Analytical Epidemiology

Statistical and Causal Inference for Public Health

Eben Kenah

January 4, 2025

Table of contents

Preface **4**
Who this book is for . . . 4
How to use this book . . . 5
Acknowledgements . . . 5

I A. One-Sample Inference for Risks and Rates **7**

1 Probability, Random Variables, and Disease Occurrence **8**
1.1 Sets, experiments, and events . . . 8
1.1.1 Experiments and events . . . 9
1.1.2 Set operations and logic . . . 10
1.1.3 Venn diagrams . . . 11
1.1.4 Sequences of events* . . . 11
1.1.5 Algebra of sets* . . . 12
1.2 Probability . . . 13
1.2.1 Probability calculations . . . 14
1.3 Random variables . . . 15
1.3.1 Indicator variables . . . 16
1.3.2 Probability distributions . . . 16
1.3.3 Mean and variance . . . 17
1.3.4 Bernoulli distribution . . . 18
1.4 Joint and marginal distributions . . . 19
1.4.1 Linear combinations* . . . 19
1.4.2 Mean and covariance* . . . 20
1.5 Probability and disease occurrence . . . 22
1.5.1 Prevalence . . . 22
1.5.2 Risk (cumulative incidence) . . . 23

2 Conditional Probability and Diagnostic Tests **24**

3 Maximum Likelihood Estimation **25**

4 Bayesian Estimation **26**

5 Longitudinal Data and Rates **27**

6 Survival Analysis **28**

II B. Two-Sample Inference and Study Design **29**

III C. Principles of Causal Inference **30**

IV D. Epidemologic and Statistical Methods for Causal Inference **31**

References **32**

Preface

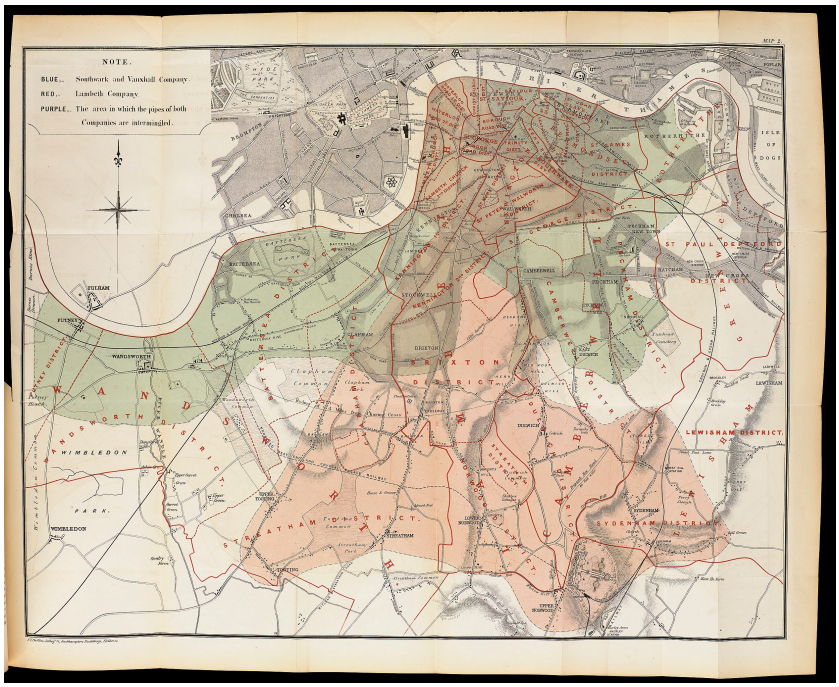

Figure 1: Areas of London supplied by the Southwark & Vauxhall (blue, now green) and Lambeth (red) water companies during the 1849-1854 cholera epidemic in London (Snow 1855). Source: Wellcome Collection via Wikimedia Commons.

One day at lunch at the Harvard School of Public Health, I overheard Professor Murray Mittleman say: "I love epidemiology. It all fits together like a diamond." As a second-year doctoral student in epidemiology, I was surprised to hear the subject described with such unstrained enthusiasm. It has taken years of study and experience for me to understand what he meant. On the way, I too have fallen in love.

Who this book is for

This book is intended primarily for two audiences:

- Epidemiologists are often protected from the mathematical foundations of their field. The long-term price of this is "dogmatism, that is, a tendency to rigidly protect a partially understood theoretical heritage" (Morabia 2004). The mathematics needed for a deeper understanding of epidemiologic methods is within reach of anyone who has come far enough to need it. Whether you master this material or just learn to approach it with more patience than fear, you will be doing a service to epidemiology and to public health.
- Biostatisticians are familiar with probability and statistical inference, but applying statistics to solve scientific problems in public health requires skills different from those needed to prove that a method works under given assumptions. Epidemiology is a living example of the interplay between theory and practice in applied statistics, and epidemiologists have shown integrity, courage, and ingenuity in confronting causal questions with statistical tools.

Beyond these audiences, I hope to explain the logic of epidemiology to any interested reader. It is one of the most important applications of statistics, and it has quite possibly already helped save your life.

How to use this book

Difficult chapters, sections, subsections, and exercises are marked with an asterisk (*). These can be skipped without harming the logical flow of the book, but none of them is beyond the reach of a determined reader. The starring is recursive: Starred sections can be skipped within a starred chapter, starred subsections can be skipped within a starred section, and so on. Footnotes offer context or hint at more advanced material. All of them can be ignored if they do not seem useful or interesting.

