = \mu_2$, the second term is $\mathrm{Var}(E(X_2 \mid X_1))$. Hence, we conclude:

$$\mathrm{Var}(X_2) \geq \mathrm{Var}(E(X_2 \mid X_1)).$$

□

Example 2.3.10

Let X_1 have the pdf $f_1(x_1) = \frac{1}{2}e^{-x_1/2}$ for $x_1 > 0$, and the conditional pdf of X_2 given $X_1 = x_1$ be

$$f(x_2 \mid x_1) = \frac{1}{x_1}, \quad \text{for } 0 \leq x_2 \leq x_1.$$

Find $E(X_2)$ and $\mathrm{Var}(X_2)$, the (unconditional) mean and variance of X_2.

Solution We begin by finding $E(X_2|x_1)$. We get that

$$E(X_2|x_1) = \int_0^{x_1} x_2 \frac{1}{x_1} dx_2 = \frac{x_1}{2}.$$

Thus then $E(X_2|X_1) = \frac{X_1}{2}$. Next We can find $\text{Var}(X_2|X_1)$,

$$\text{Var}(X_2|X_1) = E(X_2^2|X_1) - E(X_2|X_1)^2 = \int_0^{x_2} x_2^2 \cdot \frac{1}{x_1} dx_2 - \left(\frac{x_1}{2}\right)^2 = \frac{x_1^2}{12}.$$

Then using Theorem 2.3.9 we get

$$E(X_2) = E(E(X_2|X_1)) = E\left(\frac{X_1}{2}\right) = \frac{1}{2}E(X_1) = \frac{1}{2}\int_0^{\infty} \frac{1}{2} x_1 e^{-x_1/2} dx_1 = 1.$$

We also get that

$$\text{Var}(X_2) = E(\text{Var}(X_2|X_1)) + \text{Var}(E(X_2|X_1)) = E\left(\frac{X_1^2}{12}\right) + \text{Var}\left(\frac{X_1}{2}\right) = \frac{5}{3}.$$

□

2.4 Independent Random Variables

We define independence of random variables similarly how we defined independence in Chapter 1. Let X_1 and X_2 denote the random variables of the continuous type that have the joint pdf $f(x_1, x_2)$ and marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. Using the definition of the conditional pdf $f_{2|1}(x_2 \mid x_1)$, we can write the joint pdf $f(x_1, x_2)$ as

$$f(x_1, x_2) = f_{2|1}(x_2 \mid x_1) f_1(x_1).$$

Suppose that we have an instance where $f_{2|1}(x_2 \mid x_1)$ does not depend upon x_1 (not a function of x_1). Then the marginal pdf of X_2 is, for random variables of the continuous type,

$$\begin{aligned} f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 &= \int_{-\infty}^{\infty} f_{2|1}(x_2 \mid x_1) f_1(x_1)\, dx_1 \\ &= f_{2|1}(x_2 \mid x_1) \int_{-\infty}^{\infty} f_1(x_1)\, dx_1 \\ &= f_{2|1}(x_2 \mid x_1). \end{aligned}$$

Accordingly,

$$f_2(x_2) = f_{2|1}(x_2 \mid x_1) \quad \text{and} \quad f(x_1, x_2) = f_1(x_1) f_2(x_2),$$

when $f_{2|1}(x_2 \mid x_1)$ does not depend on x_1. That is, if the conditional distribution of X_2, given $X_1 = x_1$, is independent of any assumption about x_1, then $f(x_1, x_2) = f_1(x_1) f_2(x_2)$. We generalize this in the following definition.

Definition 2.4.1

- Let (X_1, X_2) be a joint discrete random variable with the joint pmf $p(x_1, x_2)$ and marginal pmfs $p_1(x_1)$ and $p_2(x_2)$, respectively. Then X_1 and X_2 are independent if and only if
$$p(x_1, x_2) = p_1(x_1)p_2(x_2)$$
for all pairs of (x_1, x_2) in their support.
- Let (X_1, X_2) be a joint continuous random variable with the joint pdf $f(x_1, x_2)$ and the marginal pdfs $f_1(x_1)$ and $f_2(x_2)$, respectively. Then X_1 and X_2 are independent if and only if
$$f(x_1, x_2) = f_1(x_1)f_2(x_2)$$
for all pairs of real numbers (x_1, x_2).

Random variables that are not independent are said to be dependent.

Example 2.4.2

Let
$$f(x_1, x_2) = 6x_1x_2^2 \quad \text{for } 0 < x_1 \leq 1,\ 0 < x_2 \leq 1,$$
and $f(x_1, x_2) = 0$ otherwise. Show that X_1 and X_2 are independent.

Solution We need to show that $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. To do we have to find the marginal pdf's of X_1 and X_2. We see that

$$f_{X_1}(x_1) = \int_0^1 6x_1x_2^2 dx_2 = 6x_1 \int_0^1 x_2^2 dx_2 = 2x_1,$$

$$f_{X_2}(x_2) = \int_0^1 6x_1x_2^2 dx_1 = 6x_2^2 \int_0^1 x_1 dx_1 = 3x_2^2.$$

Then since for every (x_1, x_2) we have that

$$f(x_1, x_2) = 6x_1x_2^2 = (2x_1)(3x_2^2) = f_1(x_1)f_2(x_2),$$

X_1 and X_2 are independent. □

Example 2.4.3

Suppose an urn contains 10 blue, 8 red, and 7 yellow balls that are the same except for color. Suppose 4 balls are drawn without replacement. Let X and Y be the number of red and blue balls drawn, respectively. Determine if X and Y are independent random variables.

Solution We need to see weather $p(x, y) = p_X(x)p_Y(y)$ for all $(x, y) \in S_{X,Y}$. We begin with finding the pmf. We need to find the probability that x red balls and y blue balls are drawn. There are $\binom{25}{4}$ total combinations we can draw 4 balls from the urn. Then we easily see that

$$p(x, y) = \frac{\binom{8}{x}\binom{10}{y}\binom{7}{4-x-y}}{\binom{25}{4}}$$

for $0 \leq x, y \leq 4$ and $x + y \leq 4$. Next we find the marginal pmf of X and Y:

$$p_X = \frac{\binom{8}{x}\binom{17}{4-x}}{\binom{25}{4}},$$

$$p_Y = \frac{\binom{10}{y}\binom{15}{4-y}}{\binom{25}{4}},$$

To show dependence we only need to show one point in the support where the equality does not hold. We find that

$$p(1,1) = 10 \cdot 8 \cdot \binom{7}{2} / \binom{25}{4} = 0.1328,$$

$$p_X(1) = 10 \cdot \binom{15}{3} / \binom{25}{4} = 0.3597,$$

$$p_X(1) = 8 \cdot \binom{17}{3} / \binom{25}{4} = 0.4300,$$

however $0.1328 \neq 0.3597 \cdot 0.4300$ so X and Y are dependent. Because we are drawing without replacement, once you "use up" one red ball or one blue ball in your sample, there are fewer balls of that color left in the urn for the remaining draws. In particular, if you happen to draw a lot of red balls, then there are fewer total balls left, and thus a smaller chance of also drawing many blues. In other words, seeing "more reds in my 4-draw sample" makes "drawing blues" somewhat less likely in the same sample of size 4 □

The following theorem allows us to check weather random variables like in Example 2.4.2 are dependent without finding the marginal pdf's.

Theorem 2.4.4

Let X_1 and X_2 have joint density function $f(x_1, x_2)$ which is positive if and only if

$$a \leq x_1 \leq b \quad \text{and} \quad c \leq x_2 \leq d,$$

for constants a, b, c, d, and $f(x_1, x_2) = 0$ otherwise. Then X_1 and X_2 are independent random variables if and only if

$$f(x_1, x_2) = g(x_1)h(x_2)$$

where $g(x_1)$ is a nonnegative function of x_1 alone and $h(x_2)$ is a nonnegative function of x_2 alone.

Proof. We prove this in the continuous case however it is also true for the discrete case which is analogous. If X_1 and X_2 are independent then the condition is satisfied trivially. Suppose that $f(x_1, x_2) = g(x_1)h(x_2)$ for some nonnegative functions g and h of x_1 and x_2 respectively. We see that

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_2 = c_1 g(x_1),$$

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1 = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_2 = c_2 h(x_2),$$

where c_1 and c_2 are constants where $c_1 \cdot c_2 = 1$ since

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)dx_1dx_2 = c_1c_2.$$

Thus we get that

$$f(x_1, x_2) = g(x_1)h(x_2) = c_1c_2g(x_1)h(x_2) = f_{X_1}(x_1)f_{X_2}(x_2).$$

□

Example 2.4.5

(a) Let X_1 and X_2 have joint density $f(x_1, x_2) = 2x_1$ for $0 < x_1 \leq 1$ and $0 < x_2 \leq 1$, and 0 otherwise. Are X_1 and X_2 independent?

(b) Let X_1 and X_2 have joint density $f(x_1, x_2) = 2$ for $0 \leq x_2 \leq x_1 \leq 1$, and 0 otherwise. Are X_1 and X_2 independent?

Solution For (a): Since $f(x_1, x_2)$ is a positive function and that can be factored into positive function and the support is of the form $[a, b] \times [c, d]$, by Theorem 2.3.4 they are independent.

For (b): The support is not a rectangular region of the form $[a, b] \times [c, d]$ but rather a triangular region so by the Theorem we can conclude the are not independent. We could also find the marginal pdf's to see that they do not factor. □

Instead of working with pdfs (or pmfs) we could have presented independence in terms of cumulative distribution functions. The following theorem shows the equivalence.

Theorem 2.4.6

Let X_1 and X_2 be two random variables with the joint cdf $F(x_1, x_2)$, and further $F_1(x_1)$ and $F_2(x_2)$ be their marginal cdfs, respectively. Then X_1 and X_2 are independent if and only if

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)$$

for all $(x_1, x_2) \in \mathbb{R}^2$.

Proof. I will prove the continuous case as the discrete case is similar. Suppose $F(x_1, x_2) = F_1(x_1)F_2(x_2)$. Then we get that

$$\frac{\partial^2}{\partial x_1 \partial x_2} F(x_1, x_2) = f_1(x_1)f_2(x_2)$$

which is the desired result. Next suppose X_1 and X_2 are independent. Then

$$\begin{aligned} F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(w_1, w_2)dw_1dw_2 &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_1(w_1)f_2(w_1)dw_1dw_2 \\ &= F_1(x_1)F_2(x_2). \end{aligned}$$

□

Naturally we then can prove a theorem that helps us determine probabilities of events that are independent.

Theorem 2.4.7

The random variables X_1 and X_2 are independent random variables if and only if the following condition holds,

$$P(a < X_1 \leq b,\ c < X_2 \leq d) = P(a < X_1 \leq b) \cdot P(c < X_2 \leq d)$$

for every $a < b$ and $c < d$, where a, b, c, and d are constants.

Solution If X_1 and X_2 are independent, then an application of the last theorem

$$\begin{aligned} P(a < X_1 \leq b,\ c < X_2 \leq d) &= F(b,d) - F(a,d) - F(b,c) + F(a,c) \\ &= F_1(b)F_2(d) - F_1(a)F_2(d) - F_1(b)F_2(c) + F_1(a)F_2(c) \\ &= [F_1(b) - F_1(a)][F_2(d) - F_2(c)], \end{aligned}$$

which is the right side of expression (2.4.2). Conversely, condition implies that the joint cdf of (X_1, X_2) factors into a product of the marginal cdfs, which in turn by Theorem 2.4.6 implies that X_1 and X_2 are independent. □

Moreover, independence also grants us with easier computations of expectations as well.

Theorem 2.4.8

Suppose X_1 and X_2 are independent and $E(u(X_1))$ and $E(w(X_2))$ exist. Then

$$E[u(X_1)w(X_2)] = E[u(X_1)]E[w(X_2)].$$

Proof. I will prove this for the continuous case as the discrete case is similar. Since X_1 and X_2 are independent then the joint pdf is $f_1(x_1)f_2(x_2)$. Then using the defintion of expectation we get

$$\begin{aligned} E[u(X_1)w(X_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1)w(x_2)f_{X_1,X_2}(x_1,x_2)dx_1dx_2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1)w(x_2)f_1(x_1)f_2(x_2) \\ &= \int_{-\infty}^{\infty} u(x_1)f_1(x_1)\int_{-\infty}^{\infty} w(x_2)f_2(x_2) \\ &= E[u(X_1)]E[w(X_2)], \end{aligned}$$

which is the desired result. □

Examples and applications of these theorems are in the end of chapter practice question answers.

2.5 The Correlation Coefficient

When looking at two random variables, a common question is whether or not each variable is associated with the other one. In turns out there are many ways we can measure this dependence between random variables. For example if one tends to increase, does the other variable do so as well? If so we say that they are correlated. In this section we introduce a parameter p of a join distribution (X, Y) that measures linear dependence between X and Y.

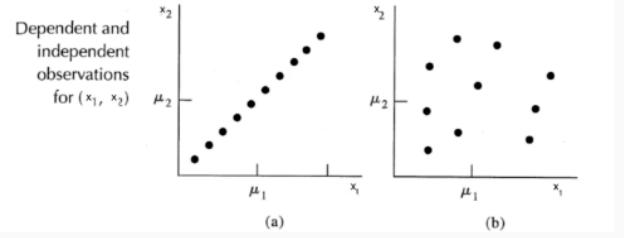


Figure 7: The figure above shows two extremes: perfect linear dependence and independence.

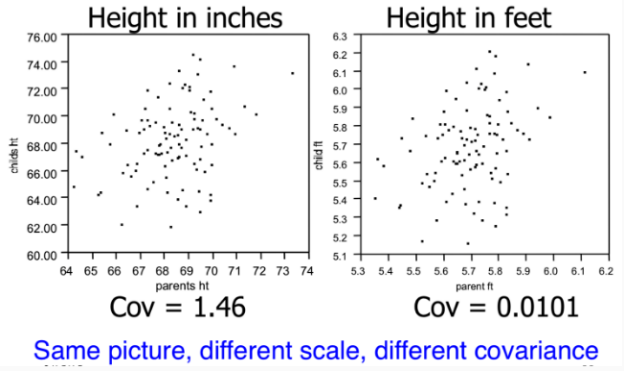


Figure 8: Two graphs representing the two same random variables with different scales and different covariance.

Definition 2.5.1

If X_1 and X_2 are random variables with means μ_1 and μ_2 respectively, then the covariance of X_1 and X_2 is defined as

$$Cov(X_1, X_2) = E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] = E(X_1 X_2) - \mu_1 \mu_2.$$

The covariance gives us the strength (magnitude) and direction (sign) of a linear relationship between two random variables. For example a strong-positive linear relationship like (a) in Figure 7 represents a large positive covariance. A weak positive linear relationship like (b) in Figure 7 represents a small positive covariance. However "small" and "large" depend on the unit of measurement of X_1 and X_2. If our units of measurement vary, such as $0 \leq X_1 \leq 1$ and $0 \leq X_2 \leq 1000$, our linear relationship on that scale can be skewed such as in Figure 8. So we would like to normalize the unit of measure before finding the covariance or instead normalize the un-normalized covariance.

Definition 2.5.2

If each of σ_1 and σ_2 is positive, then the correlation coefficient between X and Y is defined by

$$-1 \leq p = \frac{Cov(X, Y)}{\sigma_1 \sigma_2} \leq 1.$$

We can simplify the interpretation of the corelation coefficient with the following:

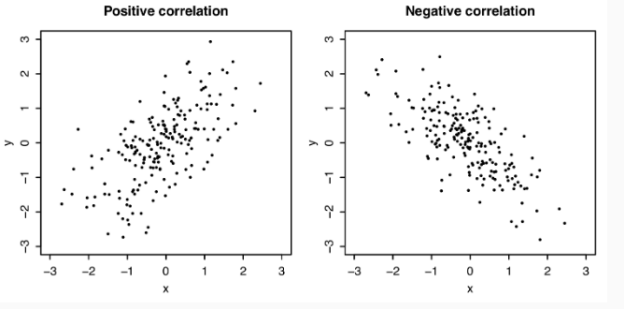


Figure 9: A positive correlation versus a negative correlation.

- A positive correlation coefficient ($\rho > 0$) means there is a positive association between X_1 and X_2.
  - As X_1 increases, X_2 also tends to increase.
- A negative correlation coefficient ($\rho < 0$) means there is a negative association between X_1 and X_2.
  - As X_1 increases, X_2 tends to decrease.
- A correlation coefficient equal to 0 ($\rho = 0$) means there is no linear association between X_1 and X_2.

Note that if the two random variables are independent then then $E(XY) = E(X)E(Y)$. Plugging this into our formula in Definition 2.5.1 we get that the covariance must be zero which implies $p = 0$. However the converse is not true in general. That is if the covariance is 0 then it need not to be true that the two random variables are independent.

Example 2.5.3

Find the covariance between X_1 and X_2, where $f(x_1, x_2) = 3x_1$ for $0 \leq x_1 \leq x_2 \leq 1$ and 0 otherwise.

Solution We begin and find the mean of X_11 and X_2.

$$E(X_1) = \int_0^1 \int_0^{x_2} x_1 3x_1 dx_1 dx_2 = 0.25,$$

$$E(X_2) = \int_0^1 \int_0^{x_2} x_2 3x_1 dx_1 dx_2 = 0.375.$$

We then find $E(X_1X_2)$:

$$E(X_1X_2) = \int_0^1 \int_0^{x_2} x_1 x_2 3x_1 dx_1 dx_2 = 0.2.$$

Thus the covariance is

$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = 0.2 - (0.25)(0.375) = 0.10625.$$

□

Before we end the section we highlight a few key things that might cause confused. If two variables X_1 and X_2 are strongly correlated, it just means they tend to move together; when one increases, the other usually increases (or decreases) too, in a linear way. But that does not mean one causes the other. For example, there could be a third factor that causes both X_1 and X_2 to move in the same direction. Suppose we find a strong correlation between ice cream sales and drowning incidents. This doesn't mean buying ice cream causes drowning. A third factor, hot weather, increases both ice cream consumption and swimming activity (which may lead to more drownings). So: high correlation, but no direct causal link.

If the correlation is zero, that means there's no linear relationship between X_1 and X_2. But they could still be related in a non-linear way. For example let X_1 be a uniform random variable and then define $X_2 = X_1^2$. These two random variables are clearly related but just not in a linear way.

Note that the textbook goes much more in depth of this section which I encourage you to read, but it is not in-scope/useful for this course. Moreover the textbook and many other text extend the notion of two joint random variables to several random variables which intuitively expands on what we learned in this chapter and honestly is quite verbose and unnecessary for the most part. These sections are skipped in this course and we end the chapter with the following section.

2.6 Linear Combinations of Random Variables

In this section we discuss many results of linear combinations of random variables. Some of these results are immediately seen to be true, such as the expectation of a linear combination of random variables is the sum of each expectation of the random variable. However there are some other results that are a little bit harder to see.

