Mathematical Statistics I
Chapter 1: Probability

Jesse Wheeler

Contents

1 Course Overview

Course Overview

- This course is the first part of a two semester introductory course on Mathematical Statistics.

- Our goal is to cover Chapters 1-10 of "Mathematical Statistics and Data Analysis", by John A. Rice.

- Topics include: Probability, Random Variables, Discrete and Continuous distributions, Order Statistics, Limit Theorems, Point and Interval Estimation, Uniformly most powerful tests, likelihood ratio tests, chi-square and F tests, and nonparameteric tests.

- Roughly speaking, 4450 and 4451 can be broken into two parts:

  - Math 4450: Probability (mathematics of randomness)
  - Math 4451: Statistics (procedures for analyzing data)

1.1 Logistics

Course Logistics

- About Me

- Course Website: https://jeswheel.github.io/4450_f25/.

- Canvas: Canvas will be used to submit assignments, view grades, and for course announcements.

- [Course Syllabus](#)

- [Homework grading rubric.](#)

1.2 Chapter I Overview

Probability: Chapter I Overview

- Probability has been around for a long time.

- Probability theory originated in the study of games of chance (i.e., dice, cards, etc.). These provide some nice introductory examples.

- More modern examples of probability in practice include:

  - Modeling mutations in genetics, playing a central role in bioinformatics.
  - Designing and analyzing computer operating systems.
  - Modeling atmospheric turbulence.
  - Probability theory is a cornerstone of the theory of finance, machine learning, and artificial intelligence.
  - Much more...

- This semester will focus on the theory of probability as a mathematical model for chance phenomena. This will be essential for building statistical theory in 4451.

2 Sample Spaces

2.1 Experiments

Sample Spaces

- Probability theory is concerned with situations in which the outcomes occur randomly. We call these situations *experiments*.

- The set of all possible outcomes is called the *sample space*.

  Example: Flipping a coin

  Flipping a coin is an *experiment*, with possible outcomes $\{H, T\}$, which defines the *sample space* of the experiment.

- An arbitrary sample space is typically denoted Ω, and an element of Ω is denoted ω.

- In the example above, $\Omega = \{H, T\}$.

Sample Space Examples

Example: Stoplights
Driving to work, a commuter passes through a sequence of three intersections with traffic lights. At each light, they either stop (s), or continues (c). The sample space Ω is the set of all possible outcomes:

$$\Omega = \{ccc, ccs, css, csc, sss, ssc, scc, scs\}$$

where csc denotes the outcome that the commuter continues at the first light, stops at the second, and continues through the third.

Example: Printing
The number of jobs in a print queue of a printing machine may be modeled as random. Here the sample space is all non-negative integers:

$$\Omega = \mathbb{N} = \{0, 1, 2, \ldots\}$$

Example: Earthquakes
We may want to model the *time* between successive earthquakes in a particular region. In this case, the experiment is the length of time between earthquakes, and our sample space is (uncountably) infinite:

$$\Omega = \{t \in \mathbb{R} | t \geq 0\}.$$

2.2   Events

Events

- As we saw, sample spaces can be comprised of many different possible outcomes.

- In probability, we are often interested in *subsets* of specific outcomes that we call *events*. For example, let A be the event that the commuter stops at the first of three lights:

$$A = \{sss, ssc, scc, scs\}.$$

- Note that *events* are sets of outcomes; any algebra you know about sets can be applied to events. That is, suppose that $B \subset \Omega$, where Ω is the sample space in the three stoplight example.

- We can then consider some event C that is the *union* or *intersection* of events A and B. E.g., $C = B \cup A$.

Common set operations
For the below definitions, we assume that Ω is the sample space, and $A, B, C \subset \Omega$ are events.

Intersection
The event (set) $A \cap B$ is the set of all outcomes ω that are in both events A and B. That is, $\omega \in A \cap B$ if and only if $\omega \in A$ and $\omega \in B$.

Union
The event (set) $A \cup B$ is the set of all outcomes ω that are in A or B; That is, $\omega \in A \cup B$ if and only if $\omega \in A$ or $\omega \in B$. Note that this definition does NOT use exclusive-or. In other words, $\omega \in A$ and $\omega \in B$ still implies that $\omega \in A \cup B$.

Compliment
The event (set) A^c (sometimes written A') is the set of all outcomes ω that are NOT in A; That is, $\omega \in A^c$ if and only if $\omega \in \Omega$, and $\omega \notin A$.

Empty Set
The *empty set* ($\emptyset$) is the set with no elements, or the event with no outcomes. If $A \cap B = \emptyset$, then A and B are *disjoint* (sometimes we would also say that A and B are mutually exlusive).

