DSCI 500B: ESSENTIAL PROBABILITY THEORY FOR DATA SCIENCE

Kristin Kuter
Saint Mary's College

LibreTexts™

Saint Mary's College

DSCI 500B Essential Probability Theory for Data Science (Kuter)

Kristin Kuter

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 10/15/2026

TABLE OF CONTENTS

Licensing

1: What is Probability?

1.1: Sample Spaces and Events

1.2: Probability Measures

1.3: Equally Likely Outcomes and Counting Techniques (Combinatorics)

2: Conditional Probability

2.1: Conditional Probability and Bayes' Rule

2.2: Independent Events

3: Discrete Random Variables

3.1: Introduction to Random Variables

3.2: Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) for Discrete Random Variables

3.3: Bernoulli and Binomial Distributions

3.4: Expected Value of Discrete Random Variables

3.5: Variance of Discrete Random Variables

4: Continuous Random Variables

4.1: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) for Continuous Random Variables

4.2: Expected Value and Variance of Continuous Random Variables

4.3: Uniform Distributions

4.4: Normal Distributions

5: Multivariate Random Variables

5.1: Joint Distributions of Discrete Random Variables

5.2: Joint Distributions of Continuous Random Variables

6: The Sample Mean and Central Limit Theorem

6.1: Functions of Normal Random Variables

6.2: Sample Mean

7: The Sample Variance and Other Distributions

7.1: Other Useful Distributions

7.2: Sample Variance

Index

Glossary

Detailed Licensing

This page titled DSCI 500B Essential Probability Theory for Data Science (Kuter) is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing***.

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by LibreTexts.

1: WHAT IS PROBABILITY?

1.1: Sample Spaces and Events
1.2: Probability Measures
1.3: Equally Likely Outcomes and Counting Techniques (Combinatorics)

This page titled 1: What is Probability? is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

1.1: Sample Spaces and Events

Introduction

We begin with a definition.

Definition $1.1.1$

Probability theory provides a mathematical model for chance (or random) phenomena.

While this is not a very informative definition, it does indicate the overall goal of this course, which is to develop a *formal, mathematical structure* for the fairly intuitive concept of probability. While most everyone is familiar with the notion of "chance" -- we informally talk about the chance of it raining tomorrow, or the chance of getting what you want for your birthday -- when it comes to *quantifying* the chance of something happening, we need to develop a mathematical model to make things precise and calculable.

Sample Spaces and Events

Before we can formally define what the mathematical model is that we will use to make probability precise, we first establish the *structure* on which the model operates: *sample spaces* and *events*.

Definition $1.1.2$

The ***sample space*** for a probability experiment (i.e., an experiment with random outcomes) is the set of all possible outcomes.

- The sample space is denoted Ω.
- An ***outcome*** is an *element* of Ω, generally denoted $\omega \in \Omega$.

Example $1.1.1$

Suppose we toss a coin twice and record the sequence of heads (h) and tails (t). A possible outcome of this experiment is then given by

$$\omega = ht$$

and the sample space is

$$\Omega = \{hh, ht, th, tt\}. \tag{1.1.1}$$

Example $1.1.2$

Suppose we record the time (t), in minutes, that a car spends waiting for a green light at a particular intersection. A possible outcome of this experiment is then given by

$$t = 1.5,$$

indicating that a particular car waited one and a half minutes for the light to turn green. The sample space consists of all non-

negative numbers, since a measurement of time cannot be negative and, in theory, there is no limit on how a long a car could wait for a green light. We can then write the sample space as follows:

$$\Omega = \{t \in \mathbb{R} \mid t \geq 0\} = [0, \infty). \tag{1.1.2}$$

Definition 1.1.3

An ***event*** is a particular subset of the sample space.

Example 1.1.3

Continuing in the context of Example 1.1.1, define A to be the event that at least one heads is recorded. We can write event A as the following subset of the sample space:

$$A = \{hh, ht, th\}.$$

Note that A is a subset of Ω given in Equation 1.1.1.

Example 1.1.4

Continuing in the context of Example 1.1.2, define B to be the event that a car waits at most 2 minutes for the light to turn green. We can write the event B as the following interval, i.e., a subset of the sample space Ω given in Equation 1.1.2:

$$B = [0, 2] = \{t \in \mathbb{R} \mid 0 \leq t \leq 2\}.$$

Set Theory: A Brief Review

As we see from the above definitions of sample spaces and events, *sets* play the primary role in the structure of probability experiments. So, in this section, we review some of the basic definitions and notation from *set theory*. We do this in the context of sample spaces, outcomes, and events.

Definition 1.1.4

1. The ***union*** of two events A and B, denoted $A \cup B$, is the set of all outcomes in A or B (*or both*).
2. The ***intersection*** of two events A and B, denoted $A \cap B$, is the set of all outcomes in both A and B.
3. The ***complement*** of an event A, denoted A^c, is the set of all outcomes in the sample space that are not in A. This may also be written as follows:

$$A^c = \{\omega \in \Omega \mid s \notin A\}.$$

4. The ***empty set***, denoted $\varnothing$, is the set containing no outcomes.
5. Two events A and B are ***disjoint*** (or ***mutually exclusive***) if their intersection is the empty set, i.e., $A \cap B = \varnothing$.

Example 1.1.5

Continuing in the context of both Examples 1.1.1 & 1.1.3, define B to be the event that exactly one heads is recorded:

$$B = \{ht, th\}.$$

Now we can apply the set operations just defined to the events A and B:

$$A \cup B = \{hh, ht, th\} = A$$

$$A \cap B = \{ht, th\} = B$$

$$A^c = \{tt\}$$

$$B^c = \{hh, tt\}$$

Note the relationship between events A and B: every outcome in B is an outcome in A. In this case, we say that B is a ***subset*** of A, and write

$$B \subseteq A.$$

Note also that events A and B are not disjoint, since their intersection is not the empty set. However, if we let C be the event that no heads are recorded, then

$$C = \{tt\},$$

and

$$A \cap C = \varnothing$$

$$B \cap C = \varnothing.$$

Thus, events A and C are disjoint, and events B and C are disjoint.

This page titled 1.1: Sample Spaces and Events is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

1.2: Probability Measures

Now we are ready to formally define probability.

Definition $1.2.1$

A ***probability measure*** on the sample space Ω is a function, denoted P, from subsets of Ω to the real numbers $\mathbb{R}$, such that the following hold:

1. $P(\Omega) = 1$
2. If A is any event in Ω, then $P(A) \geq 0$.
3. If events A_1 and A_2 are disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.
   More generally, if $A_1, A_2, \ldots, A_n, \ldots$ is a sequence of *pairwise disjoint* events, i.e., $A_i \cap A_j = \varnothing$, for every $i \neq j$, then
   $$P(A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_n) + \cdots.$$

So essentially, we are defining probability to be an ***operation*** on the events of a sample space, which assigns numbers to events in such a way that the three properties stated in Definition 1.2.1 are satisfied.

Definition 1.2.1 is often referred to as the ***axiomatic definition of probability***, where the three properties give the three **axioms** of probability. These three axioms are all we need to assume about the operation of probability in order for many other desirable properties of probability to hold, which we now state.

Properties of Probability Measures

Let Ω be a sample space with probability measure P. Also, let A and B be any events in Ω. Then the following hold.

1. $P(A^c) = 1 - P(A)$
2. $P(\varnothing) = 0$
3. If $A \subseteq B$, then $P(A) \leq P(B)$.
4. $P(A) \leq 1$
5. **Addition Law:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Exercise $1.2.1$

Can you prove the five properties of probability measures stated above using only the three axioms of probability measures stated in Definition 1.2.1?

Answer

(1) For the first property, note that by definition of the complement of an event A we have

$$A \cup A^c = \Omega \quad \text{and} \quad A \cap A^c = \varnothing.$$

In other words, given any event A, we can represent the sample space Ω as a disjoint union of A with its complement. Thus, by the first and third axioms, we derive the first property:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

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

(2) For the second property, note that we can write $\Omega = \Omega \cup \varnothing$, and that this is a disjoint union, since anything intersected with the empty set will necessarily be empty. So, using the first and third axioms, we derive the second property:

$$1 = P(\Omega) = P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + P(\varnothing)$$

$$\Rightarrow P(\varnothing) = 0$$

(3) For the third property, note that we can write $B = A \cup (B \cap A^c)$, and that this is a disjoint union, since A and A^c are disjoint. By the third axiom, we have

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c). \tag{1.2.1}$$

By the second axiom, we know that $P(B \cap A^c) \geq 0$. Thus, if we remove it from the right-hand side of Equation 1.2.1, we are left with something smaller, which proves the third property:

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

(4) For the fourth property, we will use the third property that we just proved. By definition, any event A is a subset of the sample space Ω, i.e., $A \subseteq \Omega$. Thus, by the third property and the first axiom, we derive the fourth property:

$$P(A) \leq P(\Omega) = 1 \quad \Rightarrow \quad P(A) \leq 1$$

(5) For the fifth property, note that we can write the union of events A and B as the union of the following two disjoint events:

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

in other words, the union of A and B is given by the union of all the outcomes in A with all the outcomes in B that are *not* in B. Furthermore, note that event B can be written as the union the following two disjoint events:

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

in other words, B is written as the disjoint union of all the outcomes in B that are also in A with the outcomes in B that are *not* in A. We can use this expression for B to find an expression for $P(A^c \cap B)$ to substitute in the expression for $A \cup B$ in order to derive the fifth property:

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

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

Note that the axiomatic definition (Definition 1.2.1) does not tell us how to *compute* probabilities. It simply defines a formal, mathematical behavior of probability. In other words, the axiomatic definition describes how probability should theoretically *behave* when applied to events. To compute probabilities, we use the properties stated above, as the next example demonstrates.

Example 1.2.1

Continuing in the context of Example 1.1.5, let's define a probability measure on Ω. Assuming that the coin we toss is *fair*, then the outcomes in Ω are **equally likely**, meaning that each outcome has the *same probability* of occurring. Since there are four outcomes, and we know that probability of the sample space must be 1 (first axiom of probability in Definition 1.2.1), it follows that the probability of each outcome is $\frac{1}{4} = 0.25$.

So, we can write

$$P(hh) = P(ht) = P(th) = P(tt) = 0.25.$$

The reader can verify this defines a probability measure satisfying the three axioms.

With this probability measure on the outcomes we can now compute the probability of any event in Ω by simply *counting* the number of outcomes in the event. Thus, we find the probability of events A and B previously defined:

$$P(A) = P(\{hh, ht, th\}) = \frac{3}{4} = 0.75$$

$$P(B) = P(\{ht, th\}) = \frac{2}{4} = 0.50.$$

We consider the case of equally likely outcomes further in the next section: Section 1.3.

There is another, more empirical, approach to defining probability, given by using *relative frequencies* and a version of the Law of Large Numbers.

Relative Frequency Approximation

To *estimate* the probability of an event A, repeat the random experiment several times (each repetition is called a *trial*) and count the number of times A occurred, i.e., the number of times the resulting outcome is in A. Then, we approximate the probability of A using **relative frequency**:

$$P(A) \approx \frac{\text{number of times } A \text{ occurred}}{\text{number of trials}}.$$

Law of Large Numbers

As the number of trials increases, the relative frequency approximation approaches the theoretical value of $P(A)$.

This approach to defining probability is sometimes referred to as the ***frequentist definition of probability***. Under this definition, probability represents a *long-run average*. The two approaches to defining probability are equivalent. It can be shown that using relative frequencies to define a probability measure satisfies the axiomatic definition.

This page titled 1.2: Probability Measures is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

1.3: Equally Likely Outcomes and Counting Techniques (Combinatorics)

In this section, we consider the problem of assigning specific probabilities to outcomes in a sample space. As we saw in Section 1.2, the axiomatic definition of probability (Definition 1.2.1) does not tell us how to compute probabilities. So in this section we consider the commonly encountered scenario referred to as *equally likely outcomes* and develop methods for computing probabilities in this special case.

Finite Sample Spaces

Before focusing on equally likely outcomes, we consider the more general case of *finite* sample spaces. In other words, suppose that a sample space Ω has a finite number of outcomes, which we can denote as N. In this case, we can represent the outcomes in Ω as follows:

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}.$$

Suppose further that we denote the probability assigned to each outcome in Ω as $P(\omega_i) = p_i$, for $i = 1, \dots, N$. Then the probability of any event A in Ω is given by adding the probabilities corresponding to the outcomes contained in A and we can write

$$P(A) = \sum_{\omega_i \in A} p_i. \tag{1.3.1}$$

This follows from the third axiom of probability (Definition 1.2.1), since we can write any event as a disjoint union of the outcomes contained in the event. For example, if event A contains three outcomes, then we can write $A = \{\omega_1, \omega_2, \omega_3\} = \{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\}$. So the probability of A is given by simply summing up the probabilities assigned to $\omega_1, \omega_2, \omega_3$. This fact will be useful in the special case of equally likely outcomes, which we consider next.

Equally Likely Outcomes

First, let's state a formal definition of what it means for the outcomes in a sample space to be *equally likely*.

Definition $1.3.1$

The outcomes in a sample space Ω are ***equally likely*** if each outcome has the *same probability* of occurring.

In general, if outcomes in a sample space Ω are equally likely, then computing the probability of a single outcome or an event is very straightforward, as the following exercise demonstrates. You are encouraged to first try to answer the questions for yourself, and then click "Answer" to see the solution.

Exercise $1.3.1$

Suppose that there are N outcomes in the sample space Ω and that the outcomes are equally likely.

a. What is the probability of a single outcome in Ω?
b. What is the probability of an event A in Ω?

Answer

First, note that we can represent the outcomes in Ω as follows:

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}.$$

For each outcome in Ω, note that we can denote its probability as

$$P(\omega_i) = c, \text{ for } i = 1, 2, \dots, N,$$

where c is some constant. This follows from the fact that the outcomes of Ω are equally likely and so have the same probability of occurring. With this set-up and using the axioms of probability (Definition 1.2.1), we have the following:

$$\begin{aligned} 1 = P(\Omega) &= P(\{\omega_1\} \cup \dots \cup \{\omega_N\}) \\ &= P(\omega_1) + \dots + P(\omega_N) \\ &= c + \dots + c \\ &= N \times c \\ \Rightarrow c &= \frac{1}{N}. \end{aligned}$$

Thus, the probability of a single outcome is given by 1 divided by the number of outcomes in Ω.

Now, for an event A in Ω, suppose it has n outcomes, where n is an integer such that $0 \leq n \leq N$. We can represent the outcomes in A as follows:

$$A = \{a_1, \dots, a_n\}.$$

Using Equation 1.3.1, we compute the probability of A as follows:

$$\begin{aligned} P(A) = \sum_{i=1}^{n} P(a_i) &= \sum_{i=1}^{n} \frac{1}{N} \\ &= \frac{1}{N} + \dots + \frac{1}{N} \\ &= n\left(\frac{1}{N}\right) \\ &= \frac{n}{N}. \end{aligned}$$

Thus, the probability of an event in Ω is equal to the number of outcomes in the event divided by the total number of outcomes in Ω.

We have already seen an example of a sample space with equally likely outcomes in Example 1.2.1. You are encouraged to revisit that example and connect it to the results of Exercise 1.3.1.

In general, Exercise 1.3.1 shows that if a sample space Ω has equally likely outcomes, then the probability of an event A in the sample space is given by

$$\boxed{P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega}.} \tag{1.3.2}$$

From this result, we see that in the context of equally likely outcomes calculating probabilities of events reduces to simply *counting* the number of outcomes in the event and the sample space. So, we take a break from our discussion of probability, and briefly introduce some counting techniques.

Counting Techniques

First, let's consider the general context of performing multi-step experiments. The following tells us how to count the number of outcomes in such scenarios.

Multiplication Principle

If one probability experiment has m outcomes and another has n outcomes, then there are $m \times n$ total outcomes for the two experiments.

More generally, if there are k many probability experiments with the first experiment having n_1 outcomes, the second with n_2, etc., then there are $n_1 \times n_2 \times \cdots \times n_k$ total outcomes for the k experiments.

Example $1.3.1$

To demonstrate the Multiplication Principle, consider again the example of tossing a coin twice (see Example 1.2.1). Each toss is a probability experiment and on each toss, there are two possible outcomes: h or t. Thus, for two tosses, there are $2 \times 2 = 4$ total outcomes.

If we toss the coin a third time, there are $2 \times 2 \times 2 = 8$ total outcomes.

Next we define two commonly encountered situations, *permutations* and *combinations*, and consider how to count the number of ways in which they can occur.

Definition $1.3.2$

A ***permutation*** is an *ordered* arrangement of objects. For example, "MATH" is a permutation of four letters from the alphabet.

A ***combination*** is an *unordered* collection of r objects from n total objects. For example, a group of three students chosen from a class of 10 students.

In order to count the number of possible permutations in a given setting, the Multiplication Principle is applied. For example, if we want to know the number of possible permutations of the four letters in "MATH", we compute

$$4 \times 3 \times 2 \times 1 = 4! = 24,$$

since there are four letters to select for the first position, three letters for the second, two for the third, leaving only one letter for the last. In other words, we treat each letter selection as an experiment in a multi-step process.

Counting Permutations

The number of permutations of n distinct objects is given by the following:

$$n \times (n-1) \times \cdots \times 2 \times 1 = n!$$

Counting combinations is a little more complicated, since we are not interested in the order in which objects are selected and so the Multiplication Principle does not directly apply. Consider the example that a group of three students are chosen from a class of 10. The group is the same regardless of the order in which the three students are selected. This implies that if we want to count the number of possible combinations, we need to be careful not to include permutations, i.e., rearrangements, of a certain selection. This leads to the following result that the number of possible combinations of size r selected from a total of n objects is given by binomial coefficients.

Counting Combinations

The number of combinations of r objects selected without replacement from n distinct objects is given by

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

Note that $\binom{n}{r}$, read as "n *choose* r", is also referred to as a ***binomial coefficient***, since it appears in the Binomial Theorem.

Using the above, we can compute the number of possible ways to select three students from a class of 10:

$$\binom{10}{3} = \frac{10!}{3! \times 7!} = 120$$

Example $1.3.2$

Consider the example of tossing a coin three times. Note that an outcome is a sequence of heads and tails. Suppose that we are interested in the number of outcomes with exactly two heads, not in the actual sequence. To find the number of outcomes with exactly two heads, we need to determine the number of ways to select positions in the sequence for the heads, then the remaining position will be a tails. If we toss the coin three times, there are three positions to select from, and we want to select two. Since the order that we make the selection of placements does not matter, we are counting the number of combinations of 2 positions from a total of 3 positions, i.e.,

$$\binom{3}{2} = \frac{3!}{2! \times 1!} = 3.$$

Of course, this example is small enough that we could have arrived at the answer of 3 using brute force by just listing the possibilities. However, if we toss the coin a higher number of times, say 50, then the brute force approach becomes infeasible and we need to make use of binomial coefficients.

This page titled 1.3: Equally Likely Outcomes and Counting Techniques (Combinatorics) is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

2: CONDITIONAL PROBABILITY

2.1: Conditional Probability and Bayes' Rule
2.2: Independent Events

This page titled 2: Conditional Probability is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

2.1: Conditional Probability and Bayes' Rule

In many situations, additional information about the result of a probability experiment is known (or at least assumed to be known) and given that information the probability of some other event is desired. For this scenario, we compute what is referred to as *conditional probability*.