Theorem 2.6.1

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be random variables with $E(X_i) = \mu_i$ for $i = 1, \ldots, n$ and $E(Y_j) = \gamma_j$ for $j = 1, \ldots, m$. Define

$$U_1 = \sum_{i=1}^{n} a_i X_i \quad \text{and} \quad U_2 = \sum_{j=1}^{m} b_j Y_j$$

for constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$. Then:

- $E(U_1) = \sum_{i=1}^{n} a_i \mu_i \quad \text{and} \quad E(U_2) = \sum_{j=1}^{m} b_j \gamma_j$
- $\text{Var}(U_1) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \text{Cov}(X_i, X_j)$
- Similar expression for $\text{Var}(U_2)$
- $\text{Cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j)$

Proof. The proof for the first part is immediate using the linearity of expectation. I have omitted the proofs for the rest as I did them on paper and aren't useful for our purposes. If you'd like the proof contact me and I can provide it. □

2.6.2

Let X_1, X_2, and X_3 be random variables, where

- $E(X_1) = 1, \quad E(X_2) = 2, \quad E(X_3) = -1$
- $\text{Var}(X_1) = 1, \quad \text{Var}(X_2) = 3, \quad \text{Var}(X_3) = 5$
- $\text{Cov}(X_1, X_2) = -0.4, \quad \text{Cov}(X_1, X_3) = 0.5, \quad \text{Cov}(X_2, X_3) = -0.2$

Find the expected value and variance of

$$U = X_1 - 2X_2 - X_3.$$

Solution From the Theorem we see that

$$E(U) = E(X_1) - 2E(X_2) - E(X_3) = 1 - 2 \cdot 2 - (-1) = -2.$$

Moreover see that

$$\begin{aligned}
\text{Var}(U) &= \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \text{Cov}(X_i, X_j) \\
&= \left[(1)^2 \text{Var}(X_1) + (-2)^2 \text{Var}(X_2) + (-1)^2 \text{Var}(X_3)\right] \\
&+ 2\left[(1)(-2)\text{Cov}(X_1, X_2) + (1)(-1)\text{Cov}(X_1, X_3) + (-2)(-1)\text{Cov}(X_2, X_3)\right] \\
&= (1 + 12 + 5) + 2(-2 \cdot (-0.4) - 0.5 + 2 \cdot (-0.2)) \\
&= 17.8.
\end{aligned}$$

□

Before our next definition, let $X_1, X_2, \ldots, X_n$ be random variables. We say these random variables are independent if

$$P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = P(X_1 \le x_1) \cdot P(X_2 \le x_2) \cdots P(X_n \le x_n).$$

Definition 2.6.3

If random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed, i.e. each X_i has the same distribution, then we say that these random variables constitute a random sample of size n from that common distribution. We abbreviate independent and identically distributed by iid.

Example 2.6.4

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with common mean μ and variance σ^2. The sample mean is defined by

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

Then $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$.

Solution We see that

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{n\mu}{n} = \mu. \end{aligned}$$

Next, since X_i are independent with the same variance we have

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

□

Example 2.6.5

Define the sample variance by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right\}$$

Then

$$E(S^2) = \sigma^2.$$

Solution By the linearity of expectation and the result of the previous Example, we have

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2) \\ &= \frac{1}{n-1}\left[(n\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right] \\ &= \sigma^2, \end{aligned}$$

where in the second equality we solved for $E(X_i^2)$ from the equation $\sigma^2 = E(X_i^2) - \mu^2$ and similarly for $E(\bar{X}^2)$. □

For our final theorem of this chapter we discuss the MGF of a Linear Function of independent Random Variables.

Theorem 2.6.6 (Also referred to as Theorem 2.6.1)

Suppose $X_1, X_2, \ldots, X_n$ are n mutually independent random variables. Suppose, for all $i = 1, 2, \ldots, n$, X_i has MGF $M_{X_i}(t)$, for $-h_i < t < h_i$, where $h_i > 0$. Let

$$T = \sum_{i=1}^{n} k_i X_i,$$

where $k_1, k_2, \ldots, k_n$ are constants. Then T has the MGF given by

$$M_T(t) = \prod_{i=1}^{n} M_{X_i}(k_i t), \qquad -\min_i\{h_i\} < t < \min_i\{h_i\}.$$

Proof. Assume t is in the interval $(-\min_i\{h_i\}, \min_i\{h_i\})$. Then, by independence,

$$\begin{aligned} M_T(t) &= E\left[e^{\sum_{i=1}^{n} t k_i X_i}\right] = E\left[\prod_{i=1}^{n} e^{t k_i X_i}\right] \\ &= \prod_{i=1}^{n} E\left[e^{t k_i X_i}\right] = \prod_{i=1}^{n} M_i(k_i t). \end{aligned}$$

□

Example 2.6.7

Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with common MGF

$$M(t) = e^{t^2/2}, \quad \text{for } t \in \mathbb{R}.$$

Find the MGF of $T = \sum_{i=1}^{n} X_i$.

Solution By Theorem 2.6.6 we have that

$$M_T(t) = \prod_{i=1}^{n} M_i(t) = M(t)^n = e^{nt^2/2}.$$

□

2.7 Practice Problems

These are the practice problem solutions from the list (questions are from the textbook) and tutorial problems organized by each section.

2.7.1 Section 2.1 Answers

2.1.1

Let $f(x_1, x_2) = 4x_1x_2$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, be the pdf of X_1 and X_2. Find

$$P\left(0 < X_1 < \frac{1}{2},\ \frac{1}{4} < X_2 < 1\right), \quad P(X_1 = X_2), \quad P(X_1 < X_2), \quad \text{and} \quad P(X_1 \leq X_2).$$

Solution First to find $P\left(0 < X_1 < \frac{1}{2},\ \frac{1}{4} < X_2 < 1\right)$, we simply integrate within the bounds.

$$\begin{aligned} P\left(0 < X_1 < \frac{1}{2},\ \frac{1}{4} < X_2 < 1\right) &= \int_{1/4}^{1}\int_{0}^{1/2} 4x_1x_2 dx_1 dx_2. \\ &= \int_{1/4}^{1} \frac{1}{2}x_2 dx_2 \\ &= 0.234. \end{aligned}$$

Next to find $P(X_1 = X_2)$, we see that this is finding the volume under the surface $4x_1x_2$ above the line segment $0 < x_1 = x_2 < 1$ in the xy plane. However a line has zero volume so we get that

$$P(X_1 = X_2) = 0.$$

Next to find $P(X_1 < X_2)$ we integrate from with respect to $x_1 < x_2 < 1$. We find that

$$P(X_1 < X_2) = \int_{x_2}^{1}\int_{x_1}^{x_2} 4x_1x_2 dx_1 dx_2 = \frac{1}{2}.$$

Finally since we know $X_1 = X_2$ has zero probability we can conclude that

$$P(X_1 < X_2) = P(X_1 \leq X_2).$$

□

2.1.10

Let the random variables X_1 and X_2 have the joint pmf described as follows:

(x_1, x_2)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$
$p(x_1, x_2)$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$

and $p(x_1, x_2)$ is equal to zero elsewhere.

(a) Write these probabilities in a rectangular array as in Example 2.1.4, recording each marginal pdf in the "margins."

(b) What is $P(X_1 + X_2 = 1)$?

Solution For (a):

Support of $X_1 \backslash X_2$	0	1	2	$p_{X_1}(x_1)$
0	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{7}{12}$
1	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{5}{12}$
$p_{X_2}(x_2)$	$\frac{4}{12}$	$\frac{5}{12}$	$\frac{3}{12}$	1

For (b): We can simply use table to find that

$$P(X_1 + X_2 = 1) = p(0,1) + p(1,0) = \frac{5}{12}.$$

□

2.1.11

Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = 15x_1^2 x_2, \quad 0 < x_1 < x_2 < 1,$$

zero elsewhere. Find the marginal pdfs and compute $P(X_1 + X_2 \leq 1)$.

Solution We find the marginal pdf of X_1 and X_2 easily by integrating the opposite variable. For X_1, we keep the variable x_1 fixed and integrate with respect to x_2. That is we sum out all possible values of x_2. Since $x_1 < x_2 < 1$, the bounds for integration are x_1 to 1.

$$f_{X_1}(x_1) = \int_{x_1}^{1} 15x_1^2 x_2 dx_2 = 15x_1^2 \left(\frac{1}{2} - \frac{x_1^2}{2}\right),$$

for $0 < x_1 < 1$ and zero otherwise. Similarly for X_2, we keep the x_2 variable fixed and integrate with respect to x_1. Since $0 < x_1 < x_2$ we see that

$$f_{X_2}(x_2) = \int_{0}^{x_2} 15x_1^2 x_2 dx_1 = 5x_2^4,$$

for $0 < x_2 < 1$ and zero otherwise. Next we see that with the condition $X_1 + X_2 \leq 1$ and $0 < x_1 < x_2 < 1$ implies that $x_1 < x_2 < 1 - x_1$. This then implies that $0 < x_1 < 0.5$. So we integrate with respect to these bounds and find that

$$\begin{aligned} P(X_1 + X_2 \leq 1) &= \int_0^{0.5} \int_{x_1}^{1-x_1} 15x_1^2 x_2 dx_2 dx_1 \\ &= \frac{15}{2} \int_0^{0.5} x_1^2 \left[(1 - x_1)^2 - x_1^2\right] dx_1 \\ &= \frac{5}{64}. \end{aligned}$$

□

2.1.13

Let X_1, X_2 be two random variables with the joint pmf

$$p(x_1, x_2) = \frac{x_1 + x_2}{12}, \quad \text{for } x_1 = 1, 2;\ x_2 = 1, 2,$$

zero elsewhere.
Compute $E(X_1),\ E(X_1^2),\ E(X_2),\ E(X_2^2),\ E(X_1X_2)$.
Is $E(X_1X_2) = E(X_1)E(X_2)$?
Find

$$E(2X_1 - 6X_2^2 + 7X_1X_2).$$

Solution To compute $E(X_1)$ and $E(X_2)$ we can either first find the marginal distributions and then compute the expectations or use our formula we stated in section 2.1/

$$\begin{aligned} E(X_1) &= \sum_{x_1=1}^{2}\sum_{x_2=1}^{2} x_1 p(x_1,x_2) \\ &= 1\cdot(p(1,1)+p(1,2)) + 2\cdot(p(2,1)+p(2,2)) \\ &= \frac{1}{6}+\frac{1}{4}+2\left(\frac{1}{4}+\frac{1}{3}\right) \\ &= \frac{19}{12}. \end{aligned}$$

Using the symmetry of the pmf we can see that $E(X_2) = 19/12$. Next then

$$\begin{aligned} E(X_1^2) &= \sum_{x_1=1}^{2}\sum_{x_2=1}^{2} x_1 p(x_1,x_2) \\ &= 1^2\cdot(p(1,1)+p(1,2)) + 2^2\cdot(p(2,1)+p(2,2)) \\ &= \frac{1}{6}+\frac{1}{4}+4\left(\frac{1}{4}+\frac{1}{3}\right) \\ &= \frac{11}{4}. \end{aligned}$$

Again using the symmetry of the pmf we can see that $E(X_2^2) = 11/4$. Then we see that

$$\begin{aligned} E(X_1X_2) &= \sum_{x_1=1}^{2}\sum_{x_2=1}^{2} x_1x_2 p(x_1,x_2) \\ &= 1\cdot(1\cdot p(1,1)+2\cdot p(1,2)) + 2\cdot(1\cdot(2,1)+2\cdot p(2,2)) \\ &= \frac{1}{6}+\frac{1}{2}+2\left(\frac{1}{4}+\frac{2}{3}\right) \\ &= \frac{5}{2}. \end{aligned}$$

Notice that $E(X_1X_2) = 5/2 \neq E(X_1)E(X_2) = (19/12)^2 \approx 2.507$. Thus X_1 and X_2 are not independent. Finally using the linearity of expectation we get that

$$E(2X_1 - 6X_2^2 + 7X_1X_2) = 2E(X_2) - 6E(X_2^2) + 7E(X_1X_2) = 2\cdot\frac{19}{12} - 6\cdot\frac{11}{4} + 7\cdot\frac{5}{2} = \frac{25}{6}.$$

□

2.7.2 Section 2.2 Answers

2.2.1

If

$$p(x_1,x_2) = \left(\frac{2}{3}\right)^{x_1+x_2}\left(\frac{1}{3}\right)^{2-x_1-x_2},$$

for $(x_1,x_2) = (0,0),(0,1),(1,0),(1,1)$, and zero elsewhere, is the joint pmf of X_1 and X_2, find the joint pmf of

$$Y_1 = X_1 - X_2 \quad \text{and} \quad Y_2 = X_1 + X_2.$$

Solution To find the pmf $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$, we simply find the support of Y and plug into the pmf of X_1 and X_2. We could find a closed form formula as well. We see that the support of Y_1 and Y_2 becomes

(x_1, x_2)	$Y_1 = x_1 - x_2$	$Y_2 = x_1 + x_2$
$(0,0)$	0	0
$(0,1)$	-1	1
$(1,0)$	1	1
$(1,1)$	0	2

Thus the joint pmf of Y_1 and Y_2 is

(Y_1, Y_2)	PMF Value
$(-1,1)$	$2/9$
$(0,0)$	$1/9$
$(0,2)$	$4/9$
$(1,1)$	$2/9$

or equivalently

$$p_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{9}, & (y_1,y_2) = (0,0) \\ \frac{2}{9}, & (y_1,y_2) = (-1,1) \\ \frac{2}{9}, & (y_1,y_2) = (1,1) \\ \frac{4}{9}, & (y_1,y_2) = (0,2) \\ 0, & \text{otherwise} \end{cases}$$

□

2.2.2

Let X_1 and X_2 have the joint pmf $p(x_1,x_2) = \frac{x_1 x_2}{36}$, $x_1 = 1,2,3$ and $x_2 = 1,2,3$, zero elsewhere. Find first the joint pmf of $Y_1 = X_1 X_2$ and $Y_2 = X_2$, and then find the marginal pmf of Y_1.

Solution We begin with finding the inverses. $x_1 = y_1/x_2 = y_1/y_2$ and $x_2 = y_2$. Then we see that

$$p_{Y_1,Y_2}(y_1,y_2) = p_{X_1,X_2}(y_1/y_2, y_2) = \frac{\frac{y_1}{y_2} \cdot y_2}{36} = \frac{y_1}{36}.$$

where $y_1 = 1,2,3,4,6,9$ and $y_2 = 1,2,3$ and zero elsewhere. Then we find the marginal pmf of Y_1:

$$\begin{aligned}
P(Y_1 = 1) &= P(X_1 = 1, X_2 = 1) = \frac{1}{36} \\
P(Y_1 = 2) &= P(X_1 = 1, X_2 = 2) + P(X_1 = 2, X_2 = 1) = \frac{2}{36} + \frac{2}{36} = \frac{4}{36} \\
P(Y_1 = 3) &= P(X_1 = 1, X_2 = 3) + P(X_1 = 3, X_2 = 1) = \frac{3}{36} + \frac{3}{36} = \frac{6}{36} \\
P(Y_1 = 4) &= P(X_1 = 2, X_2 = 2) = \frac{4}{36} \\
P(Y_1 = 6) &= P(X_1 = 2, X_2 = 3) + P(X_1 = 3, X_2 = 2) = \frac{6}{36} + \frac{6}{36} = \frac{12}{36} \\
P(Y_1 = 9) &= P(X_1 = 3, X_2 = 3) = \frac{9}{36}.
\end{aligned}$$

□

2.2.4

Let X_1 and X_2 have the joint pdf $h(x_1, x_2) = 8x_1x_2$, $0 < x_1 < x_2 < 1$, zero elsewhere. Find the joint pdf of $Y_1 = \frac{X_1}{X_2}$ and $Y_2 = X_2$. *Hint:* Use the inequalities $0 < y_1y_2 < y_2 < 1$ in considering the mapping from $\mathcal{S}$ onto $\mathcal{T}$.

Solution We first determine the support of the joint pdf of Y_1 and Y_2. We see that $0 < y_1 = x_1/x_2 < 1$ and $0 < y_2 = x_2 < 1$. This implies that $0 < x_1 < x_2$. If we rewrite this in terms of y_1 and y_2 we get that $0 < y_1y_2 < y_2 < 1$ is the support of the joint pdf. We can express this as $0 < y_1 < 1$ and $0 < y_2 < 1$. Since this transformation is injective we do the Jacobian method and compute the jacobian. We get that

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_! \\ 0 & 1 \end{vmatrix} = y_2$$

Together we get that

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{X_1,X_2}(y_1y_2, y_2)|J| \\ &= 8(y_1y_2)y_2|y_2| \\ &= 8y_1y_2^3. \end{aligned}$$

□

2.2.6

Suppose X_1 and X_2 have the joint pdf

$$f_{X_1,X_2}(x_1, x_2) = e^{-(x_1+x_2)}, \quad 0 < x_i < \infty, \ i = 1, 2,$$

zero elsewhere.

(a) Use formula (2.2.5) to find the pdf of $Y_1 = X_1 + X_2$.

(b) Find the mgf of Y_1.

Solution To find the pdf of $Y_1 = X_1 + X_2$, we can find the joint pdf of Y_1 and Y_2 where Y_2 is any linear one-to-one transformation and then find the marginal pdf of X_1 from the joint pdf. Let $Y_2 = x_2$. Then the inverses are $x_1 = y_1 - y_2$ and $x_2 = y_2$. Thus it is easy to see that the Jacobian $|J| = 1$. So we find that $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(y_1 - y_2, y_2)$. Then we can go ahead and find the marginal pdf Y_1

$$\begin{aligned} f_{Y_1}(y_1) &= \int_0^{y_1} f_{Y_1,Y_2}(y_1, y_2)dy_2 \\ &= \int_0^{y_1} f_{X_1,X_2}(y_1 - y_2, y_2)dy_2 \\ &= \int_0^{y_1} e^{-(y_1-y_2+y_2)}dy_2 \\ &= \int_0^{y_1} e^{-y_1}dy_2 \\ &= y_1e^{-y_1}. \end{aligned}$$

Next we find the mgf of Y_1.

$$\begin{aligned} M_{Y_1}(t) &= E(e^{tY_1}) \\ &= \int_0^\infty e^{ty_1} y_1 e^{-y_1} dy_1 \\ &= \int_0^\infty y_1 e^{-y_1+ty_1} dy_1 \\ &= \int_0^\infty y_1 e^{-(1-t)y_1} dy_1 \\ &= \frac{1}{(1-t)^2} \end{aligned}$$

for $t < 1$. □

2.7.3 Section 2.3 Answers

2.3.1

Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = x_1 + x_2, \quad 0 < x_1 < 1,\ 0 < x_2 < 1,$$

zero elsewhere. Find the conditional mean and variance of X_2, given $X_1 = x_1$, $0 < x_1 < 1$.

Solution We see that the conditional mean of X_2 given $X_1 = x_1$ is $E(X_2|x_1)$. This is equal to

$$E(X_2|x_1) = \int_0^1 x_2 f_{X_2|X_1}(x_2|x_1) dx_2.$$

We first find the conditional pdf. We see that the conditional pdf is given by

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)}.$$

We find the marginal pdf of X_1.

$$f_{X_1}(x_1) = \int_0^1 (x_1 + x_2) dx_2 = x_1 + \frac{1}{2}.$$

We then get that

$$\begin{aligned} E(X_2|x_1) &= \int_0^1 x_2 \cdot \frac{x_1 + x_2}{x_1 + \frac{1}{2}} dx_2 \\ &= \frac{1}{x_1 + \frac{1}{2}} \left[x_1 \int_0^1 x_2 dx_2 + \int_0^1 x_2^2 dx_2 \right] \\ &= \frac{1}{x_1 + \frac{1}{2}} \left(\frac{x_1}{2} + \frac{1}{3} \right) \\ &= \frac{3x + 2}{6x + 3}. \end{aligned}$$

Next, the conditional variance $\text{Var}(X_2|x_1) = E(X_2^2|x_1) - [E(X_2|x_1)]^2$. We begin and find $E(X_2^2|x_1)$.

$$\begin{aligned} E(X_2^2|x_1) &= \int_0^1 x_2^2 \cdot \frac{x_1 + x_2}{x_1 + \frac{1}{2}} dx_2 \\ &= \frac{1}{x_1 + \frac{1}{2}} \int_0^1 x_2^2 (x_1 + x_2)\, dx_2 \\ &= \frac{1}{x_1 + \frac{1}{2}} \left[x_1 \int_0^1 x_2^2 dx_2 + \int_0^1 x_2^3 dx_2 \right] \\ &= \frac{1}{x_1 + \frac{1}{2}} \left(x_1 \cdot \frac{1}{3} + \frac{1}{4} \right). \end{aligned}$$

Substituting we get that

$$\text{Var}(X_2|x_1) = \frac{\frac{1}{3}x_1 + \frac{1}{4}}{x_1 + \frac{1}{2}} - \left(\frac{\frac{1}{2}x_1 + \frac{1}{3}}{x_1 + \frac{1}{2}} \right)^2 = \frac{6x_1^2 + 6x_1 + 1}{2\,(6x_1 + 3)^2}.$$

□

2.3.4

Suppose X_1 and X_2 are random variables of the discrete type that have the joint pmf

$$p(x_1, x_2) = \frac{x_1 + 2x_2}{18}, \quad (x_1, x_2) = (1,1), (1,2), (2,1), (2,2),$$

zero elsewhere. Determine the conditional mean and variance of X_2, given $X_1 = x_1$, for $x_1 = 1$ or 2. Also, compute $E(3X_1 - 2X_2)$.

