NoBS Statistics

Calculus-based
Introductory Probability & Statistics

Jeffrey Wang

January 8, 2019 – January 13, 2019
version 2019.01.13.03:06
First edition

Contents

Author's Notes

NoBS

NoBS, short for "no bull$#!%", strives for succinct guides that use simple, smaller, relatable concepts to develop a full understanding of overarching concepts.

What NoBS Statistics covers

This guide succinctly and comprehensively covers most topics in an explanatory notes format for a college-level introductory calculus-based probability and statistics course.

Prerequisites

You don't need any prior probability or statistics experience, although the experience you had in elementary, middle, and perhaps high school would be helpful. AP Statistics does not greatly benefit you in this material matter.

However, this is a calculus-based course, so at the very least, you should have taken Calculus 2. For certain sections in this study guide, Multivariable Calculus is necessary. (In particular, joint probability distributions will use double integrals, and you should know Fubini's Theorem and the like.)

What this study guide does

It explains all the concepts to you in an intuitive way so you understand the course material better.

If you are a mathematics major, it is recommended you read a proof-based book.

If you are not a mathematics major, this study guide is intended to make your life easier.

What this study guide does not do

This study guide is not intended as a replacement for a textbook. This study guide does not teach via proofs; rather, it teaches by concepts. If you are looking for a formalized, proof-based textbook, seek other sources.

Other study resources

NoBS Statistics should by no means be your sole study material. Refer to your textbook for further studying.

Dedication

To all those that helped me in life: this is for you.

Sources

This guide has been inspired by, and in some cases borrows, certain material from the following sources, which are indicated below and will be referenced throughout this guide by parentheses and their names when necessary:

- *Probability and Statistics for Engineering and the Sciences* (9th ed.) by Jay L. Devore
- *Introduction to Probability* by Dimitri P. Bertsekas and John N. Tsitsiklis
- Some inline citations do not appear here but next to their borrowed content.

Copyright and resale

Chapter 1

Introduction and descriptive statistics

1.1 Introduction

In statistics, we deal with data. In particular, this data will probably be collected from a **sample**, which is a subset of a **population** that will hopefully be representative of the population.

Example: Let's say you want to know the percentage of people who like vanilla ice cream in a high school. If you survey 100 out of 500 students, then these 100 people's are a sample of the population.

This data, which can vary (so we call it a **variable**), can be a variation of one thing (univariate), two things (bivariate), or multiple things (multivariate).

Of course, this data will probably have some **randomness** to it, since a subset of a population never gives the whole picture. But, we can get close to it. And that is the beauty of statistics.

1.2 Mean, median, and mode

The mean is the average of a sample. Sum all of the data in your sample set and divide by the set's cardinality, or how many things are in the set.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The median is the middle value of the ordered set.

$$\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ ordered value if } n \text{ odd} \\ \text{Average of } \left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n+1}{2}\right)^{th} \text{ ordered value if } n \text{ even} \end{cases}$$

The mode is the value that appears most frequently in the sample set.

1.3 Range and variability

The range is the smallest value subtracted from the largest value. It shows how widely varying the data is.

The sample variance is denoted s^2.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

The standard deviation is the positive square root of the variance: $s = \sqrt{s^2}$. Both the variance and standard deviation are nonnegative.

1.4 Pictorial and tabular representations

Stem and leaf plots, dot plots, histograms, and boxplots (also known as box-and-whisker plots) are all ways to represent data. Each have their own benefits.

Histograms are better suited for closely-grouped and repeating data, while stem and leaf plots are better for widely varying, nonrepeating data. Dot plots are similar to stem and leaf but do not display all of the data. Boxplots are good at showing the variability of the data, but requires you to compute the first and third quartile of the data, in addition to the mean and minimum/maximum of the data.

Chapter 2

Probability

Probability takes whole populations and analyzes subsets of them. This usually includes the chance that you might get a certain subset, expressed as a decimal value between 0 and 1, inclusive.

2.1 Sample space, and events

In probability, we get subsets of populations through experiments. The **sample space** of an experiment, denoted by S, is the set of all possible outcomes of that experiment (Devore). This is somewhat of an abstract concept, but think of the ways you could flip a coin.

An **event** is any collection (subset) of outcomes contained in the sample space S. (Devore)

Example: Let H denote heads and T denote tails. Then, for two coin tosses, an event could be HH, TT, HT, or TH. Collectively, they are the sample space of the experiment.

2.2 Set theory

It is helpful to treat events as sets so that we may use set theory to manipulate the representations. Given events A and B (Devore):

1. The **complement**, denoted by A^C, is the set of all outcomes in S that are not contained in A.

2. The **union** of A and B, denoted by $A \cup B$, is the set of all outcomes either in A or B or in both.

3. The **intersection** of A and B, denoted $A \cap B$, is the set of all outcomes in both A and B at the same time.

Union and intersection can be thought of as the OR and AND operations in Boolean algebra. Indeed, you can say "A or B" instead of "A union B" when describing $A \cup B$.

We define a **null event** as the empty set, denoted $\emptyset$, which is the event consisting of no outcomes at all.