Definition $2.1.1$

For events A and B, with $P(B) > 0$, the ***conditional probability*** of A given B, denoted $P(A \mid B)$, is given by

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

In computing a conditional probability we *assume* that we know the outcome of the experiment is in event B and then, given that additional information, we calculate the probability that the outcome is also in event A. This is useful in practice given that partial information about the outcome of an experiment is often known, as the next example demonstrates.

Example $2.1.1$

Continuing in the context of Example 1.2.1, where we considered tossing a fair coin twice, define D to be the event that at least one tails is recorded:

$$D = \{ht, th, tt\}$$

Let's calculate the conditional probability of A given D, i.e., the probability that at least one heads is recorded (event A) assuming that at least one tails is recorded (event D). Recalling that outcomes in this sample space are equally likely, we apply the definition of conditional probability (Definition 2.1.1) and find

$$P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{P(\{ht, th\})}{P(\{ht, th, tt\})} = \frac{(2/4)}{(3/4)} = \frac{2}{3} \approx 0.67.$$

Note that in Example 1.2.1 we found the ***un***-conditional probability of A to be $P(A) = 0.75$. So, knowing that at least one tails was recorded, i.e., assuming event D occurred, the conditional probability of A given D decreased. This is because, if event D occurs, then the outcome hh in A cannot occur, thereby decreasing the chances that event A occurs.

Exercise $2.1.1$

Suppose we randomly draw a card from a standard deck of 52 playing cards.

a. If we know that the card is a King, what is the probability that the card is a club?
b. If we instead know that the card is black, what is the probability that the card is a club?

Answer

In order to compute the necessary probabilities, first note that the sample space is given by the set of cards in a standard deck of playing cards. So the number of outcomes in the sample space is 52. Next, note that the outcomes are equally likely, since we are *randomly* drawing the card from the deck.

For part (a), we are looking for the conditional probability that the randomly selected card is club, given that it is a King. If we let C denote the event that the card is a club and K the event that it is a King, then we are looking to compute

$$P(C \mid K) = \frac{P(C \cap K)}{P(K)}. \tag{2.1.1}$$

To compute these probabilities, we count the number of outcomes in the following events:

Misplaced '#'

Misplaced '#'

Misplaced '#'

The probabilities in Equation 2.1.1 are then given by dividing the counts of outcomes in each event by the total number of outcomes in the sample space (by the boxed Equation 1.3.2 in Section 1.3):

$$P(C \mid K) = \frac{P(C \cap K)}{P(K)} = \frac{(1/52)}{(4/52)} = \frac{1}{4} = 0.25.$$

For part (b), we are looking for the conditional probability that the randomly selected card is club, given that it is instead black. If we let B denote the event that the card is black, then we are looking to compute

$$P(C \mid B) = \frac{P(C \cap B)}{P(B)}. \tag{2.1.2}$$

To compute these probabilities, we count the number of outcomes in the following events:

Misplaced '#'

Misplaced '#'

The probabilities in Equation 2.1.2 are then given by dividing the counts of outcomes in each event by the total number of outcomes in the sample space:

$$P(C \mid B) = \frac{P(C \cap B)}{P(B)} = \frac{(13/52)}{(26/52)} = \frac{13}{26} = 0.5.$$

Remark: Exercise 2.1.1 demonstrates the following fact. For sample spaces with equally likely outcomes, conditional probabilities are calculated using

$$\boxed{P(A \mid B) = \frac{\text{number of outcomes in } A \cap B}{\text{number of outcomes in } B}.} \tag{2.1.3}$$

In other words, if we know that the outcome of the probability experiment is in the event B, then we restrict our focus to the outcomes in that event that are also in A. We can think of this as event B taking the place of the sample space, since we know the outcome must lie in that event.

Properties of Conditional Probability

As with unconditional probability, we also have some useful properties for conditional probabilities. The first property below, referred to as the *Multiplication Law*, is simply a rearrangement of the probabilities used to define conditional probability. The Multiplication Law provides a way for computing the probability of an intersection of events when the conditional probabilities

are known.

Multiplication Law

$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$

The next two properties are useful when a *partition* of the sample space exists, where a partition is a way of dividing up the outcomes in the sample space into non-overlapping sets. A partition is formally defined in the *Law of Total Probability* below. In many cases, when a partition exists, it is easy to compute the conditional probability of an event in the sample space given an event in the partition. The Law of Total Probability then provides a way of using those conditional probabilities of an event, given the partition to compute the unconditional probability of the event. Following the Law of Total Probability, we state *Bayes' Rule*, which is really just an application of the Multiplication Law. Bayes' Rule is used to calculate what are informally referred to as "reverse conditional probabilities", which are the conditional probabilities of an event in a partition of the sample space, given any other event.

Law of Total Probability

Suppose events $B_1, B_2, \dots, B_n$, satisfy the following:

1. $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n$
2. $B_i \cap B_j = \varnothing$, for every $i \neq j$
3. $P(B_i) > 0$, for $i = 1, \dots, n$

We say that the events $B_1, B_2, \dots, B_n$, **partition** the sample space Ω. Then for any event A, we can write

$$P(A) = P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_n)P(B_n).$$

Bayes' Rule

Let $B_1, B_2, \dots, B_n$, partition the sample space Ω and let A be an event with $P(A) > 0$. Then, for $j = 1, \dots, n$, we have

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{P(A)}.$$

A common application of the Law of Total Probability and Bayes' Rule is in the context of medical diagnostic testing.

Example $2.1.2$

Consider a test that can diagnose kidney cancer. The test correctly detects when a patient has cancer 90% of the time. Also, if a person does not have cancer, the test correctly indicates so 99.9% of the time. Finally, suppose it is known that 1 in every 10,000 individuals has kidney cancer. We find the probability that a patient has kidney cancer, given that the test indicates she does.

First, note that we are finding a conditional probability. If we let A denote the event that the patient tests positive for cancer, and we let B_1 denote the event that the patient actually has cancer, then we want

$$P(B_1 \mid A).$$

If we let $B_2 = B_1^c$, then we have a partition of all patients (which is the sample space) given by B_1 and B_2.

In the first paragraph of this example, we are given the following probabilities:

$$\text{test correctly detects cancer 90} \quad P(A \mid B_1) = 0.9$$

$$\text{test correctly detects no cancer 99.9} \quad P(A^c \mid B_2) = 0.999 \Rightarrow P(A \mid B_2) = 1 - P(A^c \mid B_2) = 0.001$$

$$\text{1 in every 10,000 individuals has cancer:} \quad P(B_1) = 0.0001 \Rightarrow P(B_2) = 1 - P(B_1) = 0.9999$$

Since we have a partition of the sample space, we apply the Law of Total Probability to find $P(A)$:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) = (0.9)(0.0001) + (0.001)(0.9999) = 0.0010899$$

Next, we apply Bayes' Rule to find the desired conditional probability:

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A)} = \frac{(0.9)(0.0001)}{0.0010899} \approx 0.08$$

This implies that only about 8% of patients that test positive under this particular test actually have kidney cancer, which is not very good.

This page titled 2.1: Conditional Probability and Bayes' Rule is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

2.2: Independent Events

In this section we consider a property of events that relates to conditional probability, namely *independence*. First, we define what it means for a pair of events to be independent, and then we consider collections of more than two events.

Independence for Pairs of Events

The following definition provides an intuitive definition of the concept of independence for two events, and then we look at an example that provides a computational way for determining when events are independent.

Definition $2.2.1$

Events A and B are ***independent*** if knowing that one occurs does not affect the probability that the other occurs, i.e.,

$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(B). \tag{2.2.1}$$

Using the definition of conditional probability (Definition 2.2.1), we can derive an alternate way to the Equations $2.2.1$ for determining when two events are independent, as the following example demonstrates.

Example $2.2.1$

Suppose that events A and B are independent. We rewrite Equations $2.2.1$ using the definition of conditional probability:

$$P(A \mid B) = P(A) \quad \Rightarrow \quad \frac{P(A \cap B)}{P(B)} = P(A)$$

$$\text{and}$$

$$P(B \mid A) = P(B) \quad \Rightarrow \quad \frac{P(A \cap B)}{P(A)} = P(B)$$

In each of the expressions on the right-hand side above we isolate $P(A \cap B)$:

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

$$\text{and}$$

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

Both expressions result in $P(A \cap B) = P(A)P(B)$. Thus, we have shown that if events A and B are independent, then the

probability of their intersection is equal to the product of their individual probabilities. We state this fact in the next definition.

Definition $2.2.2$

Events A and B are ***independent*** if

$$P(A \cap B) = P(A)P(B).$$

Generally speaking, Definition 2.3.2 tends to be an easier condition than Definition 2.3.1 to verify when checking whether two events are independent.

Example $2.2.2$

Consider the context of Exercise 2.2.1, where we randomly draw a card from a standard deck of 52 and C denotes the event of drawing a club, K the event of drawing a King, and B the event of drawing a black card.

Are C and K independent events? Recall that $P(C \cap K) = 1/52$, and note that $P(C) = 13/52$ and $P(K) = 4/52$. Thus, we have

$$P(C \cap K) = \frac{1}{52} = P(C)P(K) = \frac{13}{52} \times \frac{4}{52},$$

indicating that C and K **are independent**.

Are C and B independent events? Recall that $P(C \cap B) = 13/52$, and note that $P(B) = 26/52$. Thus, we have

$$P(C \cap B) = \frac{13}{52} \neq P(C)P(B) = \frac{13}{52} \times \frac{26}{52},$$

indicating that C and B **are *not* independent**.

Let's think about the results of this example intuitively. To say that C and K are independent means that knowing that one of the events occurs does not affect the probability of the other event occurring. In other words, knowing that the card drawn is a King does not influence the probability of the card being a club. The proportion of clubs in the entire deck of 52 is the same as the proportion of clubs in just the collection of Kings: $1/4$. On the other hand, C and B are not independent (AKA *dependent*) because knowing that the card drawn is club indicates that the card *must be black*, i.e., the probability that the card is black is 1. Alternately, knowing that the card drawn is black increases the probability that the card is a club, since the proportion of clubs in the entire deck is $1/4$, but the proportion of clubs in the collection of black cards is $1/2$.

Independence for 3 or More Events

For collections of 3 or more events, there are two different types of independence.

Definition $2.2.3$

Let $A_1, A_2, \ldots, A_k$, where $k \geq 3$, be a collection of events.

1. The events are ***pairwise independent*** if every pair of events in the collection is independent.
2. The events are ***mutually independent*** if every sub-collection of events, say $A_{i_1}, A_{i_2}, \ldots, A_{i_n}$, satisfy the following:

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}) = P(A_{i_1}) \times P(A_{i_2}) \times \ldots \times P(A_{i_n})$$

Mutually independent is a stronger type of independence, since it *implies* pairwise independent. But pairwise independence does NOT imply mutual independence, as the following example will demonstrate.

Example $2.2.3$

Consider again the context of Example 1.1.1, i.e., tossing a fair coin twice, and define the following events:

$$A = \text{first toss is heads}$$

$$B = \text{second toss is heads}$$

$$C = \text{exactly one head is recorded}$$

We show that this collection of events - A, B, C - is pairwise independent, but NOT mutually independent. First, we note that the individual probabilities of each event are 0.5:

$$P(A) = P(\{hh, ht\}) = 0.5$$

$$P(B) = P(\{hh, th\}) = 0.5$$

$$P(C) = P(\{ht, th\}) = 0.5$$

Next, we look at the probabilities of all pairwise intersections to establish pairwise independence:

$$P(A \cap B) = P(hh) = 0.25 = P(A)P(B)$$

$$P(A \cap C) = P(ht) = 0.25 = P(A)P(C)$$

$$P(B \cap C) = P(th) = 0.25 = P(B)P(C)$$

However, note that the three events do not have any outcomes in common, i.e., $A \cap B \cap C = \varnothing$. Thus, we have

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C),$$

and so the events are not mutually independent.

This page titled 2.2: Independent Events is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3: DISCRETE RANDOM VARIABLES

3.1: Introduction to Random Variables

3.2: Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) for Discrete Random Variables

3.3: Bernoulli and Binomial Distributions

3.4: Expected Value of Discrete Random Variables

3.5: Variance of Discrete Random Variables

This page titled 3: Discrete Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3.1: Introduction to Random Variables

Now that we have formally defined probability and the underlying structure, we add another layer: *random variables*. Random variables allow characterization of outcomes, so that we do not need to focus on each outcome specifically. We begin with the formal definition.

Definition $3.1.1$

A ***random variable*** is a function from a sample space Ω to the real numbers $\mathbb{R}$. We denote random variables with capital letters, e.g.,

$$X : \Omega \to \mathbb{R}.$$

Informally, a random variable assigns numbers to outcomes in the sample space. So, instead of focusing on the outcomes themselves, we highlight a specific characteristic of the outcomes.

Example $3.1.1$

Consider again the context of Example 1.1.1, where we recorded the sequence of heads and tails in two tosses of a fair coin. The sample space for this random experiment is given by

$$\Omega = \{hh, ht, th, tt\}.$$

Suppose we are only interested in tosses that result in heads. We can define a random variable X that tracks the number of heads obtained in an outcome. So, if outcome hh is obtained, then X will equal 2. Formally, we denote this as follows:

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto \text{ number of } h\text{'s in } \omega \end{aligned}$$

Since there are only four outcomes in Ω, we can list the value of X for each outcome individually:

$$\begin{array}{ccc} \text{inputs: } \Omega & \xrightarrow{\text{function: } X} & \text{outputs: } \mathbb{R} \\ hh & \overset{X}{\mapsto} & 2 \\ th & \overset{X}{\mapsto} & 1 \\ ht & \overset{X}{\mapsto} & 1 \\ tt & \overset{X}{\mapsto} & 0 \end{array}$$

We can also write the above as follows:

$$X(hh) = 2, \quad X(ht) = X(th) = 1, \quad X(tt) = 0.$$

The advantage to defining the random variable X in this context is that the two outcomes ht and th are both assigned a value of 1, meaning we are not focused on the actual sequence of heads and tails that resulted in obtaining one heads.

In Example 3.1.1, note that the random variable we defined only equals one of three possible values: $0, 1, 2$. This is an example of what we call a *discrete* random variable. We will also encounter another type of random variable: *continuous*. The next definitions make precise what we mean by these two types.

Definition $3.1.2$

A ***discrete random variable*** is a random variable that has only a finite or countably infinite (think integers or whole numbers) number of possible values.

Definition $3.1.3$

A ***continuous random variable*** is a random variable with infinitely many possible values (think an interval of real numbers, e.g., $[0, 1]$).

In this chapter, we take a closer look at discrete random variables, then in Chapter 4 we consider continuous random variables.

This page titled 3.1: Introduction to Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3.2: Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) for Discrete Random Variables

Since random variables simply assign values to outcomes in a sample space and we have defined probability measures on sample spaces, we can also talk about probabilities for random variables. Specifically, we can compute the probability that a discrete random variable equals a specific value (*probability mass function*) and the probability that a random variable is less than or equal to a specific value (*cumulative distribution function*).

Probability Mass Functions (PMFs)

In the following example, we compute the probability that a discrete random variable equals a specific value.

Example $3.2.1$

Continuing in the context of Example 3.1.1, we compute the *probability* that the random variable X equals 1. There are two outcomes that lead to X taking the value 1, namely ht and th. So, the probability that $X = 1$ is given by the probability of the event ht, th, which is 0.5:

Misplaced '#'

In Example 3.2.1, the probability that the random variable X equals 1, $P(X = 1)$, is referred to as the *probability mass function* of X evaluated at 1. In other words, the specific value 1 of the random variable X is associated with the probability that X equals that value, which we found to be 0.5. The process of assigning probabilities to specific values of a discrete random variable is what the probability mass function is and the following definition formalizes this.

Definition $3.2.1$

The ***probability mass function (pmf)*** (or ***frequency function***) of a discrete random variable X assigns probabilities to the possible values of the random variable. More specifically, if $x_1, x_2, \ldots$ denote the possible values of a random variable X, then the probability mass function is denoted as p and we write

$$p(x_i) = P(X = x_i) = P(\underbrace{\{\omega \in \Omega \mid X(s) = x_i\}}_{\text{set of outcomes resulting in } X=x_i}). \tag{3.2.1}$$

Note that, in Equation 3.2.1, $p(x_i)$ is *shorthand* for $P(X = x_i)$, which represents the probability of the *event* that the random variable X equals x_i.

As we can see in Definition 3.2.1, the probability mass function of a random variable X depends on the probability measure of the underlying sample space Ω. Thus, pmf's inherit some properties from the axioms of probability (Definition 1.2.1). In fact, in order for a function to be a *valid* pmf it must satisfy the following properties.

Properties of Probability Mass Functions

Let X be a discrete random variable with possible values denoted $x_1, x_2, \ldots, x_i, \ldots$. The probability mass function of X, denoted p, must satisfy the following:

1. $\sum_{x_i} p(x_i) = p(x_1) + p(x_2) + \cdots = 1$
2. $p(x_i) \geq 0$, for all x_i

Furthermore, if A is a subset of the possible values of X, then the probability that X takes a value in A is given by

$$P(X \in A) = \sum_{x_i \in A} p(x_i). \tag{3.2.2}$$

Note that the first property of pmf's stated above follows from the first axiom of probability, namely that the probability of the sample space equals 1: $P(\Omega) = 1$. The second property of pmf's follows from the second axiom of probability, which states that all probabilities are non-negative.

We now apply the formal definition of a pmf and verify the properties in a specific context.

Example $3.2.2$

Returning to Example 3.2.1, now using the notation of Definition 3.2.1, we found that the pmf for X at 1 is given by

$$p(1) = P(X = 1) = P(\{ht, th\}) = 0.5.$$

Similarly, we find the pmf for X at the other possible values of the random variable:

$$p(0) = P(X = 0) = P(\{tt\}) = 0.25$$
$$p(2) = P(X = 2) = P(\{hh\}) = 0.25$$

Note that all the values of p are positive (second property of pmf's) and $p(0) + p(1) + p(2) = 1$ (first property of pmf's). Also, we can demonstrate the third property of pmf's (Equation $3.2.2$) by computing the probability that there is at least one heads, i.e., $X \geq 1$, which we could represent by setting $A = \{1, 2\}$ so that we want the probability that X takes a value in A:

$$P(X \geq 1) = P(X \in A) = \sum_{x_i \in A} p(x_i) = p(1) + p(2) = 0.5 + 0.25 = 0.75$$

We can represent probability mass functions *numerically* with a table, *graphically* with a **histogram**, or *analytically* with a formula. The following example demonstrates the numerical and graphical representations. In the next three sections, we will see examples of pmf's defined analytically with a formula.

Example $3.2.3$

We represent the pmf we found in Example 3.2.2 in two ways below, numerically with a table on the left and graphically with a histogram on the right.

x	$p(x)$
0	0.25
1	0.5
2	0.25

Table 1: Frequency function of X

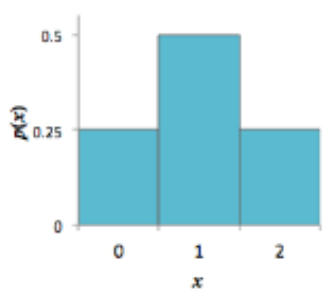


Figure 1: Histogram of X

In the histogram in Figure 1, note that we represent probabilities as *areas of rectangles*. More specifically, each rectangle in the histogram has width 1 and height equal to the probability of the value of the random variable X that the rectangle is centered over. For example, the leftmost rectangle in the histogram is centered at 0 and has height equal to $p(0) = 0.25$, which is also the area of the rectangle since the width is equal to 1. In this way, histograms provides a visualization of the *distribution* of the

probabilities assigned to the possible values of the random variable X. This helps to explain where the common terminology of "probability distribution" comes from when talking about random variables.