Solution We see that the conditional mean of X_2 given $X_1 = x_1$ is $E(X_2|x_1)$. This is equal to

$$E(X_2|x_1) = \sum_{x_2=1}^{2} x_2 p_{X_2|X_1}(x_2|x_1).$$

We first find the conditional pmf

$$p_{X_2|X_1}(x_2|x_1) = \frac{p(x_1, x_2)}{p_{X_1}(x_1)}.$$

To find this we find the marginal pdf of X_1.

$$\begin{aligned} p_{X_1}(x_1) &= \sum_{x_2=1}^{2} p(x_1, x_2) \\ &= p(x_1, 1) + p(x_1, 2) \\ &= \frac{x_1 + 2}{18} + \frac{x_1 + 4}{18} \\ &= \frac{x_1 + 3}{9}, \end{aligned}$$

for $x_1 = 1, 2$. Thus

$$p_{X_2|X_1}(x_2|x_1) = \frac{\frac{x_1 + 2x_2}{18}}{\frac{x_1 + 3}{9}} = \frac{9(x_1 + 2x_2)}{18(x_1 + 3)}.$$

Together we get that

$$\begin{aligned} E(X_2|x_1) &= \sum_{x_2=1}^{2} x_2 \cdot \frac{9(x_1+2x_2)}{18(x_1+3)} \\ &= \left(\frac{9}{18(x_1+3)}\right) \sum_{x_2=1}^{2} x_2(x_1+2x_2). \end{aligned}$$

Next to find variance we need $\text{Var}(X_2|x_1) = E(X_2^2|x_1) - (E(X_2|x_1))^2$. We find that

$$\text{Var}(X_2|x_1) = \left(\frac{9}{18(x_1+3)}\right) \sum_{x_2=1}^{2} x_2^2(x_1+2x_2) - \left[\left(\frac{9}{18(x_1+3)}\right) \sum_{x_2=1}^{2} x_2(x_1+2x_2)\right]^2.$$

We get that then $\text{Var}(X_2|x_1=1) = 15/64$ and $\text{Var}(X_2|x_1=2) = 6/25$. Finally to find $E(3X_1-2X_2)$ we find $E(X_1)$ and $E(X_2)$. We see that

$$E(X_1) = \sum_{x_1=1}^{2} x_1 p_{X_1}(x_1) = \frac{1}{9} \sum_{x_1=1}^{2} x_1(x_1+3) = \frac{14}{9}.$$

Next we first find the marginal pdf of X_2.

$$\begin{aligned} p_{X_2}(x_2) &= \sum_{x_1=1}^{2} p(x_1, x_2) \\ &= p(1, x_2) + p(2, x_2) \\ &= \frac{1+2x_2}{18} + \frac{2+2x_2}{18} \\ &= \frac{4x_2+3}{9}, \end{aligned}$$

for $x_2 = 1, 2$ and zero elsewhere. Then

$$E(X_2) = \sum_{x_2=1}^{2} x_2 p_{X_2}(x_2) = \frac{1}{18} \sum_{x_2=1}^{2} x_2(4x_2+3) = \frac{29}{18}.$$

Then

$$E(3X_1 - 2X_2) = 3E(X_1) - 2E(X_2) = \frac{13}{9}.$$

□

2.3.8

Let X and Y have the joint pdf

$$f(x, y) = 2\exp\{-(x+y)\}, \quad 0 < x < y < \infty,$$

zero elsewhere. Find the conditional mean $E(Y \mid x)$ of Y, given $X = x$.

Solution We need to find

$$E(Y|x) = \int_x^\infty y f_{Y|X}(Y|x) dy = \int_x^\infty y \cdot \frac{f(x,y)}{f_X(x)} dy.$$

We first find the marginal pdf of X

$$\begin{aligned} f_X(x) &= \int_x^\infty f(x,y)dy \\ &= \int_x^\infty 2e^{-x}e^{-y}dy \\ &= 2e^{-x}\int_x^\infty e^{-y}dy \\ &= 2e^{-x}\lim_{b\to\infty}\left[-e^b + e^{-x}\right] \\ &= 2e^{-2x}. \end{aligned}$$

Then

$$f_{Y|X}(Y|x) = \frac{2e^{-x}e^{-y}}{2e^{-2x}} = e^{x-y}$$

for $y > x$. Then

$$\begin{aligned} E(Y|x) &= \int_x^\infty y \cdot \frac{f(x,y)}{f_X(x)}dy \\ &= \int_x^\infty ye^x e^{-y}dy \\ &= e^x \lim_{b\to\infty}\left[\left(\frac{-b}{e^b} - \frac{1}{e^b}\right) - \left(-xe^{-x} - e^{-x}\right)\right] \\ &= e^x[xe^{-x} + e^{-x}] \\ &= x + 1. \end{aligned}$$

□

2.7.4 Section 2.4 Answers

2.4.1

Show that the random variables X_1 and X_2 with joint pdf

$$f(x_1,x_2) = \begin{cases} 12x_1x_2(1-x_2), & 0 < x_1 < 1,\ 0 < x_2 < 1 \\ 0, & \text{elsewhere} \end{cases}$$

are independent.

Solution To show that they are independent we have to show $f(x_1,x_2) = f_1(x_1)f_2(x_2)$. However instead of finding the marginal pdf of each random variable we can instead notice that

$$12x_1x_2(1-x_2) = (12x_1)\left[x_2 - x_2^2\right] = g(x_1)h(x_2)$$

where g and h are positive functions where $0 < x_1 < 1$ and $0 < x_2 < 1$. Then theorem 2.4.4 we have that they are independent. □

2.4.3

Let $p(x_1,x_2) = \frac{1}{16}$, $x_1 = 1,2,3,4$ and $x_2 = 1,2,3,4$, zero elsewhere, be the joint pmf of X_1 and X_2. Show that X_1 and X_2 are independent.

Solution We first find the marginal pmf of X_1.

$$p_{X_1}(x_1) = \sum_{x2=1}^{4} p(x_1, x_2) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4}.$$

Similarly for X_2,

$$p_{X_2}(x_2) = \sum_{x1=1}^{4} p(x_1, x_2) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4}.$$

Then we see that

$$p(x_1, x_2) = \frac{1}{16} = p_{X_1}(x_1)p_{X_2}(x_2).$$

Thus they are independent. □

2.4.8

Let X and Y have the joint pdf $f(x, y) = 3x$, $0 < y < x < 1$, zero elsewhere. Are X and Y independent? If not, find $E(X \mid y)$.

Solution We find the marginal pdf of X.

$$f_X(x) = \int_0^x f(x, y)dy = \int_0^x 3xdy = 3x^2$$

for $y < x < 1$ and zero elsewhere. Then the pdf for Y is

$$f_Y(y) = \int_y^1 f(x, y)dx = \int_y^1 3xdx = \frac{3}{2}(1 - y^2).$$

Then we see clearly that

$$f(x, y) = 3x \neq f_X(x)f_Y(y) = \frac{6x^2(1 - y^2)}{2},$$

which means that it is not independent. We also could have immediately noticed that the support of joint distribution is not a rectangular region of the form $[a, b] \times [c, d]$ but rather a triangular region which my Theorem 2.4.4 says they are not independent. Next we find $E(X|y)$. We get that

$$\begin{aligned} E(X|y) &= \int_y^1 x f_{X|Y}(x|y)dx \\ &= \int_y^1 x \cdot \frac{3x}{\frac{3}{2}(1 - y^2)}dx \\ &= \frac{2}{1 - y^2} \int_y^1 x^2 dx \\ &= \frac{2}{1 - y^2} \left[\frac{1}{3} - \frac{y^3}{3}\right] \\ &= \frac{2}{3} \cdot \frac{1 - y^3}{1 - y^2} \end{aligned}$$

For $0 < y < 1$. □

2.7.5 Section 2.5 Answers

2.5.1

Let the random variables X and Y have the joint pmf

(a) $p(x,y) = \frac{1}{3},\ (x,y) = (0,0),(1,1),(2,2)$, zero elsewhere.

(b) $p(x,y) = \frac{1}{3},\ (x,y) = (0,2),(1,1),(2,0)$, zero elsewhere.

(c) $p(x,y) = \frac{1}{3},\ (x,y) = (0,0),(1,1),(2,0)$, zero elsewhere.

In each case compute the correlation coefficient of X and Y.

Solution To find the correlation coefficient between X and Y we need to find $p = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$. We begin with finding the covariance. Thus for each part we will find $E(X), E(Y), E(X^2), E(Y^2), E(XY)$.

For (a):

$$E[X] = \sum x \cdot p(x,y) = \frac{1}{3}(0+1+2) = 1, \quad E[Y] = E[X] = 1$$

$$E[XY] = \frac{1}{3}(0\cdot 0 + 1\cdot 1 + 2\cdot 2) = \frac{5}{3}$$

$$E[X^2] = \frac{1}{3}(0^2 + 1^2 + 2^2) = \frac{5}{3} = E[Y^2]$$

Variances:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{5}{3} - 1 = \frac{2}{3}$$

Covariance:

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{5}{3} - 1 = \frac{2}{3}$$

Correlation:

$$\rho_{X,Y} = \frac{2/3}{\sqrt{2/3}\cdot\sqrt{2/3}} = \frac{2/3}{2/3} = \boxed{1}$$

For (b):

$$E[X] = \frac{1}{3}(0+1+2) = 1, \quad E[Y] = \frac{1}{3}(2+1+0) = 1$$

$$E[XY] = \frac{1}{3}(0\cdot 2 + 1\cdot 1 + 2\cdot 0) = \frac{1}{3}$$

$$E[X^2] = \frac{1}{3}(0+1+4) = \frac{5}{3}, \quad E[Y^2] = \frac{1}{3}(4+1+0) = \frac{5}{3}$$

Covariance:

$$\mathrm{Cov}(X,Y) = \frac{1}{3} - (1)(1) = \frac{1}{3} - 1 = -\frac{2}{3}$$

Variances:
Same as case (a): $\mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{2}{3}$
Correlation:

$$\rho_{X,Y} = \frac{-2/3}{\sqrt{2/3} \cdot \sqrt{2/3}} = -1$$

For (c):

$$E[X] = \frac{1}{3}(0+1+2) = 1, \quad E[Y] = \frac{1}{3}(0+1+0) = \frac{1}{3}$$

$$E[XY] = \frac{1}{3}(0 \cdot 0 + 1 \cdot 1 + 2 \cdot 0) = \frac{1}{3}$$

$$E[X^2] = \frac{1}{3}(0+1+4) = \frac{5}{3}, \quad E[Y^2] = \frac{1}{3}(0+1+0) = \frac{1}{3}$$

Covariance:

$$\text{Cov}(X,Y) = \frac{1}{3} - (1)\left(\frac{1}{3}\right) = 0$$

Variance:
$\text{Var}(X) = \frac{5}{3} - 1^2 = \frac{2}{3}$, $\text{Var}(Y) = \frac{1}{3} - \left(\frac{1}{3}\right)^2 = \frac{2}{9}$
Correlation:

$$\rho_{X,Y} = \frac{0}{\sqrt{2/3} \cdot \sqrt{2/9}} = 0$$

□

2.5.3

Let $f(x,y) = 2$, $0 < x < y$, $0 < y < 1$, zero elsewhere, be the joint pdf of X and Y. Show that the conditional means are, respectively, $(1+x)/2$, $0 < x < 1$, and $y/2$, $0 < y < 1$. Show that the correlation coefficient of X and Y is $\rho = \frac{1}{2}$.

Solution We begin with finding the conditional means of X and Y. First we find the marginal pdf of X and Y.

$$f_X(x) = \int_x^1 2dy = 2(1-x),$$

$$f_Y(y) = \int_0^y 2dx = 2y.$$

Then we find that

$$\begin{aligned} E(X|y) &= \int_0^y x \cdot \frac{2}{2y} dx \\ &= \frac{1}{y}\int_0^y x dx \\ &= \frac{y^2}{2y} = \frac{y}{2}. \end{aligned}$$

Similarly,

$$\begin{aligned} E(Y|x) &= \int_x^1 y \cdot \frac{2}{2(1-x)} dy \\ &= \frac{1}{1-x} \int_y^1 x dy \\ &= \frac{1}{1-x} \cdot \left(\frac{1}{2} - \frac{x^2}{2} \right) \\ &= \frac{1-x^2}{2(1-x)} = \frac{(1-x)(1+x)}{2(1-x)} = \frac{1+x}{2}. \end{aligned}$$

Next to find the covariance of X and Y we need to find $E(X)$,$E(Y)$, $E(XY)$, $\text{Var}(X)$, $\text{Var}(Y)$.

$$E(X) = \int_0^1 x f_X(x) dx = \int_0^1 2x(1-x) dx = \frac{1}{3},$$

$$E(Y) = \int_0^1 y f_Y(y) dy == \frac{1}{2} \int_0^1 y^2 dy = \frac{1}{6},$$

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^y xy f(x,y) dx dy \\ &= \frac{1}{4}. \end{aligned}$$

Thus we get that

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{36}.$$

Next to find the correlation coefficient we need the standard deviation so we need to find $\text{Var}(X)$ and $\text{Var}(Y)$. To find this we need to find $E(X^2)$ and $E(Y^2)$.

$$E(X^2) = \int_0^1 x^2 f_X(x) dx = \int_0^1 2x^2(1-x) dx = \frac{1}{6},$$

$$E(Y^2) = \int_0^1 y^2 f_Y(y) dy == \frac{1}{2} \int_0^1 y^3 dy = \frac{1}{8}.$$

We get that then

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{1}{18},$$

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = \frac{1}{18}.$$

Then we see that

$$p = \frac{Cov(X,Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}} = \frac{1/36}{1/18} = \frac{1}{2}.$$

□

2.5.4

Show that the variance of the conditional distribution of Y, given $X = x$, in Exercise 2.5.3, is $(1-x)^2/12$, $0 < x < 1$, and that the variance of the conditional distribution of X, given $Y = y$, is $y^2/12$, $0 < y < 1$.

Solution We need to show that the variance of $f_{Y|X}(y|x)$ is $(1-x)^2/12$ and similarly for X. Notice that

$$\text{Var}(Y|x) = E(Y^2|x) - (E(Y|x))^2.$$

We find that

$$E(Y^2|x) = \frac{1}{1-x}\int_x^{!} y^2 dy = \frac{1-x^3}{3(1-x)}.$$

Then

$$\text{Var}(Y|x) = \frac{1-x^3}{3(1-x)} - \left(\frac{1+x}{2}\right)^2 = \frac{(1-x)^2}{12}.$$

Similarly

$$E(X^2|y) = \int_0^y x^2 \frac{1}{y} dx = \frac{y^3}{3}.$$

Then

$$\text{Var}(X|y) = \frac{y^3}{3} - \left(\frac{y}{2}\right) = \frac{y^2}{12}.$$

□

2.5.9

Let X and Y have the joint pmf $p(x,y) = \frac{1}{7}$, $(0,0),(1,0),(0,1),(1,1),(2,1),(1,2),(2,2)$, zero elsewhere. Find the correlation coefficient ρ.

Solution We first find the marginal pmf of X and Y using the table.

$y\backslash x$	0	1	2	$p_Y(y)$
0	1/7	1/7	0	2/7
1	1/7	1/7	1/7	3/7
2	0	1/7	1/7	2/7
$p_X(x)$	2/7	3/7	2/7	

We will find $E(X), E(X^2), \text{Var}(X), \sigma_X$ and similarly for Y.

$$E(x) = \sum_{x=0}^{2} xp(x) = 1,$$

$$E(X^2) = \sum_{x=0}^{2} x^2 p(x) = \frac{11}{7},$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{4}{7},$$

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{4/7}.$$

Next due to the symmetry of the distribution we have that $E(Y) = E(X), E(Y^2) = E(X^2), \text{Var}(Y) = \text{Var}(X)$ and $\sigma_Y = \sigma_X$. Then we see that

$$E(XY) = \sum_{x=0}^{2}\sum_{y=0}^{2} xyp(x,y) = \frac{9}{7}.$$

Then

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{2}{7}.$$

Thus

$$p = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{2/7}{4/7} = \frac{1}{2}.$$

□

2.7.6 Section 2.8 Answers

2.8.2

Let X_1, X_2, X_3, X_4 be four iid random variables having the same pdf $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Find the mean and variance of the sum Y of these four random variables.

Solution Let $Y = X_1 + X_2 + X_3 + X_4$. Since they are iid we know that the mean is going to be

$$E(Y) = E(X_1 + X_2 + X_3 + X_4) = 4\mu$$

where μ is the common mean. We see that

$$\mu = \int_0^1 2x^2 dx = \frac{2}{3}.$$

Then $E(Y) = \frac{8}{3}$. Next we find the common second moment $E(X^2)$.

$$E(X^2) = \int_0^1 2x^3 dx = \frac{1}{2}.$$

So the common variance is

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{1}{18}.$$

Then using Theorem 2.6.1 we get that

$$\text{Var}(Y) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \le i \le j \le 4} Cov(X_i, X_j).$$

However notice that since these events are independent then the covariance is zero and thus

$$\text{Var}(Y) = \sum_{i=4}^{4} \text{Var}(X_i) = 4 \cdot \frac{1}{18} = \frac{2}{9}.$$

□

2.8.6

Determine the mean and variance of the sample mean $\bar{X} = 5^{-1} \sum_{i=1}^{5} X_i$, where $X_1, \ldots, X_5$ is a random sample from a distribution having pdf $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

Solution The sample mean as found in Example 2.6.4 will be the common mean μ. That is

$$E(\bar{X}) = \mu = \int_0^1 4x^4 dx = \frac{4}{5}.$$

Next the variance is then the common variance σ^2 divided by n. That is

$$\text{Var}(\bar{X}) = \frac{E(X^2) - \mu^2}{5}$$

We find that

$$E(X^2) = \int_0^1 4x^5 dx = \frac{2}{5}.$$

Then

$$\text{Var}(X) = \frac{2}{5} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}.$$

Then

$$\text{Var}(\bar{X}) = \frac{2}{375}.$$

□

2.8.7

Let X and Y be random variables with $\mu_1 = 1$, $\mu_2 = 4$, $\sigma_1^2 = 4$, $\sigma_2^2 = 6$, $\rho = \frac{1}{2}$. Find the mean and variance of the random variable $Z = 3X - 2Y$.

Solution The mean of Z is

$$E(Z) = E(3X - 2Y) = 3E(X) - 2E(Y) = 3 \cdot 1 - 2 \cdot 4 = -5.$$

Before we solve for the variance, we are going to need to the covariance of X and Y. We can solve for the covariance and see that $Cov(X, Y) = p\sigma_1\sigma_2$. Next the variance is then

$$\begin{aligned}
\text{Var}(Z) &= \sum_{i=1} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \le i \le j \le 2} a_i b_j Cov(X_i, X_j) \\
&= 3^2 \text{Var}(X_1) + (-2)^2 \text{Var}(X_2) + 2(3)(-2)p\sigma_1\sigma_2 \\
&= 9 \cdot 4 + 4 \cdot 6 + 2(3)(-2)(1/2)(\sqrt{4})(\sqrt{6}) \\
&= 30.606
\end{aligned}$$

□

2.8.10

Determine the correlation coefficient of the random variables X and Y if $\text{var}(X) = 4$, $\text{var}(Y) = 2$, and $\text{var}(X + 2Y) = 15$.