Properties

- Commutative:

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$

- Associative:

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

- Distributive:

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

3 Probability Measures

Probability Measures

- *Measure theory* is a branch of mathematics that allows us to rigorously talk about the "size" or "length" of sets in interesting ways.

- Measure theory is the basis of probability (and thereofore statistics), but a complete treatment of measure theory is outside the scope of this course.

- Instead, we will focus on a specific type of measure, called a *probability measure.*

- Roughly speaking, we can think of a probability measure P on a sample space Ω as a function that assigns real-valued weights (or probabilities) to subsets of Ω.

Definition
A probability measure P on a sample space Ω is a function that assigns real values to subsets of Ω, and must satisfy the following conditions:

Definition: Probability Measure

1. $P(\Omega) = 1$.

2. If $A \subset \Omega$, then $P(A) \geq 0$.

3. If A_1 and A_2 are disjoint, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

   More generally, if $A_1, A_2, \ldots$ are a set of mutually disjoint sets, then

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

Properties

Property A (Theorem 1.1)
$P(A^c) = 1 - P(A)$.

Proof.

Note that $\Omega = A \cup A^c$, and because $A \cap A^c = \emptyset$. Thus, by the first axiom and third axioms:

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

and therefore $P(A^c) = 1 - P(A)$.

Properties II

Property B (Corrolary 1.1)
$P(\emptyset) = 0$.

Proof.

Follows directly from Property A and the first axiom, as $\Omega^c = \emptyset$, and therefore $P(\emptyset) = 1 - P(\Omega) = 0$.

Properties II

Property C (Theorem 1.2)
If $A \subset B$, the $P(A) \leq P(B)$.

Proof.

This property states that if B occurs whenever A occurs, the $P(A) \leq P(B)$. For instance, if whenever it rains (A) it is also cloudy (B), then the probability that it rains is less than or equal to the probability that it is cloudy.
Break B into two parts: the parts of B that are in A, and the parts of B that are in A^c:

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

Because $A \subset B$, $B \cap A = A$, thus

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

Now A and A^c are disjoint, meaning they share no outcomes. Thus, because $B \cap A^c \subset A^c$, A also shares no outcomes with $B \cap A^c$, and A and $B \cap A^c$ are disjoint. Thus by the third axiom:

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

Proprties III

Property D (Theorem 1.3)
"Addition Law": $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof as an exercise. Often, it can be helpful to draw Venn diagrams. See section 1.3 of our textbook for help.

Assigning Probabilities

- Recall that the sample space Ω can consist of many types of outcomes. A probability measure can assign probabilities to subsets of Ω in a variety of ways, as long as it satisfies the axioms provided above.

- The most simple examples are finite sample spaces.

Example: Finite Spaces
Suppose that $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. A probability measure for this sample space defines a mapping such that for all $\omega_i \in \Omega$, $P(\{\omega_i\}) = p_i$, with $\sum_i p_i = 1$.
That is, each element $\omega_i \in \Omega$ gets assigned a probability value p_i.

Tossing a coin

- Consider the simple experiment of tossing a coin once. The sample space is $\Omega = \{h, t\}$. There are many possible probability measures that can be assigned to this space:

Example: Fair coin
One probability measure P assumes that the coin is fair, or that each outcome is equally as likely. Thus, $P(\{h\}) = \frac{\text{Total Number of Outcomes of} h}{\text{Total Number of Outcomes}} = 1/2$.

Example: Unfair coin
Alternatively, we could define a probability measure P' that represents an unfair coin. In this case, we could assign $P'(\{h\}) = p_h$, and $P'(\{t\}) = 1 - p_h$. p_h can be any value such that $0 \leq p_h \leq 1$, and this proposed measure P' will satisfy the axioms.

Computing probabilities

- Formally, probability measures are defined on subsets (events) of Ω, not elements of Ω, which is why we write $P(\{h\})$ rather than $P(h)$, though often we use the later notation for convenience.

- For finite spaces, we can think of each outcome (ω) as it's own event ($\{\omega\}$). Then, for more interesting events, the probability is computed by summing the disjoint events that make up the more interesting subset.

Example: Flipping two fair coins
Consider flipping a fair coin twice. The sample space is:

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

Each event is equally likely, that is, for all $\omega \in \Omega$, $P(\{\omega\}) = 1/4$. Now consider the event A, that there is at least one tails in the two coin tosses. Then,

$$P(A) = P(\{ht, th, tt\}) = P(\{ht\}) + P(\{th\}) + P(\{tt\}) = \frac{3}{4}.$$

4 Computing Probabilities

Counting Methods

- In the above example, all outcomes were equally likely.