This is a work in progress, and you may notice that some parts are obviously unfinished. Please report errors (including typos) or submit suggestions (especially good examples) at:

https://github.com/ekenah/analyticalepi/issues.

Acknowledgements

Devesh Kapur, Paul Farmer, and James H. Maguire guided me to a career in public health when I was an undergraduate. James Robins, Miguel A. Hernán, Marc Lipsitch, and Stephen P. Luby helped me become an epidemiologist, biostatistician, and epidemic modeler in graduate school. My career began under the mentorship of Ira M. Longini, Jr., and M. Elizabeth Halloran as a postdoctoral fellow at the Univerity of Washington and an assistant professor at the University of Florida. My colleagues Yang Yang, Grzegorz Rempała, Forrest Crawford, and Patrick Schnell have all provided useful comments. For their patience with early versions of this material, I am grateful to the students of STA 6177/PHC 6937 (Applied Survival

Analysis) at the University of Florida from 2013 to 2016 and PUBHEPI 8430 (Epidemiology 4) at The Ohio State University from 2019 to the present.

My parents, Chris and Kate Kenah, courageously allowed me to travel to places they had never been to and do things I had been told to avoid. These experiences in the United States, India, South Africa, and especially Bangladesh opened my eyes to the terrible importance of clear thinking in public health. My wife, Asma Aktar, and our sons Rafi, Rayhan, and Rabi remind me every day how important it is to destroy everything that stifles humanity. To that end, I hope this book is useful.

Any mistakes are my own, and God knows best) .(

Part I

A. One-Sample Inference for Risks and Rates

1 Probability, Random Variables, and Disease Occurrence

> One sees, from this essay, that probability theory is basically common sense reduced to calculation; it makes us appreciate with exactitude that which fair minds sense with a sort of instinct, often without being able to account for it. (Laplace 1820)[1]

To begin at the beginning, we will start with probability. Morabia (2004) accurately observed that "Epidemiology came late in human history because it had to wait for the emergence of probability." This is probably the most difficult chapter of the entire book, but it will make all subsequent chapters easier. You can use it as a reference and come back to the difficult parts when you need them. Learning to think clearly about probability will give you a compass to find your way through epidemiologic methods.

1.1 Sets, experiments, and events

To speak clearly about probabilities, we need some basic notation for sets. If A is a set that contains an object a, we write

$$a \in A \tag{1.1}$$

to indicate that a is an **element** of A. If A and B are sets such that every element of A is also an element of B, we write

$$A \subseteq B. \tag{1.2}$$

to indicate that A is a **subset** of B. Sets A and B are equal if and only if $A \subseteq B$ and $B \subseteq A$, which means they contain exactly the same elements. The *empty set* with no elements is denoted $\emptyset$. For any set A, it is true that $A \subseteq A$ and $\emptyset \subseteq A$.

[1] Pierre-Simone, marquis de Laplace (1749-1827) is often called the Newton of France. He proved that the solar system is stable, developed theories of ocean tides and gravitational potential, proved one of the first general versions of the central limit theorem, and pioneered the Bayesian interpretation of probability. For just six weeks in 1799, he was Minister of the Interior for France under Napoleon. His is one of the 72 names on the Eiffel Tower.

We use $\mathbb{R}$ to denote the real numbers. Intervals are subsets of $\mathbb{R}$ that take one of the following forms:

$$\begin{aligned}
[a,b] &= \{x \in \mathbb{R} : a \leq x \leq b\},\\
[a,b) &= \{x \in \mathbb{R} : a \leq x < b\},\\
(a,b] &= \{x \in \mathbb{R} : a < x \leq b\},\\
(a,b) &= \{x \in \mathbb{R} : a < x < b\}.
\end{aligned}$$

An endpoint with a square bracket is included in the interval and an endpoint with a round bracket is not. We can have $a = -\infty$ or $b = \infty$ as long as we use a round bracket for the corresponding endpoint. For example, it is true that $\mathbb{R} = (-\infty, \infty)$. However, $\mathbb{R} \neq [-\infty, \infty]$ because $\pm\infty$ are not real numbers.

1.1.1 Experiments and events

In probability, an **experiment** is any process that will produce one outcome out of a set possible outcomes. The set of possible outcomes is called the **sample space** and is traditionally denoted Ω. An experiment produces a single outcome $\omega \in \Omega$. For example, the sample space for a single coin flip is

$$\Omega = \{H, T\}, \tag{1.3}$$

where $\omega = H$ if we get heads and $\omega = T$ if we get tails.

The outcomes in the sample space must determine everything about the random outcome of the experiment. If we flip a coin twice, the sample space cannot be $\{H, T\}$ because each $\omega \in \Omega$ must specify the outcome of both coin flips. Instead,

$$\Omega = \{HH, HT, TH, TT\} \tag{1.4}$$

where $\omega = XY$ if we get X on the first flip and Y on the second. This helps us see, for example, that there are two ways to get one H and one T in two coin flips.

The purpose of probability is to summarize uncertainty about the outcomes of experiments. However, the outcomes themselves do not have probabilities. Probabilities are assigned to **events**, which are subsets of the sample space Ω. If A is an event, then A occurs if and only if the outcome ω produced by our experiment is an element of A (i.e., if and only if $\omega \in A$). If we flip a coin twice, the event that we get two heads is $\{HH\}$, the event that we get one head is $\{HT, TH\}$, and the event that we get zero heads is $\{TT\}$. By definition, the event Ω always occurs and the event $\emptyset$ never occurs.

