Analytical Epidemiology

Statistical and Causal Inference for Public Health

Eben Kenah

January 15, 2025

Table of contents

Preface

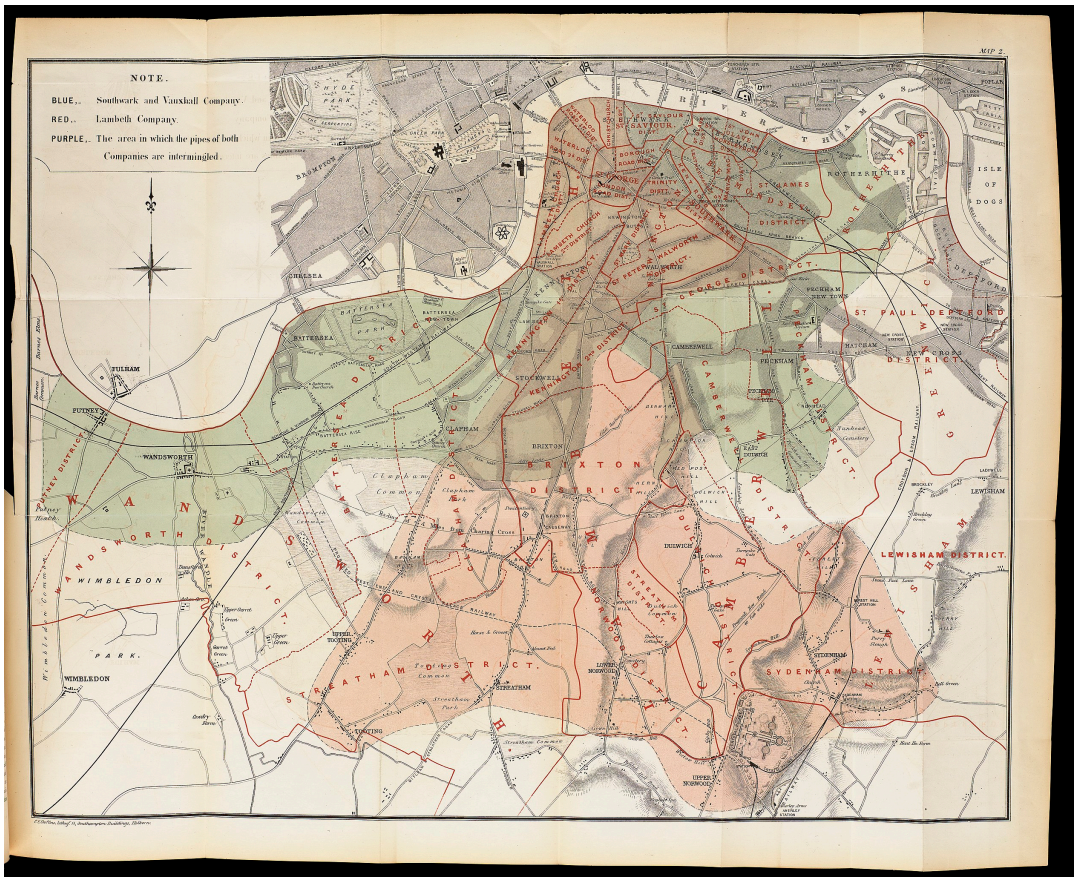

Figure 1: Areas of London supplied by the Southwark & Vauxhall (blue, now green) and Lambeth (red) water companies during the 1849-1854 cholera epidemic in London (Snow 1855). Source: Wellcome Collection via Wikimedia Commons.

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 applications in 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 possible that epidemiologic research has 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. You may find that some parts are unfinished or just bad. Please report errors (including typos) or submit suggestions (especially good examples) at:

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

Acknowledgements

This book is written in LaTeX and Quarto with calculations and figures generated in R, Python, and Inkscape. I have also included many links to Wikipedia. These are free, open-source, and publicly available thanks to the work of many contributors.

Tony Barry, 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

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 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 difficult terrain in epidemiology.

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 **element** a, we write

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

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. 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:

$$(a, b) = \{x \in \mathbb{R} : a < x < b\}, \tag{1.3}$$
$$(a, b] = \{x \in \mathbb{R} : a < x \leq b\}, \tag{1.4}$$
$$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}, \tag{1.5}$$
$$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}. \tag{1.6}$$
$$\tag{1.7}$$

An endpoint with a square bracket is included in the interval; 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

$$\Omega = \{H, T\}, \tag{1.8}$$

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,

$$\Omega = \{HH, HT, TH, TT\} \tag{1.9}$$

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 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 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

$$A^{\mathsf{C}} = \{\omega \in \Omega : \omega \notin A\}, \tag{1.10}$$

  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

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}, \tag{1.11}$$

  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

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}, \tag{1.12}$$

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

[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 nonempty interval (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, but it has become a cornerstone of modern mathematics.

[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

$$A \cap A = A \cup A = A. \tag{1.13}$$

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 42-year-old Ohio residents, 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 in Figure 1.1. The rectangle represents Ω, and the circles A and B represent events. A^{C} is everything in Ω outside the circle A, and B^{C} is everything outside the circle B. Their intersection $A \cap B$ is the area where the two circles overlap. Their union $A \cup B$ is everything contained in at least one of A or B.

[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.

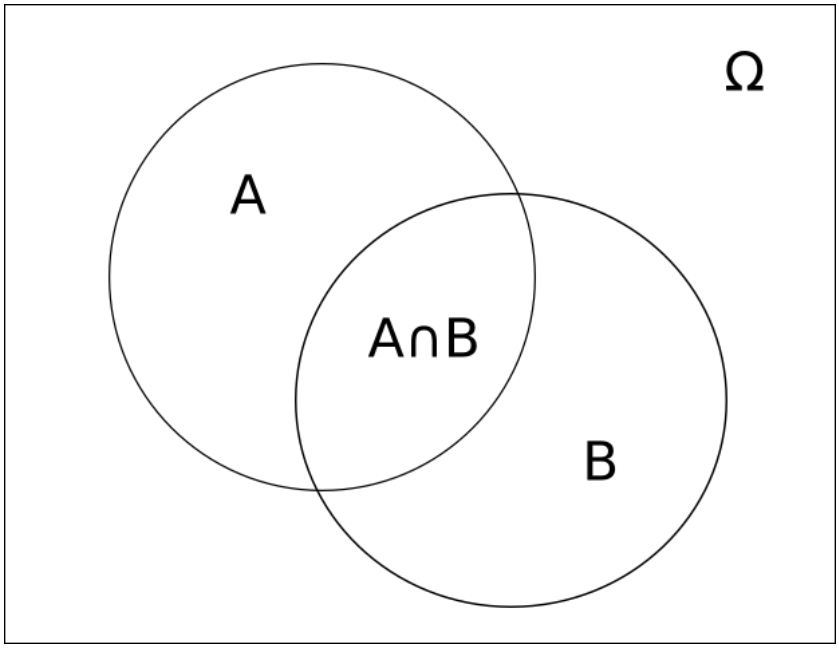

Figure 1.1: Venn diagram showing events A and B. The area contained in both events is their intersection $A \cap B$. The union $A \cup B$ is all area contained in at least one of A and B, including $A \cap B$.