- This type of situation is very common, which leads us to the first general method for computing probabilities in situations where it is not so easy to write down all possibilities.

Finite, equal probabilities
Suppose Ω has N elements, and the probability measure P assigns equal weight to all outcomes. Then, for any event $A \subset \Omega$ that can occur in n possible ways, then:

$$P(A) = \frac{\text{Number of ways } A \text{ can occur}}{\text{Total number of outcomes}} = \frac{n}{N}.$$

- Because of this, we now will focus on some counting strategies.

Multiplication Principle

Multiplication Principle
Consider an experiment that consists of two smaller experiments with sample spaces Ω_1 and Ω_2, such that $|\Omega_1| = n$, and $|\Omega_2| = m$. Then the total number of outcomes is $n \times m$.

Proof. Write the outcomes of the first experiment as $(a_1, \ldots, a_n)$, and the outcomes of the second experiment as $(b_1, \ldots, b_m)$. The outcome of the complete experiment can be expressed as pairs (a_i, b_j). We can then write the complete set of experiments as an $n \times m$ matrix, with entries the unique combinations of (a_i, b_j). This matrix has $n \times m$ elements. $\square$

Example: Student Government
In a class, a teacher would like to randomly select 1 boy and 1 girl to serve as representatives to the student government. If there are 12 boys and 18 girls, then the total number of ways she can pick students (outcomes) is $12 \times 18 = 216$.

- The multiplication principle also extends to the case where there are many experiments.

Example: Binary Numbers
An 8-bit binary number contains a sequence of 8 digits, each being 0 or 1. How many different 8-bit words are there? For each bit, there are two choices. Thus, using the multiplication principal, there are:

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8 = 256$$

digits.

Permutations

- A *permutation* is an ordered arrangement of objects.

- For instance, consider a set $A = \{a_1, a_2, \ldots, a_n\}$, and suppose that we want to choose r elements from this set and list them in order.

- How many ways can we do this?

- *Answer*: Depends on if we sample *with* or *without* replacement.

With Replacement
Suppose there are n labeled marbles of in a bag, and I want to perform the following experiment: draw out a marble, record it's label, put it back in the bag, and repeat this experiment r times.
By the multiplication principle, we will treat these as r experiments. Each experiment has n possible outcomes, so there are $n \times n \times \ldots \times n = n^r$ possible outcomes.

Without Replacement
Suppose there are n labeled marbles of in a bag, and I want to perform the following experiment: draw out a marble, record it's label, and then repeat this experiment r times without putting marbles back in the bag.
We can still use the multiplication principle, but now the experiments change. The first time, there are n possible outcomes. After taking a marble out, there are only $n - 1$ outcomes for the second experiment, and so on. Thus, the total number of outcomes is:

$$n \times (n-1) \times (n-2) \ldots \times (n-r+1).$$

- The previous examples can be used as a basic "proof sketch" to the following proposition:

Proposition 1.1: Ordering objects
For a set of size n, and a sample size of r, there are n^r different ordered samples with replacement and

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

without replacement.

Corollary 1.1.1
The number of orderings of n elements is $n(n-1)(n-2)\ldots 1 = n!$

Example: License plates
In some states, license plates have six characters: three letters followed by three numbers. How many unique plates are possible?
This is an example of sampling with replacement (as each license plate can have duplicated numbers or letters). Thus, there are 26^3 different ways to choose the letters, and 10^3 different ways to choose the numbers. Using the multiplication principle, there are $26^3 \times 10^3 = 17,576,000$ possible unique license plate numbers (using this rule).

Exercise: Birthday probabilities
Suppose that a room contains n people. Assuming birthdays are uniformly distributed for 365 days, what is the probability that *at least two* of them have a common birthday?

Solution:

Let A be the event that at least two people have a common birthday. Calculating $P(A)$ directly is surprisingly difficult, because A can happen in many different ways; however, A^c (the event that nobody has a common birthday) is a much simpler event.
To see this in practice, let's think about the people walking into the room one at a time (it doesn't really matter how the people get into the room of course, but it may help with the intuition). For event A^c to occur, the first person can have a birthday on any of the 365 days, and therefore there are 365 ways for this to happen. Once the second person enters the room, for event A^c to occur, the second person's birthday must not overlap with the first person's, and therefore there are only 364 ways for this to occur. We continue this way to see that for n people in a room, there are

$$365 \times 364 \times ... \times (365 - n + 1)$$

different ways for A^c to occur. Because there are 365^n possible outcomes, we can now calculate:

$$P(A^c) = \frac{365 \times 364 \times ... \times (365 - n + 1)}{365^n},$$

and therefore

$$P(A) = 1 - P(A^c) = 1 - \frac{365 \times 364 \times ... \times (365 - n + 1)}{365^n}.$$

- Table 1 gives some probabilities for a few values of n.