Cumulative Distribution Functions (CDFs)

There is one more important function related to random variables that we define next. This function is again related to the probabilities of the random variable equaling specific values. It provides a shortcut for calculating many probabilities at once.

Definition $3.2.2$

The ***cumulative distribution function* (*cdf*)** of a random variable X is a function on the real numbers that is denoted as F and is given by

$$F(x) = P(X \leq x), \quad \text{for any } x \in \mathbb{R}. \tag{3.2.3}$$

Before looking at an example of a cdf, we note a few things about the definition.

First of all, note that we did not specify the random variable X to be discrete. CDFs are also defined for continuous random variables (see Chapter 4) in exactly the same way.

Second, the cdf of a random variable is defined for *all* real numbers, unlike the pmf of a discrete random variable, which we only define for the possible values of the random variable. Implicit in the definition of a pmf is the assumption that it equals 0 for all real numbers that are not possible values of the discrete random variable, which should make sense since the random variable will never equal that value. However, cdf's, for *both* discrete and continuous random variables, are defined for all real numbers. In looking more closely at Equation $3.2.3$, we see that a cdf F considers an upper bound, $x \in \mathbb{R}$, on the random variable X, and assigns that value x to the probability that the random variable X is less than or equal to that upper bound x. This type of probability is referred to as a *cumulative probability*, since it could be thought of as the probability accumulated by the random variable up to the specified upper bound. With this interpretation, we can represent Equation $3.2.3$ as follows:

$$F: \underbrace{\mathbb{R}}_{\text{upper bounds on RV } X} \longrightarrow \underbrace{\mathbb{R}}_{\text{cumulative probabilities}} \tag{3.2.4}$$

In the case that X is a discrete random variable, with possible values denoted $x_1, x_2, \ldots, x_i, \ldots$, the cdf of X can be calculated using the third property of pmf's (Equation $3.2.2$), since, for a fixed $x \in \mathbb{R}$, if we let the set A contain the possible values of X that are less than or equal to x, i.e., $A = \{x_i \mid x_i \leq x\}$, then the cdf of X evaluated at x is given by

$$F(x) = P(X \leq x) = P(X \in A) = \sum_{x_i \leq x} p(x_i).$$

Example $3.2.4$

Continuing with Examples 3.2.2 and 3.2.3, we find the cdf for X. First, we find $F(x)$ for the possible values of the random variable, $x = 0, 1, 2$:

$$\begin{aligned}
F(0) &= P(X \leq 0) = P(X = 0) = 0.25 \\
F(1) &= P(X \leq 1) = P(X = 0 \text{ or } 1) = p(0) + p(1) = 0.75 \\
F(2) &= P(X \leq 2) = P(X = 0 \text{ or } 1 \text{ or } 2) = p(0) + p(1) + p(2) = 1
\end{aligned}$$

Now, if $x < 0$, then the cdf $F(x) = 0$, since the random variable X will never be negative.

If $0 < x < 1$, then the cdf $F(x) = 0.25$, since the only value of the random variable X that is less than or equal to such a value x is 0. For example, consider $x = 0.5$. The probability that X is less than or equal to 0.5 is the same as the probability that $X = 0$, since 0 is the only possible value of X less than 0.5:

$$F(0.5) = P(X \leq 0.5) = P(X = 0) = 0.25.$$

Similarly, we have the following:

$$\begin{aligned} F(x) &= F(1) = 0.75, \quad \text{for } 1 < x < 2 \\ F(x) &= F(2) = 1, \quad \text{for } x > 2 \end{aligned}$$

Exercise 3.2.1

For this random variable X, compute the following values of the cdf:

a. $F(-3)$
b. $F(0.1)$
c. $F(0.9)$
d. $F(1.4)$
e. $F(2.3)$
f. $F(18)$

Answer

a. $F(-3) = P(X \leq -3) = 0$
b. $F(0.1) = P(X \leq 0.1) = P(X = 0) = 0.25$
c. $F(0.9) = P(X \leq 0.9) = P(X = 0) = 0.25$
d. $F(1.4) = P(X \leq 1.4) = \sum_{x_i \leq 1.4} p(x_i) = p(0) + p(1) = 0.25 + 0.5 = 0.75$
e. $F(2.3) = P(X \leq 2.3) = \sum_{x_i \leq 2.3} p(x_i) = p(0) + p(1) + p(2) = 0.25 + 0.5 + 0.25 = 1$
f. $F(18) = P(X \leq 18) = P(X \leq 2) = 1$

To summarize Example 3.2.4, we write the cdf F as a *piecewise function* and Figure 2 below gives its graph:

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ 0.25 & \text{for } 0 \leq x < 1 \\ 0.75 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2. \end{cases}$$

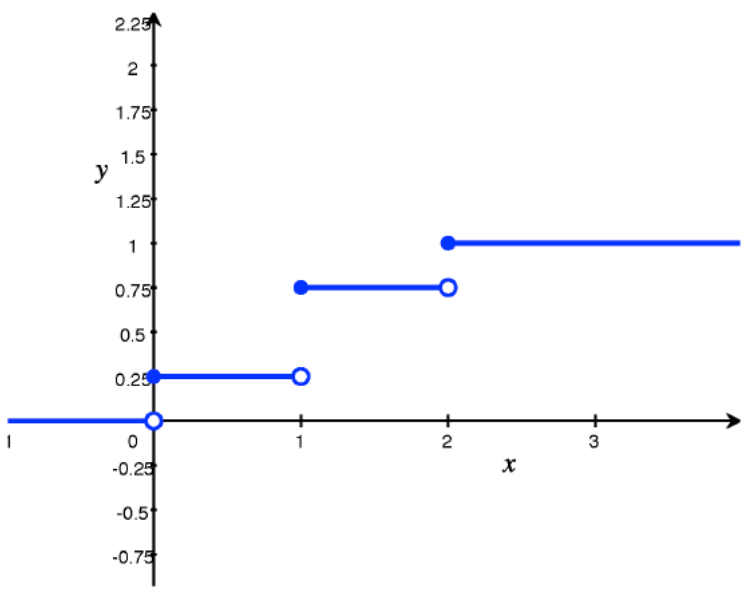


Figure 2: Graph of cdf in Example 3.2.4

Note that the cdf we found in Example 3.2.4 is a "step function", since its graph resembles a series of steps. This is the case for *all*

discrete random variables. Additionally, the value of the cdf for a discrete random variable will always "jump" at the possible values of the random variable, and the size of the "jump" is given by the value of the pmf at that possible value of the random variable. For example, the graph in Figure 2 "jumps" from 0.25 to 0.75 at $x = 1$, so the size of the "jump" is $0.75 - 0.25 = 0.5$ and note that $p(1) = P(X = 1) = 0.5$. The pmf for any discrete random variable can be obtained from the cdf in this manner.

We end this section with a statement of the properties of cdf's. The reader is encouraged to verify these properties hold for the cdf derived in Example 3.2.4 and to provide an intuitive explanation (or formal explanation using the axioms of probability and the properties of pmf's) for why these properties hold for cdf's in general.

Properties of Cumulative Distribution Functions

Let X be a random variable with cdf F. Then F satisfies the following:

1. F is non-decreasing, i.e., F may be constant, but otherwise it is increasing.
2. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

This page titled 3.2: Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) for Discrete Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3.3: Bernoulli and Binomial Distributions

In this section, we introduce two *families* of common discrete probability distributions, i.e., probability distributions for discrete random variables. We refer to these as "families" of distributions because in each case we will define a probability mass function by specifying an explicit formula, and that formula will incorporate a constant (or set of constants) that are referred to as ***parameters***. By specifying values for the parameter(s) in the pmf, we define a specific probability distribution for a specific random variable. For each family of distributions introduced, we will list a set of defining characteristics that will help determine when to use a certain distribution in a given context.

Bernoulli Distribution

Consider the following example.

Example $3.3.1$

Let A be an event in a sample space Ω. Suppose we are only interested in whether or not the outcome of the underlying probability experiment is in the specified event A. To track this we can define an **indicator random variable**, denoted I_A, given by

$$I_A(s) = \begin{cases} 1, & \text{if } s \in A, \\ 0, & \text{if } s \in A^c. \end{cases}$$

In other words, the random variable I_A will equal 1 if the resulting outcome is in event A, and I_A equals 0 if the outcome is not in A. Thus, I_A is a discrete random variable. We can state the probability mass function of I_A in terms of the probability that the resulting outcome is in event A, i.e., the probability that event A occurs, $P(A)$:

$$\begin{aligned} p(0) &= P(I_A = 0) = P(A^c) = 1 - P(A) \\ p(1) &= P(I_A = 1) = P(A) \end{aligned}$$

In Example 3.3.1, the random variable I_A is a *Bernoulli random variable* because its pmf has the form of the *Bernoulli probability distribution*, which we define next.

Definition $3.3.1$

A random variable X has a ***Bernoulli distribution*** with parameter p, where $0 \le p \le 1$, if it has only two possible values, typically denoted 0 and 1. The probability mass function (pmf) of X is given by

$$\begin{aligned} p(0) &= P(X = 0) = 1 - p, \\ p(1) &= P(X = 1) = p. \end{aligned}$$

The cumulative distribution function (cdf) of X is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases} \tag{3.3.1}$$

In Definition 3.3.1, note that the defining characteristic of the Bernoulli distribution is that it models random variables that have only two possible values. As noted in the definition, the two possible values of a Bernoulli random variable are usually 0 and 1. In the typical application of the Bernoulli distribution, a value of 1 indicates a "success" and a value of 0 indicates a "failure", where "success" refers that the event or outcome of interest. The parameter p in the Bernoulli distribution is given by the probability of a "success". In Example 3.3.1, we were interested in tracking whether or not event A occurred, and so that is what a "success" would be, which occurs with probability given by the probability of A. Thus, the value of the parameter p for the Bernoulli distribution in Example 3.3.1 is given by $p = P(A)$.

Exercise 3.3.1

Derive the general formula for the cdf of the Bernoulli distribution given in Equation 3.3.1.

Hint

First find $F(0)$ and $F(1)$.

Answer

Recall that the only two values of a Bernoulli random variable X are 0 and 1. So, first, we find the cdf at those two values:

$$\begin{aligned} F(0) &= P(X \leq 0) = P(X = 0) = p(0) = 1 - p \\ F(1) &= P(X \leq 1) = P(X = 0 \text{ or } 1) = p(0) + p(1) = (1-p) + p = 1 \end{aligned}$$

Now for the other values, a Bernoulli random variable will never be negative, so $F(x) = 0$, for $x < 0$. Also, a Bernoulli random variable will always be less than or equal to 1, so $F(x) = 1$, for $x \geq 1$. Lastly, if x is in between 0 and 1, then the cdf is given by

$$F(x) = P(X \leq x) = P(X = 0) = p(0) = 1 - p), \text{ for } 0 \leq x < 1.$$

Binomial Distribution

To introduce the next family of distributions, we use our continuing example of tossing a coin, adding another toss.

Example 3.3.2

Suppose we toss a coin three times and record the sequence of heads (h) and tails (t). Supposing that the coin is fair, each toss results in heads with probability 0.5, and tails with the same probability of 0.5. Since the three tosses are mutually independent, the probability assigned to any outcome is 0.5^3. More specifically, consider the outcome hth. We could write the probability of this outcome as $(0.5)^2(0.5)^1$ to emphasize the fact that two heads and one tails occurred. Note that there are two other outcomes with two heads and one tails: hht and thh. Recall from Example 1.3.2 in Section 1.3, that we can count the number of outcomes with two heads and one tails by counting the number of ways to select positions for the two heads to occur in a sequence of three tosses, which is given by $\binom{3}{2}$. In general, note that $\binom{3}{x}$ counts the number of possible sequences with exactly x heads, for $x = 0, 1, 2, 3$.

We generalize the above by defining the discrete random variable X to be the number of heads in an outcome. The possible values of X are $x = 0, 1, 2, 3$. Using the above facts, the pmf of X is given as follows:

$$\begin{aligned} p(0) = P(X = 0) = P(\{ttt\}) &= \frac{1}{8} = \binom{3}{0}(0.5)^0(0.5)^3 \\ p(1) = P(X = 1) = P(\{htt, tht, tth\}) &= \frac{3}{8} = \binom{3}{1}(0.5)^1(0.5)^2 \\ p(2) = P(X = 2) = P(\{hht, hth, thh\}) &= \frac{3}{8} = \binom{3}{2}(0.5)^2(0.5)^1 \\ p(3) = P(X = 3) = P(\{hhh\}) &= \frac{1}{8} = \binom{3}{3}(0.5)^3(0.5)^0 \end{aligned} \tag{3.3.2}$$

In the above, the fractions in orange are found by calculating the probabilities directly using equally likely outcomes (note that the sample space Ω has 8 outcomes, see Example 1.3.1). In each line, the value of x is highlighted in red so that we can see the pattern forming. For example, when $x = 2$, we see in the expression on the right-hand side of Equation 3.3.2 that "2" appears in the binomial coefficient $\binom{3}{2}$, which gives the number of outcomes resulting in the random variable equaling 2, and "2" also appears in the exponent on the first 0.5, which gives the probability of two heads occurring.

The pattern exhibited by the random variable X in Example 3.3.2 is referred to as the *binomial distribution*, which we formalize in the next definition.

Definition $3.3.2$

Suppose that n independent trials of the same probability experiment are performed, where each trial results in either a "success" (with probability p), or a "failure" (with probability $1-p$). If the random variable X denotes the total number of successes in the n trials, then X has a ***binomial distribution*** with parameters n and p, which we write $X \sim \text{binomial}(n,p)$. The probability mass function of X is given by

$$p(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, 1, \dots, n. \tag{3.3.3}$$

In Example 3.3.2, the independent trials are the three tosses of the coin, so in this case we have parameter $n=3$. Furthermore, we were interested in counting the number of heads occurring in the three tosses, so a "success" is getting a heads on a toss, which occurs with probability 0.5 and so parameter $p=0.5$. Thus, the random variable X in this example has a binomial$(3,0.5)$ distribution and applying the formula for the binomial pmf given in Equation 3.3.3 when $x=2$ we get the same expression on the right-hand side of Equation 3.3.2:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \Rightarrow \quad p(2) = \binom{3}{2} 0.5^2 (1-0.5)^{3-2} = \binom{3}{2} 0.5^2 0.5^1$$

In general, we can connect binomial random variables to Bernoulli random variables. If X is a binomial random variable, with parameters n and p, then it can be written as the sum of n independent *Bernoulli* random variables, $X_1, \dots, X_n$. (Note: We will formally define *independence* for random variables later, in Chapter 5.) Specifically, if we define the random variable X_i, for $i = 1, \dots, n$, to be 1 when the ith trial is a "success", and 0 when it is a "failure", then the sum

$$X = X_1 + X_2 + \cdots + X_n$$

gives the total number of success in n trials. This connection between the binomial and Bernoulli distribution will be useful in a later section.

One of the main applications of the binomial distribution is to model population characteristics as in the following example.

Example $3.3.3$

Consider a group of 100 voters. If p denotes the probability that a voter will vote for a specific candidate, and we let random variable X denote the number of voters in the group that will vote for that candidate, then X follows a binomial distribution with parameters $n=100$ and p.

This page titled 3.3: Bernoulli and Binomial Distributions is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3.4: Expected Value of Discrete Random Variables

In this section, and the next, we look at various *numerical* characteristics of discrete random variables. These give us a way of classifying and comparing random variables.

Expected Value of Discrete Random Variables

We begin with the formal definition.

Definition $3.4.1$

If X is a discrete random variable with possible values $x_1, x_2, \dots, x_i, \dots$, and probability mass function $p(x)$, then the ***expected value*** (or ***mean***) of X is denoted $\mathrm{E}[X]$ and given by

$$\mathrm{E}[X] = \sum_i x_i \cdot p(x_i). \tag{3.4.1}$$

The expected value of X may also be denoted as μ_X or simply μ if the context is clear.

The expected value of a random variable has many interpretations. First, looking at the formula in Definition 3.4.1 for computing expected value (Equation $3.4.1$), note that it is essentially a ***weighted average***. Specifically, for a discrete random variable, the expected value is computed by "weighting", or multiplying, each value of the random variable, x_i, by the probability that the random variable takes that value, $p(x_i)$, and then summing over all possible values. This interpretation of the expected value as a weighted average explains why it is also referred to as the ***mean*** of the random variable.

The expected value of a random variable is also interpreted as the ***long-run value*** of the random variable. In other words, if we repeat the underlying random experiment several times and take the average of the values of the random variable corresponding to the outcomes, we would get the expected value, approximately. (Note: This interpretation of expected value is similar to the relative frequency approximation for probability discussed in Section 1.2.) Again, we see that the expected value is related to an average value of the random variable. Given the interpretation of the expected value as an average, either "weighted" or "long-run", the expected value is often referred to as a ***measure of center*** of the random variable.

Finally, the expected value of a random variable has a graphical interpretation. The expected value gives the ***center of mass*** of the probability mass function, which the following example demonstrates.

Example $3.4.1$

Consider again the context of Example 1.1.1, where we recorded the sequence of heads and tails in two tosses of a fair coin. In Example 3.1.1 we defined the discrete random variable X to denote the number of heads obtained. In Example 3.2.2 we found the pmf of X. We now apply Equation $3.4.1$ from Definition 3.4.1 and compute the expected value of X:

$$\begin{aligned}\mathrm{E}[X] &= 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) \\ &= 0 \cdot (0.25) + 1 \cdot (0.5) + 2 \cdot (0.25) \\ &= 0.5 + 0.5 = 1.\end{aligned}$$

Thus, we expect that the number of heads obtained in two tosses of a fair coin will be 1 in the long-run or on average. Figure 1 demonstrates the graphical representation of the expected value as the center of mass of the probability mass function.

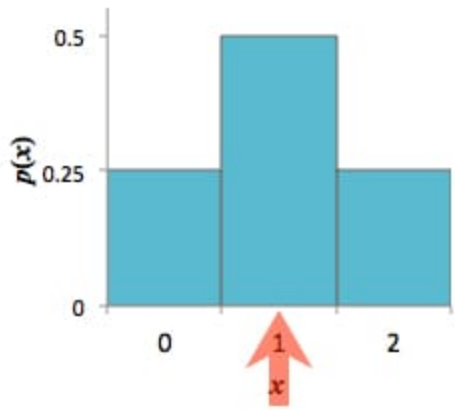


Figure 1: Histogram of X. The red arrow represents the center of mass, or the expected value of X

Example $3.4.2$

Suppose we toss a fair coin three times and define the random variable X to be our winnings on a single play of a game where

- we win \$$x$ if the first heads is on the xth toss, for $x = 1, 2, 3$,
- and we lose \$1 if we get no heads in all three tosses.

Then X is a discrete random variable, with possible values $x = -1, 1, 2, 3$, and pmf given by the following table:

x	$p(x) = P(X = x)$
-1	$\frac{1}{8}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$
3	$\frac{1}{8}$

Applying Definition 3.4.1, we find

$$\begin{aligned} \mathrm{E}[X] &= \sum_i x_i \cdot p(x_i) \\ &= (-1) \cdot \frac{1}{8} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} = \frac{5}{4} = 1.25. \end{aligned}$$

Thus, the expected winnings for a single play of the game is \$1.25. In other words, if we played the game multiple times, we expect the average winnings to be \$1.25.