Solution Since $\text{var}(X + 2Y) = 15$, we can solve for $Cov(X, Y)$.

$$\begin{aligned}
\text{var}(X + 2Y) = 15 &= \text{var}(X) + 4\text{var}(Y) + 2Cov(X, Y) \\
&= 4 + 8 + 4Cov(X, Y).
\end{aligned}$$

We then get that $Cov(X, Y) = \frac{3}{4}$. Then we see that

$$p = \frac{Cov(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{3\sqrt{2}}{16}.$$

□

2.8.11

Let X and Y be random variables with means μ_1, μ_2; variances σ_1^2, σ_2^2; and correlation coefficient ρ. Show that the correlation coefficient of $W = aX + b$, $a > 0$, and $Z = cY + d$, $c > 0$, is ρ

Solution We first find the means and variances:

$$E(W) = E(aX + b) = aE(X) + b = a\sigma_1^2 + b,$$
$$E(Z) = E(cY + d) = cE(Y) + d = c\sigma_2^2 + d,$$
$$\text{Var}(aX + b) = a^2\text{Var}(X) = a^2\sigma_1^2,$$
$$\text{Var}(cY + d) = c^2\text{Var}(Y) = c^2\sigma_2^2.$$

Then we solve for the covariance

$$\begin{aligned} Cov(W, Z) &= E[WZ] - E[W]E[Z] \\ &= E[(aX + b)(cY + d)] - [aE[X] + b][cE[Y] + d] \\ &= E[acXY + adX + bcY + bd] - (acE[X]E[Y] + adE[X] + bcE[Y] + bd) \\ &= acE[XY] + adE[X] + bcE[Y] + bd - acE[X]E[Y] - adE[X] - bcE[Y] - bd \\ &= ac\,(E[XY] - E[X]E[Y]) \\ &= ac\,\text{Cov}(X, Y). \end{aligned}$$

Putting it together we get that

$$p_{wz} = \frac{Cov(W, Z)}{\sigma_W\sigma_Z} = \frac{acCov(X, Y)}{a\sigma_1 \cdot c\sigma_2} = \frac{Cov(X, Y)}{\sigma_1\sigma_2} = p_{xy}.$$

□

2.8.15

Let X_1, X_2, and X_3 be random variables with equal variances but with correlation coefficients $\rho_{12} = 0.3$, $\rho_{13} = 0.5$, and $\rho_{23} = 0.2$. Find the correlation coefficient of the linear functions $Y = X_1 + X_2$ and $Z = X_2 + X_3$.

Solution Since all random variables have the same variance we will try to express the covariance and variances of X and Y in terms of the common variance σ^2. A key thing to notice here is that

$$p_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{\sigma^2}\sqrt{\sigma^2}} = \frac{Cov(X_i, X_j)}{\sigma^2}.$$

Solving for the covariance we get that $Cov(X_i, X_j) = p_{ij}\sigma^2$. Then we have that

$$Cov(Y, Z) = \sum_{i=1}^{2}\sum_{j=`}^{2} a_i b_j Cov(X_I, X_j)$$
$$\begin{aligned} &= Cov(X_1, X_2) + Cov(X_2, X_3) + Cov(X_1, X_3) + Cov(X_2, X_2) \\ &= p_{12}\sigma^2 + p_{23}\sigma^2 + p_{13}\sigma^2 + \sigma^2 \\ &= \sigma^2(p_{12} + p_{23} + p_{13} + 1) \\ &= 2\sigma^2. \end{aligned}$$

Next we find the variances of Y and Z.

$$\begin{aligned}\text{Var}(Y) &= \sum_{i=1}^{2} \text{Var}(X_i) + 2 \sum_{1\le i\le j\le 2} Cov(X_i, X_j) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + Cov(X_1, X_2) \\ &= \sigma^2 + \sigma^2 + 2p_{12}\sigma^2 \\ &= \sigma^2(2 + 2p_{12}) \\ &= 2.6\sigma^2,\end{aligned}$$

$$\begin{aligned}\text{Var}(Z) &= \sum_{i=1}^{2} \text{Var}(X_i) + 2 \sum_{1\le i\le j\le 2} Cov(X_i, X_j) \\ &= \text{Var}(X_2) + \text{Var}(X_3) + Cov(X_2, X_3) \\ &= \sigma^2 + \sigma^2 + 2p_{23}\sigma^2 \\ &= \sigma^2(2 + 2p_{23}) \\ &= 2.4\sigma^2.\end{aligned}$$

Putting it together we get

$$\begin{aligned}p &= \frac{Cov(Y, Z)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(X)}} \\ &= \frac{2\sigma^2}{\sqrt{2.6\sigma^2}\sqrt{2.4\sigma^2}} \\ &= \frac{2\sigma^2}{\sigma^2\sqrt{2.4 \cdot 2.6}} \\ &= \frac{2}{\sqrt{6.24}} \\ &\approx 0.801.\end{aligned}$$

□

Tutorial Week 9

Let X_1, X_2, X_3 be independent random variables with the common moment generating function $M_X(t) = e^{2t+\frac{t^2}{2}}$, $t \in \mathbb{R}$. Determine the moment generating function of $T = 2X_1 - X_2 + X_3$.

Solution We will use Theorem 2.6.1. We get that

$$\begin{aligned}M_T(t) &= \prod_{i=1}^{3} M_{X_i}(k_i t) \\ &= \prod_{i=1}^{3} M_X(k_i t) \\ &= e^{4t+\frac{4t^2}{2}} \cdot e^{-2t+\frac{t^2}{2}} \cdot e^{2t+\frac{t^2}{2}} \\ &= e^{4t+3t^2}\end{aligned}$$

for $t \in \mathbb{R}$. □

3 Some Special Distributions

3.1 The Binomial and Related Distributions

Up to now, we've explored probability distributions tailored to specific problems or scenarios. While these custom approaches are valuable, many real-world situations follow well-known patterns captured by standard probability distributions. In this section, we'll introduce some of these widely applicable distributions, starting with the binomial distribution, and see how they help simplify and unify our understanding of probability.

A Bernoulli experiment is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, for instance, success or failure (e.g., female or male, life or death, nondefective or defective). A sequence of Bernoulli trials occurs when a Bernoulli experiment is performed several independent times so that the probability of success, say p, remains the same from trial to trial. That is, in such a sequence, we let p denote the probability of success on each trial. Let X be a random variable associated with a Bernoulli trial by defining it as follows:

$$X(\text{success}) = 1, \qquad X(\text{failure}) = 0$$

That is, the two outcomes, success and failure, are denoted by one and zero, respectively. The pmf of X can be written as

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

We can show that $\mu = E(X) = p$, and $\sigma^2 = \text{Var}(X) = p(1-p)$. Further,

$$M_X(t) = (1-p) + pe^t, \quad t \in \mathbb{R}.$$

We derive these in Theorem 3.1.3. Now suppose we have completed n Bernoulli trials where the probability of success remains constant. Then this sequence of trials is a n-tuple of zeros and ones. In such a sequence of Bernoulli trials, we are often interested in the total number of successes and not in the order of their occurrence. Let X be the total number of successes in n Bernoulli trials. Then the possible values of X are $0, 1, 2, \ldots, n$. If x success occur where $x = 0, 1, 2, \ldots, n$, then $n - x$ failures occur. Then given these n Bernoulli trials trials with x successes, how many different combinations of these n trials are there so that we can have x successes and $n - x$ failures? Well there are

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

Since the trials are independent and the probabilities of success and failure on each trial are, respectively, p and $1 - p$, the probability of each of these ways is $p^x(1-p)^{n-x}$. Thus the pmf of X is the sum of the probabilities of these $\binom{n}{x}$ mutually exclusive events; that is,

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{elsewhere.} \end{cases}$$

It is clear that $p(x) \geq 0$. To verify that $p(x)$ sums to 1 over its range, recall the binomial series, expression (1.3.7) of Chapter 1, which is:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} b^x a^{n-x},$$

for n a positive integer. Thus,

$$\sum_x p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = [(1-p)+p]^n = 1.$$

Thus $p(x)$ satisfies the condition of being a valid pmf.

Definition 3.1.1

A random variable X is said to have a binomial distribution based on n trials with success probability p if and only if

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{elsewhere.} \end{cases}$$

for $x = 0, 1, 2, \ldots, n$ and $0 < p < 1$.

A binomial distribution is denoted by the symbol $Bin(n, p)$. The constants n and p are called the parameters of the binomial distribution.

Example 3.1.2

Suppose we roll a fair six-sided die 3 times. What is the probability of observing exactly 2 sixes?

Solution First I am going to answer this using what we learned in Chapter 1. We know that there $\binom{3}{2}$ ways we can have two sixes from three dice rolls. Moreover There is a $1/6$ chance of rolling any number. So we have a $(1/6)^2$ chance of rolling two sixes and then $(5/6) = 1 - 1/6$ chance that we roll any other number other than a six for the last roll. Thus the probability is $\binom{3}{2}(1/6)^2(1-1/6) = 0.0694$.

However this is a binomial distribution. Namely we have 3 independent trials and need 2 successes (where a success is rolling a six). Let X be the random variable of total number of sixes from 3 rolls of the dice. So we get that

$$p(2) = \binom{3}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{3-2} = 0.0694.$$

□

Theorem 3.1.3

Let X be a random variable with a binomial distribution. The mgf of X is

$$M(t) = [(1-p) + pe^t]^n.$$

Also we have that

$$\mu = np \quad \text{and} \quad \text{Var}(X) = np(1-p).$$

Proof. We see that

$$\begin{aligned} M(t) &= \sum_x e^{tx} p(x) \\ &= \sum_x^n e^{tx} p^x (1-p)^{n-x} \\ &= \sum_x^n \binom{n}{x} (pe^t)^X (1-p)^{n-x} \\ &= [(1-p) + pe^t]^n. \end{aligned}$$

We then see the mean and variance can be computed from the mfg. We see that

$$M'(t) = n[(1-p) + pe^t]^{n-1}(pe^t).$$

we also see that

$$M''(t) = n[(1-p) + pe^t]^{n-1}(pe^t) + n(n-1)[(1-p) + pe^t]^{n-2}(pe^t)^2.$$

It follows that

$$\mu = M'(0) = np$$

and

$$\sigma^2 = M''(0) - \mu^2 = np(1-p).$$

□

Example 3.1.4

Exxon has just bought a large tract of land in northern Quebec, with the hope of finding oil. Suppose they think that the probability that a test hole will result in oil is 0.2. Assume that Exxon decides to drill 7 test holes. What is the probability that

1. Exactly 3 of the test holes will strike oil?
2. At most 2 of the test holes will strike oil?
3. Between 3 and 5 (including 3 and 5) of the test holes will strike oil?
4. What are the mean and standard deviation of the number of test holes which strike oil?

Solution We are given that $p = 0.2$. Since this is a Bernoulli experiment we then know that the distribution of the random variable X which refers to the number of holes that strike oil from the 7 drilled test holes will be

$$p(x) = \binom{7}{x} (0.2)^x (0.8)^{7-x}.$$

For 1: We see that the probability that there are exactly 3 test holes that strike oil and 4 test holes that do not is

$$p(3) = \binom{7}{3} (0.2)^3 (0.8)^{7-3} = 0.1147.$$

Binomial Probabilities

Tabulated values are $P(Y \le a) = \sum_{y=0}^{a} p(y)$. (Computations are rounded at third decimal place.)

(a) $n = 5$

	p												
a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	a
0	.951	.774	.590	.328	.168	.078	.031	.010	.002	.000	.000	.000	.000	0
1	.999	.977	.919	.737	.528	.337	.188	.087	.031	.007	.000	.000	.000	1
2	1.000	.999	.991	.942	.837	.683	.500	.317	.163	.058	.009	.001	.000	2
3	1.000	1.000	1.000	.993	.969	.913	.812	.663	.472	.263	.081	.023	.001	3
4	1.000	1.000	1.000	1.000	.998	.990	.969	.922	.832	.672	.410	.226	.049	4

(b) $n = 10$

	p												
a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	a
0	.904	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000	.000	0
1	.996	.914	.736	.376	.149	.046	.011	.002	.000	.000	.000	.000	.000	1
2	1.000	.988	.930	.678	.383	.167	.055	.012	.002	.000	.000	.000	.000	2
3	1.000	.999	.987	.879	.650	.382	.172	.055	.011	.001	.000	.000	.000	3
4	1.000	1.000	.998	.967	.850	.633	.377	.166	.047	.006	.000	.000	.000	4
5	1.000	1.000	1.000	.994	.953	.834	.623	.367	.150	.033	.002	.000	.000	5
6	1.000	1.000	1.000	.999	.989	.945	.828	.618	.350	.121	.013	.001	.000	6
7	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.322	.070	.012	.000	7
8	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.624	.264	.086	.004	8
9	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.893	.651	.401	.096	9

Figure 10: This table shows cumulative binomial probabilities for different values of n, p and a.

For 2: The probability at most 2 of the test holes will strike oil refers to $p(0) + p(1) + p(2)$ which gives us

$$P(X \le 2) = p(0)+p(1)+p(2) = \binom{7}{0}(0.2)^0(0.8)^{7-0}+\binom{7}{1}(0.2)^1(0.8)^{7-1}+\binom{7}{2}(0.2)^2(0.8)^{7-2} = 0.8520.$$

For 3: We see that the probability that there either 3,4,5 holes strike out is

$$P(3 \le 5 \le 5) = p(3) + p(4) + p(5) = 0.1147 + \binom{7}{4}(0.2)^4(0.8)^{7-4} + \binom{7}{5}(0.2)^5(0.8)^{7-5} = 0.1477.$$

For 4: We know that the mean is np. That is $\mu = 7 \times 0.2 = 1.4$. The standard deviation is going to be $\sqrt{np(1-p)} = \sqrt{1.12} \approx 1.0583$. □

Example 3.1.5

Data show that 30% of people recover from a particular illness without treatment. Ten ill people are selected at random to receive a new treatment and 9 recover after the treatment. Assuming the medication is worthless (i.e., has no effect on the illness). What is the probability that prior to treatment we would have expected to observe that at least 9 of the 10 treated people recover?

Solution We are given that $p = 0.3$ is the probability that people recover from a particular illness without treatment. We are then asked given $n = 10$, what is the probability that at least 9 people recover. That is we need to find $P(9 \ge X) = P(X = 9) + P(X = 10)$. We get that

$$P(9 \ge X) = \binom{10}{9}(0.3)^9(1-0.3)^1 + \binom{10}{10}(0.3)^{10}(1-0.3)^0 \approx 0.0001437.$$

□

Before we move onto the next distribution take a look at Figure 10. This shows us a table of the binomial distribution for different values of n and p. Although this isn't super necessary since computation of binomial distributions are quite simple.

Imagine (like with the binomial) you have a situation with a sequence of independent trials, where there are only two outcomes (success and failure), and the trials have probability of success

p. Now, if we let X be the number of failures until the first success, then X has a geometric distribution. Note that we do not fix the number of trials in advance. For example we repeatedly flip a coin. Then the number of tails (failure) before the first heads has a geometric distribution.

Definition 3.1.6

A random variable X is said to have a geometric distribution with success probability p if and only if

$$p_X(x) = p(1-p)^x$$

for $x = 0, 1, 2 \ldots$ and $0 < p < 1$.

Example 3.1.7

Assume the probability that an engine fails during any one-hour period is constant at $p = 0.02$. Find the probability that the engine continues to operate for more than two hours.

Solution Let X be the number of hours till the first failure. We want the probability the engine operates for more than 2 hours. This means it does not fail in the first two 1-hour periods. That is $P(X \geq 2) = 1 - (p(0) + p(1))$. We get that

$$P(X \geq 2) = 1 - p(0) - p(1) = 1 - 0.02 - (0.02 \cdot 0.98).$$

We used the fact that $P(X = 1) = (1-p)p$ which is the probability that there is not a failure in the first hour □

Let X be a random variable that has a geometric distribution with success probability p. Then

$$\mu = E(X) = \frac{1-p}{p}$$

and

$$\sigma^2 = \text{Var}(X) = \frac{1-p}{p^2}$$

and

$$M_X(t) = \frac{p}{1-(1-p)e^t}, \quad \text{for } t < -\log(1-p).$$

3.1.1 Negative Binomial Distribution

Assume you have a sequence of independent Bernoulli trials with only two outcomes (success and failure) and constant probability of success p. Let the random variable X denote the total number of failures in this sequence before the r-th success, that is, $X + r$ is equal to the number of trials necessary to produce exactly r successes with the last trial as a success. Here r is a fixed positive integer. To find the pmf assume that $x \in S_X$ is an element in the support of X. Since these trials are independent then $P(X = x)$ is equal to the product of the probability of obtaining x failures $(1-p)$ multiplied by the probability of obtaining $r-1$ successes on $y+r-1$ trials because the $y+r$-th trial is the r-th success. Since the number of ways of having $r-1$ successes on the $y+r-1$ trials is $\binom{y+r-1}{r-1}$ we have that

$$p(x) = \binom{y+r-1}{r-1} p^r (1-p)^x.$$

Example 3.1.8

A geological study indicates that wells drilled in a particular region should strike oil with probability 20%. Find the probability that the third oil strike comes on the fifth well drilled.

Solution We are given a probability of success $p = 0.2$. Let X be the random variable of failures until we reach 3 oil strikes. Since we want 3 wells drilled to strike oil on the fifth well drilled, this means that $x + 3 = 5$ which means we have to find $p(2)$. So we get that

$$P(X = 2) = p(2) = \binom{2+3-1}{3-1}(0.2)^3(0.8)^2 = 0.0307.$$

□

Let X be a random variable that has a negative binomial distribution with parameters r and p. Then

$$\mu = E[X] = \frac{r(1-p)}{p}$$

and

$$\sigma^2 = \text{Var}(X) = \frac{r(1-p)}{p^2}$$

and

$$M_X(t) = \left[\frac{p}{1-(1-p)e^t}\right]^r, \quad \text{for } t < -\log(1-p).$$

3.1.2 Hypergeometric Distribution

Assume you have a finite population N and you are selecting a sample size of n without replacement. There are $\binom{N}{n}$ ways to choose a sample of size n out of a population of size N. Since all samples are equally likely then $P(E_i) = 1/\binom{N}{n}$. Now suppose D of the N elements are "red". There are $\binom{D}{x}$ ways to choose x reds. Then the there are remaining $\binom{N-d}{n-x}$ ways to choose $n - x$ non-reds from the remaining $N - D$ non-reds. Putting it all together, the probability that you get exactly x red elements in your sample of size n

$$p(x) = \frac{\binom{D}{x}\binom{N-d}{n-x}}{\binom{N}{n}}.$$

Definition 3.1.9

A random variable X has a hypergeometric distribution if and only if

$$p(x) = \frac{\binom{D}{x}\binom{N-d}{n-x}}{\binom{N}{n}}.$$

where x is an integer equaling either $0, 1, \ldots, n$ if $D \geq n$ or $0, 1, \ldots, D$ if $D < n$ with restrictions $n \leq N$, $D \leq N$ and $x \leq \min(n, D)$.

Example 3.1.10

A warehouse contains ten printing machines, four of which are defective. A company selects five of the machines at random, thinking all are in working condition. What is the probability that all five of the machines are nondefective?

Solution Here our population is $N = 10$ and we are selecting without replacement $n = 5$. Also here $D = 6$ are the machines that are nondefective which means $N - D = 4$ are defective. Thus we get that

$$p(5) = \frac{\binom{6}{5}\binom{4}{0}}{\binom{10}{5}} = \frac{1}{42}.$$

□

If X is a random variable with a hypergeometric distribution, then

$$E(X) = \frac{nD}{N}$$

and

$$\text{Var}(X) = n\left(\frac{D}{N}\right)\left(\frac{D-N}{N}\right)\left(\frac{N-n}{N-1}\right).$$

3.2 The Poisson Distribution

The Poisson distribution is used to model the number of rare events that occur in a fixed interval of time or space. When these events happen they are independent of each other, the average rate of events is constant and two events can't happen at the exact same instant. For example we can use this to model the number of emails arriving per hour, number of cars passing through a toll booth per minute , number of phone calls to a call center per hour, number of typos per page in a book, or number of earthquakes per year and etc. Notice that the intervals do not have to be in terms of time. This distribution is also related to the binomial distribution which we show later.