In experiments with a finite or countably infinite sample space,[2] the distinction between the

[2]The natural numbers $\mathbb{N} = \{0, 1, 2, ...\}$ are *countably infinite*, as are the integers $\mathbb{Z}$ and the rational numbers $\mathbb{Q}$. The real numbers $\mathbb{R}$ are *uncountably infinite*, as are the real numbers in any interval (a, b) with $a < b$ and the irrational numbers. Uncountably infinite sets are infinitely larger than countably infinite sets. This distinction was discovered in the 1870s by the German mathematician Georg Cantor (1845–1918). It was considered shocking at the time, but it has become a cornerstone of modern mathematics.

outcome ω and the event $\{\omega\}$ can be safely ignored. In more complex experiments (e.g., taking a random sample from a standard normal distribution), this distinction is important.[3] In all cases, experiments have outcomes and events have probabilities.

In epidemiology, it is often useful to think of the sample space Ω as being a population and each $\omega \in \Omega$ as an individual in this population. In this context, our experiment is to sample a person from Ω and ask them questions, take measurements, or follow them over time to ascertain disease occurrence. Events would be subpopulations of Ω, such as $\{\omega \in \Omega : \omega \text{ lives in Ohio}\}$. This event occurs if the sampled individual ω lives in Ohio, and it does not occur if they live somewhere else.

1.1.2 Set operations and logic

There are three basic set operations that take one or more given sets and define another set: complement, intersection, and union. Each operation has a simple interpretation in terms of logic.

- The **complement** of a set A is

$$A^{\mathsf{C}} = \{\omega \in \Omega : \omega \notin A\}, \tag{1.5}$$

  which can be interpreted logically as ***not*** A. If A is an event, then the event A^{C} occurs if $\omega \notin A$. For the same reason that "not not A" means "A", we have $(A^{\mathsf{C}})^{\mathsf{C}} = A$.

- The **intersection** of two sets A and B is

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}, \tag{1.6}$$

  which can be interpreted logically as A ***and*** B. If A and B are events, then the event $A \cap B$ occurs if $\omega \in A$ and $\omega \in B$.

- The **union** of two sets A and B is

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}, \tag{1.7}$$

  which can be interpreted logically as A ***or*** B as long as we use "or" in an inclusive sense (i.e., and/or). If A and B are events, then the event $A \cup B$ occurs if $\omega \in A$ or $\omega \in B$.

[3] In experiments with uncountably infinite sample spaces, the probability of an event A cannot always be calculated by adding up the probabilities of $\{\omega\}$ for all $\omega \in A$. For example: If we choose a number at uniformly at random in $[0, 1]$, the probability of getting any particular number ω is zero. The sum of the probabilities of all $\{\omega\} \subseteq A$ is zero (if A is countable) or undefined (if A is uncountable). By maintaining a distinction between outcomes and events and by limiting probability calculations to countable (i.e., finite or countably infinite) sums, we end up with something coherent and useful.

If $A \subseteq B$, then $A \cap B = A$ and $A \cup B = B$. An important special case is that

$$A \cap A = A \cup A = A. \tag{1.8}$$

For the empty set $\emptyset$, we get $A \cap \emptyset = \emptyset$ and $A \cup \emptyset = A$. For the sample space Ω, we get $A \cap \Omega = A$ and $A \cup \Omega = \Omega$.

Union and intersection are *commutative* operations like addition and multiplication, so the order of A and B does not matter:

$$A \cup B = B \cup A$$

and

$$A \cap B = B \cap A.$$

Events A and B are **disjoint** or *mutually exclusive* when $A \cap B = \emptyset$. If A and B are disjoint, then at most of one of them can occur in a single experiment. Any set and its complement are disjoint, and the empty set $\emptyset$ is disjoint with itself and all other sets.

If Ω is a population, these set operations allow us to define subpopulations in terms of multiple traits. If the event $A = \{\omega \in \Omega : \omega \text{ lives in Ohio}\}$, then its complement A^{C} contains all individuals in Ω who live outside Ohio. If the event $B = \{\omega \in \Omega : \omega \text{ is 42 years old}\}$, then the intersection $A \cap B$ contains everyone in Ω who is 42 years old and lives in Ohio. If Ω does not contain any such individual, then A and B are disjoint. The union $A \cup B$ contains everyone in Ω who lives in Ohio or is 42 years old. This could include both a 24-year-old who lives Ohio and a 42-year-old who lives Michigan.

1.1.3 Venn diagrams

A useful tool for understanding events and set operations is the **Venn diagram**.[4] An example is shown below.

1.1.4 Sequences of events*

Intersections can be written for more than two events. The intersection of $A_1, A_2, \ldots, A_n$ is

$$I_n = \bigcap_{i=1}^{n} A_i. \tag{1.9}$$

[4]Named after John Venn (1834-1923), an English logician and philosopher who was one of the pioneers of the frequentist interpretation of probability. He was ordained as an Anglican priest in 1859 but resigned from the church in 1883. He was a prize-winning gardener of roses and white carrots and a prominent supporter of women's right to vote. From 1903 until his death, he was President of Fellows in Gonville and Caius College at the University of Cambridge, where he is commemorated with a Venn diagram in a stained glass window.