Two sets A and B are **disjoint** if their intersection is the empty set.

$$A \cap B = \emptyset \iff A, B \text{ disjoint}$$

The set operations can be extended to more than two sets. For sake of clarity, it is best to use parenthesis to denote order of operations, although if there are no parentheses, then the complement has precedence over intersection, and intersection has precedence over union.

2.3 Axioms of probability

Given an experiment and a sample space S, the objective of probability is to assign to each event A a number $\mathrm{P}(A)$, called the probability of the event A, which will give a precise measure of the chance that A will occur. (Devore)

In order for us to determine probabilities, we first need to define some fundamental ground rules, called axioms.

Axiom 1: For any event A, $0 \leq \mathrm{P}(A) \leq 1$.

Axiom 2: $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(S) = 1$.

Axiom 3: If $A_1, A_2, \ldots$ is an infinite collection of *disjoint* events, then:

$$\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

What axiom 3 really means is that you can add up the probabilities of disjoint events. For instance, let's return to our example of two coin tosses. HH and TT are clearly disjoint events, so we can say the probability that $HH \cup TT$ occurs, $\mathrm{P}(HH \cup TT)$, is $\mathrm{P}(HH) + \mathrm{P}(TT)$.

2.4 Subsequent properties of probability

Axioms were fundamental, but we can now create more properties of probability from these axioms.

Complement property. For any event A, $\mathrm{P}(A) + \mathrm{P}(A^C) = 1$. Thus, we can say $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$ and vice versa.

Inclusion-exclusion principle. For any two events A and B,

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

Intuitively, you may be tempted to say that the subtraction of the intersection is not needed, but it is in fact necessary because $\mathrm{P}(A) + \mathrm{P}(B)$ double counts the elements that are intersecting. Remember, $A \cap B \subseteq A$ ($\subseteq$ means "subset of").

The inclusion-exclusion principle for three events A, B, C is:

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C) - \mathrm{P}(A \cap B) - \mathrm{P}(A \cap C) - \mathrm{P}(B \cap C) + \mathrm{P}(A \cup B \cup C)$$

It can be generalized to as many events as you want, but typically it is (painfully) derived from these lower forms.

2.5 Combinatorics

We use counting methods to enhance our probability determinations.

Product rule. Suppose we have an event A that can broken down into a sequence of events $E_1, E_2, \ldots, E_k$. First, let the number of outcomes of E_1 be n_1 and so on. Then, the number of possibilities for A can be determined by $n_1 n_2 \ldots n_k$.

Sum rule. Suppose we have an event A that can be broken down into multiple possibilities of events $E_1, E_2, \ldots, E_k$ that are all disjoint. First, let the number of outcomes of E_1 be n_1 and so on. Then, the number of possibilities for A can be determined by $n_1 + n_2 + \cdots + n_k$.

These two rules can be combined. For instance, you could have $n_1(n_{2,1} + n_{2,2})n_3$.

The number of ways to select a subset of size k from the set of size n is called a **permutation**. We denote this as $\mathrm{P}(n, k)$. It can be calculated by:

$$\mathrm{P}(n, k) = \frac{n!}{(n-k)!}$$

In permutations, the order in which we make these selections matters and all of the possible orders are included. If we do not care about order, then we should use **combinations** instead, which filter out all of the orders.

$$\mathrm{C}(n, k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

We pronounce $\binom{n}{k}$ as "n choose k".

If we make repetitions, then we get **permutations with repetition**, which is simple: $\mathrm{P}^*(n, k) = n^k$. However, **combinations with repetition** are more nuanced. They can be determined by:

$$\mathrm{C}^*(n, k) = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

This is most widely known as the stars and bars method, where there are n stars (items) separated by $k-1$ bars (effectively k boxes or partitions).

Now, it's important to note that permutations and combinations directly stem from the product rule. So, keep in mind that product rule works by decomposing an event into smaller independent events of a certain sequence, so you should aim to do the same with permutations and combinations, and use them with the regular product and sum rules.

2.6 Conditional probability

Sometimes, probabilities depend on previous events occurring. For instance, the probability that you'll be soaked is much higher when it rains than when it doesn't rain.

In this case, we'll use **conditional probability** to indicate this. Given events A and B, $\mathrm{P}(A|B)$ is the probability that A happens given that B has happened.

The conditional probability can also be thought of as the probability that both A and B occur, but out of the probability that B has occurred. This can be written as:

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

Rewriting this equation yields the **multiplication rule**:

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

We call $\mathrm{P}(B)$ the **prior probability** (or the "before" probability). Then, the **posterior probability** (a fancy word meaning "after") is $\mathrm{P}(A|B)$.

Now, it is possible for you to switch between conditional probabilities and intersections. This definition is highly useful for conditional probability and is the basis of a very important rule called Bayes' Theorem.

First, we need to generalize the posterior probability to cover multiple disjoint events $A_1, A_2, \ldots, A_k$. We can formulate the **Law of Total Probability** from this:

$$\mathrm{P}(B) = \mathrm{P}(B|A_1)\mathrm{P}(A_1) + \cdots + \mathrm{P}(B|A_k)\mathrm{P}(A_k) = \sum_{i=1}^{k} \mathrm{P}(B|A_i)\mathrm{P}(A_i)$$

This law simply extends the definition of conditional probability to breaking apart an event A into multiple disjoint events $A_1, \ldots, A_k$.