Defintion 3.2.1

A random variable X has a Poisson distribution if and only if

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k = 0, 1, 2, \ldots$ and $\lambda > 0$. Here λ is sometimes called the "rate" since it is the average number of events that occur in an interval.

To see that this is a valid distribution we see that since $\lambda > 0$ then $p(x) \geq 0$ and that

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

We not find the MGF of the Poisson distribution.

$$\begin{aligned} M(t) = \sum_{x=0}^{\infty} e^{tx} p(x) &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

for all real values of t. We find the first derivative:

$$M'(t) = e^{\lambda(e^t-1)}(\lambda e^t).$$

and the second derivative:

$$M''(t) = e^{\lambda(e^t-1)}(\lambda e^t)(\lambda e^t) + e^{\lambda(e^t-1)}(\lambda e^t).$$

We then find that the mean is

$$\mu = M'(0) = \lambda$$

and then

$$\sigma^2 = M''(0) - \mu^2 = \lambda.$$

That is the Poisson distribution has the same mean and variance.

Example 3.2.2

The number of typing errors made by a typist has a Poisson distribution with an average of four errors per page. If more than four errors appear on a given page, the typist must retype the whole page. What is the probability that a randomly selected page does not need to be retyped?

Solution We are given that the average number of errors on a page is 4. That is $\lambda = 4$. Our interval here is a single page. Thus we can let X be the number of errors on a single page by the typist and we can use the Poisson distribution. We need to find the probability that a random page does not have more than four errors on a page. That is we need to find $P(X \leq 4)$. We find that

$$P(X \leq 4) = p(0) + p(1) + p(2) + p(3) + p(4) = \frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{2!} + \frac{e^{-4}4^3}{3!} + \frac{e^{-4}4^4}{4!} = 0.6288.$$

□

We now talk about the differences between Binomial and Poisson distribution. For binomial there are a fixed number of trials whereas for Poisson we are counting the total number of events over an interval. For example the number of operational units out of n sample for binomial distribution whereas Number of defectives per unit (where unit could be a batch).

Proposition 3.2.3

Let $\lambda > 0$ and X be a random variable with Poisson distribution with rate λ. Then

$$p(x) = \lim_{n\to\infty} Bin\left(n, p = \frac{\lambda}{n}\right).$$

Proof. The Poisson distribution is the limiting form of the Binomial distribution. To make this explicit, as $n \to \infty$, if we set $\lambda = np$, then

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} &= \lim_{n\to\infty} \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{n}\right)^n \left(\frac{1-\lambda/n}{1}\right)^{-x} \frac{n(n-1)\dots(n-x+1)}{n^x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{x-1}{n}\right)
\end{aligned}$$

Now, note that

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

and all other terms to the right of it have a limit of 1.

Thus

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

□

Next I will talk about Poisson processes. A Poisson process is one of the most fundamental stochastic processes in probability theory. It models the occurrence of random events over continuous time or space, under very natural assumptions. A Poisson process with rate $\lambda > 0$ is a collection of random variables $\{X_t : t \geq 0\}$, where X_t counts the number of events that have occurred in the interval $(0, t]$. The defining characteristics of the Poisson process are: (1) $X_0 = 0$; (2) the process has independent increments meaning that the number of events occurring in non-overlapping intervals of time are independent random variables; (3) the process has stationary increments. That is the distribution of the number of events in any interval depends only on the length of the interval, not its location on the time axis; specifically, for any $t > 0$, $X_t \sim$ Poisson(λt), that is, $P(X_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$; and (4) the probability of more than one event occurring in an infinitesimally small time interval is negligible (formally, the probability of one event in an interval of length h is $\lambda h + o(h)$, while the probability of two or more events is $o(h)$). An important consequence of these assumptions is that the inter-arrival times between successive events are independent and identically distributed exponential random variables with parameter λ, i.e., the waiting times between events follow an exponential distribution with mean $1/\lambda$. The Poisson process provides an elegant model for many real world phenomena involving randomly timed events, such as radioactive decay, arrival of phone calls at a call center, photons hitting a detector, failure times of mechanical systems, or customers entering a store. Moreover, the Poisson process can be generalized in various ways, such as to non homogeneous Poisson processes where the rate $\lambda(t)$ varies with time, or to compound Poisson processes where each event may carry a random "weight" or size. Overall, the Poisson process forms the basis for the study of continuous-time Markov processes and many other areas in statistics and probability theory which is taught in future courses.

3.3 The Γ, χ^2, β and Uniform Distributions

The previous distributions are used to describe discrete random variables. We now consider some standard distributions for continuous random variables.

3.3.1 The Γ Distribution

We begin with the Γ (pronounced "gamma") distribution. The Γ-distribution is a continuous probability distribution that models waiting times or total times for multiple random events to occur for example lifetimes, failure times, or service times. The support of the distribution is the set of positive reals. Before we define the Γ-distribution, we recall the Γ-function

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

for all $\alpha > 0$. Then we see that if $\alpha = 1$ then

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

If $\alpha > 1$ then we see that through integration by parts

$$\Gamma(\alpha) = (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy = (\alpha-1)\Gamma(\alpha-1)$$

or if α is a positive integer greater than 1 then

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

Definition 3.3.1

A random variable X has a Gamma distribution with parameters $\alpha > 0$ (shape parameter) and $\lambda > 0$ (scale parameter) if and only if the pdf of X is

$$f(x) = \frac{x^{\alpha-1}e^{-x/\lambda}}{\lambda^\alpha\Gamma(\alpha)}$$

for $0 < x < \infty$ where the Gamma function is defined as $\Gamma(\alpha) = (\alpha-1)!$.

Here α is the shape parameter which controls how many events occur and λ is the scale parameter which controls how fast the events occur. We not find the mean of the $\Gamma-$distribution:

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot \frac{x^{\alpha-1}e^{-x/\lambda}}{\lambda^\alpha\Gamma(\alpha)}dx \\ &= \frac{1}{\lambda^\alpha\Gamma(\alpha)}\int_0^\infty x^\alpha e^{-x/\lambda} \end{aligned}$$

We do a change-of-variable by letting $u = x/\lambda$ then $x = \lambda u$ and $dx = \lambda du$. We then get that

$$\begin{aligned}\frac{1}{\lambda^\alpha \Gamma(\alpha)} \int_0^\infty x^\alpha e^{-x/\lambda} &= \frac{1}{\lambda^\alpha \Gamma(\alpha)} \int_0^\infty (\lambda u)^\alpha e^{-u} \lambda du \\ &= \frac{1}{\lambda^\alpha \Gamma(\alpha)} \lambda^{\alpha+1} \int_0^\infty u^\alpha e^{-u} du \\ &= \frac{\lambda}{\Gamma(\alpha)} \cdot \Gamma(\alpha+1) \\ &= \frac{\lambda}{\Gamma(\alpha)} \cdot \alpha\Gamma(\alpha) = \lambda\alpha.\end{aligned}$$

We next find $E(X^2)$ similarly following pretty much the same process with the same change-of-variable and get that $\text{Var}(X) = \alpha\lambda^2$. Moreover we find that the MGF is

$$M(t) = \frac{1}{(1-\lambda t)^\alpha}$$

for $t < 1/\lambda$.

Next we show that the Gamma distribution arises naturally from the Poisson process. For $t > 0$, let X_t denote the number of events of interest that occur in the interval $(0, t]$. Assume X_t satisfies the three assumptions of a Poisson process. Let k be a fixed positive integer and define the continuous random variable W_k to be the waiting time until the kth event occurs. Then the range of W_k is $(0, \infty)$. We now ask what is the distribution of W_k What is the distribution of W_k? In other words: If events happen at rate λ, how long do you have to wait to see k events?

Note that for $w > 0$, $W_k > w$ if and only if $X_w \leq k-1$. That is the time to wait for k events to occur can only be greater than w if the number of events that happen in the interval $(0, w]$ is less than or equal to k. Hence,

$$P(W_k > w) = P(X_w \leq k-1) = \sum_{x=0}^{k-1} P(X_w = x) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}.$$

This is the tail probability. This tells us how likely it is that the k-th event has not yet occurred by time w. Now it can be shown that

$$F_{Wk}(x) = 1 - P(W_k > x)$$

which we then can further show that (Problem 3.3.5)

$$\int_{\lambda w}^\infty \frac{z^{k-1} e^{-z}}{(k-1)!} dz = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!},$$

and for $w \leq 0$, $F_{W_k}(w) = 0$. We can conclude that in a Poisson process with rate λ the time to the k-th event, W_k, follows a Gamma distribution Gamma$(\lambda, 1/\lambda)$. This leads us to the next distribution. So far we have concluded that given a Poisson process, the waiting time until the k-th event occurs follows a gamma distribution however what we did was we summed up $k-1$ exponential distributions. That is the exponential distribution models the amount of time you wait until the next random event occurs, assuming the events are happening at a constant average rate.

Definition 3.3.2

A random variable X has an exponential distribution with parameter $\lambda > 0$ if and only if the pdf of X is

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}$$

for $0 \leq x < \infty$ where λ is the rate parameter − average number of events per unit time.

Notice that this is the Gamma distribution with parameters Gamma$(1, \lambda)$. Like above it is easy to show that $E(X) = \lambda$ and $\text{Var}(X) = \lambda^2$. Further we see that the MGF of an exponential distribution is

$$M(t) = \frac{1}{1 - \lambda t}$$

for $t < 1/\lambda$. Given the simplicity of the distribution the CDF has a closed-form

$$F_X(x) = P(X \leq x) = \int_0^x \frac{1}{\lambda} e^{-x/\lambda} = 1 - e^{-x/\lambda}$$

for $x > 0$. So in a Poisson process with rate λ: the number of events in time t is Poisson(λ). The time between successive events is Exponential(λ). Furthermore it can be shown that if $X_1, X_2, \ldots, X_n$ are independent random variables with distributions $\Gamma(\alpha_i, \beta)$ for all i and $Y = \sum_{i=1}^n X_i$ then Y has $\Gamma(\sum_{i=1}^n \alpha_i, \beta)$. We now discuss an important property of this distribution which is memoryless (Exercise 3.3.25).

If X has an exponential distribution, with parameter $\lambda > 0$, then

$$P(X > a + b | X > b) = P(X > a).$$

Meaning if you have already waited b units of time, the probability that you will need to wait another a units is exactly the same as if you had just started waiting from scratch. For example, suppose a certain type of light bulb has a lifetime that is exponentially distributed

$$X \sim \text{Exponential}(\lambda = 0.1)$$

That means on average, bulbs last $1/\lambda = 10$ hours. Now suppose you install a new bulb. It has already been working for 5 hours (without failing). What is the probability that the bulb lasts at least 3 more hours? The answer is

$$P(X > 5 + 3 \mid X > 5) = P(X > 3)$$

Even though the bulb has already been on for 5 hours, it "forgets" its past — the chance of lasting another 3 hours is just:

$$P(X > 3) = e^{-\lambda \cdot 3} = e^{-0.1 \cdot 3} \approx 0.7408$$

Example 3.3.3

The length of time between arrivals at a hospital clinic has an approximately exponential probability distribution. Suppose the mean time between arrivals for patients at a clinic is 4 minutes. What is the probability that a particular inter-arrival time (the time between the arrival of two patients) is less than 1 minute? What is the probability that the next four interarrival times are all less than 1 minute? What is the probability that an interarrival time will exceed 10 minutes?

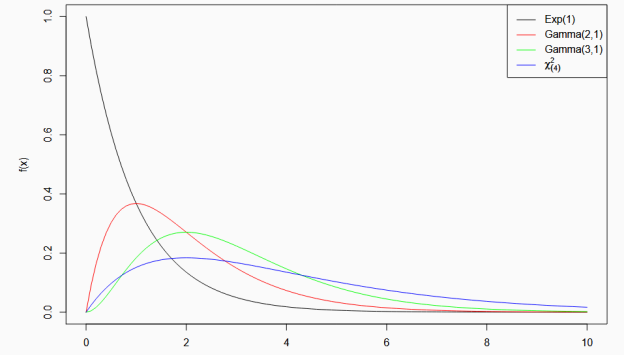


Figure 11: Gamma, exponential and Chi-squared distributions graphed

Solution We are given that the length of time between arrivals at a hospital clinic follows a exponential distribution with parameter rate $\lambda = 4$. We then are asked to find the probability that $P(X \leq 1) = 1 - e^{-1/4} = 0.221$.

Since the interarrival times are independent, we have

$$[P(X \leq 1)]^4 = 0.00239.$$

Then finnaly we find that

$$P(X > 10) = 1 - F(10) = 1 - [1 - e^{-10/4}] = 0.0821.$$

□

3.3.2 The χ^2 Distribution

We now discuss another special case of the Gamma distribution.

Definition 3.3.4

A random variable X has a Chi-squared distribution with v degrees of freedom (χ^2_v) if and only if X is a random variable with a Gamma distribution with parameters $\alpha = v/2$ and $\beta = 2$. The mean of the Chi-squared distribution is $\mu = \alpha\beta = 2 \cdot v/2 = v$ and the variance is $\sigma^2 = \alpha\beta^2 = 2v$. We also see that

$$M(t) = \frac{1}{(1-2t)^{v/2}}$$

for $t < 1/2$.

For more a more in depth history, derivation, mean and variance, and visualization you can watch this video. I recommend watching this video after learning Normal distributions however for this course this is enough. Also see Figure 11 for the shapes of various distributions.

3.3.3 The β Distribution

We now consider distributions whose supports are a bounded positive set of $\mathbb{R}$. The β-distribution is a continuous probability distribution on the interval $(0, 1)$

Definition 3.3.5

A random variable X has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if and only if the pdf of X is

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

for $0 < x < 1$ where

$$B(\alpha,\beta) = \int_0^\infty y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b).}$$

The Beta distribution models uncertainty about a probability. For example suppose you are learning about the true probability that a machine works (say p) — but you don't know what p is. You can model your uncertainty about p using a Beta distribution. We also see that if X has a beta distribution then $E(X) = \alpha/(\alpha+b)$ and

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

Note that the MGF does not exist in a closed form. Also note that the textbook goes more in depth into the derivation of the beta distribution which in my opinion is laborious and of no use for this course.

Example 3.3.6

Errors in measuring the time of arrival of a wave front from an acoustic source sometimes have an approximate beta distribution. Suppose that these errors, measured in microseconds, have approximately a beta distribution with $\alpha = 1$ and $\beta = 2$. What is the probability that the measurement error in a randomly selected instance is less that 0.5? Find the mean and standard deviation of the measurement errors.

Solution We are given that the error measured in microseconds follows a beta distribution with parameters $\alpha = 1$ and $\beta = 2$. We are asked to find $P(X \leq 0.5)$. That is we need to find

$$\begin{aligned} P(X \leq 0.5) &= \int_0^{0.5} f(x)dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{0.5} x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{(1+2-1)!}{(1-1)!(2-1)!} \int_0^{0.5} x^{1-1}(1-x)^{2-1}dx \\ &= 0.75. \end{aligned}$$

We then see that $\mu = 1/(1+2) = 1/3$ and

$$\sigma^2 = \frac{1 \cdot 2}{(1+2)^2(1+2+1)} = \frac{1}{18}$$

and so $\sigma = 1/\sqrt{18}$. □

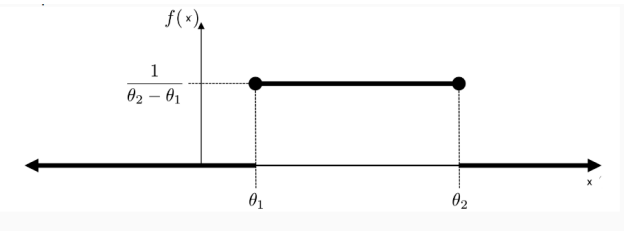


Figure 12: Graph depicting a uniform distribution.

3.3.4 Uniform Distribution

The continuous uniform distribution models a random variable that is equally likely to take any value in a given interval $[a, b]$ see Figure 12.

Definition 3.3.7

If $\theta_1 < \theta_2$, a random variable X has a Uniform distribution on the interval $[\theta_1, \theta_2]$, often denoted $U(\theta_1, \theta_2)$ if and only if the pdf of X is

$$f(x) = \frac{1}{\theta_2 - \theta_1}$$

for $\theta_1 \leq x \leq \theta_2$.

Note that if X has a uniform distribution on $[\theta_1, \theta_2]$ then the CDF of X is

$$F_X(x) = \int_{\theta_1}^{x} \frac{1}{\theta_2 - \theta_1} dt = \frac{x - \theta_1}{\theta_2 - \theta_1}$$

for $\theta_1 \leq x \leq \theta_2$ and zero elsewhere. Then we see that the mean would be $E(X) = (\theta_1 + \theta_2)/2$ and the variance would be $\text{Var}(X) = (\theta_2 - \theta_1)^2/12$. WE also see that the MGF is

$$M_X(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$$

for $t \neq 0$ and $M_X(t) = 1$ for $t = 0$. Like we said above the Uniform distribution arises when the probability of an event occurring in some fixed interval is the same over all possible intervals of that size. For example imagine the example of bus arrival, where the bus is just as likely to arrive in any two-minute interval over a 10-minute window. In experiments and surveys where you might need to select a random sample, the easiest way is to assign each entity a computer-generated uniformly distributed random number. Some continuous random variables in physical, biological, and other sciences have a uniform distribution. If the number of arrivals into some system has a Poisson distribution and we are told that exactly one event happened in the interval $(0, t)$ then the time of occurrence of the event is uniformly distributed on $(0, t)$. The Poisson process has independent increments so the event could have occurred anywhere in $(0, t)$. There is no "preference". That is no time is more likely than any other. Given that there was exactly 1 event, it could have landed uniformly anywhere

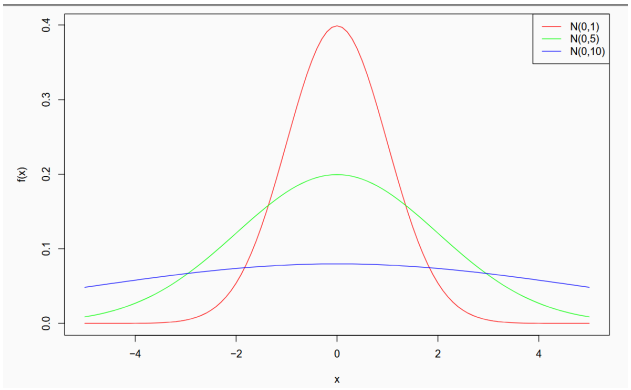


Figure 13: The Normal Distribution with mean zero and different variances.

Example 3.3.8

Delta Airlines quotes a flight time of 2 hours, 5 minutes for its flights from Cincinnati to Tampa. Suppose we believe that actual flight times are uniformly distributed between 2 hours and 2 hours, 20 minutes. What is the probability that the flight will be no more than 5 minutes late? What is the probability that the flight will be more than 10 minutes late? What is the expected flight time?

Solution We are given that in the interval $(120, 140)$ the flight times follows a uniform distribution. We are asked to find the probability that it is no more than 5 minutes late from the 2 hours and 5 minute quote they gave. That is we need to find

$$P(125 < X < 130) = \int_{125}^{130} \frac{1}{140 - 120} dt = \frac{1}{4}.$$

We are now asked to find the probability will be more than 10 minutes late. That is we need to find

$$P(X > 135) = \int_{135}^{140} \frac{1}{140 - 120} dt = \frac{1}{4}.$$

Finally we see that

$$E(X) = \frac{120 + 140}{2} = 130.$$

□

3.4 The Normal Distribution

The normal distribution (also called the Gaussian distribution) describes a continuous random variable whose values tend to cluster around the mean, with fewer and fewer values farther from the mean in a perfectly symmetric bell-shaped curve. The main motivation for the normal distribution is found in the Central Limit Theorem which is discussed later but in short: if you add up lots of small, independent random effects, their total becomes approximately normal, no matter the shape of the original distribution.