Because set intersection is commutative and associative, any ordering of $A_1, \ldots, A_n$ produces the same intersection. The event I_n occurs if and only if all of the events $A_1, \ldots, A_n$ occur. Each new event makes the intersection smaller (i.e., never larger) in the sense that

$$\bigcap_{i=1}^{n+1} A_i \subseteq I_n.$$

whenever A_{n+1} is another event.

Similarly, unions can be written for more than two events. If $A_1, A_2, \ldots, A_n$ is a set of events, then their union is

$$U_n = \bigcup_{i=1}^{n} A_i. \tag{1.10}$$

Because set union is commutative and associative, any ordering of $A_1, \ldots, A_n$ produces the same union. The event U_n occurs if and only if at least one of the events A_i occurs. Each new event makes the union bigger (i.e., never smaller) in the sense that

$$U_n \subseteq \bigcup_{i=1}^{n+1} A_i$$

whenever A_{n+1} is another event.

Both unions and intersections can be defined for infinite sequences of events.[5] To describe this, we let $n = \infty$ in the notation from Equation 1.9 or Equation 1.10. The union of any finite sequence of events can be turned into the union of an infinite sequence of events by adding an endless sequence of empty sets to the finite sequence. The new sequence is still a sequence of disjoint events, and each empty set $\emptyset$ leaves the union unchanged. If $(A_1, A_2, \ldots)$ is an infinite sequence of events such that $A_i = \emptyset$ for all $i > n$, then

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{n} A_i.$$

This turns out to be useful when we try to give a mathematically rigorous definition of probability.

1.1.5 Algebra of sets*

Unions, intersections, and complements can be combined in complex ways. Fortunately, there are a few basic principles that can be used to simplify these calculations. We have already seen

[5] In probability, we only consider unions and intersections of finite or countably infinite sets of events. Although unions and intersections can be defined for uncountably infinite sets of events, it can be impossible to assign probabilities to the resulting sets (see the Banach-Tarski paradox). As an epidemiologist, this should not keep you up at night.

that unions and intersections are commutative. Unions and intersections are also *associative*, so

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

and

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

for any sets A, B, and C.

De Morgan's laws describe how complements affect unions and intersections. If A and B are sets, then

$$(A \cap B)^{\mathsf{C}} = A^{\mathsf{C}} \cup B^{\mathsf{C}} \tag{1.11}$$

because you are outside $A \cap B$ if and only f you are outside A or outside B. Similarly,

$$(A \cup B)^{\mathsf{C}} = A^{\mathsf{C}} \cap B^{\mathsf{C}}. \tag{1.12}$$

because you are outside $A \cup B$ if and only if you are outside A and outside B. Note that each of these equations implies the other if we replace $A = (A^{\mathsf{C}})^{\mathsf{C}}$ with A^{C} and replace $B = (B^{\mathsf{C}})^{\mathsf{C}}$ with B^{C}. They are two sides of the same coin, but it is helpful to remember them both.

The *distributive properties* describe how unions and intersections interact with each other. Recall that multiplication distributes over addition, so $a(b + c) = ab + ac$. For any sets A, B, and C, we have the following distributive properties:

- Intersections distribute over unions, so

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

- Unions distribute over intersections, so

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

Intersections and unions also distribute over themselves. However, this is a consequence of commutativity, associativity, and Equation 1.8, not a separate property like the distributive rules above.

1.2 Probability

A *probability measure* is a function that takes an event $A \subseteq \Omega$ and returns a number $\Pr(A) \in [0, 1]$ in any way that conforms to the following rules:

- $\Pr(\Omega) = 1$.

- $\Pr(A) \in [0, 1]$ for any event $A \subseteq \Omega$.[6]
- The **addition rule**: If $(A_1, A_2, ...)$ is any sequence of disjoint events, then

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

  The addition rule is stated in terms of an infinite sequence of disjoint events because this implies the addition rule for any finite sequence of disjoint events (see Section 1.1.4).

It is useful to think of probability as a generalization of our intuitions about area or volume. When there is no overlap in a set of two-dimensional shapes, we can get the total area they cover by adding up the areas of the individual shapes. Similarly, we can get the total volume taken up by a set of bowling balls by adding up their individual volumes.

There is a lot of debate about the meaning of probability, but its definition does not assume any particular interpretation. Probability calculations are based on the rules above no matter what we think it all means, and any interpretation that is consistent with these rules is valid.

1.2.1 Probability calculations

Several useful properties of probability follow immediately from the definition above. A short proof follows each result. To follow the proofs, it helps to draw Venn diagrams.

Theorem 1.1. *If A is an event,* $\Pr(A^{\mathsf{C}}) = 1 - \Pr(A)$.

Proof. Because $\Omega = A \cup A^{\mathsf{C}}$ and A and A^{C} are disjoint, we have

$$\Pr(A) + \Pr(A^{\mathsf{C}}) = \Pr(\Omega) = 1$$

by the addition rule. The result follows when we subtract $\Pr(A)$ from both sides. □

Theorem 1.2. *If A and B are events such that $A \subseteq B$, then* $\Pr(A) = \Pr(B) - \Pr(B \cap A^{\mathsf{C}})$. *This implies that* $\Pr(A) \leq \Pr(B)$.