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

$$I_n = \bigcap_{i=1}^{n} A_i. \tag{1.14}$$

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

$$U_n = \bigcup_{i=1}^{n} A_i. \tag{1.15}$$

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 1.14 or Equation 1.15. 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.

[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.

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 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

$$(A \cap B)^{\mathsf{C}} = A^{\mathsf{C}} \cup B^{\mathsf{C}} \tag{1.16}$$

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

$$(A \cup B)^{\mathsf{C}} = A^{\mathsf{C}} \cap B^{\mathsf{C}}. \tag{1.17}$$

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 1.13, 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 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 class $\mathcal{F}$ 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 class $\mathcal{F}$ is called a σ-algebra or σ-field, and this restriction on the domain of probability helps avoid internal contradictions like the 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,

$$\Pr(A \cup B) = \Pr(A \cap B^{\mathsf{C}}) + \Pr(B \cap A^{\mathsf{C}}) + \Pr(A \cap B). \tag{1.18}$$

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 1.18 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. This can be see clearly in Figure 1.1. □

1.3 Random variables

The outcomes of an experiment can be anything, not just 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

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

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. When sampling from a population, we can define indicator variables for membership in different subpopulations.

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

$$\mathbb{1}_{A^\mathsf{C}} = 1 - \mathbb{1}_A. \tag{1.19}$$

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

$$\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B. \tag{1.20}$$

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

$$\mathbb{1}_{A \cup B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B) = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A \cap B}. \tag{1.21}$$

  This follows from Equation 1.17 because $A \cup B = (A^\mathsf{C} \cap B^\mathsf{C})^\mathsf{C}$.

1.4 R

1.4.1 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:

[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.

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

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

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

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

- 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 $x \in \operatorname{supp}(X)$, 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

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

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

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

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

1.4.2 Mean

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

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

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.

Indicators are an extremely useful for calculating probabilities using means. For any event A, its probability is the mean of the indicator variable $\mathbb{1}_A$:

$$\Pr(A) = 0 \Pr(\mathbb{1}_A = 0) + 1 \Pr(\mathbb{1}_A = 1) = \mathbb{E}(\mathbb{1}_A).$$

This is a common way to calculate probabilities in data analyses.

1.5 R

1.5.1 Variance

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

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

{eq-Var} 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. An equivalent expression for the variance that is often easier to use is:

$$\text{Var}(X) = \mathbb{E}(X^2) - \mu^2 \tag{1.22}$$

where $\mathbb{E}(X^2)$ is the expected value of the random variable X^2. 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.

1.5.2 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

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

Equivalently, it has the CDF

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } x \in [0, 1) \\ 1 & \text{if } x \geq 1. \end{cases}$$

[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.

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

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

and variance

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

Its standard deviation is $\sqrt{p(1-p)}$, which 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.

1.6 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. The distinction between joint and marginal distributions is extremely important in epidemiology and other applications of probability.

For example, Table 1.1 shows the joint and marginal PMFs for two binary random variables X and Y. By definition,

$$f(0,0) + f(0,1) + f(1,0) + f(1,1) = 1.$$

In the table, it is clear that the joint distribution determines the marginal distributions. However, there are many different joint distributions that are consistent with the same marginal distributions. Thus, the marginal distributions do not determine the joint distribution.[9]

[9] This becomes a fundamental insight when we discuss hypothesis tests for independence as well as confounding and selection bias.

Table 1.1: Joint and marginal PMFs for binary random variables X and Y.

	$Y = 0$	$Y = 1$	X margin
$X = 0$	$f(0,0)$	$f(0,1)$	$f_X(0) = f(0,0) + f(0,1)$
$X = 1$	$f(1,0)$	$f(1,1)$	$f_X(1) = f(1,0) + f(1,1)$
Y **margin**	$f_Y(0) = f(0,0) + f(1,0)$	$f_Y(1) = f(0,1) + f(1,1)$	1

1.7 R

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

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

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 $n = 3$. In this same case, we can talk about the joint distribution of any two variables marginalized over the third. For example,

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

For larger n, the formulas gets uglier but the ideas are the same.

1.7.1 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 $X_1, \dots, X_n$ are random variables defined on a sample space and $a_1, \dots, a_n$ 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 $X_1, \dots, X_n$. 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^{C} in Equation 1.19 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 1.21 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 1.20 is not a linear combination of $\mathbb{1}_A$ and $\mathbb{1}_B$ because we have to multiply these two variables.

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

$$\mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y). \tag{1.23}$$

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.

1.7.2 Variance and covariance*

The variance of $aX + bY$ is

$$\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y) \tag{1.24}$$

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. When X and Y are *independent* in the sense that the value of one tells us nothing about the value of the other, then $\text{Cov}(X, Y) = 0$ and $\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2 Var(Y)$.[10]

[10] Discrete random variables X and Y are independent if $\Pr(X = x \text{ and } Y = y) = \Pr(X = x)\Pr(Y = y)$ for any possible values $x \in \text{supp}(X)$ and $y \in \text{supp}(Y)$. We will discuss independence more rigorously when we discuss conditional probabilities in Chapter 2.

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*,[11] 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 (but not only if) X and Y are independent. 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.

1.8 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

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

[11] 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.

The set of individuals in Ω who have $D(\omega) = 1$ is an event in Ω, and our measure of disease occurrence is

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

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

There is an important technical detail 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}.

The time scale used to define disease onset is flexible, and this flexibility is useful. The most obvious time scale is *calendar time* or *absolute time*. Another common time scale is age, which is an important determinant of the risk of many diseases. In some cases, time since an event is a useful time scale. The event that defines time scale could be a single event (e.g., exposure to contaminated food at a party) or an event that occurs at different times for different individuals (e.g., time since menopause). In general, it is wise to choose the time scale that corresponds to the most important time-varying determinant of disease onset. The chosen time scale is often called the *analysis time scale*.

1.8.1 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 includes prevalent cases at time t_a and cases with disease onset in $(t_a, t_b]$. The period prevalence in $(t_a, t_b]$ is the point prevalence at t_a plus the risk of disease onset in $(t_a, t_b]$, to which we now turn.

1.9 R

1.9.1 Risk (cumulative incidence) and the survival function

To define **risk** or **cumulative incidence**, we first choose an nonempty time interval $(t_a, t_b]$. 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 $t^{\text{ons}} \leq t_b$. The risk is sometimes called the *incidence proportion*.

The risk depends on a specified interval $(t_a, t_b]$. We can always define our time scale so that $t_a = 0$, so the risk in $(t_a, t_b]$ on the original time scale is the same as the risk in the interval $(0, t_b - t_a]$ on the analysis time scale. On the analysis time scale, the **cumulative incidence function** $F(t)$ is the risk of disease in $(0, t]$ for any possible t. The corresponding **survival function** is

$$S(t) = 1 - F(t),$$

which is the probability of no disease onset in $(0, t]$. In practice, it is often easier to calculate the survival function than to calculate the cumulative incidence function directly. There is only one way to survive disease-free through the interval $(0, t]$, but you can have disease onset at any time.