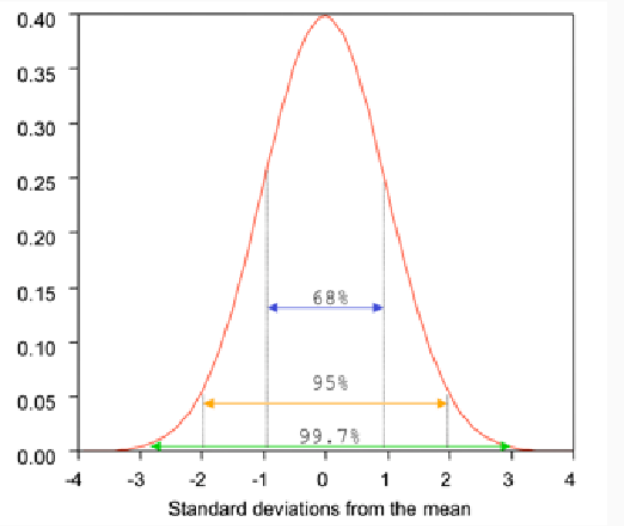


Figure 14: Graph of the Normal Distribution showing the standard deviations from the mean

Definition 3.4.1

A random variable X is said to have a Normal distribution if and only if for σ^2 and $-\infty < \mu < \infty$, the pdf of X is

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

for $-\infty < x < \infty$.

The parameters for the normal distribution are μ (location parameter) and σ (scale parameter) and it is denoted by $N(\mu, \sigma^2)$. To make sense of this I encourage you to play around with the distribution on desmos and look at Figure 13. If X is a random variable with a normal distribution then $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. The MGF is also

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

for $t \in \mathbb{R}$. We use Z to represent a random variable from a standard normal distribution with $\mu = 0$ and $\sigma = 1$. That is $Z \sim N(0, 1)$. If Z has a standard normal distribution, the CDF, $F_Z(z) = P(Z \leq z)$, is often denoted by $\Phi(z)$. Because the normal is symmetric for Z, for $\Phi(0) = 0.5$ and $\Phi(z) = 1 - \Phi(-z)$. Also note that $P(Z > z) = 1 - \Phi(z)$ and $P(z \leq Z \leq b) = \Phi(b) - \Phi(a)$. With this we can standardize any general normal distribution $N(\mu, \sigma)$ into a standard normal distribution $N(0, 1)$. We do this by doing Z-score normalization which some of you may know from ML. That is we let $Z = \frac{X-\mu}{\sigma}$. We get that

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(\frac{a-\mu}{\sigma} \leq Z\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

The empirical rule is if a dataset follows a Normal Distribution then (see Figure 14)

- 68% of the probability is within 1 standard deviation of the mean.
- 95% of the probability is within 2 standard deviation of the mean.
- 99.7% of the probability is within 3 standard deviation of the mean.

The pdf of a Normal Distribution is not directly integrable so we must use numerical integration. What we do instead is use a table for the standard normal distribution. So, either standardize then look up or think about the problem in terms of standard deviations from the mean. You can find this table by a simple google search or click here.

Example 3.4.2

A protein naturally produced in a rare tropical fruit can convert a sour taste into a sweet taste. Consequently, miraculin has the potential to be an alternative low-calorie sweetener. In Plant Science, a group of Japanese environmental scientists investigated the ability of a hybrid tomato plant to produce miraculin. For a particular generation of the tomato plant, the amount of miraculin produced (X, measured in micrograms per gram of fresh weight) had a mean of 105.3 and a standard deviation of 8.0. Assume that X is normally distributed. Find $P(X > 120)$. Find $P(100 < X < 110)$. Find the values of a for which $P(X < a) = 0.25$.

Solution Since this is not a standard normal distribution what we can do instead is standardize it by letting $Z = \frac{X-\mu}{\sigma} = \frac{X-105.3}{8}$. We get that then

$$P(X > 120) = P\left(Z > \frac{120 - 105.3}{8}\right) = 1 - P\left(Z < \frac{120 - 105.3}{8}\right) = 1 - \Phi(1.838) = 1 - 0.96638 = 0.033.$$

Similarly we get that

$$\begin{aligned} P(100 < X < 110) &= P\left(\frac{100 - 105.3}{8} < Z < \frac{110 - 105.3}{8}\right) \\ &= P(-0.663 < Z < 0.587) \\ &= \Phi(0.587) - \Phi(-0.663) \\ &= 0.71904 - 0.25463 = 0.468. \end{aligned}$$

Finally we see that

$$P(X < a) = P\left(Z < \frac{a - 105.3}{8}\right).$$

Since this is a continuous one-to-one distribution we can solve for the value of a by finding the probability 0.25 on the table. We see that -0.67449 corresponds to 0.25 and thus $\frac{a-105.3}{8} = -0.67449$ or $a = 99.904$. □

We end this section with two theorems about Normal Distributions. The first is if you standardize a normal random variable (center it by subtracting the mean and scale it by dividing by standard deviation), and then square it, the resulting variable follows a chi-squared distribution with 1 degree of freedom. This is the building block of the chi-squared distribution. If you do this with n independent standard normal variables and add up the results, you get χ^2_n. This leads to hypothesis testing and confidence intervals which is taught in later courses.

Theorem 3.4.3

If a random variable X is $N(\mu, \sigma)$ then the random variable $V = \left(\frac{X-\mu}{\sigma}\right)^2$ has a X_1^2 distribution.

The next theorem says that any linear combination of independent normal random variables is again normal. The mean of Y is the linear combination of the means. The variance of Y is the sum of the squared coefficients times the variances.

Theorem 3.4.4

Let $X_1, \ldots, X_n$ be independent random variables such that for $i = 1, \ldots, n$, X_i has a $N(\mu_i, \sigma_i^2)$ distribution. Let $Y = \sum_{i=1}^n a_i X_i$, where $a_1, \ldots, a_n$ are constants. Then the distribution of Y is

$$N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

In particular, if $Y = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ where $X_1, X_2, \ldots, X_n$ are iid with $N(\mu, \sigma^2)$ then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

3.5 Practice Problems

3.5.1 Section 3.1 Answers

3.1.4

Let the independent random variables $X_1, X_2, \ldots, X_{40}$ be iid with the common pdf $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere. Find the probability that at least 35 of the X_i's exceed $\frac{1}{2}$.

Solution We are asked to find the probability that 35 of the 40 random values are bigger than 0.5. So what we can do instead is define new random variables $Y_i = 1$ if $X_i > 0.5$ and $Y_i = 0$ otherwise. Then we let $S = Y_1 + Y_2 + \cdots + Y_{40}$ gives us the total number of random variables greater than 0.5. So now we find $P(S \geq 35)$. However what is the probability that a single X_i exceeds 0.5? This depends on the distribution $f(x) = 3x^2$ so we find that

$$p = P(X \geq 0.5) = \int_{0.5}^1 3x^2 dx = 0.875.$$

That gives us the success probability pp for a single trial. Now since the probability of a success (0.875) for one of the X_i having a probability of 0.5 or greater is constant for all trials our random variable $S \sim Bin(40, 0.875)$. We get that

$$P(S \geq 35) = 1 - P(S \leq 34) = 1 - \sum_{n=1}^{34} \text{nCr}\,(40, n)\,(0.875)^n (1 - 0.875)^{(40-n)} = 0.6162.$$

□

3.1.6

Let Y be the number of successes throughout n independent repetitions of a random experiment with probability of success $p = 1/4$. Determine the smallest value of n so that $P(1 \leq Y) \geq 0.70$.

Solution Here Y is a random variable with binomial distribution $Y \sim Bin(n, p = 1/4)$. We need to find $P(1 \leq Y) = 1 - P(Y < 1) \geq 0.70$ or equivalently $P(Y < 1) \leq 0.3$. This simplifies to

$$P(Y < 1) = \binom{n}{0}(1/4)^0(3/4)^n = (3/4)^n \leq 0.3.$$

Solving this we get that $n = 5$ is the smallest number of repetitons. □

3.1.7

Let the independent random variables X_1 and X_2 have binomial distribution with parameters $n_1 = 3$, $p = \frac{2}{3}$ and $n_2 = 4$, $p = \frac{1}{2}$, respectively. Compute $P(X_1 = X_2)$. *Hint:* List the four mutually exclusive ways that $X_1 = X_2$ and compute the probability of each.

Solution The support of X_1 is $\{0, 1, 2, 3\}$ and the support for X_2 is $\{0, 1, 2, 3, 4\}$. We then see that

$$P(X_1 = X_2) = \sum_{k=0}^{3} P(X_1 = k) \cdot P(X_2 = k)$$

Computing this we get

$$P(X_1 = X_2) = \frac{1}{27} \cdot \frac{1}{16} + \frac{2}{9} \cdot \frac{1}{4} + \frac{4}{9} \cdot \frac{3}{8} + \frac{8}{27} \cdot \frac{1}{4} = 0.2986.$$

□

3.1.15

Let X have the pmf $p(x) = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^x$, $x = 0, 1, 2, 3, \ldots$, zero elsewhere. Find the conditional pmf of X given that $X \geq 3$.

Solution we see that

$$P(X = x | X \geq 3) = \frac{P(X = x)}{P(X \geq 3)} = \frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^x}{P(X \geq 3)}.$$

Then we see that

$$P(X \geq 3) = \sum_{x=3}^{\infty} \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^x = \left(\frac{1}{3}\right) \sum_{x=3}^{\infty} \left(\frac{2}{3}\right)^x = \frac{8}{27}$$

Putting together we get that

$$P(X = x | X \geq 3) = \frac{P(X = x)}{P(X \geq 3)} = \frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^x}{8/27} = \frac{9}{8}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^x$$

for $x \geq 3$. □

3.1.27

Let X have a geometric distribution. Show that

$$P(X \geq k+j \mid X \geq k) = P(X \geq j),$$

where k and j are nonnegative integers. Note that we sometimes say in this situation that X is memoryless.

Solution Since X has a geometric distribution then that means the distribution is $p(x) = p(1-p)^x$. We need to show that the probability that $k+j$ failures occurred given that k failures already happened before the first success is equal to the probability that there were j before the first success. In other words we need to show that the geometric distribution is memoryless.

We begin on the left-hand side and use the definition of conditional probability. Let A denote the event where $X \geq k+j$ and B denote the event where $X \geq k$. We see that

$$P(X \geq k+j \mid X \geq k) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$$

where in the second equality we used the fact that $B \subseteq A$. Also note that $P(X \geq n) = (1-p)^n$ for any nonnegative integer n.

We then see that

$$P(X \geq k+j \mid X \geq k) = \frac{(1-p)^{k+j}}{(1-p)^k} = (1-p)^j = P(X \geq j)$$

as required. □

3.1.30

Consider a shipment of 1000 items into a factory. Suppose the factory can tolerate about 5% defective items. Let X be the number of defective items in a sample without replacement of size $n = 10$. Suppose the factory returns the shipment if $X \geq 2$.

(a) Obtain the probability that the factory returns a shipment of items that has 5% defective items.

(b) Suppose the shipment has 10% defective items. Obtain the probability that the factory returns such a shipment.

(c) Obtain approximations to the probabilities in parts (a) and (b) using appropriate binomial distributions.

Solution For (a): If 5% of the items are defective which means there are 50 defective items. Thus X has a hypergeometric distribution of the form

$$p(x) = \frac{\binom{50}{x}\binom{950}{10-x}}{\binom{1000}{10}}$$

where $x = 0, 1, 2, \ldots, 10$. We need to find the probability of $P(X \geq 2) = 1 - P(X < 2) = 1 - p(0) - p(1)$. We get that

$$P(X \geq 2) = 1 - \frac{\binom{50}{0}\binom{950}{10}}{\binom{1000}{10}} - \frac{\binom{50}{1}\binom{950}{9}}{\binom{1000}{10}} = 0.0853.$$

For (b): Using the same logic since 10% of the items are defective we have that 100 items are defective. Then X has a distribution

$$p(x) = \frac{\binom{100}{x}\binom{900}{10-x}}{\binom{1000}{10}}.$$

We then find that

$$P(X \geq 2) = 1 - \frac{\binom{100}{0}\binom{900}{10}}{\binom{1000}{10}} - \frac{\binom{100}{1}\binom{900}{9}}{\binom{1000}{10}} = 0.2637$$

For (c): We will approximate the hypergeometric distribution with the binomial distribution. We will approximate part (a) by $X \approx Bin(10, 0.05)$. We get that

$$P(X \geq 2) = 1 - \binom{10}{0}(0.95)^{10} - \binom{10}{1}(0.05)^1(0.95)^9 = 0.08613.$$

Similarly for part (b) we approximate it by $X \approx Bin(10, 0.1)$. We get that

$$P(X \geq 2) = 1 - \binom{10}{0}(0.90)^{10} - \binom{10}{1}(0.1)^1(0.90)^9 = 0.2639.$$

□

3.5.2 Section 3.2 Answers

3.2.1

If the random variable X has a Poisson distribution such that $P(X = 1) = P(X = 2)$, find $P(X = 4)$.

Solution IF X has a poisson distribution then it has a pmf of the form

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with rate $\lambda > 0$. We are given that

$$P(X = 1) = \lambda e^{-\lambda} = \frac{\lambda^2 e^{-\lambda}}{2} = P(X = 2).$$

We now solve for the rate. We begin and see that

$$\begin{aligned} \lambda e^{-\lambda} &= \frac{\lambda^2 e^{-\lambda}}{2} \\ 2\lambda e^{-\lambda} &= \lambda^2 e^{-\lambda} \\ 0 &= \lambda^2 e^{-\lambda} - 2\lambda e^{-\lambda} \\ &= \lambda(\lambda e^{-\lambda} - 2e^{-\lambda}) \end{aligned}$$

Since $\lambda > 0$ we see that the only solution is

$$\lambda e^{-\lambda} - 2e^{-\lambda} = 0$$

which gives us $\lambda = 2$. Thus we see that

$$P(X = 4) = \frac{2^4 e^{-2}}{4!} = 0.09.$$

□

3.2.2

The mgf of a random variable X is $e^{4(e^t-1)}$. Show that

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.931.$$

Solution We being with finding the mean by finding $M'(0)$. We see that

$$M'(t) = 4e^{4(e^t-1)}e^t.$$

Then we see that

$$\mu = M'(0) = 4.$$

We now find $M''(t)$.

$$M''(t) = 16e^{4(e^t-1)}e^{2t} + 4e^{4(e^t-1)}e^t.$$

WE then get that $E(X^2) = M''(0) = 20$. Then we get that $\sigma^2 = 20 - 4^2 = 4$. Thus we need to find $P(4 - 2 \cdot 2 < X < 4 + 2 \cdot 2) = P(0 < X < 8)$. Since our MGF is the MGF of the form of a Poisson distribution with rate $\lambda = 4$ we see that

$$P(-4 < X < 12) = \sum_{k=1}^{7} \frac{4^k e^{-4}}{k!} = 0.931.$$

□

3.2.3

In a lengthy manuscript, it is discovered that only 13.5 percent of the pages contain no typing errors. If we assume that the number of errors per page is a random variable with a Poisson distribution, find the percentage of pages that have exactly one error.

Solution Let X be the random variable of the number of errors per page with rate $\lambda > 0$. We are given that $P(X = 0) = 0.135$. That is we get

$$P(X = 0) = e^{-\lambda} = 0.135.$$

Solving for the rate we get that $\lambda = -\ln(0.135) \approx 2$. We now find the probability that the number of errors per page is 1.

$$P(X = 1) = 2e^{-2} = 0.2706$$

□

3.5.3 Section 3.3 Answers

3.3.6

Let $X_1, X_2,$ and X_3 be iid random variables, each with pdf $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere.

(a) Find the distribution of $Y = \min(X_1, X_2, X_3)$.
Hint:
$$P(Y \leq y) = 1 - P(Y > y) = 1 - P(X_i > y, i = 1, 2, 3).$$

(b) Find the distribution of $Y = \max(X_1, X_2, X_3)$.

Solution For (a): Using the hint we get

$$F_Y(y) = 1 - P(X_1 > y, X_2 > y, X_3 > y).$$

Since all three random variables are independent we get that

$$F_Y(y) = 1 - P(X_1 > y) \cdot P(X_2 > y) \cdot P(X_3 > y) = 1 - (P(X_1 > y))^3.$$

We find that

$$P(X_i > y) = \int_y^\infty e^{-x} dx = e^{-y}.$$

Thus we find that

$$F_Y(y) = 1 - e^{-3y}$$

and the pdf is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 3e^{-3y}.$$

For (b): We see that

$$P(Y \leq y) = P(X_1 \leq y, X_2 \leq y, X_3 \leq y) = (P(X_1 \leq y))^3.$$

We find that

$$P(X_i \leq y) = \int_0^y e^{-x} dx = 1 - e^{-y}.$$

Thus we find that

$$F_Y(y) = (1 - e^{-y})^3$$

and the pdf is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 3e{-}y(1 - e^{-y})^2.$$

□

3.3.14

In a warehouse of parts for a large mill, the average time between requests for parts is about 10 minutes.

(a) Find the probability that in an hour there will be at least 10 requests for parts.

(b) Find the probability that the 10th request in the morning requires at least 2 hours of waiting time.

Solution Let X denote the number of requests in an hour. We are asked to find $P(X \geq 10)$. This is a poisson distribution since we are finding the total number of parts within an interval. Since the average time between requests is 10 minutes then in an hour the average number of requests is $\lambda = 6$. We then get that

$$p(x) = \frac{6^x e^{-6}}{x!}.$$

We then see that

$$\begin{aligned} P(X \geq 10) &= 1 - P(X \leq 9) \\ &= 1 - \sum_{x=0}^{9} \frac{6^x e^{-6}}{x!} \\ &= 0.0839. \end{aligned}$$

Next, we are asked to find the probability that the waiting time for the 10th request is greater than 2 hours. Let W_{10} denote the waiting time for the 10th request. We need to find $P(W_{10} \geq 120$. We know that this follows a gamma distribution with parameters $\alpha = 10$ and $\beta = 1/10$. We get that

$$P(W_{10} \geq 2) = 1 - \text{pgamma}(2, \alpha = 10, \beta = 1/10) \approx 02424.$$

□

3.3.16

Let X and Y have the joint pmf $p(x, y) = e^{-2}/[x!(y - x)!]$, $y = 0, 1, 2, \ldots$, $x = 0, 1, \ldots, y$, zero elsewhere.

(a) Find the mgf $M(t_1, t_2)$ of this joint distribution.

(b) Compute the means, the variances, and the correlation coefficient of X and Y.

(c) Determine the conditional mean $E(X|y)$.
Hint: Note that

$$\sum_{x=0}^{y} [\exp(t_1 x)] y!/[x!(y - x)!] = [1 + \exp(t_1)]^y.$$

Why?

Solution Skipped □

3.3.19

Determine the constant c in each of the following so that each $f(x)$ is a β pdf:

(a) $f(x) = cx(1 - x)^3$, $0 < x < 1$, zero elsewhere.

(b) $f(x) = cx^4(1 - x)^5$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = cx^2(1 - x)^8$, $0 < x < 1$, zero elsewhere.

Solution Note that the β distribution is of the form

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

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. For (a): Here $\alpha = 2$ and $\beta = 4$. We then find the value of

$$c = \frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{(4+2-1)!}{(2-1)!(4-1)!} = 20.$$

For (b): We again find the value of the constant with parameters $\alpha = 5$ and $beta = 6$.

$$c = \frac{(5+6-1)!}{(5-1)!(6-1)!} = 1260.$$

For (c): The parameters are $\alpha = 3$ and $\beta = 9$. We see that

$$c = \frac{(3+9-1)!}{(3-1)!(9-1)!} = 495.$$

□

3.3.24

Let X_1, X_2 be two independent random variables having gamma distributions with parameters $\alpha_1 = 3$, $\beta_1 = 3$ and $\alpha_2 = 5$, $\beta_2 = 1$, respectively.

(a) Find the mgf of $Y = 2X_1 + 6X_2$.

(b) What is the distribution of Y?