- These days, a computer can very quickly calculate these probabilities. You just need to be careful not to run into numeric overflow when calculating $n!$.

n	$P(A)$
4	.016
16	.284
23	.507
32	.753
40	.891
56	.988

Table 1: Some birthday probabilities for various numbers of n.

- Specifically, don't try $n!/(n-r)!$, because the numerator and denominator will be so large for even small values of n that the computer will just say $n! = \text{inf}$. Instead, reduce the fraction, then compute.

- For this class, we have 11 people (including myself). So the probability that at least one of us shares a birthday can be approximated in R by:

$$\texttt{1 - (prod(365:355)) / (365\^{}11)} \approx 0.141$$

- I said "approximate"... Why?

Combinations

- So far, we have considered the case where we care about the order. What if the order doesn't matter? Instead, when just want to know what makes up a sample and don't care about the order in which they are obtained.

- *Question*: If r objects are taken from a set of n objects without replacement (and disregarding order), how many different samples are possible?

- Probably a few ways to think about this, but let's use the theorems / corollaries / propositions that we have already derived.

- First, consider how many unique ordered samples there are (without replacement). Then, how many times are we counting the same sample? That is, how many unique ways can we arrange the sample?

- First, there are $nPr(n,r)$ ordered samples, which is equal to the number of unique samples (the thing we want, let's call it $\binom{n}{r}$), times the number of ways to order each unique sample.

- From Corollary 1.1.1, the latter value is $r!$.

- Thus: $\binom{n}{r} = nPr(n,r)/r! = \frac{n!}{(n-r)!r!}$.

Proposition 1.2: Binomial Coefficient
The number of unordered samples of r objects selected from n objects without replacement is:

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

We call this "*n choose r*", or sometimes the binomial coefficients.

- The term *binomial coefficient* comes from the Binomial Thereom, which states:

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

- In particular, this implies that $2^n = \sum_{k=0}^{n} \binom{n}{k}$, which can be interpreted as the number of subsets of a set of n objects; that is, it is the sum of the number of subsets of size 0 (which, is taken to be 1), the number of subsets of size 1, the number of subsets of size 2, etc.

- The set of all possible subsets is known as the *power set*, and for finite sets of size n, this set has size 2^n based on this calculation.

Example: Lottery

- Suppose to win a jackpot for a given lottery, you must correctly choose 6 numbers from 1 to 53, where the order doesn't matter.

- How many possible combinations of numbers are there, and if you play once, what is the probability that you win (these numbers come from California lottery in the 90's)

- *Answer*: There are $\binom{53}{6} = 22,957,480$ possible combinations. If you play once, your probability of winning is $1/22,957,480 \approx 0.00000004$.

Sketch.

Example: Quality Control

- Suppose you are tasked with quality control of a manufacturing process, and that there are n total items, and k defective items.

- If you randomly sample r items, what is the probability that you find exactly m defective items in your sample (this question is relevant because it can be used to design effective sample schemes).

Solution:

Let A be the event that we sample exactly m defective items in our sample of r items. The total number of samples of size r is $\binom{n}{r}$, which will be our denominator. The total number of ways that A can occur can be found using the multiplication principle. There are $\binom{k}{m}$ ways to choose m defective samples, and $\binom{n-k}{r-m}$ ways to choose the remaining (non defective) samples. That is, there are a total of $\binom{n}{r} \times \binom{n-k}{r-m}$ ways for A to occur. Thus,

$$P(A) = \frac{\binom{k}{m}\binom{n-k}{r-m}}{\binom{n}{r}}.$$

Extending the binomial coefficient

Proposition 1.3: Multinomials
The number of ways that n objects can be grouped into r classes of size n_i, $i = 1, \ldots, r$ is

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

where $n = \sum_{i=1}^{r} n_i$.

- *Proof Sketch:* There are $\binom{n}{n_1}$ ways to choose the objects for the first class.

- Having done that, there are $\binom{n-n_1}{n_2}$ ways of choosing the objects for the second class, etc. Thus:

$$\binom{n}{n_1 \, n_2 \, \cdots \, n_r} = \frac{n!}{n_1!(n-n_1)!} \frac{(n-n_1)!}{(n-n_1-n_2)!n_2!} \frac{(n-n_1-n_2)!}{(n-n_1-n_2-n_3)n_3!} \cdots$$

- Canceling out factors, we get the desired result.