1.10 R

The survival function has several important properties:

- $S(0) = 1$ because $(0, 0]$ is an empty interval where no one can have disease onset.
- Because $S(t)$ is a probability, $S(t) \in [0, 1]$ for all t.
- $S(t)$ monotonically decreases (i.e., never increases) with increasing t. If $t_a < t_b$, then the time interval $(0, t_a]$ is contained $(0, t_b]$. Everyone who survives disease-free through $(0, t_b]$ must have survived disease-free through $(0, t_a]$, but some people who survived through $(0, t_a]$ might not make it all the way through $(0, t_b]$. Thus, $S(t_a) \geq S(t_b)$ whenever $t_a < t_b$.
- If the disease or event occurs eventually for all individuals in our population Ω (e.g., death), then $S(t) \to 0$ as $t \to \infty$.

Each of these probabilities follows directly from the definition of $S(t)$. Similarly, the cumulative incidence function F has $F(0) = 0$ and $F(t) \in [0, 1]$, and it is monotonically increasing (i.e., never decreasing) with increasing t. If the disease or event occurs eventually in all individuals, then $F(t) \to 1$ as $t \to \infty$. Figure 1.2 shows the survival and cumulative hazard curves for the data generated in the prevalence example above.

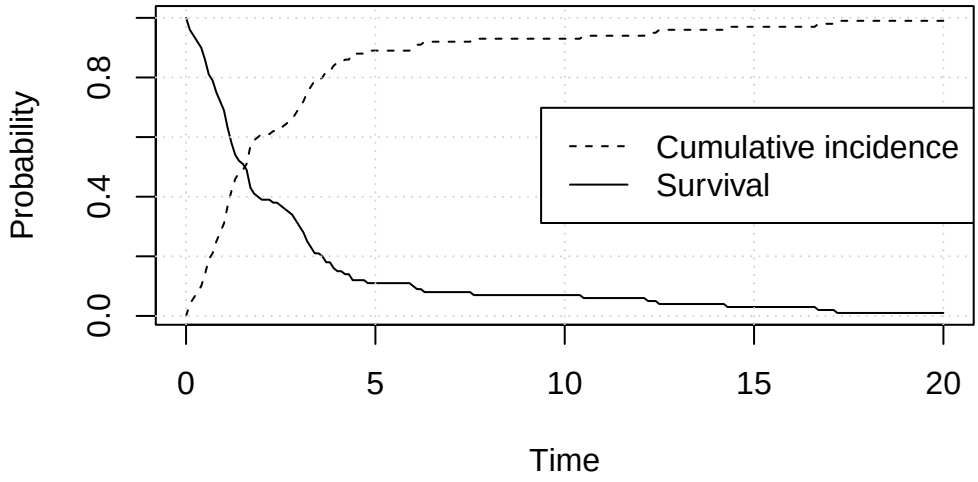

Figure 1.2: Survival and cumulative incidence curves for the data from the prevalence example.

Here, I will generally use the word "risk" to refer to the probability of disease onset in a specified interval. When there is possible confusion about the meaning of "risk", I will use "cumulative incidence" instead. The terms "cumulative incidence function" and "survival function" are standard in survival analysis, which is the branch of statistics that studies times to events. The creative use of "risk" in public health and medicine should not make you shy away from using the word correctly.

1.10.1 Prevalence and the duration of disease

Point and period prevalence are both affected by the duration of disease. Both measures will increase if the duration of disease increases. A simple illustration of this is given in Figure 1.3. For a fixed set of onset times, the point prevalence of disease at any time t either stays the same or increases when the duration of disease increases. The prevalence at time $t = 5$ is $\frac{2}{5} = 0.4$ under the shorter duration of disease but $\frac{3}{5} = 0.6$ under the longer duration of disease. Period prevalence over any interval $(t_a, t_b]$ is affected by the duration of disease because it is the point prevalence at t_a (which is affected by disease duration) plus the risk of disease onset over $(t_a, t_b]$. In a given population, the relationship between prevalence, frequency of disease onset (incidence), and the duration of disease can be complex (Freeman and Hutchison 1980; Preston 1987; Keiding 1991; Alho 1992). The risk of disease in any given interval is not affected by the duration of disease.

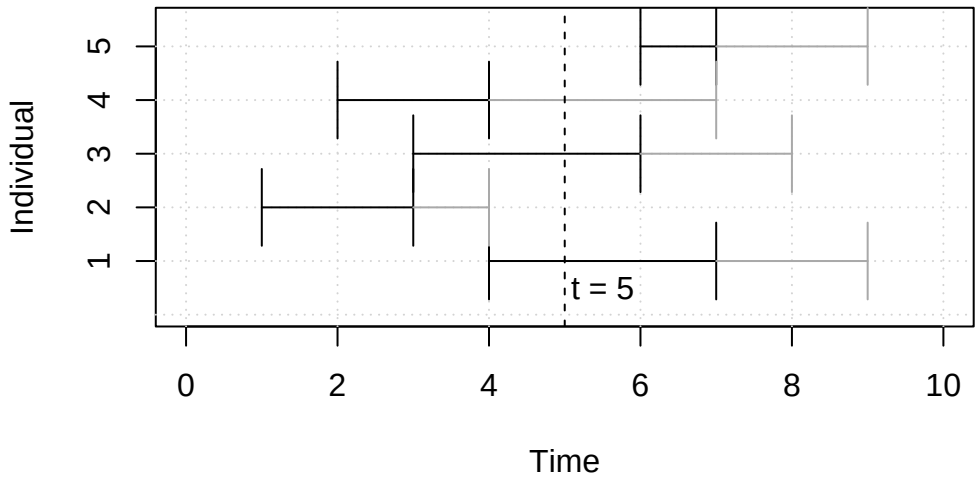

Figure 1.3: Each black horizonal line shows the onset of disease and recovery from disease in a single individual. The gray lines show recoveries from disease if the disease duration increases.

1.10.2 Descriptive and analytic epidemiology

Prevalence is often a useful measure for **descriptive epidemiology**, which measures the distribution of disease over person, place, and time. Because prevalence depends on both incidence and duration of disease, a change in the prevalence of disease can generally be explained several different ways (MacMahon and Terry 1958; Dunn Jr 1962). For example, an increase in prevalence of human immunodeficiency virus (HIV) infection could be cause by an increase in the incidence of HIV infection (which is bad) or an increase in the life expectancy of HIV-infected people (which is good).

Risk (cumulative incidence) is generally more useful than prevalence for **analytic epidemiology**, which attempts to identify the causes of a disease. 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) and for outcomes that remove individuals from the population (e.g., emigration or death). Prevalence is not a useful measure of the public health impact of these events.