Given this law, we can now formulate Bayes' Theorem, which gives the posterior probability of a subevent A_j given that B has occurred. The simple form is $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$, but when we have a disjoint series of events, use the form using the Law of Total Probability:

Bayes' Theorem. Let $A_1, A_2, \ldots, A_k$ be a collection of k-many mutually exclusive and exhaustive events with prior probabilities $\mathrm{P}(A_i)$ where $i = 1, \ldots, k$. Then, for any other event B for which $\mathrm{P}(B) > 0$, the posterior probability of A_j given that B has occurred is: (Devore)

$$\mathrm{P}(A_j|B) = \frac{\mathrm{P}(B|A_j)\mathrm{P}(A_j)}{\sum_{i=1}^{k} \mathrm{P}(B|A_i)\mathrm{P}(A_i)} \text{ where } j \in [1, k]$$

2.7 Independence

Some events happen independently of each other, and this gives them special properties that make them easier to deal with in probability and statistics.

For instance, the probability that you will eat pizza for lunch outdoors and the probability that it's snowing in Australia are independent. However, the probability that you will eat pizza for lunch and the probability you get acid reflux are not independent, since pizza is a food that may cause acid reflux.

So, how can we define independence? By using conditional probability, the definition of independence says that the condition that an event will occur should not be affected by another event happening first.

$$A, B \text{ independent} \iff \mathrm{P}(A|B) = \mathrm{P}(A) \iff \mathrm{P}(B|A) = \mathrm{P}(B)$$

From this, we can also derive a **multiplication rule for independent events**.

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B) \iff A, B \text{ independent}$$

This means that if two events are independent, we can simply multiply their probabilities together to get their intersection's probability, which is very convenient.

We can extend the definition of independence to multiple events, meaning they will be **mutually independent** if for every $k \in [2, n]$ and every subset of indices $i_1, i_2, \ldots, i_k$:

$$\mathrm{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathrm{P}(A_{i_1})\mathrm{P}(A_{i_2}) \ldots \mathrm{P}(A_{i_k})$$

Chapter 3

Random variables

3.1 Introduction to random variables

A **random variable**, shortened to **rv**, is a variable that can take on random values, usually the results of an experiment. Unlike in regular math, where we know the value of the variable is fixed, we do not know what the value of the random variable is. However, we can predict a pattern for it, which we will discuss later, not necessarily limiting its range but identifying areas of high probability of occurrence.

Definition of random variable (from Devore): For a given sample space S of some experiment, a **random variable (rv)** is any rule that associates a number with each outcome in S. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

Random variables are denoted by uppercase letters, like X, Y, or Z, as opposed to lowercase letters, which are reserved for fixed variables.

There are two kinds of random variables: discrete and continuous.

Discrete random variables are reserved for cases where only countable integer possibilities exist. For instance, the number of coin flips required to get 4 tails is a discrete rv. It is not possible to need 13.75 coin flips for this to happen because coin flips are measured by discrete, countable numbers.

On the other hand, **continuous random variables** are for anything that isn't countable. The lifetime of a lightbulb, for instance, is a continuous rv, because it is possible the lightbulb works for exactly 2.3235842384824 years.

These random variables tend to fit distributions that have been created by mathematicians, which can model random variables very well depending on what the rv is for. Discrete and continuous distributions will be discussed in the next two chapters.

3.2 Expected values

It's frustrating to use random variables because they don't have a specified value; they can be anything! Hence, we want to introduce some order into the randomness. It turns out if we identify a random variable's distribution, we can 'expect' a certain value from it. The expected value can be thought of as the mean value of the rv's distribution. The calculation for this depends upon whether the rv is discrete or continuous, but the purpose and concept are the same for both. We will explore how to mathematically find the expected values in many ways.

An expected value of some random variable X is usually denoted by $\mathrm{E}(X)$. Sometimes, we can also let the Greek letter mu denote expected value, with the subscript being the rv: μ_X.

Expected value of functions

Sometimes, we need the expected value of not just a rv by itself but a function using a rv. We'll denote this general kind of function by $h(X)$.

Expected value of a linear function

Let's say $h(X)$ is a linear function. Then,

$$\mathrm{E}(aX + b) = a \cdot \mathrm{E}(X) + b$$

where a is a coefficient of the rv X and b is a constant additive factor.

Expected value of a quadratic function

This will actually depend on another concept called variance. But, it would be calculated by:

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

3.3 Variance

While the expected value gives a probable mean value for a certain rv, the variance of the possible values resulting from this rv is also really important. If all of the data varies, then the expected value is not very useful. So, we need to determine the variance of the rv. Again, the manner of determination varies between discrete and continuous rv distributions.

Given the random variable X, the variance is denoted by $\mathrm{Var}(X)$; alternatively, σ_X^2. (σ_X is the standard deviation.)

However, we can determine the variance using the expected value.