Solution The mgf of Y is

$$M_Y(t) = M_{X_1}(3t) \cdot M_{X_2}(6t).$$

So we begin by finding the MGF of X_1 and X_2. We see that the MGF of X_1

$$M_{X_1}(t) = \frac{1}{(1-3t)^3}.$$

The MGF of X_2 is then

$$M_{X_2}(t) = \frac{1}{(1-t)^5}.$$

We then get that

$$M_Y(t) = \frac{1}{(1-6t)^3} \cdot \frac{1}{(1-6t)^5} = \frac{1}{(1-6t)^8}$$

for $t < 1/6$.From this we can see that the distribution of Y follows a gamma distribution with parameters of $\alpha = 8$ and $\beta = 6$. That is

$$f_Y(y) = \frac{x^7 e^{-x/6}}{6^8(7)!}.$$

□

3.3.25

Let X have an exponential distribution.

(a) For $x > 0$ and $y > 0$, show that

$$P(X > x + y \mid X > x) = P(X > y).$$

Hence, the exponential distribution has the **memoryless** property. Recall from Exercise 3.1.9 that the discrete geometric distribution has a similar property.

(b) Let $F(y)$ be the cdf of a continuous random variable Y. Assume that $F(0) = 0$ and $0 < F(y) < 1$ for $y > 0$. Suppose memoryless property holds for Y. Show that $F_Y(y) = 1 - e^{-\lambda y}$ for $y > 0$.
Hint: Show that $g(y) = 1 - F_Y(y)$ satisfies the equation

$$g(y + z) = g(y)g(z),$$

Solution For (a): Just like we did with the geometric distribution we will use the definition of conditional probability so simplify the left-hand side. We get

$$P(X > x + y \mid X > x) = \frac{P(A \cap B)}{P(B)}$$

Where A denotes the event where $X > x + y$ and B denotes the event where $X > x$. Again notice that $A \cap B = A$ since x and y are positive. We then see that

$$\begin{aligned} P(X > x + y \mid X > x) &= \frac{P(A)}{P(B)} \\ &= \frac{\frac{1}{\lambda}\int_{x+y}^{\infty} e^{-t/\lambda} dt}{\frac{1}{\lambda}\int_{x}^{\infty} e^{-t/\lambda} dt} \end{aligned}$$

Where $\lambda > 0$ is the rate parameter. We simplify the numerator and get

$$\frac{1}{\lambda}\int_{x+y}^{\infty} e^{-t/\lambda} dt = e^{-(x+y)/\lambda}$$

and the denominator becomes

$$\frac{1}{\lambda}\int_{x}^{\infty} e^{-t/\lambda} dt = e^{-x/\lambda}.$$

We then get

$$P(X > x + y \mid X > x) = \frac{e^{-(x+y)/\lambda}}{e^{-x/\lambda}} = e^{-y/\lambda} = P(X > y).$$

For (b): We are given that Y is memoryless. Then

$$P(Y > y + z \mid Y > y) = \frac{g(y + z)}{g(y)} = g(z)$$

this implies that $g(y + z) = e^{-(y+z)/\lambda} = e^{-y/\lambda}e^{-z\lambda} = g(y)g(z)$. Thus since $F_Y(y) = 1 - g(y)$ then

$$F_Y(y) = 1 - e^{-y\lambda}.$$

□

3.5.4 Section 3.4 Answers

3.4.1

If

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

show that $\Phi(-z) = 1 - \Phi(z)$.

Solution Let $f(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw$. Note this is the standard normal distribution pdf and it is even. That is $f(-w) = f(w)$. We begin and see that

$$\Phi(z) + \Phi(-z) = \int_{-\infty}^{z} f(w)dw + \int_{-\infty}^{-z} f(w)dw.$$

Using basic calc we see that

$$\Phi(-z) = \int_{\infty}^{z} f(w)dw = \int_{z}^{\infty} f(w)dw.$$

Together we get that

$$\Phi(-z) = \int_{z}^{\infty} f(w)dw = 1 - \int_{-\infty}^{z} f(w)dw.$$

□

3.4.4

Let X be $N(\mu, \sigma^2)$ so that $P(X < 89) = 0.90$ and $P(X < 94) = 0.95$. Find μ and σ^2.

Solution We convert this problem into the standard normal distribution. We let $Z = \frac{X-\mu}{\sigma}$. Then we see that

$$P(X < 89) = P(Z < \frac{89-\mu}{\sigma}) = 0.90 \quad \text{and} \quad P(X < 94) = P(Z < \frac{94-\mu}{\sigma}) = 0.95.$$

Using the standard normal distribution table we can find which values of x give us a probability of 0.90 and 0.95. We then see that $\Phi^{-1}(0.90) \approx 1.3$ and $\Phi^{-1}(0.95) \approx 1.65$. We then get the equations

$$\frac{80-\mu}{\sigma} = 1.3 \quad \text{and} \quad \frac{94-\mu}{\sigma} = 1.65.$$

Solving this system of equations we get that $\mu \approx 71$ and $\sigma^2 \approx 189$. □

3.4.6

If X is $N(\mu, \sigma^2)$, show that $E(|X - \mu|) = \sigma\sqrt{2/\pi}$.

Solution We shift this to the standard normal distribution by letting $Z = \frac{X-\mu}{\sigma}$. We then get

$$E(|X - \mu|) = E(|\sigma(\frac{X-\mu}{\sigma})|) = E(\sigma|Z|) = \sigma E(|Z|).$$

So we only need to compute $E(|Z|)$. We get that

$$E(|Z|) = 2\int_0^\infty z \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Solving this we get that

$$E(|Z|) = 2 \cdot \frac{1}{\sqrt{2\pi}} \cdot 1 = \sqrt{\frac{2}{\pi}}.$$

Thus

$$E(|X-\mu|) = \sigma\sqrt{\frac{2}{\pi}}.$$

□

3.4.9

Determine the 90th percentile of the distribution, which is $N(65, 25)$.

Solution We need to find the value p such that $P(X \leq p) = 0.90$. We can do this easily by just looking at the standard normal distribution but first we have to convert normalize the distribution. Let $Z = \frac{X-65}{5}$. Then we need to find the value of $P(Z \leq \frac{p-65}{5}) = 0.90$. We find that $\Phi^{-1}(0.90) \approx 1.3$. Thus we get that

$$p = 5 \cdot 1.3 + 65 = 71.5.$$

□

3.4.10

If e^{3t+8t^2} is the mgf of the random variable X, find $P(-1 < X < 9)$.

Solution This MGF resembles the MGF of a normal distribution with mean and variance $\mu = 3$ and $\sigma^2 = 16$. Thus we can find $P(-1 < X < 9)$ by the standard normal distribution. We get that

$$P(-1 < X < 9) = P(\frac{-1-3}{4} < X < \frac{9-3}{4}) = \Phi(1.5) - \Phi(-1) = 0.97725 - 0.15866 = 0.774.$$

□

3.4.13

If X is $N(1, 4)$, compute the probability $P(1 < X^2 < 9)$.

Solution We need to find

$$P(1 < X^2 < 9) = P(X^2 < 9) - P(X^2 \leq 1) = P(-3 < X < 3) - P(-1 \leq X \leq 1).$$

Using standard normal distribution we need to find

$$P(1 < X^2 < 9) = P(\frac{-3-1}{2} < X < \frac{3-1}{2}) - P(\frac{-1-1}{2} \leq X \leq \frac{1-1}{2}).$$

Using the standard normal table we get that the answer is

$$P(1 < X^2 < 9) = 0.477.$$

□

3.4.30

Compute $P(X_1 + 2X_2 - 2X_3 > 7)$ if X_1, X_2, X_3 are iid with common distribution $N(1, 4)$.

Solution Using Theorem 3.4.4 we know that a linear combination of independent random variables with Normal distribution is still normal with parameters $N(1+2-2, 4+16+16) = N(1, 32)$. Using the standard normal distribution we see that $P(X_1 + 2X_2 - 2X_3 > 7) = 0.158$. □

4 Consistency and Limiting Distributions

The theory behind statistical inference procedures often depends on the distribution of a pivot random variable. When we make conclusions from data (like estimating parameters or testing hypotheses), we often rely on a random variable called a pivot. This variable has a known distribution that helps us build things like confidence intervals or test statistics. But in many real situations, especially with small samples, the exact distribution of this pivot might be messy or unknown. That's where approximation helps.

As your sample size n grows large, the distribution of your pivot variable (like a sample mean, sample proportion, etc.) often approaches a known distribution (like the normal distribution). This is what lets you use familiar tools like z-scores even when the underlying distribution isn't exactly normal. To talk about these approximations precisely, statisticians use types of convergence. We will discuss the two most important types: Convergence in probability and Convergence in distribution.

4.1 Convergence in Probability

In this section we formalize a way of saying that a sequence of random variables $\{X_n\}$ is getting close to another random variable X as $n \to \infty$.

Definition 4.1.1

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. We say that X_n converges in probability to X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0$$

or equivalently,

$$\lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1$$

If so, we write $X_n \xrightarrow{P} X$.

If $X_n \xrightarrow{P} X$ we say that the mass of the difference $X_n - X$ is converging to 0 in probability. One way of showing convergence in probability is to use Chebyshev's Theorem. We formalize this in the following theorem:

Theorem 4.1.2 : Weak Law of Large Numbers

Let $\{X_n\}$ be a sequence of iid random variables having common μ and variance $\sigma^2 < \infty$. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$\overline{X} \xrightarrow{P} \mu.$$

Proof. We know that that $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. By Chebyshev's inequality we have that

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

And if we take the limit as $n \to \infty$ the right-hand side goes to zero so we can conclude that $\overline{X} \xrightarrow{P} \mu$. □

This theorem says that all the mass of the distribution of $\bar{X}_n$ is converging to μ as $n \to \infty$.

Example 4.1.3

Let the random variable X have a distribution that is $Bin(n, p)$. Show that X_n/n converges in probability to p.

Solution Since $X_n \sim Bin(n, p)$ then $E(X_n) = np$ and $\text{Var}(X_n) = np(1-p)$. Let $Y_n = X_n/n$. Then $E(Y_n) = E(X_n/n) = p$ and $\text{Var}(Y_n) = (1/n^2) \cdot np(1-p) = p(1-p)/n$. Using Chebyshev's inequality we get that

$$P(|Y_n - p| > \epsilon) \leq \frac{\text{Var}(Y_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}.$$

Thus as $n \to \infty$ the right hand side goes to zero and $X_n/n \to p$ in probability. Easy □

The next theorem gives us some useful results for convergence in probability.

Theorem 4.1.4

et $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables.

- Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n + Y_n \xrightarrow{P} X + Y$.
- Suppose $X_n \xrightarrow{P} X$ and a is a constant. Then $aX_n \xrightarrow{P} aX$.
- Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_nY_n \xrightarrow{P} XY$.
- Suppose that $X_n \xrightarrow{P} a$ and the real function g is continuous at a. Then $g(X_n) \xrightarrow{P} g(a)$.

Proof. Will finish later. □

Example 4.1.5

Let the random variable X_n have an exponential distribution with the mean β. Prove that $\bar{X}_n^2$ converges to β^2 in probability.

Solution By theorem 4.1.2 we know that $\bar{X}_n$ converges in probability to $\mu = \beta$. By using Theorem 4.1.4 part 4, let $g(x) = x^2$. Since this is real continuous function at β we see that $g(\bar{X}_n) = \bar{X}_n{}^2 \to g(\beta) = \beta^2$ as needed. □

Example 4.1.6

Show that

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

converges in probability to σ^2.

Solution Will finish later □

We are interested in estimating some unknown parameter θ of a population (for example, the population mean μ, or variance σ^2, etc.). You observe a random sample $X_1, X_2, \ldots, X_n$ drawn from the population (so all X_i have the same distribution as X). Now, you define an estimator T_n, which is just some function of the sample for example, the sample mean:

$$T_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We say that T_n is a consistent estimator of θ if:

$$T_n \xrightarrow{P} \theta$$

that is, converges in probability to θ as $n \to \infty$.

Definition 4.1.7

Let X be a random variable with cdf $F(x, \theta)$, $\theta \in \Omega$. Let $X_1, \ldots, X_n$ be a sample from the distribution of X and let T_n denote a statistic. We say T_n is a consistent estimator of θ if

$$T_n \xrightarrow{P} \theta.$$

For example, if $X_1, \ldots, X_n$ are drawn from a distribution with finite mean μ and variance $\sigma^2 < \infty$ then by the Weak Law of Large Numbers, the sample mean converges in probability to the population mean:

$$\overline{X} \xrightarrow{P} \mu$$

So, $\overline{X}$ is a consistent estimator of μ.

4.2 Convergence in Distribution

In the previous section, you learned about convergence in probability, which tells us that a statistic (like the sample mean) gets close to a parameter (like the true mean) as the sample size increases. However, convergence in probability doesn't say much about the distribution of the statistic just that it concentrates around a value. In contrast, convergence in distribution allows us to make inferences about the distributional behavior of an estimator, which is useful for quantifying uncertainty, building confidence intervals, conducting hypothesis tests, etc.

Definition 4.2.1

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of X_n and X. Let $\mathcal{C}(F_X)$ denote the set of all points where F_X is continuous. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F(x), \quad \text{for all } x \in \mathcal{C}(F_X).$$

We denote this convergence by $X_n \xrightarrow{D} X$.

The following theorem gives us some important results of convergence in distribution. The first result tells us that adding a small (vanishing in probability) "noise" Y_n to a sequence that already converges in distribution doesn't disturb the limiting distribution. The second result tells us if a sequence of random variables converges in distribution, and you apply a continuous function to the sequence, then the transformed sequence also converges in distribution to the transformed limit. The final result tells us random coefficients converging in probability can be "plugged in" safely when combined with convergence in distribution.

Theorem 4.2.2

- Suppose X_n converges to X in distribution and Y_n converges in probability to 0. Then $X_n + Y_n$ converges to X in distribution.
- *(Continuous Mapping Theorem)* Suppose X_n converges to X in distribution and g is a continuous function on the support of X. Then $g(X_n)$ converges to $g(X)$ in distribution.
- *(Slutsky's Theorem)* Let X_n, X, A_n and B_n be random variables and let a and b be constants. If $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$ and $B_n \xrightarrow{P} b$, then

$$A_n + B_n X_n \xrightarrow{D} a + bX.$$

Proof. I will prove the first two parts beginning with the first part. The key idea is that $Y_n \to 0$ in probability allows us to trap $X_n + Y_n$ between X_n shifted by an arbitrarily small ε; the limits of the resulting probabilities are governed solely by the convergence of X_n in distribution. Denote the distribution functions

$$F_n(x) = P\big(X_n + Y_n \le x\big), \qquad F(x) = P(X \le x),$$

and let $C(F)$ be the set of continuity points of F. Fix $x \in C(F)$ and an arbitrary $\varepsilon > 0$. We begin by finding an upper bound for $F_n(x)$ Split the probability:

$$F_n(x) = P(X_n + Y_n \le x) = P\big(X_n + Y_n \le x,\ |Y_n| \le \varepsilon\big) + P\big(X_n + Y_n \le x,\ |Y_n| > \varepsilon\big).$$

Because $Y_n \xrightarrow{P} 0$, $P(|Y_n| > \varepsilon) \to 0$. On $\{|Y_n| \le \varepsilon\}$ we have $X_n + Y_n \le x \Rightarrow X_n \le x + \varepsilon$. Hence

$$F_n(x) \le P(X_n \le x + \varepsilon) + P(|Y_n| > \varepsilon).$$

Taking lim sup and using $X_n \xrightarrow{D} X$,

$$\limsup_{n\to\infty} F_n(x) \le \lim_{n\to\infty} P(X_n \le x + \varepsilon) = F(x + \varepsilon). \quad (1)$$

Now we find the lower bound for $F_n(x)$

Similarly, on $\{|Y_n| \le \varepsilon\}$,

$$X_n \le x - \varepsilon \implies X_n + Y_n \le x,$$

so

$$F_n(x) \ge P(X_n \le x - \varepsilon) - P(|Y_n| > \varepsilon).$$

Taking lim inf,

$$\liminf_{n\to\infty} F_n(x) \ge F(x - \varepsilon). \quad (2)$$

From (1)–(2),

$$F(x - \varepsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x + \varepsilon).$$

Because x is a continuity point of F, letting $\varepsilon \downarrow 0$ forces

$$\lim_{n\to\infty} F_n(x) = F(x).$$

The equality above holds for every $x \in C(F)$; hence, by definition of convergence in distribution,

$$X_n + Y_n \xrightarrow{D} X.$$

as required.

For the next part, fix a point y at which G is continuous and an arbitrary $\varepsilon > 0$. Because g is continuous, the pre-images

$$A_{y-\varepsilon} = g^{-1}\big((-\infty, y - \varepsilon]\big), \qquad A_{y+\varepsilon} = g^{-1}\big((-\infty, y + \varepsilon]\big)$$

are closed subsets of $\mathbb{R}$. Notice that

$$A_{y-\varepsilon} \subset g^{-1}\big((-\infty, y]\big) \subset A_{y+\varepsilon}. \quad (1)$$

Again we find an upper bound for $G_n(y)$.By the nesting in (1),

$$G_n(y) = P\big(X_n \in g^{-1}((-\infty, y])\big) \le P\big(X_n \in A_{y+\varepsilon}\big) = F_n^*(A_{y+\varepsilon}),$$

where F_n^* denotes the inner probability.

Because $A_{y+\varepsilon}$ is closed, the portmanteau equivalence for cdf-convergence (or directly the definition with continuity points) gives

$$\limsup_{n\to\infty} G_n(y) \le \limsup_{n\to\infty} P(X_n \in A_{y+\varepsilon}) \le P(X \in A_{y+\varepsilon}) = G(y + \varepsilon). \quad (2)$$

Now for the lower bound for $G_n(y)$. Again from (1),

$$G_n(y) \ge P\big(X_n \in A_{y-\varepsilon}\big).$$

Because $A_{y-\varepsilon}$ is closed,

$$\liminf_{n\to\infty} G_n(y) \ge \liminf_{n\to\infty} P(X_n \in A_{y-\varepsilon}) \ge P(X \in A_{y-\varepsilon}) = G(y - \varepsilon). \quad (3)$$

Combine (2) and (3):

$$G(y-\varepsilon) \;\le\; \liminf_{n\to\infty} G_n(y) \;\le\; \limsup_{n\to\infty} G_n(y) \;\le\; G(y+\varepsilon).$$

Since G is continuous at y, letting $\varepsilon \downarrow 0$ forces

$$G(y-\varepsilon) \;\longrightarrow\; G(y), \qquad G(y+\varepsilon) \;\longrightarrow\; G(y),$$

and the squeeze yields

$$\lim_{n\to\infty} G_n(y) = G(y).$$

Because y was an arbitrary continuity point of G, this proves that $g(X_n)$ converges in distribution to $g(X)$ according to the cdf definition. □

To find the limiting distribution function of a random variable X_n by using the definition, it obviously requires that we know $F_{X_n}(x)$ for each positive integer n. But it is often difficult to obtain $F_{X_n}(x)$ in closed form. Fortunately, if it exists, the mgf that corresponds to the cdf $F_{X_n}(x)$ often provides a convenient method of determining the limiting cdf.

Theorem 4.2.3

Let $\{X_n\}$ be a sequence of random variables with mgf $M_{X_n}(t)$ that exists for $-h < t < h$ for all n. Let X be a random variable with mgf $M(t)$, which exists for $|t| \le h_1 \le h$. If

$$\lim_{n\to\infty} M_{X_n}(t) = M(t) \quad \text{for } |t| \le h_1,$$

then $X_n \xrightarrow{D} X$.

This is useful as it gives us a way to prove convergence of distribution of X_n and also find the limiting distribution.

Example 4.1.11

Let X_n have a distribution that is $Bin(n,p)$. Suppose that the mean $\mu = np$ is the same for every n; i.e., $p = \frac{\mu}{n}$, where μ is a constant. Find the limiting distribution of the binomial distribution, when $p = \frac{\mu}{n}$, by finding the limit of $M_{X_n}(t)$.