Listing 1.1 `indicators.R`

```r
## Indicator variables for events A and B, etc.

# Setting the seed ensures that everyone gets the same random samples.
# Functions are called using parentheses (round brackets).
# The function rbinom() is a random sample from a binomial distribution.
set.seed(42)
n <- 100
dat <- data.frame(A = rbinom(n, 1, 0.3))
dat$B <- rbinom(n, 1, 0.6)

# inspecting a data frame
names(dat)  # variables in the data frame
nrow(dat)   # number of rows (individuals)
ncol(dat)   # number of columns (variables)
dim(dat)    # rows and columns in the data frame
str(dat)    # summary of the data frame structure (variables and types)

# inspecting columns of a data frame (or vectors)
# Our sample space or population consists of 100 individuals.
# Square brackets are used for indices, which can be numbers or TRUE/FALSE.
dat$A                 # indicator for A for all 100 individuals
dat$A[10]             # indicator for A in individual 10
dat$A[2:6]            # indicator variables for individuals 2 to 6
dat$A[c(10, 20, 30)]  # A indicators for individuals 10, 20, and 30
which(dat$A == 1)     # which individuals are in event A
which(dat$A == 0)     # which individuals are not in event A

# indicator variable for A complement
# In R (and many other languages), "!" means "not".
# The function as.integer() changes TRUE/FALSE to 1/0.
dat$Acomp <- as.integer(!dat$A)

# indicator variable for A intersection B
# In R (and many other languages), "&" means "and".
dat$ABintersect <- as.integer(dat$A & dat$B)

# indicator variable for A union B
# In R (and many other languages), "|" means "or".
dat$ABunion <- as.integer(dat$A | dat$B)

# save the data frame as a CSV file
# The file argument can be a path (e.g., "./data/indicators.csv" in Linux).
write.csv(dat, file = "indicators.csv", row.names = FALSE)
```

Listing 1.2 `probabilities.R`

```
## Indicator variables and probability calculations

# read in CSV file with indicator variables using the function read.csv()
# The argument can be a path (e.g., "./data/indicators.csv" in Linux).
dat <- read.csv("indicators.csv")

# calculate probabilities from indicator variables using the function mean()
# This will also work with TRUE/FALSE (i.e., logical) variables, which are
# converted to TRUE = 1 and FALSE = 0 in calculations.
prob_A <- mean(dat$A)
prob_B <- mean(dat$B)
prob_Acomp <- mean(dat$Acomp)
prob_ABintersect <- mean(dat$ABintersect)
prob_ABunion <- mean(dat$ABunion)

# Pr(A complement) = 1 - Pr(A)
prob_Acomp
1 - prob_A

# Pr(A union B) = Pr(A) + Pr(B) - Pr(A intersect B)
prob_ABunion
prob_A + prob_B - prob_ABintersect

# Beware of numerical error when comparing floating-point numbers!
# This example is from The R Inferno by Patrick Burns.
# https://www.burns-stat.com/pages/Tutor/R_inferno.pdf
0.1 == 0.3 / 3
sprintf("%.20f", 0.1)
sprintf("%.20f", 0.3 / 3)

# math can be more accurate than computers (which is not their fault)
prob_ABunion == prob_A + prob_B - probABintersect
sprintf("%.20f", prob_ABunion)
sprintf("%.20f", prob_A + prob_B - prob_ABintersect)
```

Listing 1.3 `jointdist.R`

```r
## Joint and marginal distributions of indicators for events A and B

# read indicator variable data from the CSV file
dat <- read.csv("indicators.csv")
n <- nrow(dat)

# tables of counts
# Putting "<name> = " before the vector creates a label.
table(A = dat$A)
table(B = dat$B)

# joint table of counts
# In table(), the first argument defines rows and the second defines columns.
# The addmargins() functions adds the row, column, and overall sums.
table(A = dat$A, B = dat$B)
addmargins(table(A = dat$A, B = dat$B))

# tables of probabilities
# Table margins match the distributions of A (rows) and B (columns).
table(Adist = dat$A) / n    # marginal distribution of A indicator
table(Bdist = dat$B) / n    # marginal distribution of B indicator
addmargins(table(A = dat$A, B = dat$B)) / n   # joint distribution
```

Listing 1.4 `prevalence.R`

```r
## Point and period prevalence

# generate onset and recovery data for 100 individuals
# Setting the seed ensures that everyone gets the same random numbers,
# but it is strictly optional.
# The function rexp() randomly samples from an exponential distribution.
set.seed(42)
cohort <- data.frame(onset = rexp(100, rate = 0.4))
cohort$duration <- rexp(100, rate = 2)
cohort$recovery <- cohort$onset + cohort$duration

# statistical summaries (mean, quartiles, range)
summary(cohort$onset)
summary(cohort$duration)
summary(cohort$recovery)

# highest and lowest recovery times
# The function sort() sorts the vector from lowest to highest.
# head() returns the first 6 values of a vector; tails() returns the last 6.
min(cohort$onset)
head(sort(cohort$onset))    # lowest 6 values (first 6 in the sorted vector)
tail(sort(cohort$onset))    # highest 6 values (last 6 in the sorted vector)
max(cohort$onset)

# With a long vector, sorting repeatedly can be slow.
# You can also control the number of elements returned by head() or tail().
onset_ordered <- sort(cohort$onset)
head(onset_ordered, n = 10)
tail(onset_ordered, n = 10)

# seeing rows and columns of the data frame
cohort[1:10, c("onset", "duration", "recovery")]
cohort[c(10, 20, 50), c("onset", "recovery")]
cohort[which(cohort$recovery < 1), c("onset", "recovery")]
cohort[, c("onset", "recovery")]    # all rows
cohort[c(2, 3, 5, 7, 11), ]         # all columns

# point prevalence
prev <- function(t) {
  # vector of TRUE/FALSE for prevalent cases at time t
  prevalent <- cohort$onset <= t & cohort$recovery > t
  mean(prevalent)
}

prev(0)
prev(1)
prev(2)
prev(6)

# period prevalence
# The parentheses around the logical tests are just for readability.
```

Listing 1.5 `risk.R`

```r
## Risk, survival function, and cumulative incidence function

# read data from CSV file
# Change or remove ".R/" in the path as needed to locate the cohort.csv file.
# You can also re-generate the data as in prevalence.R using the same seed.
cohort <- read.csv("./R/cohort.csv")

# risk (cumpulative incidence)
risk <- function(t) {
  # vector of TRUE/FALSE for incident cases in (0, t]
  incident <- cohort$onset <= t
  mean(incident)
}

risk(0)
risk(1)
risk(2)
risk(6)

# cumulative incidence function
# Vectorize() takes a function like risk() that takes a single number as input
# and creates a function that can take a number or vector as input.
cuminc <- Vectorize(risk)
cuminc(c(0, 1, 2, 6))

# survival function
# A simple function can be put on one line.
# It takes the same input as cuminc(), so it can take a vector
surv <- function(t) 1 - cuminc(t)
surv(c(0, 1, 2, 6))

# plot the survival and cumulative incidence functions
t <- seq(0, 20, by = 0.1)
plot(t, surv(t), type = "l",
     xlab = "Time", ylab = "Probability")
lines(t, cuminc(t), lty = "dashed")
grid()
legend("right", bg = "white", lty = c("dashed", "solid"),
       legend = c("Cumulative incidence", "Survival"))
```