For many of the common probability distributions, the expected value is given by a parameter of the distribution. The expected value may not be exactly equal to a parameter of the probability distribution, but rather it may be a function of the parameters. The following table gives the expected value for each of the common discrete distributions, including the Bernoulli and binomial distributions we introduced previously.

Expected Values for Discrete Distributions

Distribution	**Expected Value**
Bernoulli(p)	p

binomial(n, p)	np
hypergeometric(N, n, m)	$\frac{nm}{N}$
geometric(p)	$\frac{1}{p}$
negative binomial(r, p)	$\frac{r}{p}$
Poisson(λ)	λ

Expected Value of Functions of Random Variables

In many applications, we may not be interested in the value of a random variable itself, but rather in a function applied to the random variable or a collection of random variables. For example, we may be interested in the value of X^2. The following theorems, which we state without proof, demonstrate how to calculate the expected value of *functions of random variables*.

Theorem $3.4.1$

Let X be a random variable and let g be a real-valued function. Define the random variable $Y = g(X)$.

If X is a discrete random variable with possible values $x_1, x_2, \ldots, x_i, \ldots$, and probability mass function $p(x)$, then the expected value of Y is given by

$$\mathrm{E}[Y] = \sum_i g(x_i) \cdot p(x_i).$$

To put it simply, Theorem 3.4.1 states that to find the expected value of a function of a random variable, just apply the function to the possible values of the random variable in the definition of expected value. Before stating an important special case of Theorem 3.4.1, a word of caution regarding order of operations. Note that, in general,

$$\mathrm{E}[g(X)] \neq g\left(\mathrm{E}[X]\right)! \tag{3.4.2}$$

For example, $\mathrm{E}[X^2] \neq (\mathrm{E}[X])^2$, in general. However, as the next theorem states, there are exceptions to Equation $3.4.2$.

Special Case of Theorem 3.4.1

Let X be a random variable. If g is a *linear* function, i.e., $g(x) = ax + b$, then

$$\mathrm{E}[g(X)] = \mathrm{E}[aX + b] = a\mathrm{E}[X] + b.$$

The above special case is referred to as the ***linearity*** of expected value, which implies the following properties of the expected value.

Linearity of Expected Value

Let X be a random variable, c, c_1, c_2 constants, and g, g_1, g_2 real-valued functions. Then expectiation $\mathrm{E}[\,\cdot]$ satisfies the following:

1. The expected value of a constant is constant:

$$\mathrm{E}[c] = c$$

2. Constants can be factored out of expected values:

$$\mathrm{E}[cg(X)] = c\mathrm{E}[g(X)]$$

3. The expected value of a sum is equal to the sum of expected values:

$$\mathrm{E}[c_1 g_1(X) + c_2 g_2(X)] = c_1 \mathrm{E}[g_1(X)] + c_2 \mathrm{E}[g_2(X)]$$

This page titled 3.4: Expected Value of Discrete Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

3.5: Variance of Discrete Random Variables

We now look at our second numerical characteristic associated to random variables.

Definition $3.5.1$

The ***variance*** of a random variable X is given by

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

where μ denotes the expected value of X. The ***standard deviation*** of X is given by

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

In words, the variance of a random variable is the average of the squared deviations of the random variable from its mean (expected value). Notice that the variance of a random variable will result in a number with units squared, but the standard deviation will have the same units as the random variable. Thus, the standard deviation is easier to interpret, which is why we make a point to define it.

The variance and standard deviation give us a ***measure of spread*** for random variables. The standard deviation is interpreted as a measure of how "spread out" the possible values of X are with respect to the mean of X, $\mu = \mathrm{E}[X]$.

Example $3.5.1$

Consider the two random variables X_1 and X_2, whose probability mass functions are given by the histograms in Figure 1 below. Note that X_1 and X_2 have the same mean. However, in looking at the histograms, we see that the possible values of X_2 are more "spread out" from the mean, indicating that the variance (and standard deviation) of X_2 is larger.

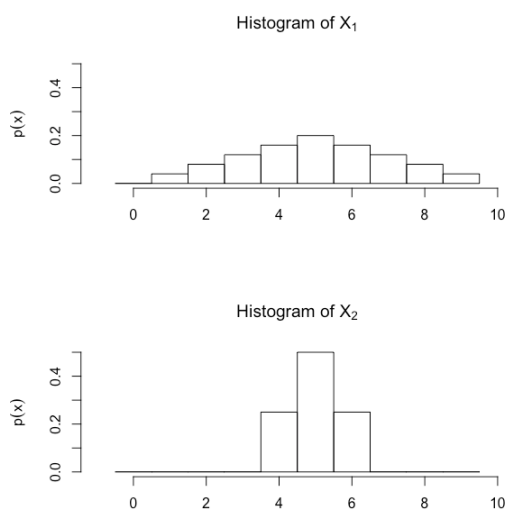


Figure 1: Histograms for random variables X_1 and X_2, both with same expected value different variance.

Theorem 3.4.1 actually tells us how to compute variance, since it is given by finding the expected value of a *function* applied to

the random variable. First, if X is a discrete random variable with possible values $x_1, x_2, \ldots, x_i, \ldots$, and probability mass function $p(x)$, then the variance of X is given by

$$\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 \cdot p(x_i).$$

The above formula follows directly from Definition 3.5.1. However, there is an alternate formula for calculating variance, given by the following theorem, that is often easier to use.

Theorem 3.5.1

Let X be any random variable, with mean μ. Then the variance of X is

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2. \tag{3.5.1}$$

Proof

By the definition of *variance* (Definition 3.5.1) and the linearity of expectation, we have the following:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2] \\
&= \mathrm{E}[X^2 + \mu^2 - 2X\mu] \\
&= \mathrm{E}[X^2] + \mathrm{E}[\mu^2] - \mathrm{E}[2X\mu] \\
&= \mathrm{E}[X^2] + \mu^2 - 2\mu\mathrm{E}[X] \quad \text{(Note: since } \mu \text{ is constant, we can take it out from the expected value)} \\
&= \mathrm{E}[X^2] + \mu^2 - 2\mu^2 \\
&= \mathrm{E}[X^2] - \mu^2
\end{aligned}$$

Example 3.5.2

Continuing in the context of Example 3.4.1, we calculate the variance and standard deviation of the random variable X denoting the number of heads obtained in two tosses of a fair coin. Using the alternate formula for variance, we need to first calculate $E[X^2]$, for which we use Theorem 3.4.1:

$$E[X^2] = 0^2 \cdot p(0) + 1^2 \cdot p(1) + 2^2 \cdot p(2) = 0 + 0.5 + 1 = 1.5.$$

In Example 3.4.1, we found that $\mu = E[X] = 1$. Thus, we find

$$\begin{aligned}
\mathrm{Var}(X) &= E[X^2] - \mu^2 = 1.5 - 1 = 0.5 \\
\Rightarrow \ \mathrm{SD}(X) &= \sqrt{\mathrm{Var}(X)} = \sqrt{0.5} \approx 0.707
\end{aligned}$$

Exercise 3.5.1

Consider the context of Example 3.4.2, where we defined the random variable X to be our winnings on a single play of game involving flipping a fair coin three times. We found that $\mathrm{E}[X] = 1.25$. Now find the variance and standard deviation of X.

Answer

First, find $\mathrm{E}[X^2]$:

$$\begin{aligned}
\mathrm{E}[X^2] &= \sum_i x_i^2 \cdot p(x_i) \\
&= (-1)^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{8} = \frac{11}{4} = 2.75
\end{aligned}$$

Now, we use the alternate formula for calculating variance:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 = 2.75 - 1.25^2 = 1.1875 \\ \Rightarrow \mathrm{SD}(X) &= \sqrt{1.1875} \approx 1.0897\end{aligned}$$

Given that the variance of a random variable is defined to be the expected value of *squared* deviations from the mean, variance is not linear as expected value is. We do have the following useful property of variance though.

Theorem $3.5.2$

Let X be a random variable, and a, b be constants. Then the following holds:

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

Exercise $3.5.2$

Prove Theorem 3.5.2.

Answer

First, let $\mu = \mathrm{E}[X]$ and note that by the linearity of expectation we have

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b = a\mu + b.$$

Now, we use the alternate formula for variance given in Theorem 3.5.1 to prove the result:

$$\begin{aligned}\mathrm{Var}(aX + b) &= \mathrm{E}\left[(aX + b)^2\right] - (\mathrm{E}[aX + b])^2 \\ &= \mathrm{E}[a^2X^2 + 2abX + b^2] - (a\mu + b)^2 \\ &= a^2\mathrm{E}[X^2] + 2ab\mathrm{E}[X] + b^2 - a^2\mu^2 - 2ab\mu - b^2 \\ &= a^2\mathrm{E}[X^2] - a^2\mu^2 = a^2(\mathrm{E}[X^2] - \mu^2) = a^2\mathrm{Var}(X)\end{aligned}$$

Theorem 3.5.2 easily follows from a little algebraic modification. Note that the "$+ b$" disappears in the formula. There is an intuitive reason for this. Namely, the "$+ b$" corresponds to a *horizontal shift* of the probability mass function for the random variable. Such a transformation to this function is not going to affect the *spread*, i.e., the variance will not change.

As with expected values, for many of the common probability distributions, the variance is given by a parameter or a function of the parameters for the distribution.

Variance for Discrete Distributions

Distribution	Expected Value
Bernoulli(p)	$p(1-p)$
binomial(n, p)	$np(1-p)$
hypergeometric(N, n, m)	$\frac{n(m/N)(1-m/N)(N-n)}{N-1}$
geometric(p)	$\frac{1-p}{p^2}$
negative binomial(r, p)	$\frac{r(1-p)}{p^2}$
Poisson(λ)	λ

This page titled 3.5: Variance of Discrete Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

4: CONTINUOUS RANDOM VARIABLES

4.1: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) for Continuous Random Variables

4.2: Expected Value and Variance of Continuous Random Variables

4.3: Uniform Distributions

4.4: Normal Distributions

This page titled 4: Continuous Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

4.1: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) for Continuous Random Variables

Probability Density Functions (PDFs)

Recall that continuous random variables have uncountably many possible values (think of intervals of real numbers). Just as for discrete random variables, we can talk about probabilities for continuous random variables using *density functions*.

Definition $4.1.1$

The ***probability density function*** **(pdf)**, denoted f, of a continuous random variable X satisfies the following:

1. $f(x) \geq 0$, for all $x \in \mathbb{R}$
2. f is piecewise continuous
3. $\int\limits_{-\infty}^{\infty} f(x)\,dx = 1$
4. $P(a \leq X \leq b) = \int\limits_{a}^{b} f(x)\,dx$

The first three conditions in the definition state the properties necessary for a function to be a ***valid pdf*** for a continuous random variable. The fourth condition tells us how to use a pdf to calculate probabilities for continuous random variables, which are given by *integrals* the continuous analog to sums.

Example $4.1.1$

Let the random variable X denote the time a person waits for an elevator to arrive. Suppose the longest one would need to wait for the elevator is 2 minutes, so that the possible values of X (in minutes) are given by the interval $[0, 2]$. A possible pdf for X is given by

$$f(x) = \begin{cases} x, & \text{for } 0 \leq x \leq 1 \\ 2 - x, & \text{for } 1 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

The graph of f is given below, and we verify that f satisfies the first three conditions in Definition 4.1.1:

1. From the graph, it is clear that $f(x) \geq 0$, for all $x \in \mathbb{R}$.
2. Since there are no holes, jumps, asymptotes, we see that $f(x)$ is (piecewise) continuous.
3. Finally we compute:

$$\int\limits_{-\infty}^{\infty} f(x)\,dx = \int\limits_{0}^{2} x\,dx = \int\limits_{0}^{1} x\,dx + \int\limits_{1}^{2} (2 - x)\,dx = 1$$

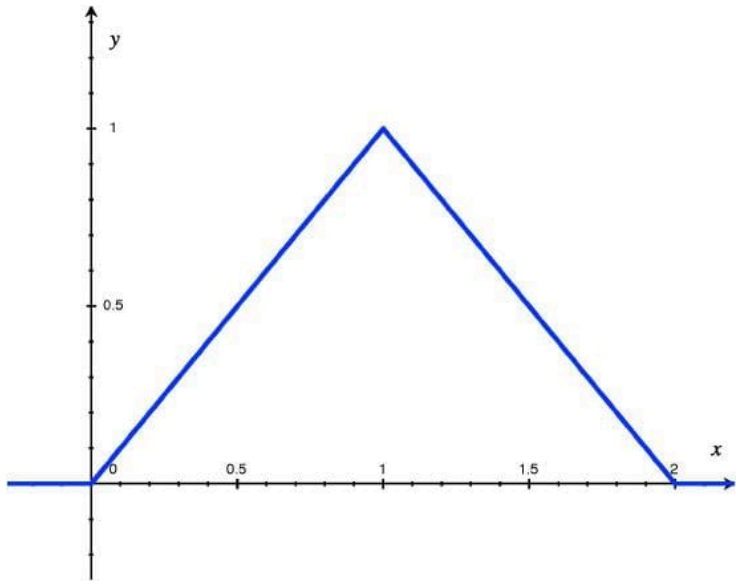


Figure 1: Graph of pdf for X, $f(x)$

So, if we wish to calculate the probability that a person waits less than 30 seconds (or 0.5 minutes) for the elevator to arrive, then we calculate the following probability using the pdf and the fourth property in Definition 4.1.1:

$$P(0 \leq X \leq 0.5) = \int_0^{0.5} f(x)\,dx = \int_0^{0.5} x\,dx = 0.125$$

Note that, unlike discrete random variables, continuous random variables have ***zero point probabilities***, i.e., the probability that a continuous random variable equals a single value is always given by 0. Formally, this follows from properties of integrals:

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)\,dx = 0.$$

Informally, if we realize that ***probability for a continuous random variable is given by areas under pdf's***, then, since there is no area in a line, there is no probability assigned to a random variable taking on a single value. This does not mean that a continuous random variable will never equal a single value, only that we do not assign any probability to single values for the random variable. For this reason, we only talk about the probability of a continuous random variable taking a value in an INTERVAL, not at a point. And whether or not the endpoints of the interval are included does not affect the probability. In fact, the following probabilities are all equal:

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

Cumulative Distribution Functions (CDFs)

Recall Definition 3.2.2, the definition of the cdf, which applies to both discrete and continuous random variables. For continuous random variables we can further specify how to calculate the cdf with a formula as follows. Let X have pdf f, then the cdf F is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \quad \text{for } x \in \mathbb{R}.$$

In other words, the cdf for a continuous random variable is found by *integrating* the pdf. Note that the **Fundamental Theorem of**

Calculus implies that the pdf of a continuous random variable can be found by *differentiating* the cdf. This relationship between the pdf and cdf for a continuous random variable is incredibly useful.

Relationship between PDF and CDF for a Continuous Random Variable

Let X be a continuous random variable with pdf f and cdf F.

- By definition, the cdf is found by *integrating* the pdf:

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

- By the Fundamental Theorem of Calculus, the pdf can be found by *differentiating* the cdf:

$$f(x) = \frac{d}{dx}\left[F(x)\right]$$

Example $4.1.2$

Continuing in the context of Example 4.1.1, we find the corresponding cdf. First, let's find the cdf at two possible values of X, $x = 0.5$ and $x = 1.5$:

$$F(0.5) = \int_{-\infty}^{0.5} f(t)\,dt = \int_{0}^{0.5} t\,dt = \frac{t^2}{2}\bigg|_0^{0.5} = 0.125$$

$$F(1.5) = \int_{-\infty}^{1.5} f(t)\,dt = \int_{0}^{1} t\,dt + \int_{1}^{1.5} (2-t)\,dt = \frac{t^2}{2}\bigg|_0^{1} + \left(2t - \frac{t^2}{2}\right)\bigg|_1^{1.5} = 0.5 + (1.875 - 1.5) = 0.875$$

Now we find $F(x)$ more generally, working over the intervals that $f(x)$ has different formulas:

$$\text{for } x < 0: \quad F(x) = \int_{-\infty}^{x} 0\,dt = 0$$

$$\text{for } 0 \le x \le 1: \quad F(x) = \int_{0}^{x} t\,dt = \frac{t^2}{2}\bigg|_0^{x} = \frac{x^2}{2}$$

$$\text{for } 1 < x \le 2: \quad F(x) = \int_{0}^{1} t\,dt + \int_{1}^{x} (2-t)\,dt = \frac{t^2}{2}\bigg|_0^{1} + \left(2t - \frac{t^2}{2}\right)\bigg|_1^{x} = 0.5 + \left(2x - \frac{x^2}{2}\right) - (2 - 0.5) = 2x - \frac{x^2}{2} - 1$$

$$\text{for } x > 2: \quad F(x) = \int_{-\infty}^{x} f(t)\,dt = 1$$

Putting this altogether, we write F as a piecewise function and Figure 2 gives its graph:

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{x^2}{2}, & \text{for } 0 \le x \le 1 \\ 2x - \frac{x^2}{2} - 1, & \text{for } 1 < x \le 2 \\ 1, & \text{for } x > 2 \end{cases}$$

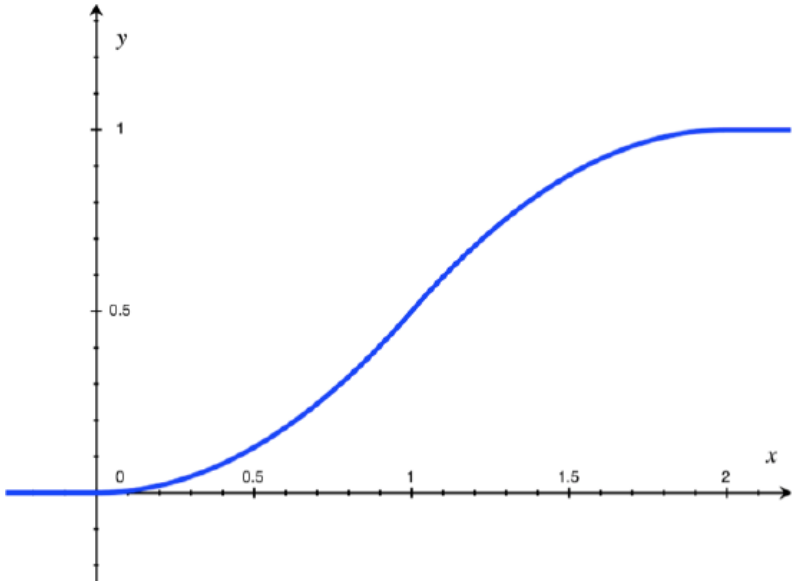


Figure 2: Graph of cdf in Example 4.1.2

Recall that the graph of the cdf for a discrete random variable is always a step function. Looking at Figure 2 above, we note that the cdf for a continuous random variable is always a *continuous* function.

Percentiles of a Distribution

Definition $4.1.2$

The ***(100p)th percentile*** $(0 \leq p \leq 1)$ of a probability distribution with cdf F is the value π_p such that

$$F(\pi_p) = P(X \leq \pi_p) = p.$$

To find the percentile π_p of a continuous random variable, which is a possible value of the random variable, we are specifying a cumulative probability p and solving the following equation for π_p:

$$\int_{-\infty}^{\pi_p} f(t)dt = p$$

Special Cases: There are a few values of p for which the corresponding percentile has a special name.