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

If you look back, this is how the expected value of a quadratic function can be determined.

Variance of a linear function

From this, we can also determine the variance of a linear function:

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

Note that the constant factor disappears. This is because variance does not depend on any constant sliding factors, it only depends on the coefficient by which it varies.

Chapter 4

Discrete random variables and their distributions

As previously established, discrete random variables are those whose range only extends through the integers rather than the reals ($X \in \mathbb{Z}$). Therefore, discrete rv's are usually used for countable applications.

Discrete rv's tend to fit several general categories of probability distributions, which we will go over in this chapter.

4.1 Probability mass functions (pmf) and discrete cumulative distribution functions (cdf)

Probability mass functions (pmf)

We can usually use a function to model the probabilities that the random variable X will be equal to a certain value x in the range. We write this as

$$p_X(x) = \mathrm{P}(X = x)$$

$p_X(x)$ is known as the **probability mass function (pmf)**. (From Devore:) In other words, for every possible value of x of the rv, the pmf specifies the probability of observing that value when the experiment is performed.

Cumulative distribution functions (cdf)

The **cumulative distribution function** is basically the summation of the pmf, a discrete integral of sorts.

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{y=-\infty}^{x} p_X(y)$$

Parameters

A probability distribution may depend on another fixed value in order to calculate the pmf. Anything other than x is known as a **parameter**. For instance, later, you will see there is a

parameter for a Bernoulli distribution that determines the probability of a success. Since this might not make any sense, revisit this concept once you've seen a distribution.

Calculating discrete expected value and variance

The expected value is calculated as a weighted average. For discrete distributions, this is usually done through a summation of products.

Let X be a discrete rv with the set of possible values called D and pmf $p_X(x)$. Then, the expected value or mean value of X is: (Devore)

$$\mathrm{E}(X) = \sum_{x \in D} x \cdot p_X(x)$$

For the expected value of a function $h(X)$, it is slightly different:

$$\mathrm{E}(h(X)) = \sum_{D} h(x) \cdot p_X(x)$$

The variance is also calculated in a similar way and can be linked to the expected value function. Let X have pmf $p(x)$ and expected value $\mathrm{E}(X)$. Then the variance of X is:

$$\mathrm{Var}(X) = \sum_{D} [x - \mathrm{E}(X)]^2 \cdot p_X(x) = \mathrm{E}([X - \mathrm{E}(X)]^2)$$

The standard deviation of some random variable X, σ_X, is simply the square root of the variance: $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

The computational formula for calculating variance was shared in the previous chapter and is an immediate consequence of $\mathrm{E}([X - \mathrm{E}(X)]^2)$.

4.2 Bernoulli distribution

A Bernoulli distribution is when we expect the random variable to have two results, usually known as a success or a failure. A failure is modeled as $\mathrm{P}(X = 0)$ while a success is modeled as $\mathrm{P}(X = 1)$. This can be rewritten as $p(0)$ and $p(1)$ respectively. For instance, a single coin toss's results will be a Bernoulli distribution, because it will either be heads (which we can consider a success) or a tails (could be a failure, but we could reverse it, it's all up to you).

The Bernoulli has a parameter p, which is the probability of a success. The probability of failure is also sometimes known as q, but $q = 1 - p$, so we usually just need p. So a Bernoulli distribution is represented by:

$$X \sim \text{Bernoulli}(x; p)$$

pmf: $p^x(1-p)^{1-x}$
Parameters: p, the probability of success.
Domain of x: $x = 0, 1$
Expected value/mean: $\mathrm{E}(X) = p$
Variance: $\mathrm{Var}(X) = p(1-p)$

4.3 Binomial distribution

Binomial distributions arise when you have series of trials with only two results: success and failure (like in a Bernoulli distribution). However, the series of trials is the distinction here, and they must be independent of each other (the result of one trial should not affect the other). For instance, a series of coin flips could be represented by a binomial distribution. It uses the concept of combinations back from probability to establish the pmf.

The binomial distribution has two parameters: p, the probability of success, and n, the number of trials performed.

Here are the formal requirements for a distribution to qualify as binomial (Devore):

1. The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes: success or failure.
3. The trials are independent; the outcome of one trial doesn't affect any of the others.
4. The probability of success p is constant from trial to trial.

A binomial distribution is represented as:

$$X \sim \text{Bin}(x; n, p)$$

pmf: $\mathrm{C}(n, x) \cdot p^x (1-p)^{n-x}$
Parameters: p, the probability of success; n, the number of trials.
Domain of x: $x = 0, 1, 2, \ldots, n$
Expected value/mean: $\mathrm{E}(X) = np$
Variance: $\mathrm{Var}(X) = np(1-p)$

4.4 Hypergeometric and negative binomial distributions

Hypergeometric distribution

The hypergeometric distribution is related to the binomial distribution. Whereas the binomial distribution involved choosing a certain item in each trial and replacing it for the next trial, the hypergeometric distribution does not replace the chosen item. It is a way to use the results of small samples to represent results of a large population.

A hypergeometric distribution is represented as:

$$X \sim h(x; n, M, N)$$

pmf: $\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

Parameters: n, the sample size; M, the number of successes in a population; N, the population size.