Listing 1.6 `surv-fig.R`

```r
## Plot of survival and cumulative incidence functions

# read data from CSV file
# Change or remove ".R/" in the path as needed to locate the cohort.csv file.
# You can also re-generate the data as in prevalence.R using the same seed.
cohort <- read.csv("./R/cohort.csv")

# risk (cumpulative incidence)
risk <- function(t) {
  # vector of TRUE/FALSE for incident cases in (0, t]
  incident <- cohort$onset <= t
  mean(incident)
}

# cumulative incidence function
cuminc <- Vectorize(risk)

# survival function
surv <- function(t) 1 - cuminc(t)

# plot the survival and cumulative incidence functions
t <- seq(0, 20, by = 0.1)
plot(t, surv(t), type = "l",
     xlab = "Time", ylab = "Probability")
lines(t, cuminc(t), lty = "dashed")
grid()
legend("right", bg = "white", lty = c("dashed", "solid"),
       legend = c("Cumulative incidence", "Survival"))
```

Listing 1.7 `prevdur-fig.R`

```
## R code for prevalence and duration plot
plot(0, 0, type = "n", xlim = c(0, 10), ylim = c(0, 5.5),
     xlab = "Time", ylab = "Individual", yaxt = "n")
Axis(side = 2, at = 1:5, labels = 1:5)
grid()
start <- c(4, 1, 3, 2, 6)
stop1 <- c(7, 3, 6, 4, 7)
stop2 <- c(9, 4, 8, 7, 9)
arrows(x0 = start, y0 = 1:5, x1 = stop1, code = 3, length = 0.2, angle = 90)
arrows(x0 = stop1, y0 = 1:5, x1 = stop2, code = 2, length = 0.2, angle = 90,
       col = "darkgray")
abline(v = 5, lty = "dashed")
text(5.5, 0.5, label = "t = 5")
```

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]

Suppose we know that an event A occurred and want calculate the probability that B also occurred. The **conditional probability** of B given A is

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}. \tag{2.1}$$

Note that this is well-defined only if $\Pr(A) > 0$. Conditional probabilities given A are just probabilities where the original sample space Ω has been replaced with an event $A \subseteq \Omega$. Everything we have learned about probabilities applies to all of the conditional probabilities given the same event A. Conditional probability is arguably the most important mathematical tool in epidemiology.

2.1 Contingency tables

In statistics, a **contingency table** classifies individuals by two discrete variables, one that defines the rows and one that defines the columns. Each cell in the table contains the number of individuals who are in the intersection of the corresponding categories of the row and column variables. These numbers are called *cell counts*. The margins of the table contain row or column totals.

[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, and it has had a profound effect on modern statistics.

Table 2.1: 2x2 table of exposure (X) and disease (D).

	$D = 1$	$D = 0$	Total
$X = 1$	a	b	$r_1 = a + b$
$X = 0$	c	d	$r_0 = c + d$
Total	$k_1 = a + c$	$k_0 = b + d$	$n = a + b + c + d$

2.1.1 2x2 tables

In epidemiology, a **2x2 table** is a contingency table based on a binary exposure variable and a binary disease outcome. We denote exposure by $X = 1$ and no exposure by $X = 0$, and we denote disease by $D = 1$ and no disease by $D = 0$. The precise definition of "disease" depends on context. In descriptive epidemiology, $D_i = 1$ might mean that person i is a prevalent case of disease. In analytic epidemiology, $D_i = 1$ might mean that person i had an onset of disease in an interval $(t_{\text{start}}, t_{\text{stop}}]$ on a relevant time scale. We put exposure in the rows and disease in the columns,[2] and the exposure and disease categories are ordered so that individuals with $X = 1$ and $D = 1$ go in the top left corner. This is the most common arrangement in epidemiologic research, but it is not universal.

Table 2.1 shows an example of a 2x2 table. There are a individuals with both exposure and disease, b individuals with exposure but not disease, c individuals with disease but no exposure, and d individuals with neither. In the rows, there are $r_1 = a + b$ exposed individuals and $r_0 = c + d$ unexposed individuals. In the columns, there are $k_1 = a + c$ individuals who had a disease onset and $k_0 = b + d$ individuals who did not. The row and column totals are called the *margins* of the table. The total number of individuals is $n = a + b + c + d$.

2.1.2 Joint and marginal probabilities

Here, we assume that Table 2.1 represents our entire population Ω and our experiment is to randomly sample an individual $\omega \in \Omega$ and measure their exposure status $X(\omega)$ and their disease status $D(\omega)$. Probabilities involving both X and D are called **joint probabilities**, and they can be calculated using the cell counts. In Table 2.1, the four joint probabilities are

$$\begin{aligned}
\Pr(X = 1 \text{ and } D = 1) &= a/n, \\
\Pr(X = 1 \text{ and } D = 0) &= b/n, \\
\Pr(X = 0 \text{ and } D = 1) &= c/n, \\
\Pr(X = 0 \text{ and } D = 0) &= d/n.
\end{aligned}$$

[2] This is partly to respect the linear algebra convention that rows come before columns in matrix indices, so M_{ij} is the entry in row i and column j of the matrix M. In analytic epidemiology, exposure must occur before any disease that it causes, so we let the exposure define the rows.

Together, these probabilities defined the joint distribution of the random variables X and D via their joint probability mass function (PMF).

Probabilities involving X or D alone are called **marginal probabilities** because they are calculated using the margins of the table. In Table 2.1, the marginal probabilities for exposure X are

$$\Pr(X = 1) = r_1/n,$$
$$\Pr(X = 0) = r_0/n.$$

Together, these define the marginal distribution of X, which is Bernoulli(r_1/n). The marginal probabilities for disease D are

$$\Pr(D = 1) = k_1/n,$$
$$\Pr(D = 0) = k_0/n.$$

Together, these define the marginal distribution of D, which is Bernoulli(k_1/n).

2.1.3 Conditional probabilities

Joint and marginal probabilities can be used to calculate conditional probabilities, which have a joint probability in the numerator and a marginal probability in the denominator. As before, we assume that Table 2.1 represents our entire population Ω and our experiment is to randomly sample an individual $\omega \in \Omega$ and measure $X(\omega)$ and $D(\omega)$. In Table 2.1, the conditional probability of disease given exposure is

$$\Pr(D = 1 \,|\, X = 1) = \frac{Pr(D = 1 \text{ and } X = 1)}{\Pr(X = 1)} = \frac{a/n}{r_1/n} = \frac{a}{r_1},$$

and the conditional probability of disease given no exposure is

$$\Pr(D = 1 \,|\, X = 0) = \frac{Pr(D = 1 \text{ and } X = 0)}{\Pr(X = 0)} = \frac{c/n}{r_0/n} = \frac{c}{r_0},$$

Similarly, the conditional probability of exposure given disease is

$$\Pr(X = 1 \,|\, D = 1) = \frac{\Pr(X = 1 \text{ and } D = 1)}{\Pr(D = 1)} = \frac{a/n}{k_1/n} = \frac{a}{k_1},$$

and the conditional probabilty of exposure given no disease is

$$\Pr(X = 1 \,|\, D = 0) = \frac{\Pr(X = 1 \text{ and } D = 0)}{\Pr(D = 0)} = \frac{b/n}{k_0/n} = \frac{b}{k_0}.$$

In all cases, the table total cancels out and we get a calculation in one row (for conditional probabilities given X) or one column (for conditional probabilities given D).