- ***Median*** or 50th percentile: $\pi_{.5} = \mu = Q_2$, separates probability (area under pdf) into two equal halves
- ***1st Quartile*** or 25th percentile: $\pi_{.25} = Q_1$, separates 1st quarter (25%) of probability (area) from the rest
- ***3rd Quartile*** or 75th percentile: $\pi_{.75} = Q_3$, separates 3rd quarter (75%) of probability (area) from the rest

Example $4.1.3$

Continuing in the context of Example 4.1.2, we find the median and quartiles.

- **median:** find $\pi_{.5}$, such that $F(\pi_{.5}) = 0.5 \Rightarrow \pi_{.5} = 1$ (from graph in Figure 1)
- **1st quartile:** find $Q_1 = \pi_{.25}$, such that $F(\pi_{.25}) = 0.25$. For this, we use the formula and the graph of the cdf in Figure 2:

$$\frac{\pi_{.25}^2}{2} = 0.25 \Rightarrow Q_1 = \pi_{.25} = \sqrt{0.5} \approx 0.707$$

- **3rd quartile:** find $Q_3 = \pi_{.75}$, such that $F(\pi_{.75}) = 0.75$. Again, use the graph of the cdf:

$$2\pi_{.75} - \frac{\pi_{.75}^2}{2} - 1 = 0.75 \;\Rightarrow\; \text{(using Quadratic Formula) } Q_3 = \pi_{.75} = \frac{4-\sqrt{2}}{2} \approx 1.293$$

This page titled 4.1: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) for Continuous Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

4.2: Expected Value and Variance of Continuous Random Variables

We now consider the expected value and variance for continuous random variables. Note that the interpretation of each is the same as in the discrete setting, but we now have a different method of calculating them in the continuous setting.

Definition $4.2.1$

If X is a continuous random variable with pdf $f(x)$, then the ***expected value*** (or ***mean***) of X is given by

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

The formula for the expected value of a continuous random variable is the continuous analog of the expected value of a discrete random variable, where instead of *summing* over all possible values we *integrate* (recall Sections 3.4 & 3.5).

For the ***variance*** of a continuous random variable, the definition is the same and we can still use the alternative formula given by Theorem 3.5.1, only we now integrate to calculate the value:

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx \right) - \mu^2$$

Example $4.2.1$

Consider again the context of Example 4.1.1, where we defined the continuous random variable X to denote the time a person waits for an elevator to arrive. The pdf of X was given by

$$f(x) = \begin{cases} x, & \text{for } 0 \le x \le 1 \\ 2 - x, & \text{for } 1 < x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

Applying Definition 4.2.1, we compute the expected value of X:

$$\mathrm{E}[X] = \int_0^1 x \cdot x\, dx + \int_1^2 x \cdot (2 - x)\, dx = \int_0^1 x^2\, dx + \int_1^2 (2x - x^2)\, dx = \frac{1}{3} + \frac{2}{3} = 1.$$

Thus, we expect a person will wait 1 minute for the elevator on average. Figure 1 demonstrates the graphical representation of the expected value as the center of mass of the pdf.

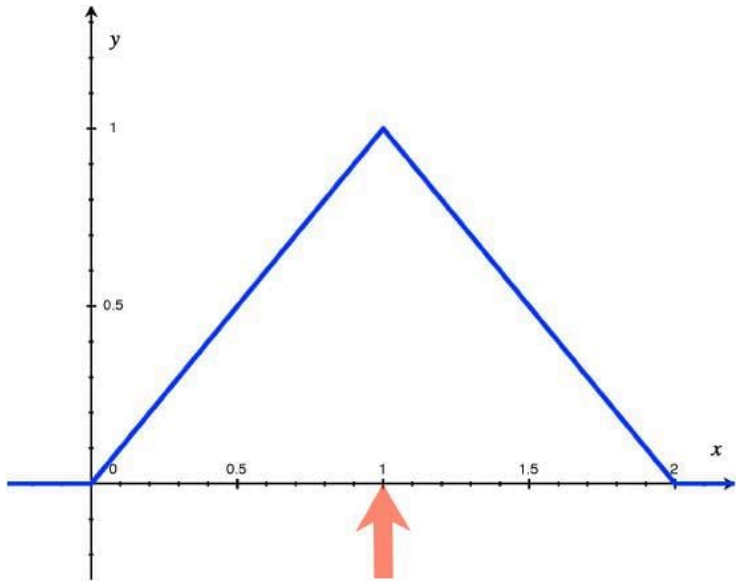


Figure 1: The red arrow represents the center of mass, or the expected value, of X.

Now we calculate the variance and standard deviation of X, by first finding the expected value of X^2.

$$\mathrm{E}[X^2] = \int_0^1 x^2 \cdot x\,dx + \int_1^2 x^2 \cdot (2-x)\,dx = \int_0^1 x^3\,dx + \int_1^2 (2x^2 - x^3)\,dx = \frac{1}{4} + \frac{11}{12} = \frac{7}{6}.$$

Thus, we have

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[X^2] - \mu^2 = \frac{7}{6} - 1 = \frac{1}{6} \\ \Rightarrow\ \mathrm{SD}(X) &= \sqrt{\mathrm{Var}(X)} = \frac{1}{\sqrt{6}} \approx 0.408 \end{aligned}$$

This page titled 4.2: Expected Value and Variance of Continuous Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

4.3: Uniform Distributions

Definition $4.3.1$

A random variable X has a ***uniform distribution*** on interval $[a, b]$, write $X \sim \text{uniform}[a, b]$, if it has pdf given by

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

The uniform distribution is also sometimes referred to as the ***box distribution***, since the graph of its pdf looks like a box. See Figure 1 below.

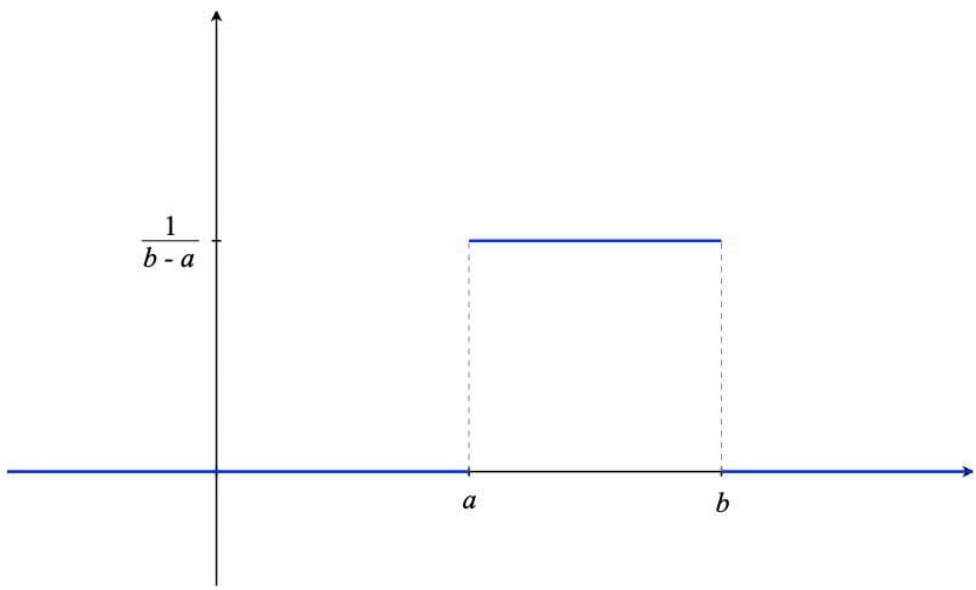


Figure 1: Graph of the pdf for a uniform distribution on interval $[a, b]$

Exercise $4.3.1$

Verify that the uniform pdf is a valid pdf, i.e., show that it satisfies the first three conditions of Definition 4.1.1.

Answer

1. In looking either at the formula in Definition 4.3.1 or the graph in Figure 1, we can see that the uniform pdf is always non-negative, i.e., $f(x) \geq 0$, for all $x \in \mathbb{R}$.
2. Given that the uniform pdf is a piecewise constant function, it is also piecewise continuous.
3. Finally, we need to verify that the area under the uniform pdf is equal to 1. This is quickly seen from the graph in Figure 1, since we calculate the area of rectangle with width $(b - a)$ and height $1/(b - a)$. Thus, the area is

$$(b - a) \times \frac{1}{(b - a)} = 1.$$

A typical application of the uniform distribution is to model randomly generated numbers. In other words, it provides the probability distribution for a random variable representing a randomly chosen number between numbers a and b.

The uniform distribution assigns equal probabilities to intervals of equal lengths, since it is a constant function, on the interval it

is non-zero $[a, b]$. This is the continuous analog to equally likely outcomes in the discrete setting.

We close the section by finding the expected value of the uniform distribution.

Example $4.3.1$

If X has a uniform distribution on the interval $[a, b]$, then we apply Definition 4.2.1 and compute the expected value of X:

$$\mathrm{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{b^2 - a^2}{2} \cdot \frac{1}{b-a} = \frac{(b-a)(b+a)}{2} \cdot \frac{1}{b-a} = \frac{b+a}{2}.$$

Thus, the expected value of the uniform$[a, b]$ distribution is given by the average of the parameters a and b, or the midpoint of the interval $[a, b]$. This is readily apparent when looking at a graph of the pdf in Figure 1 and remembering the interpretation of expected value as the center of mass. Since the pdf is constant over $[a, b]$, the center of mass is simply given by the midpoint.

This page titled 4.3: Uniform Distributions is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

4.4: Normal Distributions

Definition $4.4.1$

A random variable X has a ***normal distribution***, with parameters μ and σ, write $X \sim \text{normal}(\mu, \sigma)$, if it has pdf given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad \text{for } x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$.

If a continuous random variable X has a normal distribution with parameters μ and σ, then $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. These facts can be derived using Definition 4.2.1; however, the integral calculations require many tricks. Note that the normal case is why the notation μ is often used for the expected value, and σ^2 is used for the variance. So, μ gives the center of the normal pdf, and its graph is symmetric about μ, while σ determines how spread out the graph is. Figure 1 below shows the graph of two different normal pdf's.

Example $4.4.1$

Suppose $X_1 \sim \text{normal}(0, 2^2)$ and $X_2 \sim \text{normal}(0, 3^2)$. So, X_1 and X_2 are both normally distributed random variables with the same mean, but X_2 has a larger standard deviation. Given our interpretation of standard deviation, this implies that the possible values of X_2 are more "spread out" from the mean. This is easily seen by looking at the graphs of the pdf's corresponding to X_1 and X_2 given in Figure 1.

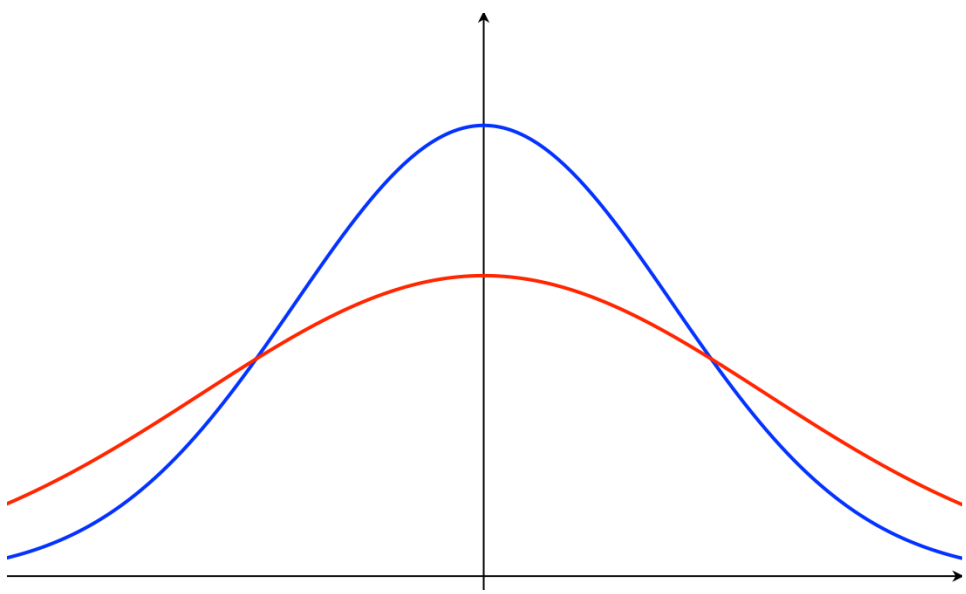


Figure 1: Graph of normal pdf's: $X_1 \sim \text{normal}(0, 2^2)$ in **blue**, $X_2 \sim \text{normal}(0, 3^2)$ in **red**

The normal distribution is arguably the most important probably distribution. It is used to model the distribution of population characteristics such as weight, height, and IQ. The pdf is terribly tricky to work with, in fact integrals involving the normal pdf cannot be solved exactly, but rather require numerical methods to approximate. Because of this, there is no closed form for the corresponding cdf of a normal distribution. Given the importance of the normal distribution though, many software programs have built in normal probability calculators. There are also many useful properties of the normal distribution that make it easy to work with. We state these properties without proof below. Note that we also include the connection to expected value and variance

given by the parameters.

Properties of the Normal Distribution

1. If $X \sim \text{normal}(\mu, \sigma)$, then $aX + b$ also follows a normal distribution with parameters $a\mu + b$ and $a\sigma$.
2. If $X \sim \text{normal}(\mu, \sigma)$, then $\dfrac{X - \mu}{\sigma}$ follows the ***standard normal distribution***, i.e., the normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

The first property says that any linear transformation of a normally distributed random variable is also normally distributed. The second property is a special case of the first, since we can re-write the transformation on X as

$$\frac{X - \mu}{\sigma} = \left(\frac{1}{\sigma}\right) X - \frac{\mu}{\sigma}.$$

This transformation, subtracting the mean and dividing by the standard deviation, is referred to as ***standardizing*** X, since the resulting random variable will *always* have the standard normal distribution with mean 0 and standard deviation 1. In this way, standardizing a normal random variable has the effect of removing the units. Before the prevalence of calculators and computer software capable of calculating normal probabilities, people would apply the standardizing transformation to the normal random variable and use a table of probabilities for the standard normal distribution.

This page titled 4.4: Normal Distributions is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

5: MULTIVARIATE RANDOM VARIABLES

5.1: Joint Distributions of Discrete Random Variables
5.2: Joint Distributions of Continuous Random Variables

This page titled 5: Multivariate Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

5.1: Joint Distributions of Discrete Random Variables

In this chapter we consider two or more random variables defined on the same sample space and discuss how to model the probability distribution of the random variables *jointly*. We will begin with the discrete case by looking at the joint probability mass function for two discrete random variables. In the following section, we will consider continuous random variables.

Definition $5.1.1$

If discrete random variables X and Y are defined on the same sample space Ω, then their ***joint probability mass function*** **(joint pmf)** is given by

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

where (x,y) is a pair of possible values for the pair of random variables (X,Y), and $p(x,y)$ satisfies the following conditions:

- $0 \leq p(x,y) \leq 1$
- $\sum\limits_{(x,y)}\sum p(x,y) = 1$
- $P\left((X,Y) \in A)\right) = \sum\limits_{(x,y)\in A}\sum p(x,y)$

Note that conditions #1 and #2 in Definition 5.1.1 are required for $p(x,y)$ to be a valid joint pmf, while the third condition tells us how to use the joint pmf to find probabilities for the pair of random variables (X,Y).

In the discrete case, we can obtain the ***joint cumulative distribution function*** **(joint cdf)** of X and Y by *summing* the joint pmf:

$$F(x,y) = P(X \leq x \text{ and } Y \leq y) = \sum_{x_i \leq x}\sum_{y_j \leq y} p(x_i, y_j),$$

where x_i denotes possible values of X and y_j denotes possible values of Y. From the joint pmf, we can also obtain the individual probability distributions of X and Y *separately* as shown in the next definition.

Definition $5.1.2$

Suppose that discrete random variables X and Y have joint pmf $p(x,y)$. Let $x_1, x_2, \ldots, x_i, \ldots$ denote the possible values of X, and let $y_1, y_2, \ldots, y_j, \ldots$ denote the possible values of Y. The ***marginal probability mass functions*** **(marginal pmf's)** of X and Y are respectively given by the following:

$$p_X(x) = \sum_j p(x, y_j) \quad \text{(fix a value of } X \text{ and sum over possible values of } Y)$$
$$p_Y(y) = \sum_i p(x_i, y) \quad \text{(fix a value of } Y \text{ and sum over possible values of } X)$$

Link to Video: Overview of Definitions 5.1.1 & 5.1.2

Example $5.1.1$

Consider again the probability experiment of Example 3.3.2, where we toss a fair coin three times and record the sequence of heads (h) and tails (t). Again, we let random variable X denote the number of heads obtained. We also let random variable Y denote the winnings earned in a single play of a game with the following rules, based on the outcomes of the probability experiment (this is the same as Example 3.4.2):

- player wins \$1 if first h occurs on the first toss
- player wins \$2 if first h occurs on the second toss
- player wins \$3 if first h occurs on the third toss
- player loses \$1 if no h occur

Note that the possible values of X are $x = 0, 1, 2, 3$, and the possible values of Y are $y = -1, 1, 2, 3$. We represent the joint pmf using a table:

Table 1: joint pmf of X and Y

$p(x,y)$	X			
Y	**0**	**1**	**2**	**3**
-1	1/8	0	0	0
1	0	1/8	2/8	1/8
2	0	1/8	1/8	0
3	0	1/8	0	0

The values in Table 1 give the values of $p(x,y)$. For example, consider $p(0,-1)$:

$$p(0,-1) = P(X = 0 \text{ and } Y = -1) = P(ttt) = \frac{1}{8}.$$

Since the outcomes are equally likely, the values of $p(x,y)$ are found by counting the number of outcomes in the sample space Ω that result in the specified values of the random variables, and then dividing by 8, the total number of outcomes in Ω. The sample space is given below, color coded to help explain the values of $p(x,y)$:

$$\Omega = \{ttt, htt, tht, tth, hht, hth, thh, hhh\}$$

Given the joint pmf, we can now find the marginal pmf's. Note that the marginal pmf for X is found by computing sums of the *columns* in Table 1, and the marginal pmf for Y corresponds to the *row* sums. (Note that we found the pmf for X in Example 3.3.2 as well, it is a binomial random variable. We also found the pmf for Y in Example 3.4.2.)

Table 2: marginal pmf's for X and Y

x	$p_X(x)$	y	$p_Y(y)$
0	1/8	-1	1/8
1	3/8	1	1/2
2	3/8	2	1/4
3	1/8	3	1/8

Finally, we can find the joint cdf for X and Y by summing over values of the joint frequency function. For example, consider $F(1,1)$:

$$F(1,1) = P(X \leq 1 \text{ and } Y \leq 1) = \sum_{x \leq 1} \sum_{y \leq 1} p(x,y) = p(0,-1) + p(0,1) + p(-1,1) + p(1,1) = \frac{1}{4}$$

Again, we can represent the joint cdf using a table:

Table 3: joint cdf of X and Y

$F(x,y)$	X			
Y	**0**	**1**	**2**	**3**
-1	1/8	1/8	1/8	1/8
1	1/8	1/4	1/2	5/8
2	1/8	3/8	3/4	7/8
3	1/8	1/2	7/8	1

Link to Video: Walkthrough of Example 5.1.1

Expectations of Functions of Jointly Distributed Discrete Random Variables

We now look at taking the expectation of jointly distributed discrete random variables. Because expected values are defined for a single quantity, we will actually define the expected value of a combination of the pair of random variables, i.e., we look at the expected value of a function applied to (X,Y).