[6] Technically, we assign probabilities only to events in a set $\mathcal{F}$ of subsets of Ω that is required to contain Ω and to be closed under complements and countable unions. "Closed under complements" means that $A^{\mathsf{C}} \in \mathcal{F}$ whenever $A \in \mathcal{F}$. For example, $\emptyset = \Omega^{\mathsf{C}}$ must be in $\mathcal{F}$ because $\Omega \in \mathcal{F}$. "Closed under countable unions" means that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ whenever $(A_1, A_2, ...)$ is a sequence of events in $\mathcal{F}$. The set $\mathcal{F}$ is called a σ-algebra, and this restriction on the domain of probability helps avoid internal contradictions like the mind-blowing Banach-Tarski paradox.

Proof. Each element of B either is or is not in A, so

$$B = (B \cap A) \cup (B \cap A^{\mathsf{C}}) = A \cup (B \cap A^{\mathsf{C}}).$$

where the second equality follows from the fact that $B \cap A = A$ because $A \subseteq B$. The two sets on the right-hand side are disjoint, so we have

$$\Pr(B) = \Pr(A) + \Pr(B \cap A^{\mathsf{C}})$$

by the addition rule. The result follows if we subtract $\Pr(B \cap A^{\mathsf{C}})$ from both sides. This implies that $\Pr(A) \leq \Pr(B)$ because $\Pr(B \cap A^{\mathsf{C}}) \geq 0$. □

Theorem 1.3. *If A and B are events,* $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Proof. We can break $A \cup B$ into three disjoint sets: elements of A and not B, elements of B and not A, and elements of both A and B. In set notation, this is

$$A \cup B = (A \cap B^{\mathsf{C}}) \cup (B \cap A^{\mathsf{C}}) \cup (A \cap B).$$

By the addition rule,

$$\Pr(A \cup B) = \Pr(A \cap B^{\mathsf{C}}) + \Pr(B \cap A^{\mathsf{C}}) + \Pr(A \cap B). \tag{1.13}$$

By Theorem 1.2, we have

$$\Pr(A \cap B^{\mathsf{C}}) = \Pr(A) - \Pr(A \cap B),$$

because $A \cap B \subseteq A$ and

$$\Pr(B \cap A^{\mathsf{C}}) = \Pr(B) - \Pr(A \cap B).$$

because $A \cap B \subseteq B$. The result follows from substituting these back into Equation 1.13 and collecting terms involving $\Pr(A \cap B)$. Intuitively, $\Pr(A) + \Pr(B)$ includes the overlap $\Pr(A \cap B)$ twice, so we have to subtract out one of them. □

1.3 Random variables

The outcomes of an experiment are not necessarily numbers. A **random variable** is a real-valued function defined on a sample space Ω. In other words, a random variable X is a function that takes an *argument* $\omega \in \Omega$ as input and returns a *value* $X(\omega) \in \mathbb{R}$. Traditionally, random variables are written as capital letters and possible values are written as lower-case letters, so $\Pr(X = x)$ denotes the probability of the event

$$\{\omega \in \Omega : X(\omega) = x\}.$$

For simplicity, random variables are usually written without the argument ω.

The distinction between outcomes and random variables is useful because we can define multiple random variables on the same sample space. For example, the height, weight, and age of an individual ω sampled from a population Ω are different random variables defined on the same sample space.

1.3.1 Indicator variables

The simplest random variables are **indicator variables**. For an event A, the indicator variable

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

When sampling from a population, we could define indicator variables for membership in different subpopulations. Indicator variables are **binary** random variables, which take exactly two values. In practice, these values should be zero and one unless there is a specific reason to do otherwise.

All of the basic set operations above can be expressed in terms of indicator variables for sets.

- The indicator function for the complement of A is

$$\mathbb{1}_{A^\mathsf{C}} = 1 - \mathbb{1}_A. \tag{1.14}$$

- If B is another event and $\mathbb{1}_B$ is its indicator variable, then the indicator variable for the intersection A and B is

$$\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B, \tag{1.15}$$

- The indicator variable for the union $A \cup B$ is

$$\mathbb{1}_{A \cup B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B) = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A \cap B}. \tag{1.16}$$

  This follows from Equation 1.12 because $A \cup B = (A^\mathsf{C} \cap B^\mathsf{C})^\mathsf{C}$.

1.3.2 Probability distributions

The set of possible values of a random variable X is called the *support* of X and denoted $\text{supp}(X)$.[7] For example, the support of an indicator variable is $\{0, 1\}$. In this section, we will focus on **discrete** random variables, which have a support on a finite or countably infinite set. There are two standard ways to describe the distribution of a discrete random variable:

- The **probability mass function** (PMF) of a discrete random variable X is

$$f(x) = \begin{cases} \Pr(X = x) > 0 & \text{if } x \in \text{supp}(X), \\ 0 & \text{if } x \notin \text{supp}(X). \end{cases}$$

[7] Technically, the support of X is the smallest closed set S_X such that $\Pr(X \in S_X) = 1$. For a discrete random variable with support on a finite set, it is just the set of possible values. For a discrete random variable with support on a countably infinite set, it can include points whose probability mass is zero—a pathological case that we can safely ignore. For a continuous random variable, it can include values whose probability density is zero—a case that is not unusual or pathological.