2.2 Multiplication of conditional probabilities

Equation 2.1 can be rearranged into

$$\Pr(A \cap B) = \Pr(B \mid A) \Pr(A), \tag{2.2}$$

exactly as described by Bayes at the beginning of this chapter (if we let A be the "1st event" and B be the "2d"). This depends only on the definition of conditional probability in Equation 2.1, not on any assumptions about the relationship between the events A and B. This multiplication rule for conditional probabilities extends to any number of events. For three events A, B, and C such that $B \cap C$ and C have probabilities greater than zero, we have

$$\begin{aligned} \Pr(A \cap B \cap C) &= \Pr(A \mid B \cap C) \Pr(B \cap C) && (2.3) \\ &= \Pr(A \mid B \cap C) \Pr(B \mid C) \Pr(C). && (2.4) \end{aligned}$$

To ensure that all of these conditional probabilities are well-defined, we need $B \cap C$ and C to have probabilities greater than zero. In practice, $\Pr(A \mid B \cap C)$ is usually written $\Pr(A \mid B, C)$.

2.2.1 Decision trees

Figure 2.1 shows an example of a **decision tree**. The *root* of the tree is on the left and the *leaves* of the tree are on the right. Each node where two or more branches meet represents a decision. In the example, the root represents the decision A or A^C (i.e., not A). The two nodes connected to the root each represent the decision B or B^C (i.e., not B). Each branch of the tree is labeled with the conditional probability of the branch given the event that it branches out from. Because of the multiplication rule for conditional probabilities, the probability of each leaf is equal to the product of the probabilities along the branches connecting it to the root.

2.2.2 Independence of events

The events A and B are **independent** if

$$\Pr(A \cap B) = \Pr(A) \Pr(B). \tag{2.5}$$

When two events are independent, the occurrence (or not) of one event tells us nothing about whether the other event occurred: If $\Pr(A) > 0$, equation Equation 2.5 is equivalent to $\Pr(B \mid A) = \Pr(B)$. If $\Pr(B) > 0$, it is equivalent to $\Pr(A \mid B) = \Pr(A)$. If A and B are not independent, the occurrence of A contains information about the occurrence of B and vice versa.

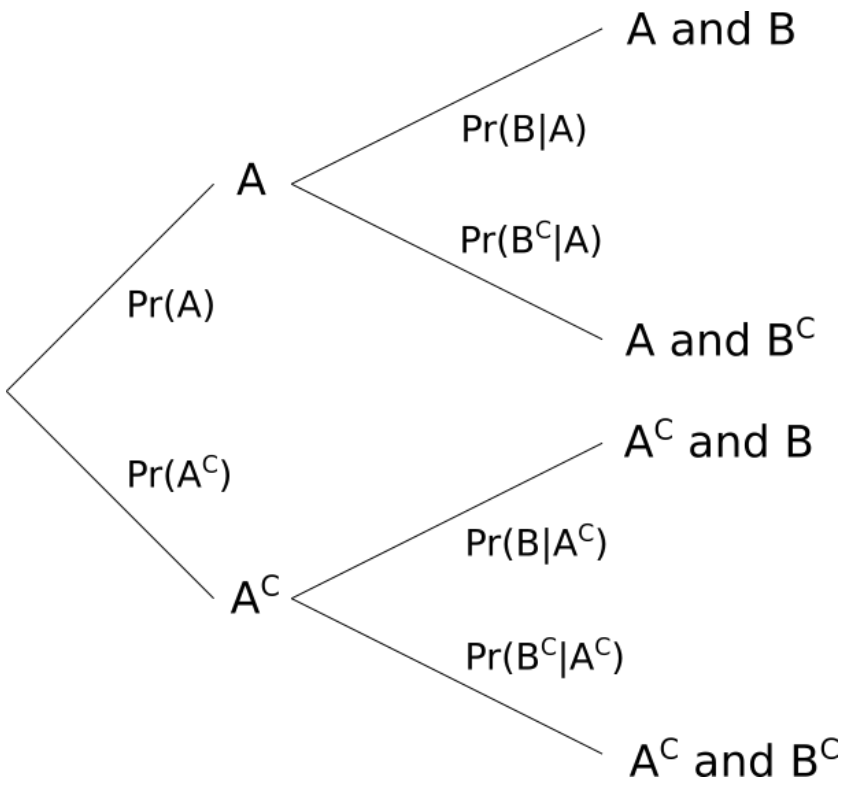

Figure 2.1: A decision tree for events A and B. The probability of each leaf is found by multiplying the probabilities along the branches leading from the leaf back to the root.

Table 2.2: Disease status (D^+/D^-) and test result (T^+/T^-).

	T^+	T^-
D^+	True positive	False negative
D^-	False positive	True negative

Independence of events A and B implies that the events A and B^{C} are also independent:

$$\begin{aligned}\Pr(A \cap B^{\mathsf{C}}) &= \Pr(A) - \Pr(A \cap B) \\ &= \Pr(A) - \Pr(A)\Pr(B) \\ &= \Pr(A)(1 - \Pr(B)) \\ &= \Pr(A)\Pr(B^{\mathsf{C}}).\end{aligned}$$

A similar argument shows that A^{C} and B are independent. Because $A^{\mathsf{C}} \cap B^{\mathsf{C}} = (A \cup B)^{\mathsf{C}}$ by DeMorgan's laws (see Section 1.1.5),

$$\begin{aligned}\Pr(A^{\mathsf{C}} \cap B^{\mathsf{C}}) &= 1 - \Pr(A \cup B) \\ &= 1 - \Pr(A) - \Pr(B) - \Pr(A \cap B) \\ &= 1 - \Pr(A) - \Pr(B) - \Pr(A)\Pr(B) \\ &= (1 - \Pr(A))(1 - \Pr(B)) \\ &= \Pr(A^{\mathsf{C}})\Pr(B^{\mathsf{C}}).\end{aligned}$$

Therefore, independence of two events implies independence between any combination of themselves or their complements.

2.3 Sensitivity and specificity

In the epidemiology of screening and diagnostic tests, several of the most important concepts are conditional probabilities. If we classify disease status into diseased (D^+) and nondiseased (D^-) and the test result into positive (T^+) and negative (T^-), we have the four possible combinations Table 2.2.

The **sensitivity** of a test is the conditional probability that the test is positive given that the individual tested has the disease:

$$\text{sens} = \Pr(T^+ \mid D^+).$$

The **specificity** of a test is the conditional probability that the test is negative given that the individual tested does not have the disease:

$$\text{spec} = \Pr(T^- \mid D^-).$$

In both cases, we are conditioning on the disease status of the individual being tested. These concepts were introduced by Yerushalmy (1947) in a comparison of different types of chest X-rays for tuberculosis case detection.