Theorem $5.1.1$

Suppose that X and Y are jointly distributed discrete random variables with joint pmf $p(x,y)$.

If $g(X,Y)$ is a function of these two random variables, then its expected value is given by the following:

$$\mathrm{E}[g(X,Y)] = \sum_{(x,y)} \sum g(x,y)p(x,y).$$

Example $5.1.2$

Consider again the discrete random variables we defined in Example 5.1.1 with joint pmf given in Table 1. We will find the expected value of three different functions applied to (X,Y).

1. First, we define $g(x,y) = xy$, and compute the expected value of XY:

$$
\begin{aligned}
\mathrm{E}[XY] = \sum_{(x,y)}\sum xy \cdot p(x,y) = (0)(-1)\left(\frac{1}{8}\right) \\
&+ (1)(1)\left(\frac{1}{8}\right) + (2)(1)\left(\frac{2}{8}\right) + (3)(1)\left(\frac{1}{8}\right) \\
&+ (1)(2)\left(\frac{1}{8}\right) + (2)(2)\left(\frac{1}{8}\right) \\
&+ (1)(3)\left(\frac{1}{8}\right) \\
&= \frac{17}{8} = 2.125
\end{aligned}
$$

2. Next, we define $g(x) = x$, and compute the expected value of X:

$$
\begin{aligned}
\mathrm{E}[X] = \sum_{(x,y)}\sum x \cdot p(x,y) = (0)\left(\frac{1}{8}\right) \\
&+ (1)\left(\frac{1}{8}\right) + (2)\left(\frac{2}{8}\right) + (3)\left(\frac{1}{8}\right) \\
&+ (1)\left(\frac{1}{8}\right) + (2)\left(\frac{1}{8}\right) \\
&+ (1)\left(\frac{1}{8}\right) \\
&= \frac{12}{8} = 1.5
\end{aligned}
$$

Recall that $X \sim \text{binomial}(n = 3, p = 0.5)$, and that the expected value of a binomial random variable is given by np. Thus, we can verify the expected value of X that we calculated above using Theorem 5.1.1 using this fact for binomial distributions: $\mathrm{E}[X] = np = 3(0.5) = 1.5$.

3. Lastly, we define $g(x, y) = y$, and calculate the expected value of Y:

$$
\begin{aligned}
\mathrm{E}[Y] = \sum_{(x,y)}\sum y \cdot p(x,y) = (-1)\left(\frac{1}{8}\right) \\
&+ (1)\left(\frac{1}{8}\right) + (1)\left(\frac{2}{8}\right) + (1)\left(\frac{1}{8}\right) \\
&+ (2)\left(\frac{1}{8}\right) + (2)\left(\frac{1}{8}\right) \\
&+ (3)\left(\frac{1}{8}\right) \\
&= \frac{10}{8} = 1.25
\end{aligned}
$$

Again, we can verify this result by reviewing the calculations done in Example 3.4.2.

Link to Video: Walkthrough of Example 5.1.2

Independent Random Variables

In some cases, the probability distribution of one random variable will not be affected by the distribution of another random variable defined on the same sample space. In those cases, the joint distribution functions have a very simple form, and we refer to the random variables as independent.

Definition 5.1.3

Discrete random variables $X_1, X_2, \dots, X_n$ are ***independent*** if the joint pmf factors into a product of the marginal pmf's:

$$p(x_1, x_2, \dots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdots p_{X_n}(x_n). \tag{5.1.1}$$

It is equivalent to check that this condition holds for the cumulative distribution functions.

Recall the definition of independent *events* (Definition 2.2.2): A and B are independent events if $P(A \cap B) = P(A)\ P(B)$. This is the basis for the definition of independent random variables because we can write the pmf's in Equation 5.1.1 in terms of events as follows:

$$p(x, y) = P(X = x \text{ and } Y = y) = P(\{X = x\} \cap \{Y = y\}) = P(X = x)P(Y = y) = p_X(x)p_Y(y)$$

In the above, we use the idea that if X and Y are independent, then the event that X takes on a given value x is independent of the event that Y takes the value y.

Example 5.1.3

Consider yet again the discrete random variables defined in Example 5.1.1. According to the definition, X and Y are independent if

$$p(x, y) = p_X(x) \cdot p_Y(y),$$

for all pairs (x, y). Recall that the joint pmf for (X, Y) is given in Table 1 and that the marginal pmf's for X and Y are given in Table 2. Note that, for $(x, y) = (0, -1)$, we have the following

$$p(0, -1) = \frac{1}{8}, \quad p_X(0) = \frac{1}{8}, \quad p_Y(-1) = \frac{1}{8} \quad \Rightarrow \quad p(0, -1) \neq p_X(0) \cdot p_Y(-1).$$

Thus, X and Y are ***not*** independent, or in other words, X and Y are ***dependent***. This should make sense given the definition of X and Y. The winnings earned depend on the number of heads obtained. So the probabilities assigned to the values of Y will be affected by the values of X.

We also have the following very useful theorem about the expected value of a product of independent random variables, which is simply given by the product of the expected values for the individual random variables.

Theorem 5.1.2

If X and Y are independent random variables, then $\mathrm{E}[XY] = \mathrm{E}[X]\ \mathrm{E}[Y]$.

Proof

Assume X and Y are independent random variables. If we let $p(x, y)$ denote the joint pmf of (X, Y), then, by Definition 5.1.3, $p(x, y) = p_X(x)p_Y(y)$, for all pairs (x, y). Using this fact and Theorem 5.1.1, we have

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{(x,y)}\sum xy \cdot p(x, y) = \sum_{(x,y)}\sum xy \cdot p_X(x)p_Y(y) \\
&= \sum_x \sum_y xyp_(x)P_Y(y) = \sum_x xp_X(x)\left(\sum_y p_Y(y)\right) = \sum_x xp_X(x)\mathrm{E}[Y] \\
&= \mathrm{E}[Y]\sum_x xp_X(x) = \mathrm{E}[Y]\ \mathrm{E}[X].
\end{aligned}$$

Theorem 5.1.2 can be used to show that two random variables are ***not*** independent: if $\mathrm{E}[XY] \neq \mathrm{E}[X]\ \mathrm{E}[Y]$, then X and Y ***cannot*** be independent. However, beware using Theorem 5.1.2 to show that random variables are independent. Note that Theorem 5.1.2 ***assumes*** that X and Y are independent and then the property about the expected value follows. The other direction does not hold. In other words, if $\mathrm{E}[XY] = \mathrm{E}[X]\ \mathrm{E}[Y]$, then X and Y ***may or may not*** be independent.

Link to Video: Independent Random Variables

This page titled 5.1: Joint Distributions of Discrete Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

5.2: Joint Distributions of Continuous Random Variables

Having considered the discrete case, we now look at joint distributions for continuous random variables.

Definition $5.2.1$

If continuous random variables X and Y are defined on the same sample space Ω, then their ***joint probability density function*** (***joint pdf***) is a piecewise continuous function, denoted $f(x,y)$, that satisfies the following.

1. $f(x,y) \geq 0$, for all $(x,y) \in \mathbb{R}^2$
2. $\displaystyle\iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1$
3. $P((X,Y) \in A) = \displaystyle\iint_A f(x,y)\,dx\,dy$, for any $A \subseteq \mathbb{R}^2$

The first two conditions in Definition 5.2.1 provide the requirements for a function to be a valid joint pdf. The third condition indicates how to use a joint pdf to calculate probabilities. As an example of applying the third condition in Definition 5.2.1, the ***joint cdf*** for continuous random variables X and Y is obtained by integrating the joint density function over a set A of the form

$$A = \{(x,y) \in \mathbb{R}^2 \mid X \leq a \text{ and } Y \leq b\},$$

where a and b are constants. Specifically, if A is given as above, then the joint cdf of X and Y, at the point (a,b), is given by

$$F(a,b) = P(X \leq a \text{ and } Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\,dx\,dy.$$

Note that probabilities for continuous jointly distributed random variables are now *volume*s instead of areas as in the case of a single continuous random variable.

As in the discrete case, we can also obtain the individual, *maginal* pdf's of X and Y from the joint pdf.

Definition $5.2.2$

Suppose that continuous random variables X and Y have joint density function $f(x,y)$. The ***marginal pdf's*** of X and Y are respectively given by the following.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \quad (\text{fix a value of } X, \text{ and integrate over all possible values of } Y)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \quad (\text{fix a value of } Y, \text{ and integrate over all possible values of } X)$$

Link to Video: Overview of Definitions 5.2.1 & 5.2.2

Example $5.2.1$

Suppose a radioactive particle is contained in a unit square. We can define random variables X and Y to denote the x- and y -coordinates of the particle's location in the unit square, with the bottom left corner placed at the origin. Radioactive particles follow completely random behavior, meaning that the particle's location should be uniformly distributed over the unit square. This implies that the joint density function of X and Y should be constant over the unit square, which we can write as

$$f(x,y) = \begin{cases} c, & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

where c is some unknown constant. We can find the value of c by using the first condition in Definition 5.2.1 and solving the following:

$$\iint\limits_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1 \quad \Rightarrow \quad \int\limits_0^1\int\limits_0^1 c\,dx\,dy = 1 \quad \Rightarrow \quad c\int\limits_0^1\int\limits_0^1 1\,dx\,dy = 1 \quad \Rightarrow \quad c = 1$$

We can now use the joint pdf of X and Y to compute probabilities that the particle is in some specific region of the unit square. For example, consider the region

$$A = \{(x,y) \mid x - y > 0.5\},$$

which is graphed in Figure 1 below.

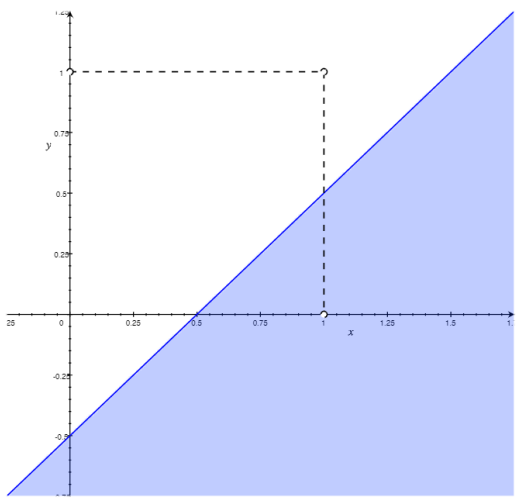


Figure 1: Graph of region A, shaded in blue.

If we want the probability that the particle's location is in the lower right corner of the unit square that intersects with the region A, then we integrate the joint density function over that portion of A in the unit square, which gives the following probability:

$$P(X - Y > 0.5) = \iint\limits_A f(x,y)\,dx\,dy = \int_0^{0.5}\int_{y+0.5}^1 1\,dx\,dy = 0.125$$

Lastly, we apply Definition 5.2.2 to find the marginal pdf's of X and Y.

$$f_X(x) = \int_0^1 1\,dy = 1, \quad \text{for } 0 \le x \le 1$$

$$f_Y(y) = \int_0^1 1\,dx = 1, \quad \text{for } 0 \le y \le 1$$

Note that both X and Y are individually uniform random variables, each over the interval $[0,1]$. This should not be too surprising. Given that the particle's location was uniformly distributed over the unit square, we should expect that the individual coordinates would also be uniformly distributed over the unit intervals.

Example $5.2.2$

At a particular gas station, gasoline is stocked in a bulk tank each week. Let random variable X denote the proportion of the tank's capacity that is *stocked* in a given week, and let Y denote the proportion of the tank's capacity that is *sold* in the same week. Note that the gas station cannot sell more than what was stocked in a given week, which implies that the value of Y cannot exceed the value of X. A possible joint pdf of X and Y is given by

$$f(x,y) = \begin{cases} 3x, & \text{if } 0 \le y \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that this function is only nonzero over the triangular region given by $\{(x,y) \mid 0 \le y \le x \le 1\}$, which is graphed in Figure 2 below:

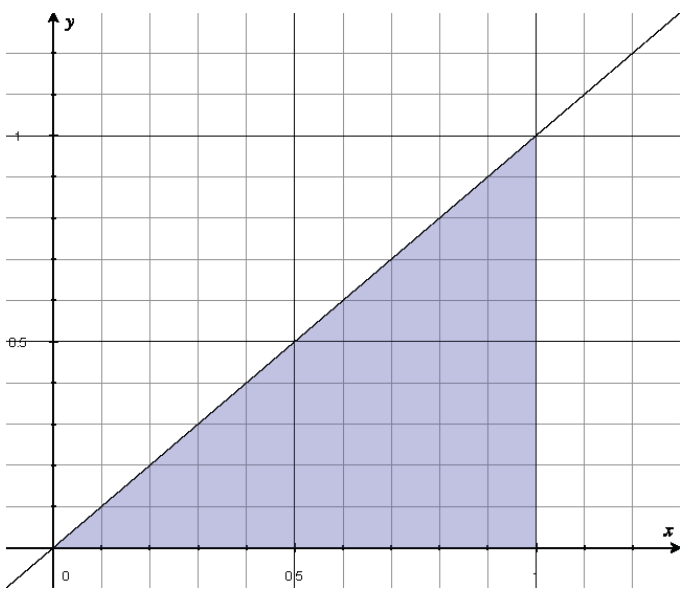


Figure 2: Region over which joint pdf $f(x,y)$ is nonzero.

Link to Video: Marginal PDFs for Example 5.2.2

We find the joint cdf of X and Y at the point $(x,y) = (1/2, 1/3)$:

$$\begin{aligned} F\left(\frac{1}{2}, \frac{1}{3}\right) = P\left(X \le \frac{1}{2} \text{ and } Y \le \frac{1}{3}\right) &= \int_0^{1/3} \int_y^{0.5} 3x\,dxdy \\ &= \int_0^{1/3} \left(\frac{3}{2}x^2\Big|_y^{0.5}\right) dy = \int_0^{1/3} \left(\frac{3}{8} - \frac{3}{2}y^2\right) dy \\ &= \frac{3}{8}y - \frac{1}{2}y^3\Big|_0^{1/3} \approx 0.1065 \end{aligned}$$

Thus, there is a 10.65% chance that less than half the tank is stocked and less than a third of the tank is sold in a given week. Note that in finding the above integral, we look at where the region given by $\{(x,y) \mid x \leq 1/2, y \leq 1/3\}$ intersects the region over which the joint pdf is nonzero, i.e., the region graphed in Figure 2. This tells us what the limits of integration are in the double integral. Figure 3 below is a graph of the intersection made on desmos.com:

Figure 3: Intersection of $\{(x,y) \mid x \leq 1/2, y \leq 1/3\}$ with the region over which joint pdf $f(x,y)$ is nonzero.

Next, we find the probability that the amount of gas sold is less than half the amount that is stocked in a given week. In other words, we find $P(Y < 0.5X)$. In order to find this probability, we need to find the region over which we will integrate the joint pdf. To do this, look for the intersection of the region given by $\{(x,y) \mid y < 0.5x\}$ with the region in Figure 2, which is graphed in Figure 4 below:

Figure 4: Intersection of $\{(x,y) \mid y < 0.5x\}$ with the region over which joint pdf $f(x,y)$ is nonzero.

The calculation is as follows:

$$\begin{aligned} P(Y < 0.5X) &= \int_0^1 \int_0^{0.5x} 3x \, dy dx \\ &= \int_0^1 \left(3xy \Big|_0^{0.5x} \right) \, dx \\ &= \int_0^1 \left(\frac{3}{2}x^2 - 0 \right) \, dx = \frac{1}{2}x^3 \Big|_0^1 \\ &= \frac{1}{2} \end{aligned}$$

Thus, there is a 50% chance that the amount of gas sold in a given week is less than half of the gas stocked.

Independent Random Variables

We can also define independent random variables in the continuous case, just as we did for discrete random variables.

Definition $5.2.3$

Continuous random variables $X_1, X_2, \dots, X_n$ are ***independent*** if the joint pdf factors into a product of the marginal pdf's:

$$f(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

It is equivalent to check that this condition holds for the cumulative distribution functions.

Example $5.2.3$

Consider the continuous random variables defined in Example 5.2.1, where the X and Y gave the location of a radioactive particle. We will show that X and Y are independent and then verify that Theorem 5.1.2 also applies in the continuous setting.

Recall that we found the marginal pdf's to be the following:

$$\begin{aligned} f_X(x) &= 1, \text{ for } 0 \leq x \leq 1 \\ f_Y(y) &= 1, \text{ for } 0 \leq y \leq 1 \end{aligned}$$

So, for (x,y) in the unit square, i.e., $0 \leq x \leq 1$ and $0 \leq y \leq$, we have

$$f(x,y) = 1 = 1 \cdot 1 = f_X(x) f_Y(y),$$

and outside the unit square, at least one of marginal pdf's will be 0, so

$$f(x,y) = 0 = f_X(x) f_Y(y).$$

We have thus shown that $f(x,y) = f_X(x)\ f_Y(y)$, for all $(x,y) \in \mathbb{R}^2$, and so by Definition 5.2.3, X and Y are independent.

Now let's look at the expected value of the product of X and Y:

$$\mathrm{E}[XY] = \iint_{\mathbb{R}^2} xy \cdot f(x,y)\, dxdy = \int_0^1 \int_0^1 xy \cdot 1\, dxdy = \int_0^1 \left(\frac{x^2}{2} y \Big|_0^1 \right)\, dy = \frac{1}{4}$$

Note that both X and Y are uniform on the interval $[0,1]$. Therefore, their expected values are both 1/2, the midpoint of $[0,1]$. Putting this all together, we have

$$\mathrm{E}[XY] = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathrm{E}[X]\ \mathrm{E}[Y],$$

which is the conclusion to Theorem 5.1.2.

Link to Video: Independent Continuous Random Variables

This page titled 5.2: Joint Distributions of Continuous Random Variables is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

6: THE SAMPLE MEAN AND CENTRAL LIMIT THEOREM

6.1: Functions of Normal Random Variables

6.2: Sample Mean

This page titled 6: The Sample Mean and Central Limit Theorem is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

6.1: Functions of Normal Random Variables

In addition to considering the probability distributions of random variables simultaneously using joint distribution functions, there is also occasion to consider the probability distribution of *functions* applied to random variables. In this section we consider the special case of applying functions to normally distributed random variables, which will be very important in the following section. We begin by deriving the probability distribution of the square of a standard normal random variable.

Before we jump into the example, note that one approach to finding the probability distribution of a function of a random variable relies on the relationship between the pdf and cdf for a continuous random variable:

$$\frac{d}{dx}[F(x)] = f(x) \qquad \text{"derivative of cdf} = \text{pdf"}$$

It is often easier to find the cdf of a function of a continuous random variable, and then use the above relationship to derive the pdf.

Example $6.1.1$

Let Z be a standard normal random variable, i.e., $Z \sim N(0,1)$. We find the pdf of $Y = Z^2$.