Because $\Pr(\Omega) = 1$, we always have

$$\sum_{x \in \text{supp}(X)} f(x) = 1.$$

- The **cumulative distribution function** (CDF) of X is

  $$F(x) = \Pr(X \leq x).$$

  $F(x)$ is monotonically increasing in x, which means that $F(a) \leq F(b)$ whenever $a < b$. It has a jump upward of size $f(x)$ at each $x \in \text{supp}(X)$, and its value at each such x is the value that it jumps to—not the value that it jumps up from. For sufficiently small x, $F(x)$ can be made arbitrarily close to zero. For sufficiently large x, $F(x)$ can be made arbitrarily close to one. More formally, we say that $\lim_{x \downarrow -\infty} F(x) = 0$ and $\lim_{x \uparrow \infty} F(x) = 1$.

The PMF and CDF provide equivalent descriptions of the distribution of X in the sense that either of these functions can be used to calculate the other. Given the PMF f, the CDF is defined by

$$F(x) = \sum_{\substack{v \in \text{supp}(X): \\ v \leq x}} f(v).$$

where the sum is taken over all $u \in \text{supp}(X)$ such that $u \leq x$. Given the CDF F, the PMF is defined by

$$f(x) = F(x) - \max_{v < x} F(v)$$

where the maximum is $F(v)$ for the largest $v \in \text{supp}(X)$ such that $v < x$.

1.3.3 Mean and variance

The **mean** or *expected value* of a random variable X is

$$\mathbb{E}(X) = \sum_{x \in \text{supp}(X)} x \Pr(X = x) = \sum_{x \in \text{supp}(X)} x f(x),$$

where f is the PMF of X. The mean is often written μ, and it is often described as a measure of the "location" or "central tendency" of X.

If X has $\mathbb{E}(X) = \mu$, then $(X - \mu)^2$ is another random variable. The **variance** of X is the expected value of $(X - \mu)^2$:

$$\text{Var}(X) = \mathbb{E}\left[(X - \mu)^2\right] = \sum_{x \in \text{supp}(X)} (x - \mu)^2 f(x).$$

Because $(x-\mu)^2 \geq 0$ with equality if and only if $x = \mu$, we always have $\text{Var}(X) \geq 0$. We have $\text{Var}(X) = 0$ if and only if $X = \mu$ with probability one. The variance is often written σ^2, and it is often described as a measure of the dispersion of X around the mean.

The square root of the variance is called the **standard devation**, which is often written σ. If a random variable X has units (e.g., length, weight, or time), the mean and the standard deviation have the same units as X. For example, the mean and standard deviation of a length in meters both have units of meters, but the variance has units of meters^2.

1.3.4 Bernoulli distribution

The distribution of an indicator variable is called the **Bernoulli distribution**.[8] A random variable with the Bernoulli(p) distribution has the PMF

$$f(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1. \end{cases}$$

If a random variable X has a Bernoulli(p) distribution, we write $X \sim \text{Bernoulli}(p)$. The indicator variable for an event A has a Bernoulli distribution with $p = \Pr(A)$.

If $X \sim \text{Bernoulli}(p)$, then it has mean

$$\mathbb{E}(X) = 0 \times (1-p) + 1 \times p \tag{1.17}$$
$$= p \tag{1.18}$$

and variance

$$\text{Var}(X) = (0-p)^2(1-p) + (1-p)^2 p \tag{1.19}$$
$$= p(1-p). \tag{1.20}$$

The standard deviation is $\sqrt{p(1-p)}$. This standard deviation is greater than zero unless $p = 0$ or $p = 1$. If $p = 0$, then $X = 0$ with probability one. If $p = 1$, then $X = 1$ with probability one.

[8] Named after Jacob Bernoulli (1655-1705), a Swiss mathematician who derived the first version of the law of large numbers and discovered the constant $e \approx 2.718281828$, which is the base for natural logarithms. He and his younger brother Johann Bernoulli (1667-1748) were some of the first mathematicians to try to understand and apply calculus, but their relationship eventually curdled into a jealous rivalry. A lunar impact crater called Bernoulli is named jointly after them.

1.4 Joint and marginal distributions

If X and Y are random variables defined on the same probability space, then their **joint** probability mass function is

$$f(x, y) = \Pr(X = x \text{ and } Y = y) = \Pr\left(\{\omega : X(\omega) = x \text{ and } Y(\omega) = y\}\right).$$

The **marginal** probability mass functions are the PMFs of X or Y individually, which can be calculated from the joint PMF. The marginal PMF of X is

$$f_X(x) = \sum_{y \in \text{supp}(Y)} f(x, y),$$

and the marginal PMF of Y is

$$f_Y(y) = \sum_{x \in \text{supp}(X)} f(x, y).$$

These are called marginal distributions by analogy to the margins of a table.

Joint distributions can be defined for more than two random variables. If $X_1, X_2, \ldots, X_n$ are random variables defined on the same sample space, then their joint PMF is

$$f(x_1, x_2, \ldots, x_n) = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

The marginal distribution of each X_i can be found by adding up the PMF over the support of all the other random variables. For example,

$$f_{X_2}(x_2) = \sum_{x_1 \in \text{supp}(X_1)} \sum_{x_3 \in \text{supp}(X_3)} f(x_1, x_2, x_3).$$

when $n = 3$. For larger n, the formula gets uglier but the idea is the same.

The distinction between joint and marginal distributions is extremely important in epidemiology (and other applications of probability). The joint distribution can be used to find any of the the marginal distributions, but the marginal distributions cannot be used to find the joint distribution—unless we make additional assumptions.