2.4 R

Listing 2.1 `sensspec.R`

```r
## Sensitivity and specificity

# generate diagnostic testing data
set.seed(42)
n <- 500
dtdat <- data.frame(disease = rbinom(n, 1, 0.5))
dtdat$testpos <- ifelse(dtdat$disease,
                        rbinom(n, 1, 0.85), rbinom(n, 1, 0.05))

# prevalence
mean(dtdat$disease)
# Pr(T+)
mean(dtdat$testpos)

# sensitivity
mean(dtdat$testpos[dtdat$disease == TRUE])
sum(dtdat$disease & dtdat$testpos) / sum(dtdat$disease)

# specificity
1 - mean(dtdat$testpos[dtdat$disease == FALSE])
mean(!dtdat$testpos[dtdat$disease == FALSE])
```

Maximizing either sensitivity or specificity alone does not necessarily lead to good screening or diagnostic test: A test where everyone tests positive has perfect sensitivity but zero specificity, and a test where everyone tests negative has perfect specificity but zero sensitivity. There is almost always a tradeoff where higher sensitivity leads to lower specificity and vice versa.

2.4.1 Example: Diabetes testing

Remein and Wilkerson (1961) describe an early study of diabetes screening conducted by the United States Public Health Service in Boston City Hospital between 1954 and 1957. They

Table 2.3: Sensitivity and specificity of the Somogyi-Nelson blood glucose test for diabetes where T^+ corresponds to a concentration above 130 mg/dL.

	T^+	T^-	Sensitivity and specificity
Before meal			
D^+	31	39	sens $= 31/70 \approx 0.443$
D^-	5	505	spec $= 505/510 \approx 0.990$
One hour after meal			
D^+	55	15	sens $= 55/70 \approx 0.786$
D^-	48	462	spec $= 462/510 \approx 0.906$
Two hours after meal			
D^+	45	25	sens $= 45/70 \approx 0.643$
D^-	16	494	spec $= 494/510 \approx 0.969$
Three hours after meal			
D^+	34	36	sens $= 34/70 \approx 0.486$
D^-	1	509	spec $= 509/510 \approx 0.998$

recruited early-morning patients who were not febrile or acutely ill. Those willing to participate gave urine and blood samples. Next, they were given a meal meant to approximate an average breakfast or light lunch (a sandwich, 5 grams of butter, 60 grams of cheese, and three filled cookies). After the meal, they gave further urine and blood samples at one, two, and three hours after eating. The samples were analyzed using four different blood tests and six different urine tests. Participants returned for a follow-up visit between 3 and 21 days after the screening tests, where a definitive diagnosis of diabetes was made using an oral glucose tolerance test and a physical examination according to criteria established by a group of experts.

A total of 595 participants completed both visits. Table 2.3 is a reconstruction of the data for the Somogyi-Nelson blood test based on the 580 participants (70 with diabetes and 510 without) who took the test at all four time points. In the table, a positive test is defined as a blood glucose concentration above 130 mg/dL (milligrams per deciliter).

2.5 R

2.5.1 Receiver operating characteristic (ROC) curves*

The tradeoff between sensitivity and sensitivity in choosing a clinical measurement cutoff to distinguish positive and negative tests can be seen using a **receiver operating characteristic** (ROC) curve (Lusted 1971a, 1971b; Swets 1988; Zweig and Campbell 1993). These curves were

Listing 2.2 `RWtable.R`

```
## Table 2 from Remein and Wilkerson (Journal of Chronic Disease, 1961)

# function to generate numbers based on sensitivity and specificity
RWtable <- function(sens, spec, n1=70, n0=510) {
  # arguments:  sensitivity, specificity,
  #             n1 is number of diabetics, n0 is number of nondiabetics
  tp <- round(sens * n1)
  fp <- round((1 - spec) * n0)
  tn <- round(spec * n0)
  fn <- round((1 - sens) * n1)
  return(c(truepos = tp, falsepos = fp, trueneg = tn, falseneg = fn))
}

RWtable(0.443, 0.990)   # before meal
RWtable(0.786, 0.906)   # one hour after
RWtable(0.643, 0.969)   # two hours after
RWtable(0.486, 0.998)   # three hours after
```

originally used in World War II to analyze the performance of radar systems locating ships and airplanes. They were applied to diagnostic tests in the late 1950s in the first attempt to automate the classification of Pap smears to detect cervical cancer (Bostrom, Sawyer, and Tolles 1959; Lusted 1984; Bengtsson and Malm 2014).

Each combination of a clinical measurement and a cutoff between positive and negative tests defines a diagnostic or screening test that has a sensitivity sens $\in [0, 1]$ and a specificity spec $\in [0, 1]$. The horizontal axis of an ROC curve plots

$$1 - \text{spec} = \Pr(T^+ \mid D^-),$$

and its vertical axis plots $\text{sens} = \Pr(T^+ \mid D^+)$. The test corresponds to a point $(1 - \text{spec}, \text{sens})$ in the unit square $[0, 1] \times [0, 1]$. The best tests correspond to points close to the top left corner $(0, 1)$, which represents a test with perfect specificity (so $1 - \text{spec} = 0$) and perfect sensitivity.

For a sequence of cutoffs, a given clinical measurement produces a curve connecting the points produced by the tests based on it. Figure 2.2 shows four ROC curves based on data from Remein and Wilkerson (1961): one for the Somogyi-Nelson blood glucose measurement before the meal and one each for the measurements one, two, and three hours after the meal. For all four measurements, the curves are based on the combinations of sensitivity and specificity for glucose concentration cutoffs from 70 mg/dL to 200 mg/dL. In these tests, using a higher glucose concentration cutoff to define a positive test leads to lower sensitivity and higher specificity.