Multinomial Theorem

- Like the binomial theorem, the numbers $\binom{n}{n_1 \, n_2 \, \cdots \, n_r}$ are called *multinomial coefficients*.

- They appear in the expansion:

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

- where the sum is over all nonnegative integers $n_1, n_2, \ldots, n_r$ that satisfy $n_1 + n_2 + \ldots + n_r = n$.

Multinomial Examples

Example: Subcommittees
How many ways can a committee of seven members be divided into three subcommittees of sizes three, two, and two, respectively?

$$\binom{7}{3 \, 2 \, 2} = \frac{7!}{3!2!2!} = 210.$$

Example: Genomics
In how many ways can the set of nucleotides $\{A, A, G, G, G, G, C, C, C\}$ be arranged in a sequence of nine letters?
To answer, let's re-frame the question. How many ways can nine positions be divided into subgroups of sizes two, four, and three (i.e., the locations of the letters A, G, and C?)

$$\binom{9}{2 \, 4 \, 3} = \frac{9!}{2!4!3!} = 1260$$

5 Conditional Probability

Introduction

- *Conditional probability* is an extremely important concept in Statistics, many scientists spend most of their time with data dealing with conditional probabilities, or conditional expectations.

- An example of this is *regression*. The idea is to model one (or many) variables conditioned on one (or many) other variables. Ideas related to conditional probability help us learn about the connections between variables, or make inference / predictions on hard-to-measure variables using easy-to-measure variables.

Conditional probabilities Motivation

- Following the textbook (Rice, 2007), we will introduce conditional probabilities using an example.

- *Digitalis therapy* (Rahimtoola, 2004) is sometimes used for patients who have suffered congestive heart failure, but there is a risk of intoxication, a serious side effect that is difficult to diagnose.

- To improve the chances of a correct diagnosis, the concentration of digitalis in the blood can be measured. Beller *et al.* (1971) conducted a study of the relation of the concentration of digitalis in the blood to digitalis intoxication in 135 patients.

- Their results are simplified slightly in the following table.

	$D+$	$D-$	Total
$T+$	25	14	39
$T-$	18	78	96
Total	43	92	135

Table 2: Taken from (Rice, 2007, Chapter 1).

	$D+$	$D-$	Total
$T+$	.185	.104	.289
$T-$	.133	.578	.711
Total	.318	.682	1.00

Table 3: Taken from (Rice, 2007, Chapter 1).

$$T+ = \text{high blood concentration (positive test)}$$
$$T- = \text{low blood concentration (negative test)}$$
$$D+ = \text{toxicity (disease present)}$$
$$D- = \text{no toxicity (disease absent)}$$

- For now, we will assume that the relative frequencies in the study roughly hold in some larger population of patients[1]. Converting the frequencies in the preceding table into proportions gives Table 3

- Suppose a patient gets a test done, and there was a positive result ($T+$, meaning there was a high blood concentration of digitalis).

- What is the probability of disease given these test results?

- In the table, we can restrict ourselves to the first row, as we know there was a positive test ($T+$).

- Of the 39 patients that tested positive, 25 suffered the disease (64.1%).

- We call he probability that a patient shows toxicity given that the test is positive the *conditional probability* of $D+$ given $T+$, denoted $P(D+|T+)$

$$P(D+|T+) = \frac{25}{39} = 0.640.$$

- Equivalently, the same conditional probability can be calculated as:

$$P(D+|T+) = \frac{P(D+\cap T+)}{P(T+)} = \frac{.185}{.289} = 0.640.$$

- Note that the *unconditional* probability of the event $D+$ is 0.318, whereas the conditional probability of $D+$ given $T+$ is 0.640. Therefore, knowing that the test is positive suggests that toxicity is twice as likely.

- This example leads to our definition of conditional probability:

---

[1]Making inference about frequencies for a larger population is a *statistics* problem (4451), not probability (4450).

Definition: Conditional Probability

Let A and B be two events with $P(B) \neq 0$. The conditional probability of A given B is defined to be:

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

- Some intuition behind this definition is the following: If we *know* that event B occurred, then the relevant sample space for other events becomes B, not Ω.

- Thus, the conditional probability is a probability measure on B. It can be shown that this new probability measure on B rather than Ω satisfies all of the axioms to be considered a probability measure.