1.4.1 Linear combinations*

If a and b are constants, then $aX + bY$ is another random variable on Ω. It is called a *linear combination* of X and Y. Linear combinations can be defined for more than two random variables. If $X_1, \ldots, X_n$ are random variables defined on a sample space and $a_1, \ldots, a_n$ are constants, then

$$\sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

is a linear combination of $X_1, \ldots, X_n$. The constants can be any real numbers, including one and zero.

Section 1.3.1 contains both examples and non-examples of linear combinations of random variables.

- The indicator function for $A^\complement$ in Equation 1.14 is a linear combination of $\mathbb{1}_A$ and the random variable $\mathbb{1}_\Omega$, which equals one for all $\omega \in \Omega$.
- The indicator function for $A \cup B$ in Equation 1.16 is linear combination of the indicator variables $\mathbb{1}_A$, $\mathbb{1}_B$, and $\mathbb{1}_{A \cap B}$.
- The indicator function for $A \cap B$ in Equation 1.15 is not a linear combination of $\mathbb{1}_A$ and $\mathbb{1}_B$ because we have to multiply these two variables.

1.4.2 Mean and covariance*

If X and Y are random variables defined on the same sample space and a and b are constants, the mean of the linear combination $aX + bY$ is

$$\mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y).$$

This is a direct consequence of the definition of expected value:

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_{x \in \text{supp}(X)} \sum_{y \in \text{supp}(Y)} (ax + by) f(x, y) \\
&= a \sum_{x \in \text{supp}(X)} \left(x \sum_{y \in \text{supp}(Y)} f(x, y) \right) + b \sum_{y \in \text{supp}(Y)} \left(y \sum_{x \in \text{supp}(X)} f(x, y) \right) \\
&= a \sum_{x \in \text{supp}(X)} x f_X(x) + b \sum_{y \in \text{supp}(Y)} y f_Y(y).
\end{aligned}$$

The algebra is not pretty, but the logic is straightforward. We split up the sum into parts depending only on x and only on y outside the joint PMF. In each part, we factor out a constant and find the marginal PMF. This same logic extends to a linear combination of any number of random variables.

The variance of $aX + bY$ is

$$\text{Var}(aX + bY) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(Y) + 2ab \,\text{Cov}(X, Y)$$

where

$$\text{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big]$$

is called the **covariance** of X and Y. Note that $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. Because $\text{Var}(X) = \text{Cov}(X, X)$, variance is a special case of covariance.

The joint distribution of X and Y has a **covariance matrix** which is

$$\begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{bmatrix}$$

The variances are along the diagonal of the matrix, and the covariances appear off the diagonal. Because $\text{Cov}(X,Y) = \text{Cov}(Y,X)$, covariance matrices are always symmetric (i.e., symmetric across the diagonal). Covariance matrices are an extremely useful tool for calculating the variances of linear combinations of random variables. For example:

$$\text{Var}(aX + bY) = \begin{pmatrix} a & b \end{pmatrix} \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{bmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

in matrix and vector notation from linear algebra. This logic extends to linear combinations of any number of random variables.

The covariance is the numerator of the *Pearson correlation coefficient*,[9] which is

$$\rho_{XY} = \rho_{YX} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

Because of the Cauchy-Schwarz inequality, it turns out that $\rho_{XY} \in [-1, 1]$.

- We get $\rho_{XY} = -1$ if and only if $Y = cX$ for some negative constant c.
- We get $\rho_{XY} = 1$ if and only if $Y = cX$ for some positive constant c. For example, $\rho_{XX} = 1$ for any random variable X.
- We get $\rho_{XY} = 0$ if X and Y are *independent* in the sense that the value of one tells us nothing about the value of the other.[10] However, it is possible to have $\rho_{XY} = 0$ when X and Y are not independent.

If we divide each entry $\text{Cov}(X,Y)$ in a covariance matrix by $\sqrt{\text{Var}(X)\,\text{Var}(Y)}$, when we get a *correlation matrix*. Any correlation matrix is symmetric, and the entries along its diagonals are all ones.

[9] Named after Karl Pearson (1857-1936), an English mathematician who founded the modern discipline of mathematical statistics. In 1911, he started the world's first university department of statistics at University College London. He was an outspoken socialist and supporter of women's rights, but he was also a vocal proponent of social Darwinism and eugenics who opposed Jewish immigration into Britain.

[10] We will define independence of random variables more rigorously when we discuss conditional probabilities in Chapter 2.

1.5 Probability and disease occurrence

In epidemiology, there are two fundamental measures of disease occurrence that are probabilities: **prevalence** and **risk**. In both cases, our experiment is to sample an individual ω from a population Ω. The **disease outcome** is a binary random variable

$$D(\omega) = \begin{cases} 1 & \text{if } \omega \text{ has the disease outcome,} \\ 0 & \text{otherwise.} \end{cases}$$

The set of individuals in Ω who have $D(\omega) = 1$ is an event in Ω, and the prevalence or risk is its probability

$$\Pr(\{\omega \in \Omega : D(\omega) = 1\}).$$

The most important difference between prevalence and risk is the role of time in the definition of D.

There is an important technicality to remember when we talk about disease onset and recovery. When a person has disease onset at time t^{ons} and recovers at time t^{rec}, they have disease for each $t \in [t^{\text{ons}}, t^{\text{rec}})$. We assume that $t^{\text{rec}} > t^{\text{ons}}$ so this interval is nonempty. We let the onset and recovery times for person i be t_i^{ons} and t_i^{rec}, respectively. If a person has multiple episodes of the disease, each episode has its own t^{ons} and t^{rec}. For example, the j^{th} episode in person i would have onset time t_{ij}^{ons} and recovery time t_{ij}^{rec}.