Let Φ denote the cdf of Z, i.e., $\Phi(z) = P(Z \leq z) = F_Z(z)$. We first find the cdf of $X = Z^2$ in terms of Φ (recall that there is no closed form expression for Φ):

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(Z^2 \leq y) \\ &= P\left(-\sqrt{y} \leq Z \leq \sqrt{y}\right), \text{ for } y \geq 0 \\ &= \Phi\left(\sqrt{y}\right) - \Phi(-\sqrt{y}) \end{aligned}$$

Note that if $y < 0$, then $F_Y(y) = 0$, since it is not possible for $Y = Z^2$ to be negative. In other words, the possible values of $Y = Z^2$ are $y \geq 0$.

Next, we take the derivative of the cdf of Y to find its pdf. Before doing so, we note that if Φ is the cdf for Z, then its derivative is the pdf for Z, which is denoted φ. Since Z is a standard normal random variable, we know that

$$\varphi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for } z \in \mathbb{R}.$$

Using this, we now find the pdf of Y:

$$\begin{aligned} f_Y(y) = \frac{d}{dy}[F_Y(y)] &= \frac{d}{dy}\left[\Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)\right] \\ &= \frac{d}{dy}\left[\Phi\left(\sqrt{y}\right)\right] - \frac{d}{dy}\left[\Phi\left(-\sqrt{y}\right)\right] \\ &= \varphi\left(\sqrt{y}\right) \cdot \frac{1}{2\sqrt{y}} + \varphi\left(-\sqrt{y}\right) \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \cdot \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{1}{\sqrt{y}}. \end{aligned}$$

In summary, if $Y = Z^2$, where $Z \sim N(0,1)$, then the pdf for Y is given by

$$f_Y(y) = \frac{y^{-1/2}}{\sqrt{2\pi}} e^{-y/2}, \text{ for } y \geq 0.$$

Note that the pdf for Y is a *gamma* pdf with $\alpha = \lambda = \frac{1}{2}$. This is also referred to as the ***chi-square distribution***, denoted χ^2. See below for a video walkthrough of this example.

There is another approach to finding the probability distribution of functions of random variables, which involves *moment-generating functions*, which we now define.

Definition $6.1.1$

The ***moment-generating function*** **(mgf)** of a random variable X is given by

$$M_X(t) = E[e^{tX}], \quad \text{for } t \in \mathbb{R}.$$

The mgf of a random variable has many theoretical properties that are very useful in the study of probability theory. One of those properties is the fact that when the derivative of the mgf is evaluated for $t = 0$, the result is equal to the expected value of the random variable:

$$\frac{d}{dt}[M_X(t)]_{t=0} = E[X]$$

This result can be extended to higher order derivatives producing higher order moments, which we will not go into. Instead, we state the following properties that are useful in the context of determining the distribution of a function of random variables. The first result below indicates that mgf's are unique in the sense that if two random variables have the same mgf, then they necessarily have the same probability distribution. The next two properties provide ways of manipulating mgf's in order to find the mgf of a function of a random variable.

Theorem $6.1.1$

The mgf $M_X(t)$ of random variable X uniquely determines the probability distribution of X. In other words, if random variables X and Y have the same mgf, $M_X(t) = M_Y(t)$, then X and Y have the same probability distribution.

Theorem $6.1.2$

Let X be a random variable with mgf $M_X(t)$, and let a, b be constants. If random variable $Y = aX + b$, then the mgf of Y is given by

$$M_Y(t) = e^{bt} M_X(at).$$

Theorem $6.1.3$

If $X_1, \dots, X_n$ are independent random variables with mgf's $M_{X_1}(t), \dots, M_{X_n}(t)$, respectively, then the mgf of random variable $Y = X_1 + \cdots + X_n$ is given by

$$M_Y(t) = M_{X_1}(t) \cdots M_{X_n}(t).$$

Theorem 6.1.1 states that mgf's are unique, and Theorems 6.1.2 & 6.1.3 combined provide a process for finding the mgf of a linear combination of random variables. All three theorems provide a *Moment-Generating-Function technique* for finding the probability distribution of a function of random variable(s), which we demonstrate with the following examples involving the normal distribution.

Example $6.1.2$

Suppose that $X \sim N(\mu, \sigma)$. It can be shown that the mgf of X is given by

$$M_X(t) = e^{\mu t + (\sigma^2 t^2/2)}, \quad \text{for } t \in \mathbb{R}.$$

Using this mgf formula, we can show that $Z = \dfrac{X-\mu}{\sigma}$ has the standard normal distribution.

1. Note that if $Z \sim N(0,1)$, then the mgf is $M_Z(t) = e^{0t + (1^2 t^2/2)} = e^{t^2/2}$
2. Also note that $\dfrac{X-\mu}{\sigma} = \left(\dfrac{1}{\sigma}\right) X + \left(\dfrac{-\mu}{\sigma}\right)$, so by Theorem 6.1.2,

$$M_{\frac{1}{\sigma}X - \frac{\mu}{\sigma}}(t) = e^{-\frac{\mu t}{\sigma}} M_X\left(\frac{t}{\sigma}\right) = e^{t^2/2}.$$

Thus, we have shown that Z and $\dfrac{X-\mu}{\sigma}$ have the same mgf, which by Theorem 6.1.1, says that they have the same distribution.

Now suppose $X_1, \dots, X_n$ are each independent normally distributed with means $\mu_1, \dots, \mu_n$ and sd's $\sigma_1, \dots, \sigma_n$, respectively.

Let's find the probability distribution of the sum $Y = a_1 X_1 + \cdots + a_n X_n$ ($a_1, \dots, a_n$ constants) using the mgf technique:

By Theorem 6.1.2, we have

$$M_{a_i X_i}(t) = M_{X_i}(a_i t) = e^{\mu_i a_i t + (\sigma_i)^2 (a_i)^2 t^2/2}, \quad \text{for } i = 1, \cdots, n,$$

and then by Theorem 6.1.3 we get the following:

$$\begin{aligned} M_Y(t) &= M_{a_1 X_1}(t) \cdot M_{a_2 X_2}(t) \cdots M_{a_n X_n}(t) \\ &= e^{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2} e^{\mu_2 a_2 t + \sigma_2^2 a_2^2 t^2/2} \dots e^{\mu_n a_n t + \sigma_n^2 a_n^2 t^2/2} \\ &= e^{(\mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n) t + (\sigma_1^2 a_1^2 + \sigma_2^2 a_2^2 + \cdots + \sigma_n^2 a_n^2) \frac{t^2}{2}} \\ \Rightarrow M_Y(t) &= e^{\mu_y t + \sigma_y^2 t^2/2} \end{aligned}$$

Thus, by Theorem 6.1.1, $Y \sim N(\mu_y, \sigma_y)$.

The second part of Example 6.1.2 proved the following result, which we will use in the next section.

Sums of Independent Normal Random Variables

If $X_1, \dots, X_n$ are mutually independent normal random variables with means $\mu_1, \dots, \mu_n$ and standard deviations $\sigma_1, \dots, \sigma_n$, respectively, then the linear combination

$$Y = a_1 X_1 + \cdots + c_n X_n = \sum_{i=1}^{n} a_i X_i,$$

is normally distributed with the following mean and variance:

$$\mu_Y = a_1 \mu_1 + \cdots + a_n \mu_n = \sum_{i=1}^{n} a_i \mu_i \qquad \sigma_Y^2 = a_1^2 \sigma_1^2 + \cdots + a_n \sigma_n^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

Video: Functions of Normal Random Variables

This page titled 6.1: Functions of Normal Random Variables is shared under a license and was authored, remixed, and/or curated by LibreTexts.

6.2: Sample Mean

Suppose we are interested in understanding the mean of some population of values, but do not have full information about the entire population. One approach to solving this problem is to obtain a random sample of a subset of values from the population and consider the mean of the sample. The mean of the sample is referred to as the *sample mean*. Since the sample is randomly selected, the sample mean may be thought of as a function applied to a collection of random variables.

Example $6.2.1$

Suppose we want to know the average SAT math score for girls in Indiana. We could randomly select seniors from high schools in the South Bend School Corporation as a *sample* from all IN girls, and use the mean SAT math score for the South Bend girls as an estimate of the overall mean for IN girls.

The mean of SB girls depends on which sample we randomly select, therefore the *sample mean is a random variable*.

The probability distribution of the sample mean is referred to as the ***sampling distribution*** of the sample mean. The following result, which is a corollary to *Sums of Independent Normal Random Variables*, indicates how to find the sampling distribution when the population of values follows a normal distribution.

Corollary $6.2.1$

If $X_1, \dots, X_n$ represent the values of a random sample from a $N(\mu, \sigma)$ population, then the *sample mean*

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{1}{n} X_i,$$

is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. In other words, we can write

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}).$$

Proof

1. Sample observations are independent when randomly selected. Furthermore, each observation has same distribution as population. $X_1, \cdots, X_n$ represent the observations in the random sample $\implies X_1, \cdots, X_n$ are independent and each $X_i \sim N(\mu, \sigma)$
2. $\bar{X}$ is the sum of independent normally distributed random variables:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n} = \sum_{i=1}^{n} \frac{1}{n} X_i \implies a_i = \frac{1}{n}, \text{ for } i = 1, \cdots, n$$

3. By *Sums of Independent Normal Random Variables*: $\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}})$, where

$$\mu_{\bar{X}} = \sum_{i=1}^{n} \frac{1}{n}\mu = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = n\frac{1}{n}\mu = \mu$$
$$\sigma^2_{\bar{X}} = \sum_{i=1}^{n} (\frac{1}{n})^2 \sigma^2 = n\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$
$$\implies \sigma_{\bar{X}} = \sqrt{\sigma^2_{\bar{X}}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

So $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Example $6.2.2$

Suppose that SAT math scores for girls in Indiana are assumed to be $N(549, 24)$.

Find and compare the sampling distributions for the sample means from a sample of size $n = 16$ and a sample of size $n = 36$.

For $n = 16$: sample mean $\bar{X} \sim \mathrm{N}\left(549, \frac{24}{\sqrt{16}} = 6\right)$

For $n = 36$: sample mean $\bar{Y} \sim \mathrm{N}\left(549, \frac{24}{\sqrt{36}} = 4\right)$

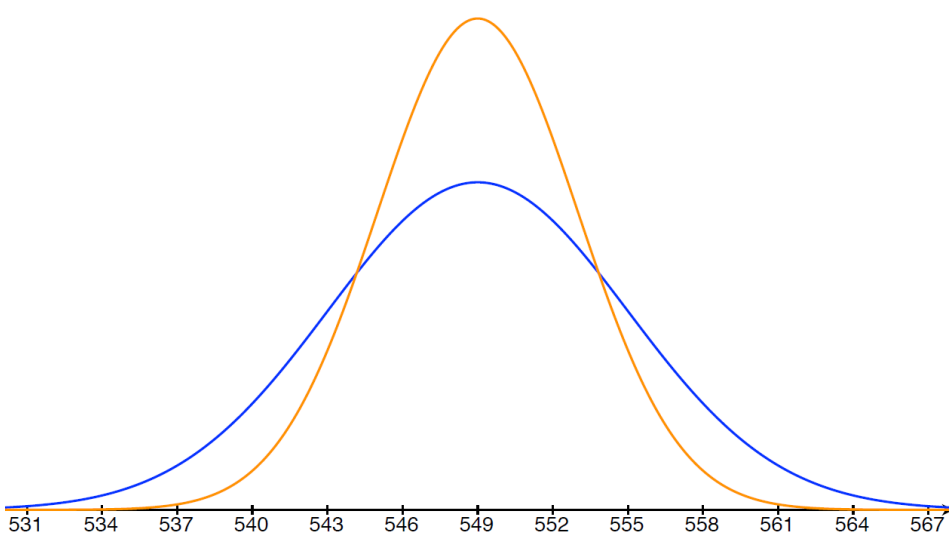
Let's find the probability that each sample mean will be within 10 points of actual population mean ($\mu = 549$):

$\bar{X} : P(|\bar{X} - \mu| \leq 10) = P(539 \leq \bar{X} \leq 559) = \text{normalcdf}(539, 559, 549, 6) = 0.9044$

$\bar{Y} : P(539 \leq \bar{Y} \leq 559) = \text{normalcdf}(539, 559, 549, 4) = 0.9876$

Note that the "normalcdf" in the above equations refers to the built-in function on many graphing calculators used to evaluate probabilities for the normal distribution. If you do not have a graphing calculator, do not worry! Many programming languages, such as Python and R, offer built-in functions to evaluate normal probabilities. However, an even simpler option is to use the online normal distribution calculator available at this link.

The following figure gives the plot of the pdf's for the sampling distributions of $\bar{X}$(blue) and $\bar{Y}$(yellow). Note that the spread of the pdf for $\bar{X}$ is *larger* than for $\bar{Y}$. This is due to the fact that the sample size that $\bar{X}$ is based on is *smaller* than the sample size for $\bar{Y}$. In other words, the sd of the sample mean is inversely related to the sample size, which can be seen in the formula provided by Corollary 6.2.1 where we see that the sample size occurs in the denominator.

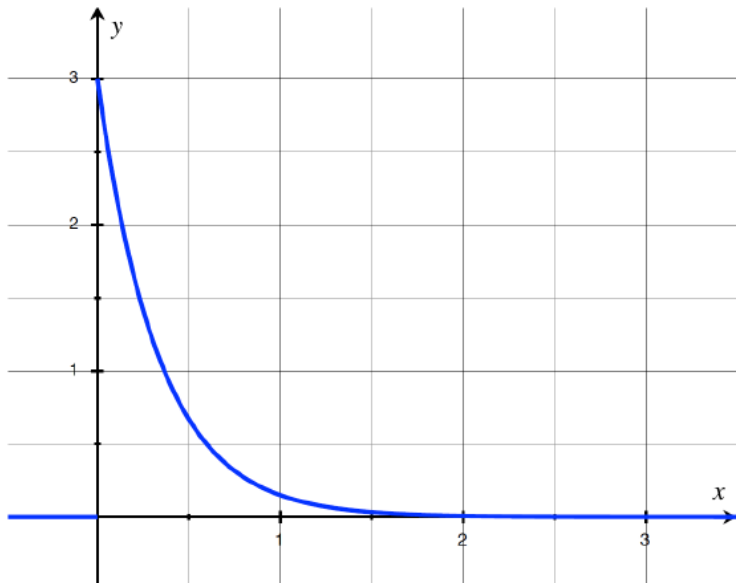


The Central Limit Theorem

We saw that when "sampling" from a normally distributed population, the sampling distribution of the sample mean is also normal. But what if the population does not follow a normal distribution? What if it is *skewed* or *uniform*?

Example $6.2.3$

Suppose we are interested in the lifetime of a radioactive particle. The probability distribution of such lifetimes can be modeled with an exponential distribution. If $\lambda = 3$, for example, then the pdf is *skewed right,* because there is a tail of values with very low probabilities off to the right.



Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from *any* probability distribution with mean μ and sd σ. Then as the sample size $n \to \infty$, the probability distribution of the sample mean approaches the normal distribution. We write:

$$\bar{X} \xrightarrow{d} N(\mu, \sigma/\sqrt{n}), \quad \text{as } n \to \infty \tag{6.2.1}$$

In other words, if n is *sufficiently large*, we can *approximate* the sampling distribution of the sample mean as $N(\mu, \sigma/\sqrt{n})$.

Furthermore,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1) \text{ as } n \to \infty \tag{6.2.2}$$

The d above the arrow in Equation $6.2.1$ above stands for *distribution* and indicates that, as the sample size increases without bound, the limit of the probability distribution of $\bar{X}$ is given by the $N(\mu, \sigma/\sqrt{n})$ distribution. This is referred to as *convergence in distribution*.

What's "sufficiently large"?

- If the distribution of the X_i is *symmetric, unimodal or continuous*, then a sample size n as small as 4 or 5 yields an adequate

approximation.

- If the distribution of the X_i is *skewed*, then a sample size n of at least 25 or 30 yields an adequate approximation.
- If the distribution of the X_i is *extremely skewed*, then you may need an even larger n.

The following website provides a simulation of sampling distributions and demonstrates the Central Limit Theorem (link available).

Example $6.2.4$

Continuing in the context of Example 6.2.3, suppose we sample $n = 25$ such radioactive particles. Then the sampling distribution of the mean of the sample is approximated as follows.

Letting $X_1, \ldots, X_{25}$ denote the random sample, we have that each $X_i \sim \text{exponential}(\lambda = 3)$. By the Properties of Exponential Distributions, we know that the mean of an exponential(3) distribution is given by $\mu = \frac{1}{\lambda} = \frac{1}{3}$ and the sd is also $\sigma = \frac{1}{\lambda} = \frac{1}{3}$. Thus, the sampling distribution of the sample mean is

$$\bar{X} \sim N\left(\frac{1}{3}, \frac{1/3}{\sqrt{25}}\right) \Rightarrow N\left(\frac{1}{3}, \frac{1}{15}\right).$$

What is the use of the Central Limit Theorem if we don't know μ, the mean of the population? We can use the CLT to approximate ***estimation error probabilities***:

$$P(|\bar{x} - \mu| \leq \varepsilon), \tag{6.2.3}$$

the probability that $\bar{X}$ is within ε units of μ. By the Central Limit Theorem and Equation $6.2.2$, we know

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1). \tag{6.2.4}$$

From this fact, we can isolate μ in the inequality in Equation $6.2.3$ as follows:

$$P(|\bar{X} - \mu| \leq \varepsilon) = P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \approx P\left(|Z| \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) = P\left(-\frac{\varepsilon}{\sigma/\sqrt{n}} \leq Z \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \tag{6.2.5}$$

Example $6.2.5$

Now suppose that we do not know the rate at which the radioactive particle of interest decays, i.e., we do not know the mean lifetime of such particles. We can develop a method for *approximating* the probability that the mean of a sample of size $n = 25$ is within 1 unit of the mean lifetime.

In other words, we want $P(|\bar{X} - \mu| \leq 1)$.

By the Central Limit Theorem and Equation $6.2.2$, we know that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{25}} = \frac{\bar{X} - \mu}{\sigma/5} \approx N(0, 1).$$

From this we derive a formula for the desired probability:

$$P\left(\frac{\bar{X} - \mu}{\sigma/5} \leq \frac{1}{\sigma/5}\right) \approx P\left(|Z| \leq \frac{5}{\sigma}\right)$$

This page titled 6.2: Sample Mean is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

7: THE SAMPLE VARIANCE AND OTHER DISTRIBUTIONS

7.1: Other Useful Distributions

7.2: Sample Variance

This page titled 7: The Sample Variance and Other Distributions is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

7.1: Other Useful Distributions

In this section, we introduce three more continuous probability distributions: the chi-squared, t, and F distributions. All three of these are very useful in the study of statistics. And we will see that each are built from the normal distribution in some way.

Chi-Squared Distributions

Definition $7.1.1$

If $Z \sim N(0,1)$, then the probability distribution of $U = Z^2$ is called the ***chi-squared distribution*** with 1 degree of freedom (df) and is denoted χ_1^2.

This definition of the chi-squared distribution with 1 df is stated in terms of a standard normal random variable, which we can relate to any non-standard normal random variable as follows. Let $X \sim N(\mu, \sigma^2)$, then

$$\frac{X-\mu}{\sigma} \sim N(0,1) \quad \Rightarrow \quad \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2.$$

We can also extend the definition of a chi-squared distribution to have more than one degree of freedom by simply summing any number of independent χ_1^2 distributed random variables.

Definition $7.1.2$

If $U_1, U_2, \dots, U_n$ are independent χ_1^2 random variables, then the probability distribution of $V = U_1 + U_2 + \cdots + U_n$ is called the ***chi-squared distribution*** with n degrees of freedom (df) and is denoted χ_n^2.