Multiplication Law

- In some situations, $P(A|B)$ and $P(B)$ can be found rather easily, and we can use the definition of conditional probability to obtain the joint probability

Multiplication Law

Let A and B be events, and assume $P(B) \neq 0$. Then

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

Example: colored balls

An urn contains three red balls, one blue ball. Two balls are selected without replacement. What is the probability that they are both red?

Let R_1 and R_2 denote the events that a red ball is drawn on the first and second trials, respectively. Then

$$P(R_1 \cap R_2) = P(R_1)P(R_2|R_1).$$

$P(R_1) = 3/4$, and the conditional probability is calculated assuming one red ball has been removed, hence $P(R_2|R_1) = 2/3$. Therefore

$$P(R_1 \cap R_2) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

Partitions

- We can extend the idea of conditional probability and the multiplication law to more than one set, but to do so we will first introduce a new definition.

Definition: Partition

Let Ω be our sample space. Consider non-empty sets $B_1, B_2, \ldots, B_n$ such that for all $i \neq j$, $B_i \cap B_j = \emptyset$, and $\cup_{i=1}^n B_i = \Omega$. The sets $B_1, B_2, \ldots, B_n$ are said to form a partition of Ω, and the set $P = \{B_1, B_2, \ldots, B_n\}$ is a partition of Ω.

- In other words, a partition is a set of sets; each element of the superset is a non-empty set, and that every element of Ω is contained in one and only one of these elemental sets.

- E.g.: $\Omega = \{1, 2, 3, 4, 5\}$, $P = \big\{\{1, 4\}, \{3\}, \{2, 5\}\big\}$.

Total Probability

Law of Total Probability

Let $B_1, B_2, \ldots, B_n$ form a partition of Ω, such that $P(B_i) > 0$ for all i. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

That is, to find the probability of an event A, we sume the conditional probabilities of A given B_i, weighted by $P(B_i)$.

Proof:

Note that because the B_i are a partition, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and therefore $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ for all $i \neq j$. Also, because $A \subset \Omega$, we have $A \cap \Omega = A$. Thus,

$$\begin{aligned}
P(A) &= P(A \cap \Omega) \\
&= P\Big(A \cap \big(\cup_{i=1}^{n} B_i\big)\Big) \\
&= P\big(\cup_{i=1}^{n} (A \cap B_i)\big) \\
&= \sum_{i=1}^{n} P(A \cap B_i) \\
&= \sum_{i=1}^{n} P(A|B_i)P(B_i).
\end{aligned}$$

Example: Colored Balls (Part II)

Returning to the colored ball example, what is the probability that a red ball is selected on the second draw?

$$\begin{aligned}
P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|R_1^c)P(R_1^c) \\
&= \frac{2}{3} \times \frac{3}{4} + 1 \times \frac{1}{4} = \frac{3}{4}
\end{aligned}$$

- It may (or may not) be surprising that $P(R_1) = P(R_2) = 3/4$.

- One way to think about this is symmetry: without any knowledge about what happens with the first draw, both the first and second draws have the same probability of being Red.

Example: Class mobility

- Glass (2013) compiled some statistics regarding movement between social classes in England and Wales in the 1950s.

- Let (U, M, L) denote the upper, middle, and lower classes, resp., and use subscripts $1, 2$ to distinguish between father and son generations (i.e., M_1 is the event that the father belongs to the middle class, L_2 is the event the son is in the lower class).

- Then, *transition* probabilities compiled from this study can be summarized as

- Table 4 is called a *matrix of transition probaiblities*, or a *transition matrix*. It provides conditional probabilities for generational social class changes.

Table 4: Transition probabilities for class mobility.

	U_2	M_2	L_2
U_1	0.45	0.48	0.07
M_1	0.05	0.70	0.25
L_1	0.01	0.50	0.49

- For example, the probability that the son is upper class, given the father middle class, is

$$P(U_2|M_1) = 0.05.$$

- *Question*: Suppose that in the Father's generation, 10% are upper class, 40% are middle class, and 50% are lower class. What is the probability that a son is in U?

Solution:

$$P(U_2) = P(U_2|U_1)P(U_1) + P(U_2|M_1)P(M_1) + P(U_2|L_1)P(L_1)$$
$$= 0.45 \times 0.10 + 0.05 \times 0.40 + 0.01 \times 0.50 \qquad = 0.07.$$

5.1   Bayes Rule

Bayes Rule: Introduction

- At times, we may want to "reverse" the order of a conditional probability.

- For example, suppose we know a son belongs to social class U_2, and we want to find the probability that his father was in social class U_1.

- In this case, we are doing an *inverse* problem. That is, we are given an "effect", and we want to find the probability of a particular "cause".