1.5.1 Prevalence

For prevalence, the disease outcome is defined by choosing a time t and letting

$$D(\omega) = \begin{cases} 1 & \text{if } \omega \text{ has disease at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, it is the proportion of the population Ω that disease at time t. This includes individuals who have disease onset at time $t^{\text{ons}} = t$ but not individuals who recover from disease at time $t^{\text{rec}} = t$. This is often called the **point prevalence** at time t.

Another version of prevalence is **period prevalence**. For period prevalence, we choose a nonempty time interval $(t_a, t_b]$ and define

$$D(\omega) = \begin{cases} 1 & \text{if } \omega \text{ has disease at any time } t \in (t_a, t_b], \\ 0 & \text{otherwise.} \end{cases}$$

In other words, it is the proportion of the population that has disease at any time in the interval $(t_a, t_b]$. This can include people who had disease onset at a time $t^{\text{ons}} \leq t_a$ as long as $t^{\text{rec}} > t_a$. In particular, it includes all individuals who have disease onset at time t_a, who must have $t^{\text{rec}} > t_a$.

1.5.2 Risk (cumulative incidence)

To define **risk** or **cumulative incidence**, we first choose an nonempty time interval $(t_a, t_b]$ and let the population Ω consist only of individuals who do not have disease at time t_a and are at risk of disease onset after that. The disease outcome is defined as

$$D(\omega) = \begin{cases} 1 & \text{if } \omega \text{ has } t^{\text{ons}} \in (t_a, t_b], \\ 0 & \text{otherwise.} \end{cases}$$

In the population that is disease-free and at risk of disease at time t_a, it is the proportion who have disease onset at $t^{\text{ons}} \leq t_b$.

Point and interval prevalence are affected by both the frequency of disease onset and the duration of disease, the risk in any time interval depends only on the frequency of disease onset (see Figure~**??**). This makes risk a more useful measure for identifying the causes of disease, which is the primary goal of *analytic epidemiology*.

Another advantage of risk is that it can be used for outcomes that begin and end very quickly (e.g., traffic accidents or being hit by lightning) or outcomes that remove individuals from the population (e.g., emigration or death). Point prevalence would not be a useful measure of the impact of these outcomes on public health, and period prevalence would be nearly identical to the risk over the same period.

2 Conditional Probability and Diagnostic Tests

> The probability that two subsequent events will happen is a ratio compounded of the probability of the 1st and the probability of the 2d on supposition the 1st happens. (Bayes 1763)[1]

[1]Thomas Bayes (1701-1761) was an English Presbyterian minister from a family of Nonconformists (i.e., Protestants who did not observe the rules of the Church of England). He studied logic and theology at the University of Edinburgh and served as a minister in Tunbridge Wells near Kent, England. He was elected a Fellow of the Royal Society in 1742 for his defense of Newton's calculus against a 1734 book called *The Analyst: A Discourse Addressed to an Infidel Mathematician* by Bishop George Berkeley (1685-1753). Late in life, Bayes became interested in probability and "inverse probability" (statistics). This essay was published posthumously. See https://en.wikipedia.org/wiki/Thomas_Bayes.

3 Maximum Likelihood Estimation

> Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise. (Tukey 1962)[1]

[1] John Tukey (1915-2000) was an American mathematician and statistician who worked at Bell Labs and Princeton University. He developed the box plot, Tukey's range test for multiple comparisons, and the fast Fourier transform. In 1947, he coined the term "bit" as shorthand for "binary digit". See https://en.wikipedia.org/wiki/John_Tukey.

4 Bayesian Estimation

> In the null hypothesis schema we are trying only to nullify something: "The null hypothesis is never proved or established but is possibly disproved in the course of experimentation." But ordinarily evidence does not take this form. With the *corpus delicti* in front of you, you do not say, "Here is evidence against the hypothesis that no one is dead." You say, "Evidently someone has been murdered." (Berkson 1942)[1]

[1] Joseph Berkson (1899–1982) was an American physician and statistician at the Mayo Clinic in Rochester, Minnesota. He helped develop and popularize the use of logistic regression for binary outcomes, coining the term "logit" for the log odds in 1944. He also pioneered the study of selection bias, a special case of which is called "Berkson's bias". Later, he became a prominent opponent of the idea that smoking causes lung cancer. See https://en.wikipedia.org/wiki/Joseph_Berkson.

5 Longitudinal Data and Rates

6 Survival Analysis

Part II

B. Two-Sample Inference and Study Design

Part III

C. Principles of Causal Inference

Part IV

D. Epidemologic and Statistical Methods for Causal Inference

References

Bayes, Thomas. 1763. "LII. An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, FRS Communicated by Mr. Price, in a Letter to John Canton, AMFRS." *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Berkson, Joseph. 1942. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37 (219): 325–35.

Laplace, Pierre Simon. 1820. *Théorie Analytique Des Probabilités*. Vol. 7. Courcier.

Morabia, Alfredo. 2004. "Epidemiology: An Epistemological Perspective." In *A History of Epidemiologic Methods and Concepts*, edited by Alfredo Morabia, 3–125. Springer.

Snow, John. 1855. *On the Mode of Communication of Cholera*. Second edition. John Churchill. https://wellcomecollection.org/works/uqa27qrt.

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1): 1–67.