Note that we could have stated Definition 7.1.2 using a collection of independent, standard normal random variables $Z_1, Z_2, \dots, Z_n$, since by Definition 7.1.1 the square of each Z_i^2 is a χ_1^2 distributed random variable. In other words, the following is another possible definition of the chi-squared distribution with n degrees of freedom:

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_n^2$$

Definition 7.1.2 also leads to the following useful property of the chi-squared distribution. Namely, if $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ are independent random variables, then $X + Y \sim \chi_{m+n}^2$. This easily follows from the definition, since every chi-squared distributed random variable is just a sum of independent χ_1^2 random variables, so summing two chi-squared random variables is just one big sum of many χ_1^2 random variables, where the number is given by the sum of the respective degrees of freedom.

While we will not have much reason to use it, the following theorem (stated without proof) provides the explicit formula for the pdf of the chi-squared distribution.

Theorem $7.1.1$

Let the random variable V have a chi-squared distribution with n degrees of freedom. Then V has pdf given by

$$f(v) = \begin{cases} \dfrac{1}{\Gamma(n/2)2^{n/2}} v^{n/2-1} e^{-v/2}, & \text{for } v \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(n/2)$ is a function (referred to as the gamma function) given by the following integral:

$$\Gamma(n/2) = \int_0^\infty t^{n/2-1} e^{-t} dt.$$

Figure 1: Graph of pdf for $\chi^2(1)$ distribution.

The chi-squared distributions are a special case of the gamma distributions with $\alpha = \frac{n}{2}, \lambda = \frac{1}{2}$, which can be used to establish the following properties of the chi-squared distribution.

Properties of Chi-Squared Distributions

If $V \sim \chi^2_n$, then V has the following properties.

1. The mean of V is $\mathrm{E}[V] = n$, i.e., the degrees of freedom.
2. The variance of V is $\mathrm{Var}(V) = 2n$, i.e., twice the degrees of freedom.

Note that there is no closed form equation for the cdf of a chi-squared distribution in general. But programming languages, such as Python and R, have a built-in function to compute chi-squared probabilities.

t Distributions

Definition $7.1.3$

If $Z \sim N(0,1)$ and $U \sim \chi^2_n$, and Z and U are independent, then the probability distribution of

$$T = \frac{Z}{\sqrt{U/n}}$$

is called the ***t* distribution** with n degrees of freedom and is denoted t_n.

Like the chi-squared distributions, the t distributions depend on a parameter referred to as the degrees of freedom.

The t distribution behaves in a similar manner to the standard normal distribution, with one main distinction. To see this, consider the graph provided in Figure 2 below, which gives the pdf's for the standard normal distribution and four different t distributions. Note that just like the standard normal distribution, all t distributions have a similarly bell-shaped pdf curve which is centered and symmetric about 0. In fact, the expected value or mean of any random variable with a t distribution is always equal to 0. However, the t distribution does not have the same variance as the standard normal distribution, in fact all t distributions have larger variance than the standard normal, which can be seen in the graph by looking at the "tails" of the pdf's, i.e., the extreme regions far away from the mean. For instance, consider the regions less than -2 and greater than 2. Note that in these regions, there is more area under each of the t distribution pdf curves because each of these curves is above the pdf of the standard normal distribution. Given that there is more area under the t distribution pdf curves in these extreme regions, this implies that there is a higher probability that the value of a random variable with a t distribution would be this far away from the center compared to a standard normal random variable, meaning that there is greater spread in the values from the mean and hence a larger variance for the t distribution. Stated another way, the "tails" of all t distribution pdf's are ***thicker*** than the "tails" of the standard normal. We will see in the next section that this aspect of the t distribution is what makes it so useful in the study of statistics.

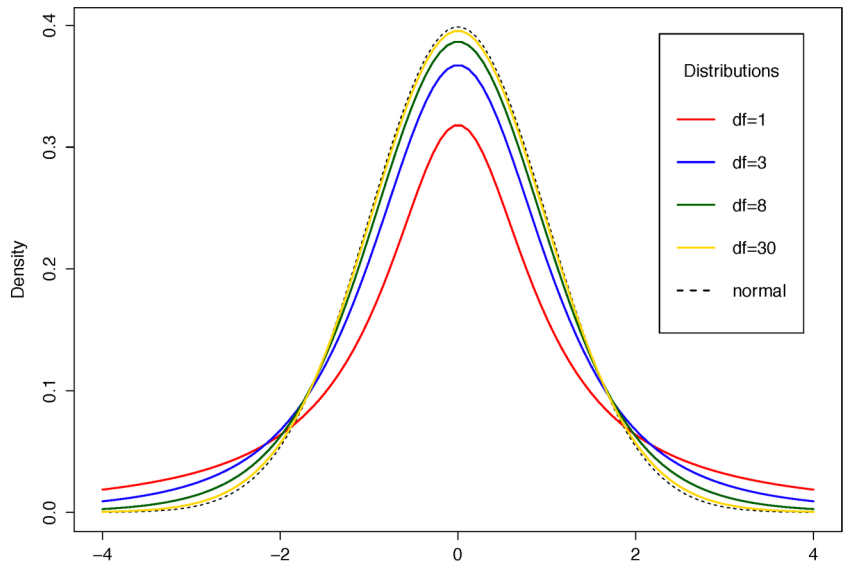


Figure 2: Comparison of t distributions to $N(0,1$ distribution.

One final aspect of the t distribution is to note that as the degrees of freedom increase, the pdf curves approach the standard normal pdf. In Figure 2, consider the red curve, which corresponds to the pdf of a t_1 distribution. This pdf is rather different than the standard normal pdf curve (which is the black dashed line): the middle peak is lower and the tails are much thicker. However, if we consider the yellow curve, which corresponds to the pdf of a t_{30} distribution, we see very little difference from the standard normal pdf. Indeed, it is very hard to distinguish those two pdf's in Figure 2. In fact, the t distributions approach the standard normal distribution in the limit as the degrees of freedom approach infinity:

$$t_n \to N(0,1) \quad \text{as } n \to \infty$$

Again, for completeness more than practical use, we state the following theorem (without proof) which provides the explicit formula for the pdf of the t distribution.

Theorem $7.1.2$

Let the random variable T have a t distribution with n degrees of freedom. Then T has pdf given by

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad \text{for } t \in \mathbb{R},$$

where Γ denotes the gamma function defined in Theorem 7.1.1 above.

F Distributions

Definition $7.1.4$

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent random variables, then the probability distribution of

$$W = \frac{U/m}{V/n}$$

is called the ***F distribution*** with m and n degree of freedom (df) and is denoted $F_{m,n}$.

Given that the F distribution is built from the ratio of two independent chi-squared distributions, the F distribution has two sets of degrees of freedom. The first set, denoted m in Definition 7.1.4, is referred to as the ***numerator degrees of freedom***, and the second set n is referred to as the ***denominator degrees of freedom***.

The following theorem connects the t distribution to the F distribution.

Theorem $7.1.3$

If $T \sim t_n$, then $T^2 \sim F_{1,n}$.

Proof

By Definition 7.1.3, we know that T can be written as $Z/\sqrt{U/n}$, for independent random variables $Z \sim N(0,1)$ and $U \sim \chi^2_n$. This gives

$$T^2 = \frac{Z^2}{U/n}.$$

Now notice that Definition 7.1.1 states that $Z^2 \sim \chi^2_1$, and so T^2 is equal to the ratio of two independent chi-squared random variables, each divided by their degrees of freedom. Thus, by Definition 7.1.4, T^2 has an F distribution with 1 numerator df and n denominator df.

And finally, we state the explicit formula for the pdf of the F distribution for completeness.

Theorem $7.1.4$

Let the random variable W have a F distribution with m and n degrees of freedom. Then W has pdf given by

$$f(w) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{\frac{m}{2}-1} \left(1 + \frac{m}{n} w\right)^{-(m+n)/2}, \quad \text{for } w \geq 0,$$

where Γ denotes the gamma function defined in Theorem 7.1.1 above.

This page titled 7.1: Other Useful Distributions is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.

7.2: Sample Variance

In Section 6.2, we introduced the sample mean $\bar{X}$ as a tool for understanding the mean of a population. In this section, we formalize this idea and extend it to define the *sample variance*, a tool for understanding the *variance* of a population.

Estimating μ and σ^2

Up to now, μ denoted the mean or expected value of a random variable. In other words, it represented a parameter of a probability distribution. In the context of statistics, the main focus is more generally a population of objects, where the objects could be actual individuals and we are interested in a certain characteristic of the individuals, e.g., height or IQ. Oftentimes, we can model the distribution of values in a population using a certain probability distribution, and so it makes sense that the mean of a population is also denoted with μ. However, we may not always have a specific probability distribution in mind when considering the values of a population. Thus, we can generalize the interpretation of μ as the mean of a population as provided in the following definition. Note that the definition also provides a more general interpretation of σ^2 the variance of a population.

Definition $7.2.1$

Suppose that a population has N elements, denoted $x_1, x_2, \dots, x_N$. Then the ***population mean*** μ is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{7.2.1}$$

and the ***population variance*** σ^2 is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2. \tag{7.2.2}$$

As we saw in Section 6.2, we can collect a random sample from a population and use the sample mean to estimate the population mean. More formally, let $X_1, \dots, X_n$ be a collection of independent random variables representing a random sample of observations drawn from a population of interest. Then the sample mean, given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{7.2.3}$$

can be used to estimate the value of the population mean μ.

Note the use of lower case letters "x_i" in Definition 7.2.1 for the elements in the population. This is in contrast to the upper case letters "X_i" used to denote the elements of the random sample. Because the values in a population are fixed, though unknown in practice, it would not be appropriate to represent them with capital letters which are reserved for random variables per convention.

We argue that the sample mean $\bar{X}$ is the "obvious" estimate of the population mean μ because the population elements in Equation $7.2.1$ are simply replaced by the corresponding sample elements in Equation $7.2.3$. In addition to being the natural choice for estimating μ, $\bar{X}$ has another desirable property, which has to do with the following result, stated in Corollary 6.2.1 for normally distributed populations.

Theorem $7.2.1$

For a random sample of size n from a population with mean μ and variance σ^2, it follows that

$$\mathrm{E}[\bar{X}] = \mu,$$
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof

Let $X_1, \ldots, X_n$ denote the elements of the random sample. Then $X_1, \ldots, X_n$ are independent random variables each having the same distribution as the population. In other words, we know that $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, for $i = 1, \ldots, n$. Given this, and using the linearity of expected value and the independence of the sample elements, we have the following:

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu$$
$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Theorem 7.2.1 provides formulas for the expected value and variance of the sample mean, and we see that they both depend on the mean and variance of the population. The fact that the expected value of the sample mean is exactly equal to the population mean indicates that the sample mean is an ***unbiased*** estimator of the population mean. This is because on average, we expect the value of $\bar{X}$ to equal the value of μ, which is precisely the value it is being used to estimate. This is a very desirable property for estimators to have as it lends more confidence to using their values in understanding the unknown population characteristic. We will keep the goal of using unbiased estimators as we now consider estimating the population variance.

Before we tackle the problem of estimating population variance, we again point out that the variance of the sample mean depends on the population variance. Thus, if we are interested in using the variance of $\bar{X}$ to quantify its accuracy in estimating the population mean, we need to know the value of σ^2, which is unlikely. (We talked about this at the end of Section 6.2 in the context of computing error probabilities. See Example 6.2.5.) So we have a specific need for an estimate of σ^2, not just for understanding the distribution of the population better.

Given Equation 7.2.2 in Definition 7.2.1, an "obvious" estimate of σ^2 is given by simply replacing the population elements by the corresponding sample elements, as we did for estimating μ. This gives the following formula for $\hat{\sigma}^2$ (note the "hat" ^), which is our first attempt at estimating σ^2:

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

The problem with this "obvious" estimate is that it is not unbiased. The following theorem (stated without proof) gives the expected value of $\hat{\sigma}^2$.

Theorem 7.2.2

For a random sample of size n from a population with mean μ and variance σ^2, it follows that

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \sigma^2\left(\frac{n-1}{n}\right).$$

As we can see in Theorem 7.2.2, the expected value of $\hat{\sigma}^2$ does not equal σ^2, so it is not an unbiased estimator. Furthermore, note that $\hat{\sigma}^2$ actually ***underestimates*** the value of σ^2, on average, since its expected value is multiplied by a factor less than 1: $(n-1)/n < 1$. This is not good. Putting this again in the context of using the variance of the sample mean to quantify its accuracy in estimating the population mean, if we use $\hat{\sigma}^2$ to estimate σ^2, we would consistently report *higher* accuracy than what is actually being obtained, since smaller variance means less spread or greater confidence in our estimate of μ. This would make our analysis unreliable and misleading.

We can find an *unbiased* estimate of σ^2 by modifying our first attempt in $\hat{\sigma}^2$. The modification is to simply multiply by the reciprocal of the factor on σ^2 in the expected value of $\hat{\sigma}^2$. In doing this, we note that expected value of the modification will equal σ^2, following from the linearity of expected value:

$$\mathrm{E}\left[\left(\frac{n}{n-1}\right)\hat{\sigma}^2\right] = \left(\frac{n}{n-1}\right)\mathrm{E}\left[\hat{\sigma}^2\right] = \left(\frac{n}{n-1}\right)\sigma^2\left(\frac{n-1}{n}\right) = \sigma^2$$

We can simplify the modification of $\hat{\sigma}^2$ algebraically as follows:

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

This leads to the following definition of the ***sample variance***, denoted S^2, our unbiased estimator of the population variance:

$$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The next theorem provides a sampling distribution for the sample variance *in the case that the population is normally distributed*.

Theorem $7.2.3$

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then, it follows that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Theorem 7.2.3 states that the distribution of the sample variance, when sampling from a normally distributed population, is chi-squared with $(n-1)$ degrees of freedom. Note that without knowing that the population is normally distributed, we are not able to say anything about the distribution of the sample variance, not even approximately. There is no "CLT-like" result for the sample variance. Further note that it is not the distribution of S^2 alone, but rather, we multiply by one less than the sample size and divide by the population variance to get the result. This may not seem like a very useful result, given that it is the distribution of a quantity involving both the estimator S^2 and the parameter it is estimating σ^2. But you will see in the study of statistics how this can be utilized to quantify the error in using S^2 to estimate σ^2. But we can use Theorem 2.7.3 to help us in the context of understanding the accuracy of our estimate given by the sample mean. Before we state the result, we need two additional properties regarding the probabilistic qualities of $\bar{X}$ and S^2 as random variables. We state these properties without proof.

Theorem $7.2.4$

1. $\bar{X}$ is independent of the collection of random variables given by $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$.
2. $\bar{X}$ and S^2 are independent.

Note that the second property in Theorem 7.2.4 follows immediately from the first one given our definition of S^2. Using these properties, we can prove the following result.

Theorem $7.2.5$

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then, it follows that

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1} \tag{7.2.4}$$

Proof

We rewrite the quotient by dividing top and bottom by the quantity $\sqrt{\sigma^2/n}$:

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} = \frac{\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right)}{\sqrt{\frac{S^2/n}{\sigma^2/n}}} = \frac{\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right)}{\sqrt{\frac{S^2}{\sigma^2}}} \tag{7.2.5}$$

Note that the quantity in the numerator is the standardization of a normally distributed random variable, thus it has the standard normal distribution. For the denominator, we can further modify the expression under the square root by multiplying top and bottom by the quantity $(n-1)$:

$$\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}} = \sqrt{\left(\frac{(n-1)S^2}{\sigma^2}\right)\frac{1}{n-1}}$$

We know from Theorem 7.2.3 that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and so the denominator in Equation 7.2.5 is the square root of a chi-squared distributed random variable divided by its degrees of freedom. Also note from Theorem 7.2.4 that the numerator and denominator in Equation 7.2.4 are independent random variables, since they are functions of $\bar{X}$ and S^2, respectively. Thus, we have shown that the quantity we started with in Equation 7.2.4 is equal to a random variable with a standard normal distribution divided by the square root of an independent random variable with a chi-squared distribution divided by its degrees of freedom. This is precisely the definition of the t distribution given in Definition 7.1.3.

Notice what the result of Theorem 7.2.5 says: when sampling from a normally distributed population, if we take the sample mean and subtract its expected value μ and divide by its standard deviation *where the population variance σ^2 is estimated by the sample variance S^2*, then the resulting random variable has a t distribution with $(n-1)$ degrees of freedom. The distribution is no longer the standard normal distribution because we have now estimated the population variance, which has the effect of increasing the overall variability in the quantity given in Equation 7.2.4. To account for that increased variability, we need a distribution with *thicker tails*, which is precisely what the t distribution provides. Notice also that the degrees of freedom of the t distribution that models the quantity in Equation 7.2.4 is one less than the sample size because we lose a degree of freedom by using the sample variance to estimate the population variance. This result provides the foundation for many statistical inference techniques.

This page titled 7.2: Sample Variance is shared under a license and was authored, remixed, and/or curated by LibreTexts.

INDEX

No index terms found.

This page titled Index is shared under a license and was authored, remixed, and/or curated by LibreTexts.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a license and was authored, remixed, and/or curated by LibreTexts.

DETAILED LICENSING

Overview

Title: DSCI 500B Essential Probability Theory for Data Science (Kuter)

Webpages: 37

All licenses found:

- Undeclared: 100% (37 pages)

By Page

- DSCI 500B Essential Probability Theory for Data Science (Kuter) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: What is Probability? - *Undeclared*
    - 1.1: Sample Spaces and Events - *Undeclared*
    - 1.2: Probability Measures - *Undeclared*
    - 1.3: Equally Likely Outcomes and Counting Techniques (Combinatorics) - *Undeclared*
  - 2: Conditional Probability - *Undeclared*
    - 2.1: Conditional Probability and Bayes' Rule - *Undeclared*
    - 2.2: Independent Events - *Undeclared*
  - 3: Discrete Random Variables - *Undeclared*
    - 3.1: Introduction to Random Variables - *Undeclared*
    - 3.2: Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) for Discrete Random Variables - *Undeclared*
    - 3.3: Bernoulli and Binomial Distributions - *Undeclared*
    - 3.4: Expected Value of Discrete Random Variables - *Undeclared*
    - 3.5: Variance of Discrete Random Variables - *Undeclared*
  - 4: Continuous Random Variables - *Undeclared*
    - 4.1: Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs) for Continuous Random Variables - *Undeclared*
    - 4.2: Expected Value and Variance of Continuous Random Variables - *Undeclared*
    - 4.3: Uniform Distributions - *Undeclared*
    - 4.4: Normal Distributions - *Undeclared*
  - 5: Multivariate Random Variables - *Undeclared*
    - 5.1: Joint Distributions of Discrete Random Variables - *Undeclared*
    - 5.2: Joint Distributions of Continuous Random Variables - *Undeclared*
  - 6: The Sample Mean and Central Limit Theorem - *Undeclared*
    - 6.1: Functions of Normal Random Variables - *Undeclared*
    - 6.2: Sample Mean - *Undeclared*
  - 7: The Sample Variance and Other Distributions - *Undeclared*
    - 7.1: Other Useful Distributions - *Undeclared*
    - 7.2: Sample Variance - *Undeclared*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*

This page titled DSCI 500B Essential Probability Theory for Data Science (Kuter) is shared under a license and was authored, remixed, and/or curated by Kristin Kuter.