- For example, calculate the probability $P(U_1|U_2)$.

- We can use the basic rules from conditional probability to do this:

$$P(U_1|U_2) = \frac{P(U_1 \cap U_2)}{P(U_2)}$$
$$= \frac{P(U_2|U_1)P(U_1)}{P(U_2|U_1)P(U_1) + P(U_2|M_1)P(M_1) + P(U_2|L_1)P(L_1)}$$
$$= \frac{0.045}{0.07} = 0.64.$$

Bayes Rule

Bayes' Rule
Let $A \subset \Omega$ be an event, and $B_1, B_2, \ldots, B_n$ form a partition of Ω such that $P(B_i) > 0$ for all i. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

- The proof of Bayes' rule follows directly from our previous example.

- That is, it is a direct result of the definition of conditional probabilities (gives the quotient), the multiplication law for conditional probabilities (numerator), and the law of total probability (denominator).

Failing Intuition

- For many people, Bayes theorem can give results that may seem counter-intuitive; this is true not just among opinions of statisticians (Bayesian vs Frequentist), but also for highly trained professionals.

- Consider the study by Eddy (1982), which showed that the vast majority of physicians do not assimilate evidence using Bayes' rule.

- One hundred physicians were presented with the following information:

- Without any special information, the probability that a woman in the study has breast cancer is 1%.

- If the patient has breast cancer, the probability that the radiologist will correctly diagnose it is 80%.

- If the patient has a benign lesion (no cancer), the probability that the radiologist will incorrectly diagnose it as cancer is 10%.

The question was asked: "What is the probability that a patient with a positive mammogram actually has breast cancer?"
95 out of 100 physicians estimated the probability to be about 75%.
Actual Probability:

Let A be the event that a patient has cancer, and B be the event of a positive test. From the information given, we have $P(A) = 0.01$, $P(B|A) = 0.8$, and $P(B|A^c) = 0.1$. We are wanting to know $P(A|B)$, which we can calculate using Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{(0.01)(0.8)}{(0.01)(0.8) + (0.99)(0.1)} \approx 0.075$$

- It may be surprising initially that the probability that a patient has breast cancer given a positive test is so low (even most physicians got this wrong).

- It is a result of the fact that breast cancer was so rare among study participants, that a positive result is more likely to be an incorrect diagnosis rather than an actual case of breast cancer.

Example: Polygraph tests

- Polygraphs (lie-detectors) are routinely administered in cop shows and in some real-world scenarios. The idea is that the test should help tell if a suspect (or prospective employee) is lying.

- Let $+$ denote the event that the polygraph reading is positive ($-$ for negative test), which suggests that the subject is lying (or telling the truth if $-$). Furthermore, let T denote the event that the subject is telling the truth, and L the event the subject is lying[2].

- According to studies of polygraph reliability (Gastwirth, 1987), $P(+|L) = .88$, and $P(-|T) = .86$. From these probabilities, we can calculate $P(-|L) = .12$, and $P(+|T) = .14$

- Now suppose that for a particular question, the vast majority of subjects have no reason to lie; for instance, we'll assume that $P(T) = 0.99$, whereas $P(L) = 0.01$.

---

[2]Note that the event $- = +^c$, and $L = T^c$, but we introduce new variables for interpretability

Question: Suppose that a subject produces a positive response (indicating a lie). What is the probability that the polygraph is incorrect, and that the subject is telling the truth?
Answer:

We want $P(T|+)$. Using Bayes rule,

$$P(T|+) = \frac{P(+|T)P(T)}{P(+|T)P(T) + P(+|L)P(L)} = \frac{(.14)(.99)}{(.14)(.99) + (.88)(.01)} = .94$$

That is, in screening a population of largely innocent people (say 99% innocent of the crime), 94% of the positive polygraph readings will be in error! Most people who are placed under suspicion because of a polygraph result will, in fact, be innocent.

6   Independence

Independence: Introduction

- Our final topic in this chapter is *independence.*

- Independence is useful because it usually makes calculating probabilities easier.

Independence
Let A, B be two events in the sample space Ω. We say that A and B are *independent* if

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

- Intuitively, the term *independent* suggests that knowing that one event occurs gives us no information about whether or not the other event occurs. That is, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

- This idea is compatible with our definition of independence.