Solution Now

$$M_{X_n}(t) = E(e^{tX_n}) = \left[(1-p) + pe^t\right]^n = \left[1 + \frac{\mu(e^t-1)}{n}\right]^n$$

for $t \in \mathbb{R}$. Hence,

$$\lim_{n\to\infty} M_{X_n}(t) = e^{\mu(e^t-1)}, \quad \text{for } t \in \mathbb{R}.$$

Because the mgf is unique, and the resulting mgf belongs to the Poisson distribution, we conclude that X_n converges in distribution to the Poisson distribution with parameter μ. □

4.3 Central Limit Theorem

We know that if $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then the random variable $\overline{X}$ (sample mean), has a normal distribution with parameters μ and variance $\frac{\sigma^2}{n}$. In other words,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, $N(0,1)$. What if the random sample does not have a normal distribution? Can we make any statement about the distribution of $\overline{X}$? Yes. If sample size is sufficiently large, under some conditions, the distribution of $\overline{X}$ converges in distribution to a normal distribution. This is known as the **Central Limit Theorem** (CLT). This is powerful because it allows us to use normal distribution-based inference (like confidence intervals, hypothesis tests) even when the population isn't normal, as long as the sample size is large.

4.3.1 : Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and positive variance σ^2. Then the random variable

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

converges in distribution to a random variable that has a normal distribution with mean zero and variance 1.

Proof. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with finite mean μ and variance σ^2. Define the sample mean as

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

We want to show that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

converges in distribution to a standard normal $N(0,1)$. Define

$$S_n = \sum_{i=1}^n X_i.$$

Then,

$$E(S_n) = n\mu, \quad \text{Var}(S_n) = n\sigma^2.$$

The standardized sum is

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Note that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Let $M_X(t)$ be the MGF of X_i. Then, the MGF of S_n is

$$M_{S_n}(t) = (M_X(t))^n .$$

Consider

$$t_n = \frac{t}{\sigma\sqrt{n}}.$$

Then,

$$M_{S_n - n\mu}(t_n) = e^{-t_n n\mu} \left(M_X(t_n)\right)^n .$$

For small t_n, we can write

$$M_X(t_n) = 1 + \mu t_n + \frac{\sigma^2 t_n^2}{2} + o(t_n^2).$$

So,

$$(M_X(t_n))^n \approx \left(1 + \mu t_n + \frac{\sigma^2 t_n^2}{2}\right)^n .$$

Using the approximation $(1 + \frac{a}{n})^n \approx e^a$, we get

$$\left(1 + \mu t_n + \frac{\sigma^2 t_n^2}{2}\right)^n \approx e^{n\mu t_n + \frac{n\sigma^2 t_n^2}{2}}.$$

Thus,

$$M_{S_n - n\mu}(t_n) \approx e^{-t_n n\mu} \cdot e^{n\mu t_n + \frac{n\sigma^2 t_n^2}{2}} = e^{\frac{n\sigma^2 t_n^2}{2}}.$$

Replacing $t_n = \frac{t}{\sigma\sqrt{n}}$, we get

$$\frac{n\sigma^2 t_n^2}{2} = \frac{n\sigma^2 (t^2)}{\sigma^2 n \cdot 2} = \frac{t^2}{2}.$$

Thus, we get

$$M_{S_n - n\mu}(t_n) \approx e^{t^2/2}.$$

This is the MGF of a standard normal distribution $N(0,1)$. Since the MGF converges to $e^{t^2/2}$ for all t, it follows that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Thus,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

as required. □

Example 4.3.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2, where μ and σ^2 are unknown. Let $\overline{X}$ and $S = \sqrt{S^2}$ be the sample mean and sample standard deviation, respectively. Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \xrightarrow{D} N(0,1)$$

Solution To see this, we start with the statistic

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

We can rewrite it as

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \left(\frac{\sigma}{S}\right)\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right).$$

By the Central Limit Theorem, we know that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Also, it is known that the sample standard deviation S converges in probability to the true standard deviation σ, that is,

$$S \xrightarrow{P} \sigma.$$

Consequently,

$$\frac{\sigma}{S} \xrightarrow{P} 1.$$

By Slutsky's theorem, if one sequence converges in distribution and another converges in probability to a constant, then their product converges in distribution to the product of the limits. Thus,

$$\left(\frac{\sigma}{S}\right)\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{d} N(0,1).$$

Therefore, we conclude that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1).$$

□

Example 4.3.3

Let $\overline{X}$ denote the mean of a random sample of size 128 from a Gamma distribution with $\alpha = 2$ and $\beta = 4$. Approximate $P(7 < \overline{X} < 9)$.

Solution Since this is a gamma distribution we know then $\mu = \alpha\beta = 8$ and $\sigma^2 = \alpha\beta^2 = 32$. Thus as we know from the Central limit theorem the sample mean is normally distributed so we can approximate it with the standard normal distribution

$$P(7 < \overline{X} < 9) = P\left(\frac{7-8}{\sqrt{32/128}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{9-8}{\sqrt{32/128}}\right)$$

$$= P(-2 < Z < 2) = \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 2\Phi(2) - 1$$

$$= 0.9545$$

□

4.3.1 Normal Approximation to the Binomial Distribution

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution that is $Ber(p)$. Here $\mu = p$, $\sigma^2 = p(1-p)$, and $M(t)$ exists for all real values of t. If $Y_n = X_1 + \ldots + X_n$, then we can show that Y_n is $Bin(n, p)$.

When dealing with large n, we often want to approximate $Y_n$using a normal distribution. The standardized version of Y_n is:

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

This is called standardization, which centers the variable by subtracting its mean np and scales it by its standard deviation $np(1-p)$. The sample mean is $\bar{X} = Y_n/n$ so we can write

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{n\bar{X} - np}{\sqrt{p(1-p)}} = \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}}$$

By the Central Limit Theorem, Z_n has a limiting distribution that is a standard normal distribution with mean zero and variance 1. This is useful because when n is large, even though Y_n is discrete (Binomial), we can approximate probabilities using the normal distribution. For example:

$$P(a \le Y_n \le b) \approx P\left(\frac{a - np}{\sqrt{np(1-p)}} \le Z \le \frac{b - np}{\sqrt{np(1-p)}}\right),$$

where $Z \sim N(0, 1)$.

4.3.2 Continuity Correction

When we approximate a discrete distribution (like a binomial) using a continuous distribution (like a normal), we use something called a continuity correction to improve the accuracy of the approximation.

A discrete variable takes integer values (like $0, 1, 2, \ldots$), while a continuous variable can take any real value. If we want to approximate, for example, $P(X \le k)$ for a discrete X, using a continuous normal distribution, we need to account for the fact that the binomial probability is concentrated at points (bars), not spread smoothly like a normal curve. The correction is done by adjusting the value by 0.5. For example:

$$P(X \le k) \approx P(Y \le k + 0.5),$$

where Y is the continuous normal approximation. Note that the area of a rectangle centered at k with width 1 (from $k - 0.5$ to $k + 0.5$) in the discrete case is approximately equal to the area

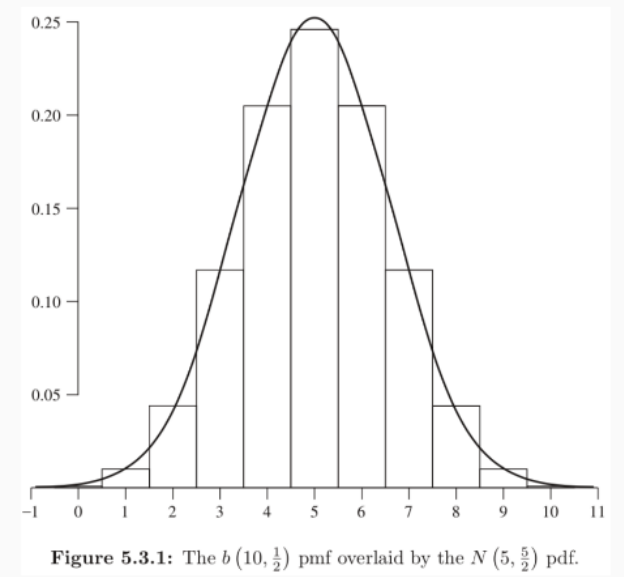


Figure 5.3.1: The $b\left(10, \frac{1}{2}\right)$ pmf overlaid by the $N\left(5, \frac{5}{2}\right)$ pdf.

Figure 15: Graph showing the overlay of a Binomial and Normal distribution approximation.

under the normal curve over $(k - 0.5, k + 0.5)$ (see Figure 15). This correction ensures that the discrete point mass at each integer value is more accurately represented when approximating with the smooth normal curve.

4.3.4

Suppose that $X_1, \ldots, X_{100}$ is a random sample from $Ber(0.5)$. Let $Y = \sum_{i=1}^{100} X_i$. We wish to approximate $P(48 \leq Y \leq 52)$.

Solution First note that Y has $Bin(100, 0.5)$ distribution. Here we have $\mu = np = 50$ and $\sigma = \sqrt{np(1-p)} = 5$. Hence, by normal approximation of the binomial distribution from the CLT, and the continuity correction, we have

$$P(47.5 < Y < 52.5) = P\left(\frac{47.5 - 50}{5} < \frac{Y - 50}{5} < \frac{52.5 - 50}{5}\right)$$

$$\approx P(-0.5 < Z < 0.5) = P(Z < 0.5) - P(Z < -0.5)$$

$$= 1 - 2\Phi(-0.5)$$

$$= 0.3829$$

□

4.4 Practice Problems

4.4.1 Section 4.1 Answers

5.1.2

Let the random variable Y_n have a distribution that is $b(n, p)$.

(a) Prove that Y_n/n converges in probability to p. This result is one form of the weak law of large numbers.

(a) Prove that $1 - Y_n/n$ converges in probability to $1 - p$.

(a) Prove that $(Y_n/n)\,(1 - Y_n/n)$ converges in probability to $p(1 - p)$.

Solution For (a): We want to show that Y_n/n converges in probability to p. We know that $E[Y_n/n] = p$ and $\text{Var}(Y_n/n) = \frac{p(1-p)}{n}$. Since $\text{Var}(Y_n/n) \to 0$ as $n \to \infty$, by Chebyshev's inequality, for any $\epsilon > 0$,

$$P\big(|Y_n/n - p| \geq \epsilon\big) \leq \frac{\text{Var}(Y_n/n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \to 0.$$

For (b): We want to show that $1 - Y_n/n$ converges in probability to $1 - p$. Since $Y_n/n \xrightarrow{P} p$, by continuous mapping theorem,

$$1 - Y_n/n \xrightarrow{P} 1 - p.$$

For (c): We want to show that $(Y_n/n)(1 - Y_n/n)$ converges in probability to $p(1 - p)$. We know $Y_n/n \xrightarrow{P} p$, and from (b), $1 - Y_n/n \xrightarrow{P} 1 - p$. Then, by continuous mapping theorem (since the product is continuous),

$$(Y_n/n)(1 - Y_n/n) \xrightarrow{P} p(1 - p).$$

□

5.1.7

Let $X_1, \ldots, X_n$ be iid random variables with common pdf

$$f(x) = \begin{cases} e^{-(x-\theta)} & x > \theta,\ -\infty < \theta < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

This pdf is called the shifted exponential. Let $Y_n = \min\{X_1, \ldots, X_n\}$. Prove that $Y_n \to \theta$ in probability by first obtaining the cdf of Y_n.

Solution The cdf of each X_i is

$$F(x) = \begin{cases} 0 & x \leq \theta, \\ 1 - e^{-(x-\theta)} & x > \theta. \end{cases}$$

Then,

$$P(Y_n > x) = P(X_1 > x, \ldots, X_n > x) = [P(X_1 > x)]^n = [1 - F(x)]^n = [e^{-(x-\theta)}]^n = e^{-n(x-\theta)}, \quad x > \theta.$$

Thus, the cdf of Y_n is

$$F_{Y_n}(x) = 1 - e^{-n(x-\theta)}, \quad x > \theta.$$

For $x \leq \theta$, $F_{Y_n}(x) = 0$.
To show $Y_n \xrightarrow{P} \theta$, we check for any $\epsilon > 0$,

$$P(|Y_n - \theta| > \epsilon) = P(Y_n - \theta > \epsilon) = P(Y_n > \theta + \epsilon) = 1 - F_{Y_n}(\theta + \epsilon) = e^{-n\epsilon}.$$

As $n \to \infty$, $e^{-n\epsilon} \to 0$.
Hence,

$$P(|Y_n - \theta| > \epsilon) \to 0,$$

which proves that $Y_n \xrightarrow{P} \theta$. □

4.4.2 Section 4.2 Answers

5.2.1

Let $\overline{X}_n$ denote the mean of a random sample of size n from a distribution that is $N(\mu, \sigma^2)$. Find the limiting distribution of $\overline{X}_n$.

Solution
We are given that $\overline{X}_n$ is the sample mean of a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. We know that if $X_i \sim N(\mu, \sigma^2)$, then:

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

The mean and variance of $\overline{X}_n$ are:

$$E(\overline{X}_n) = \mu, \quad \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Moreover, because the normal distribution is closed under averaging (a linear combination of normals is normal), we have:

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

When $n \to \infty$, the variance $\frac{\sigma^2}{n}$ goes to 0, so $\overline{X}_n$ converges in probability to μ. However, if we are asked for the limiting distribution, we look at $\overline{X}_n$ directly as $n \to \infty$:

$$\overline{X}_n \xrightarrow{D} N(\mu, 0),$$

which is a degenerate distribution at μ.
Alternatively, if we standardize it:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

But as stated, the limiting distribution of $\overline{X}_n$ itself is $N\left(\mu, \frac{\sigma^2}{n}\right)$ for finite n, and converges to a point mass at μ as $n \to \infty$. □

5.2.7

Let X_n have a gamma distribution with parameter $\alpha = n$ and β, where β is not a function of n. Let $Y_n = X_n/n$. Find the limiting distribution of Y_n.

Solution We are given that

$$X_n \sim \text{Gamma}(\alpha = n, \beta).$$

This means:

$$E(X_n) = \alpha\beta = n\beta, \quad \text{Var}(X_n) = \alpha\beta^2 = n\beta^2.$$

We define:

$$Y_n = \frac{X_n}{n}.$$

Then,

$$E(Y_n) = \frac{E(X_n)}{n} = \frac{n\beta}{n} = \beta,$$

$$\text{Var}(Y_n) = \frac{\text{Var}(X_n)}{n^2} = \frac{n\beta^2}{n^2} = \frac{\beta^2}{n}.$$

As $n \to \infty$, $\text{Var}(Y_n) \to 0$. Thus, Y_n converges in probability to β. However, we are asked for the limiting distribution. By the Central Limit Theorem for Gamma distributions with large shape, we know:

$$\frac{X_n - n\beta}{\sqrt{n}\beta} \xrightarrow{D} N(0,1).$$

Equivalently,

$$\frac{Y_n - \beta}{\beta/\sqrt{n}} \xrightarrow{D} N(0,1).$$

Thus, for large n, Y_n is approximately normal with mean β and variance $\frac{\beta^2}{n}$. Therefore, the limiting distribution of Y_n is a degenerate distribution at β, and if we standardize, we get convergence to $N(0,1)$. □

5.2.8

Let Z_n be $\chi^2(n)$ and let $W_n = Z_n/n$. Find the limiting distribution of W_n.

Solution We are given:

$$Z_n \sim \chi^2(n), \quad W_n = \frac{Z_n}{n}.$$

Properties of Z_n:

$$E(Z_n) = n, \quad \text{Var}(Z_n) = 2n.$$

Then,

$$E(W_n) = \frac{E(Z_n)}{n} = 1,$$

$$\text{Var}(W_n) = \frac{\text{Var}(Z_n)}{n^2} = \frac{2n}{n^2} = \frac{2}{n}.$$

As $n \to \infty$, $\text{Var}(W_n) \to 0$. Thus, W_n converges in probability to 1. However, if we want to find the limiting distribution, we consider a standardized version instead

$$\sqrt{n}(W_n - 1) = \sqrt{n}\left(\frac{Z_n}{n} - 1\right) = \frac{Z_n - n}{\sqrt{n}}.$$

Now, it is known that if $Z_n \sim \chi^2(n)$, then as $n \to \infty$:

$$\frac{Z_n - n}{\sqrt{2n}} \xrightarrow{D} N(0,1).$$

So,

$$\frac{Z_n - n}{\sqrt{n}} = \sqrt{2}\frac{Z_n - n}{\sqrt{2n}} \xrightarrow{D} N(0,2).$$

Thus,

$$\sqrt{n}(W_n - 1) \xrightarrow{D} N(0,2).$$

□

4.4.3 Section 4.3 Answers

5.3.1

Let $\overline{X}$ denote the mean of a random sample of size 100 from a distribution that is $\chi^2(50)$. Compute an approximate value of $P(49 < \overline{X} < 51)$.

Solution We are given:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad n = 100, \quad X_i \sim \chi^2(50).$$

The mean and variance of X_i is

$$E(X_i) = 50, \quad \text{Var}(X_i) = 2 \times 50 = 100.$$
$$E(\overline{X}) = 50, \quad \text{Var}(\overline{X}) = \frac{100}{100} = 1.$$

Approximate distribution of $\overline{X}$:** By the Central Limit Theorem, for large n,

$$\overline{X} \approx N(50, 1).$$

Standardizing this we get that

$$P(49 < \overline{X} < 51) = P\left(\frac{49-50}{\sqrt{1}} < Z < \frac{51-50}{\sqrt{1}}\right) = P(-1 < Z < 1).$$

Thus

$$P(-1 < Z < 1) \approx 0.6826.$$

□

5.3.3

Let Y be $Bin(72, \frac{1}{3})$. Approximate $P(22 \leq Y \leq 28)$.

Solution We are given:

$$Y \sim Bin(72, \tfrac{1}{3}).$$

The mean and variance are

$$\mu = np = 72 \times \tfrac{1}{3} = 24,$$

$$\sigma^2 = np(1-p) = 72 \times \tfrac{1}{3} \times \tfrac{2}{3} = 16,$$

$$\sigma = \sqrt{16} = 4.$$

For large n, $Y \approx N(\mu, \sigma^2)$. Using continuity correction

$$P(22 \leq Y \leq 28) \approx P(21.5 < Y < 28.5).$$

Standardizing this we get that

$$P\left(\frac{21.5-24}{4} < Z < \frac{28.5-24}{4}\right) = P(-0.625 < Z < 1.125).$$

Find probabilities using normal table:

$$P(Z < 1.125) \approx 0.869,$$

$$P(Z < -0.625) \approx 0.266.$$

Then,

$$P(-0.625 < Z < 1.125) \approx 0.869 - 0.266 = 0.603.$$

□

5.3.10

Forty-eight measurements are recorded to several decimal places. Each of these 48 numbers is rounded off to the nearest integer. The sum of the original 48 numbers is approximated by the sum of these integers. If we assume that the errors made by rounding off are iid and have a uniform distribution over the interval $(-1/2, 1/2)$, compute approximately the probability that the sum of the integers is within two units of the true sum.

Solution The mean and variance of a single rounding error X

$$E(X) = 0,$$

$$\text{Var}(X) = \frac{(b-a)^2}{12} = \frac{(1)^2}{12} = \frac{1}{12}.$$

The sum of 48 measurements

$$S = \sum_{i=1}^{48} X_i.$$

$$E(S) = 48 \times 0 = 0,$$

$$\mathrm{Var}(S) = 48 \times \frac{1}{12} = 4,$$

$$\sigma_S = \sqrt{4} = 2.$$

We need to find

$$P(-2 < S < 2).$$

We now approximate it using normal distribution $S \approx N(0, 4)$. Standardizing we get that

$$P(-2 < S < 2) = P\left(\frac{-2}{2} < Z < \frac{2}{2}\right) = P(-1 < Z < 1).$$

Thus we get that the normal probabilities are

$$P(Z < 1) \approx 0.8413,$$

$$P(Z < -1) \approx 0.1587,$$

$$0.8413 - 0.1587 = 0.6826.$$

□