Domain of x: $x = [\max(0, n-N+M), \min(n, M)]$
Expected value/mean: $\mathrm{E}(X) = n \cdot \frac{M}{N}$
Variance: $\mathrm{Var}(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$

Negative binomial distribution

The negative binomial distribution measures how likely it will take for r successes to occur in a Bernoulli trial, with probability of success in each trial being fixed at p. For instance, how likely is it I will need to buy 10 cereal boxes to collect all 3 different cereal box prizes, assuming each cereal box has one and only one prize?

$$X \sim nb(x; r, p)$$

pmf: $\binom{x-1}{r-1} p^r (1-p)^{x-r}$
Parameters: r, the number of successes needed; p, the probability of a success occurring.
Domain of x: $x = 1, 2, \ldots$
Expected value/mean: $\mathrm{E}(X) = \frac{r(1-p)}{p}$
Variance: $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

Geometric distribution

The geometric distribution is a special case of the negative binomial distribution where we only need one success (i.e. $r = 1$).

$$X \sim \text{Geometric}(x; p)$$

pmf: $p(1-p)^{x-1}$
Parameters: p, the probability of a success occurring.
Domain of x: $x = 1, 2, \ldots$
Expected value/mean: $\mathrm{E}(X) = \frac{1}{p}$
Variance: $\mathrm{Var}(X) = \frac{1-p}{p^2}$

4.5 Poisson distribution

Poisson distributions are unlike the other distributions, but are best for modeling the number of events occurring in a timeframe, usually rare. The events occur independently, and the rate at which the events occur is constant. For instance, how many bridges will fail in a year? This can be modeled with a Poisson distribution.

$$X \sim f(x; \mu)$$

We use f to denote a Poisson distribution because poisson is the French word for fish. μ is the Poisson parameter, usually determined through methods that are nontrivial, and which is the expected value and variance of the rv. We'll discuss how to find a possible μ parameter later in the point estimation chapter.

pmf: $\frac{e^{-\mu} \cdot \mu^x}{x!}$
Parameters: μ, the Poisson parameter.
Domain of x: $x = 0, 1, 2, \ldots$
Expected value/mean: $\mathrm{E}(X) = \mu$
Variance: $\mathrm{Var}(X) = \mu$

4.6 Summary of discrete probability distributions

Name	**Params**	**pmf** $p_X(x)$	x	$\mathrm{E}(X)$	$\mathrm{Var}(X)$
Bernoulli	p	$p^x(1-p)^{1-x}$	$x = 0, 1$	p	$p(1-p)$
Binomial	n, p	$\mathrm{C}(n,x)\, p^x(1-p)^{n-x}$	$x = 0, 1, 2, \ldots, n$	np	$np(1-p)$
Hypergeometric	n, M, N	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$x = [\max(0, n-N+M), \min(n, M)]$	$n \cdot \frac{M}{N}$	$\left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$
Negative binomial	r, p	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	$x = 1, 2, \ldots$	$r(1-p)/p$	$r(1-p)/p^2$
Geometric	p	$p(1-p)^{x-1}$	$x = 1, 2, \ldots$	$1/p$	$(1-p)/p^2$
Poisson	λ	$\frac{\lambda^x e^{-\lambda}}{x!}$	$x = 0, 1, 2, \ldots$	λ	λ
Uniform	n	$1/n$	$x = 1, 2, \ldots, n$	$(n+1)/2$	$(n+1)(n-1)/12$

Chapter 5

Continuous random variables and their distributions

Continuous random variables are those whose range extends throughout all of the reals rather than just the integers (X $\in \mathbb{R}$). Therefore, continuous rv's are usually used for uncountable applications.

Continuous rv's tend to also fit several general categories of probability distributions.

5.1 Probability distribution functions (pdf) and continuous cumulative distribution functions (cdf)

Probability distribution functions (pdf)

When dealing with discrete rv's, pmfs provided probabilities for one single discrete value. However, with continuous rv's, the main difference is that probabilities are spread across infinitely many possibilities, since the range is the reals. So, at a single point, $P(X = x) = 0$ for any continuous distribution, since that single point is infinitesimally small.

Instead, we define a probability distribution using some function $f_X(x)$, which is the **probability density function** (pdf). From the pdf, we can find probabilities for ranges using integrals.

Definition from Devore: Let X be a continous rv. Then a pdf (probability density function) of X is a function $f_X(x)$ such that for any two numbers a and b with $a \leq b$,

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

Continuous cumulative distribution functions (cdf)

A continuous probability distribution's cdf has the same concept as a cdf, but it's calculated differently. Instead of using a sum, we use an integral and integrate from negative infinity to the point a, where $f_X(x)$ is the pdf.

$$\mathrm{P}(X \leq a) = \int_{-\infty}^a f_X(x)\,\mathrm{d}x$$

Calculating continuous expected values and variances

These are again conceptually similar to their discrete version but have different formulae to calculate them.

$$\mathrm{E}(X) = \int_a^b x \cdot f_X(x)\ \mathrm{d}x$$

$$\mathrm{Var}(X) = \int_a^b x^2 \cdot f_X(x)\ \mathrm{d}x$$

5.2 Uniform distribution

A **uniform distribution** is one with a flat pdf defined as follows:

$$X \sim U(x; A, B)$$

The uniform distribution has a smaller per-area probability the larger the spread is. When $B - A$ is big, the individual pdf regions will be small, and the opposite is true too.