- Using the definition of conditional probabilities, we have for instance

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

Example: Checking Independence
A card is selected randomly from a deck. Let A denote the event that it is an ace, and D the event that it is a diamond. Intuitively, knowing that the card is an Ace gives us no information about the suit, and we suspect the events are independent.
Knowing that only one card is the ace of diamonds, $P(A \cap D) = 1/52$. However, we note that $P(A) = 4/52$, and $P(D) = 13/52$, and since $P(A)P(D) = \frac{4}{52} \times \frac{13}{52} = 1/52$, we see that $P(A \cap D) = P(A)P(D)$, and A and D are independent.

Example: System Reliability
A system is designed so that it fails only if a unit and a backup unit both fail. Say the probability of failure of both units is $p = 0.1$. If the status of backup is designed to be independent of the primary, the probability that the system fails is $p^2 = 0.01$.

Mutual vs Pairwise Independence

- Now we extend the idea of independence to more than 2 events. Let $A_1, A_2, \ldots, A_n$ be n events. We introduce the following definitions:

Pairwise independence

The events $A_1, \ldots, A_n$ are said to be *pairwise independent* if for all $i \neq j$, A_i is independent of A_j. That is,
$$P(A_i \cap A_j) = P(A_i)P(A_j).$$

Mutual Independence

The events $A_1, \ldots, A_n$ are said to be *mutually independent* if for any sub-collection $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$, where $m \geq 2$,
$$P(A_{i_1} \cap \cdots \cap A_{i_m}) = P(A_{i_1}) \cdots P(A_{i_m}).$$

- Note that these two definitions are *not* the same. Mutual independence implies pairwise independence, but pairwise independence does not necessarily imply mutual independence.

Example: Tossing Coins

A fair coin is tossed twice. Let A denote the event of heads on the first toss, B the event of heads on the second toss, and C the event that exactly one head is thrown.

A and B are clearly independent. Note that $P(C) = 0.5$ (as there are 4 total outcomes of the two coin flips, and C is the event $\{ht, th\}$).

Now note that A and C are independent, as $P(C|A) = P(C) = 0.5$, and similarly for the events B and C (by symmetry), However,

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

and therefore the events A, B, and C are not mutually independent.

Example: AIDS transmission

Exercise: AIDS transmission

AIDS is transmitted via sexual contacts. Following several studies on AIDS, an article in the *Los Angeles Times* (August 24, 1987) suggests that the average risk of contracting AIDS per sexual contact with someone infected with AIDS is about 1/500.

Assuming that virus transmissions are independent events, calculate the probability of contracting AIDS if an individual has sexual contacts with n infected individuals.

Solution:

Because there are many different ways for the outcome to occur, it is easier to instead find the probability of the complement of this event (similar to the birthday problem).

Let C_i denote the event that the virus transmission does not occur during the ith sexual contact with an infected individual. Then, the probability of no infections in n such contacts is:

$$P(C_1 \cap C_2 \cap \ldots \cap C_n) = \left(1 - \frac{1}{500}\right)^n.$$

- It is a very common mistake to confuse "Mutually Independent" with "Mutually Disjoint" (or "Mutually Exclusive"), which is a topic we covered previously.

- Recall mutually disjoint events $A_1, \ldots, A_n$ are such that $A_i \cap A_j = \emptyset$. If you made this mistake with the previous example, you may try saying that the probability of AIDS infection are mutually disjoint (not the case), and that the probability would be calculated by letting A_i be the event of transmission, and then the probability of getting infected would be

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i).$$

- This is incorrect, as the events A_i are NOT mutually exclusive. See Textbook Example B in Chapter 1.3 for more details.

To summarize the above caution:

- *IF* the events are mutually disjoint (also called mutually exclusive), then calculating *UNIONS* (often "or" statements) is easy, because we can sum the probabilities.

- *IF* the events are mutually independent (not pairwise), then calculating *INTERSECTIONS* (often "and" statements) is easy, because we can take the product.

The above summaries are meant to be a guide on how to calculate difficult probabilities, not a rule that must be followed.

Acknowledgments

References

Beller GA, Smith TW, Abelmann WH, Haber E, Hood Jr WB (1971). "Digitalis intoxication: a prospective clinical study with serum level correlations." *New England Journal of Medicine*, **284**(18), 989–997. 28

Eddy D (1982). "Probabilistic reasoning in clinical medicine: Problems and opportunities." *Judgment under uncertainty: Heuristics and biases*, pp. 249–267. 35

Gastwirth JL (1987). "The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data." *Statistical Science*, **2**(3), 213–222. 36

Glass DV (2013). *Social mobility in Britain*. Routledge. 32

Rahimtoola SH (2004). "Digitalis Therapy for Patients in Clinical Heart Failure." *Circulation*, **109**(24), 2942–2946. 28

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 28, 2, 3