Since the pdf and cdf are piecewise functions, do any integrations for pdf/cdfs by adjusting the bounds of the integral to be between A and B.

pdf: $\begin{cases} \frac{1}{B-A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$

cdf: $\begin{cases} \frac{x}{B-A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$

Parameters: A, the lower bound; B, the upper bound.

Expected value/mean: $\mathrm{E}(X) = \frac{B+A}{2}$

Variance: $\mathrm{Var}(X) = \frac{(B-A)^2}{12}$

Mode: none

5.3 Normal distribution

A **normal distribution**, also known as a **Gaussian distribution**, is the most common continuous probability distribution. It's shaped like a bell so the most probable value is going to be its center, with its length determined by the variance of the data. When we don't know what kind of distribution a data is, we usually assume it is a normal distribution first. (This will be evident through the central limit theorem, which will be covered later.)

The two parameters of a normal distribution are μ (the expected value) and σ^2 (the variance). There are no complicated formulas to use to calculate the expected value; they *are* the parameters.

$$X \sim N(x; \mu, \sigma^2)$$

pdf: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Parameters: μ, the expected value; σ^2, the variance.

Expected value/mean: $\mathrm{E}(X) = \mu$

Variance: $\mathrm{Var}(X) = \sigma^2$
Mode: μ

Standard normal distribution

The cdf of the normal distribution is really difficult to calculate, so we use lookup tables if we ever need them. However, the cdf function changes for every combination of μ and σ^2. Consequently, we need a standard normal distribution to go off of. We want the mean to be centered in the middle, so we set it to be 0. Then, we want our variance to be balanced between too tight and too far, so we pick it to be 1. This is the **standard normal distribution**, $N(0, 1)$. Its cdf function is called Φ_X. Whenever you want to get $P(X \leq x) = F_X(x)$, you can get it from the phi function. However, how do we get any arbitrary normal distribution to be standard? You *standardize* it by subtracting the nonstandard μ from the x value and dividing it by the nonstandard σ^2:

$$\Phi_X(\frac{x-\mu}{\sigma^2})$$

This will result in $\mu = 0, \sigma^2 = 1$, and you'll be able to use the phi function.

5.4 Exponential, gamma, and beta distributions

The gamma function is used in the pdfs of all three functions. For $\alpha > 0$, the **gamma function** is:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$$

Gamma distribution

The standard gamma distribution has $\beta = 1$.

$$X \sim \text{Gamma}(X; \alpha, \beta)$$

pdf: $\begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1}e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$
cdf: $\begin{cases} \int_0^{x/\beta} \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\,\mathrm{d}y & x > 0 \\ 0 & \text{otherwise} \end{cases}$
Parameters: α, the shape parameter; β, the scale parameter.
Expected value/mean: $\mathrm{E}(X) = \frac{\alpha}{\beta}$
Variance: $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$
Mode: $\frac{\alpha-1}{\beta}$

The gamma distribution is sometimes rewritten with β being replaced by $\frac{1}{\beta}$, which is valid but it's just another parameterization of the function.

Beta distribution

Commonly used to model proportions, the beta distribution uses the gamma distribution and has the special property of having positive density over a finite interval of length (Devore).

$$X = \text{Beta}(x; \alpha, \beta, A, B)$$

pdf: $\begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$

Parameters: α, the shape parameter; β, the scale parameter; A, the lower bound; B, the upper bound.

Expected value/mean: $\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$

Variance: $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Mode: $\frac{\alpha-1}{\alpha+\beta-2}$

The standard beta distribution is when $A = 0$ and $B = 1$.

Chi-squared distribution (χ^2)

This distribution is widely used for statistical inference, due to it representing the sum of squares of ν independent standard normal random variables (Wikipedia). ν (the Greek letter nu, not a v), in this case, is the number of degrees of freedom of the random variable.

The chi-squared distribution is a special case of the gamma distribution if $\alpha = \nu/2$ and $\beta = 2$.

$$X \sim \chi^2(x; \nu)$$

pdf: $\begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Parameters: ν, the number of degrees of freedom.

Expected value/mean: $\mathrm{E}(X) = \frac{\nu}{4}$

Variance: $\mathrm{Var}(X) = \frac{\nu}{8}$

Exponential distribution

The exponential distribution is a special case of the gamma distribution (where $\alpha = 1$ and $\beta = \frac{1}{\lambda}$) is used for cases similar to the Poisson distribution, but when the data is continuous. It is important to note that the exponential distribution is memoryless. For instance, if we want to find the probability a lightbulb will burn out in 10 years given that it has been working for 5 years already, it's the same as finding the probability that a new lightbulb will burn out in 5 years.

$$X \sim \text{Exponential}(X; \lambda)$$

pdf: $\begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

cdf: $\begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$

Parameters: λ, the scale parameter ($\lambda > 0$).

Expected value/mean: $\mathrm{E}(X) = \frac{1}{\lambda}$

Variance: $\mathrm{Var}(X) = \frac{1}{\lambda^2}$

Mode: 0

5.5 Summary of common continuous probability distributions

Name	**Params**	**pdf** $f_x(x)$	x	**Mean/Expected value**	**Variance**	**Mode**
Uniform	a, b	$\frac{1}{b-a}$	$a \leq x \leq b$	$(b+a)/2$	$(b-a)^2/12$	none
Normal	μ, σ^2	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$-\infty < x < \infty$	μ	σ^2	μ
Exponential	λ	$\lambda e^{-\lambda x}$	$0 < x < \infty$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0
Gamma	α, β	$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$	$0 < x < \infty$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
Beta	α, β	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$0 \leq x \leq 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\alpha-1}{\alpha+\beta-2}$

Chapter 6

Joint probability distributions

So far, we have only dealt with one random variable at a time. There are many times when multiple random variables need to be taken into account. This section covers what to do with two random variables, both of which are either discrete or continuous (no mixing).

6.1 Two joint random variables

When we have two random variables, the joint pmf considers how much probability mass is placed on each possible pair of values (x, y). (Devore)

We define a joint pmf to be:

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

Any joint pmf function must be greater than zero, and its sum must be 1, just like a regular univariate pmf.

Now, in case we need some subset of the pmf A, we can sum up all of the values that fall within A.

For discrete rv's, we sum up each case that fits within our bounds.

$$\mathrm{P}((X, Y) \in A) = \sum_{(x,y)\in A} \sum p_{X,Y}(x, y)$$

For continuous rv's, we use a double integral over the appropriate region.

$$\mathrm{P}((X, Y) \in A) = \int\int_{A} p_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

Remember that the order of integration can be swapped as long as it is a valid use of Fubini's Theorem.

Now, say you wanted to get the pmf for just one variable, X. Then, for every possible pmf value of X (let's call it x_i), all of the pmf values of Y where $X = x_i$ are summed together ($y_1 + y_2 + \cdots + y_k$, where k is the number of Y combinations). Represented as an equation, we get the **marginal pmf of** X:

$$p_X(x) = \sum_{i=0}^{k} p_{X,Y}(x, y_i)$$

Similarly, the marginal pmf of Y would be:

$$p_Y(y) = \sum_{j=0}^{k} p_{X,Y}(x_j, y)$$

These are sometimes just shortened to the term "marginal of X" or simply "the marginal" if it's obvious which one we're referring to.

For continuous random variables, we simply replace the sums with integrals. The variable of integration is the one that you're getting rid of.

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\,\mathrm{d}y$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\,\mathrm{d}x$$

These rules can be extended for a quantity of random variables more than two.

6.2 Independent random variables

Using the definition of independence established in the probability chapter, we can say that two random variables X and Y are independent if, for every pair of values, the following is true:

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

If this doesn't hold true for every single pair of values, then X and Y are dependent.

6.3 Conditional distributions

Distributions can be conditional on one another if they are dependent. This can be mathematically defined with a conditional pmf/pdf.

Let X and Y be two continuous rv's with joint pmf/pdf $p_{X,Y}(x, y)$ and marginal X pmf/pdf $p_X(x)$. Then for any X value x for which $f_X(x) > 0$, the conditional pmf/pdf of Y given $X = x$ is: (Devore)

$$f_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

(for continuous cases, the bound is $-\infty < y < \infty$.)

This parallels the definition of conditional probability, $\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}$.

6.4 Expected values for joint random variables

The joint expected value can be found by simply replacing the univariate pmf/pdf $p_X(x)$ with the joint pmf/pdf $p_{X,Y}(x, y)$ in the definition of univariate expected values, and making sure the bounds are respective of both X and Y.

The following is true for any two rv's X and Y:

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

Furthermore, the following are valid for independent rv's X and Y:

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

As a corollary, we can say that:

$$\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y)$$

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

The difference between two independent normally distributed rv's (not necessarily having the same μ and σ^2 parameters) is also normally distributed.

These rules can be extended to any length linear combination of rv's.

6.5 Covariance

Covariance is the measure of variability between two rv's, defined by their expected values:

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

where $\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, p_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$ for continuous rv's. (For discrete rv's, replace the double integrals with double summations.)

The **correlation coefficient** is defined as such:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

where σ_X is the standard deviation of X, or $\sqrt{\mathrm{Var}(X)}$, and similarly for σ_Y as well.

When the correlation coefficient is 0, then the two variables are considered uncorrelated. In fact, if X and Y are independent, then $\rho_{X,Y} = 0$ but the reverse is not necessarily true.

6.6 Conditional expected values, variances, and covariances

Law of iterated expectations

Conditional expected values sometimes appear. For instance, $\mathrm{E}(Y|X)$. Since Y depends on X, $\mathrm{E}(Y|X)$ is actually a random variable instead of a fixed value. In order to calculate the expected value of solely Y, we must get the expected value of the random variable $\mathrm{E}(Y|X)$. This double expected value leads to an iterated chain of expected values, hence we get the **law of iterated expectations**.

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X))$$

Law of iterated variances

The same principle applies to variances, leading to the **law of iterated variances** (also known as the **law of conditional variances**).

$$\operatorname{Var}(Y) = \operatorname{E}(\operatorname{Var}(Y|X)) + \operatorname{Var}(\operatorname{E}(Y|X))$$

Law of total covariance

The law of total covariance parallels the structure of iterated variances. In this case, $\operatorname{E}(X|Z)$ and $\operatorname{E}(Y|Z)$ are random variables that depend on Z.

$$\operatorname{Cov}(X, Y) = \operatorname{E}(\operatorname{Cov}(X, Y|Z)) + \operatorname{Cov}(\operatorname{E}(X|Z), \operatorname{E}(Y|Z))$$

Chapter 7

Derived distributions

A **derived distribution** is a distribution that arises from a random variable being passed through some function $g(\cdot)$. For instance, in the equation $Y = g(X)$, we know that Y is derived from X by it being the result of the function $g(\cdot)$. Our goal is to determine the pmf/pdf and cdf of these derived distributions based off of information from the original random variable's distribution.

7.1 Determination of derived distributions

For instance, let $X \sim U(x; 0, 1)$ be our original rv. As we've just specified, it's uniformly distributed from 0 to 1. Then, let $Y = 2X$. How do we determine what the pdf of Y is?

We know that the pdf of X is $f_X(x) = \frac{1}{B-A} = \frac{1}{1-0} = 1$, by the definition of a uniform distribution. Thus, the cdf of X is $F_X(x) = \int_{-\infty}^{x} f_X(x')\,\mathrm{d}x' = x$ (from 0 to 1).

However, how can we relate X and Y given that $Y = 2X$? The trick is that the sum of both cdf's will be equal to 1 due to the fundamental probability law: all possibilities of a random variable must sum to 1. So, we can set the cdf's equal to each other, make substitutions, and determine Y's cdf.

Thus, we start from Y's cdf. We know that $F_Y(y) = P(Y \leq y)$. Here, we replace Y with $2X$ based on the definition that we have been given. Hence, we get $P(2X \leq y)$. From here, we make this $P(X \leq \frac{y}{2})$, which equals $= F_X(\frac{y}{2}) = \frac{y}{2}$, so we get $F_Y(y) = \frac{y}{2}$.

Now that we have the Y cdf in terms of x, we need to differentiate to get the pdf of Y. This yields us $\frac{\mathrm{d}F_Y}{\mathrm{d}y}(y) = f_Y(y) = \frac{1}{2}$. Therefore, the pdf in this case is $\frac{1}{2}$.

Why couldn't we have just dealt with the pdf's without using the cdf's? Because whenever the bounds change, then we run into an issue. Be sure to keep bounds in mind when dealing with these derived distributions. In this case, the bounds were the same, but for more complicated cases, the bounds may change, such as if we had $Y = \frac{1}{X}$.

7.2 Linear derived distributions

If X is a continuous rv with pdf f_X and $Y = aX + b$ ($a \neq 0$), then: (Bertsekas)

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

Chapter 8

Limit theorems

8.1 Central limit theorem

The **central limit theorem** argues that all large data pools of independent identically distributed random variables $X_1, X_2, \ldots, X_n$ of *any distribution* tend to resemble a normal distribution, where the likelihood centers at a certain point with tails in both directions. This means we can treat all large data pools as normal distributions. What their mean and variance are, we do not know, but for most purposes, we just assume the variance is known.

8.2 Markov's inequality

For nonnegative random variables, Markov's inequality bounds the probability of large values to be very low the lower the expected value is. In other words, the smaller the mean is, the smaller the chance the random variable will take on a large value.

$$\mathrm{P}(X \geq a) = \frac{\mathrm{E}(X)}{a}$$

where $a > 0$.

8.3 Chebyshev's inequality

Chebyshev's inequality has a similar premise as Markov's inequality. The smaller the variance of a rv is, the smaller the probability that it will fall outside a certain range from the mean/expected value. The inequality provides a strict numerical bound of the probability for some given range distance a.

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

The probability's absolute value may trip up some; it's really just a compact way to write $\mathrm{P}(\mathrm{E}(X) - X \leq a \cup \mathrm{E}(X) + X \geq a)$, or $\mathrm{P}(\mathrm{E}(X) - X \leq a) + \mathrm{P}(\mathrm{E}(X) + X \geq a)$. (However, it would be better to keep it in union form for sake of computing a.)

Special care should be taken when converting $\leq, \geq$ into $<, >$ when dealing with discrete rv's.

Chapter 9

Point estimations

When we were dealing with probability distributions, they had parameters. In statistics, we sometimes want to find ways to maximize the parameters to find the most probable result for a distribution.

9.1 Method of moments estimation

9.2 Maximum likelihood estimation

Chapter 10

Classical inferential statistics

10.1 Confidence intervals

10.2 Critical value method

10.3 P-value method

Chapter 11

Bayesian inferential statistics