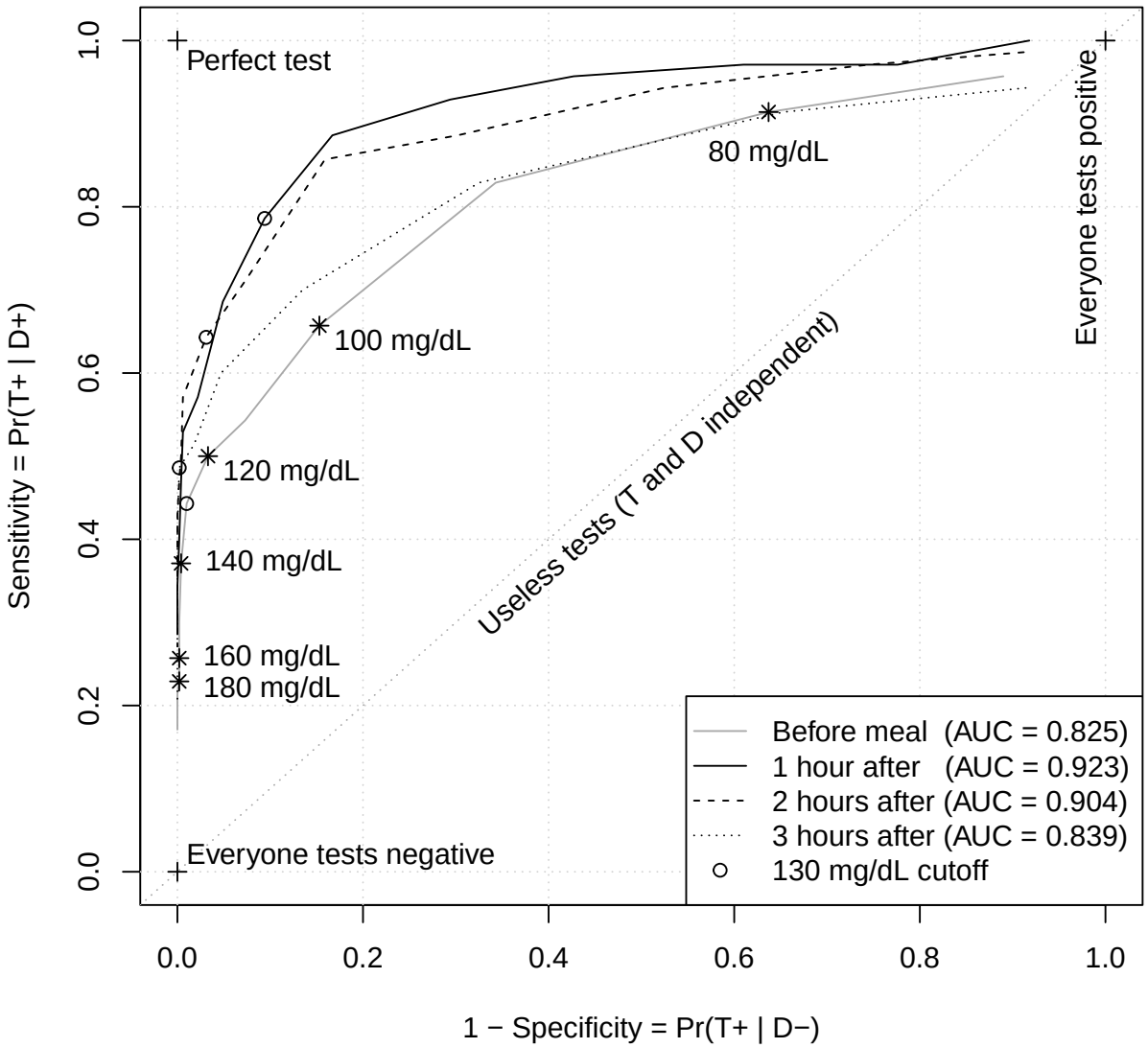

Figure 2.2: ROC curves for Somogyi-Nelson blood tests conducted before the meal and at 1-3 hours after the meal. Cutoff values for the before-meal curve are labeled, and the points corresponding to the 130 mg/dL cutoff along the curve for each blood glucose measurement are circled.

ROC curves for different clinical measurements can be compared using the area under the curve (AUC), which is the area between the x-axis $[0, 1]$ and the ROC curve. Greater AUC corresponds to a measurement that is better able to distinguish between disease and no disease (Bamber 1975; Hanley and McNeil 1982). For a test that is positive when a clinical measurement is above a given cutoff, the AUC is the probability that a person with disease

has a higher value than a person without disease.[3] In this example, it is the probability that a true diabetic has a higher blood glucose concentration than a true nondiabetic at the time blood glucose concentration is measured. A measurement that was always higher (or always lower) for individuals with disease than individuals without disease would have AUC = 1. The AUCs in Figure 2.2 show clearly that the tests one and two hours after the meal, which have curves above and to the left of the other two curves, better distinguish between diabetics and nondiabetics than the tests before and three hours after the meal. This is biologically plausible: Before the meal, there is no glucose load. Three hours after the meal, the glucose from the meal has largely been absorbed.

2.6 R

The test one hour after the meal with a 130 mg/dL cutoff has a good combination of sensitivity and specificity. It is near the top left corner, where perfect tests live. If a diagnostic test was completely useless, the test results (T^+ or T^-) would be independent of disease status (D^+ or D^-). In that case,

$$\Pr(T^+ \,|\, D^+) = \Pr(T^+ \,|\, D^-) = \Pr(T^+).$$

Thus, the ROC curve for a useless test follows the diagonal line from the lower left corner $(0, 0)$ to the upper right corner $(1, 1)$, and it has an AUC of 0.5. Tests below the diagonal on an ROC curve are worse than useless: the definitions of positive and negative should be reversed.

The sensitivity and specificity of a test tell us how accurate it is with a given definition of positive and negative. The ROC curve shows us how this accuracy depends on the cutoff between positive and negative tests, and the area under the curve shows us how well the underlying clinical measurement (e.g., blood glucose concentration) can distinguish between people with and without disease. However, the best cutoff for a test depends on its purpose, the population to be tested, and the benefit of identifying a true positive or negative versus the harm of a false positive or negative (Blumberg 1957; Kessel 1962).

2.7 Law of total probability

Suppose $A_1, \ldots, A_n$ are disjoint events such that their union is Ω. This is called a **partition** of Ω. An important special case is when we partition Ω into A and A^{C}.

[3] For a test that is positive when a clinical measurement is below a given cutoff, it is the probability that a person with disease has a lower value than a person without disease. Bamber (1975) showed that the AUC is closely related to the Wilcoxon rank sum statistic for the null hypothesis that the diseased and nondiseased have the same distribution for the measurement on which the test is based.

Let B be another event. Every $\omega \in B$ is in exactly one of the A_i. For each i, $B \cap A_i$ is the part of B that is contained in A_i. The event B is the union of these subsets:

$$B = \bigcup_{i=1}^{n} (B \cap A_i).$$

Because A_i are disjoint, so are the subsets $B \cap A_i$. By the addition rule for probabilities of disjoint sets, we have

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap A_i)$$

which is the sum of the $\Pr(B \cap A_i)$.[4] Using the multiplication rule for conditional probabilities in Equation 2.2 on each $\Pr(B \cap A_i)$, we get

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \mid A_i) \Pr(A_i).$$

This is called the **law of total probability**.

2.7.1 Example: probability of a positive or negative test

We can use the law of total probability to calculate the probability of a positive or negative test based on the sensitivity and specificity of the test and the prevalence of disease. Because all individuals either do or do not have the disease,[5] we have

$$T^+ = (T^+ \cap D^+) \cup (T^+ \cap D^-).$$

These two groups are mutually exclusive, so

$$\Pr(T^+) = \Pr(T^+ \cap D^+) + \Pr(T^+ \cap D^-).$$

We can calculate each probability on the right-hand side using the multiplication rule in Equation 2.2:

$$\begin{aligned}
\Pr(T^+ \cap D^+) &= \Pr(T^+ \mid D^+) \Pr(D^+) = \text{sensitivity} \times \text{prevalence}, \\
\Pr(T^+ \cap D^-) &= \Pr(T^+ \mid D^-) \Pr(D^-) = (1 - \text{specificity}) \times (1 - \text{prevalence}).
\end{aligned}$$

Putting everything together, we get

$$\begin{aligned}
\Pr(T^+) &= \Pr(T^+ \mid D^+) \Pr(D^+) + \Pr(T^+ \mid D^-) \Pr(D^-) \\
&= \text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence}).
\end{aligned} \tag{2.6}$$

[4] The symbol Σ, which is an upper-case Greek letter σ (sigma), stands for a sum. For products, we use Π, which is an upper-case Greek letter π (pi).

[5] Many diseases are complex processes (Rothman 1981), making any binary classification of disease status somewhat arbitrary. Here, we assume that we have an operational definition of disease status that allows a reasonable binary classification.

A similar chain of reasoning shows that

$$\Pr(T^-) = (1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence}),$$

which equals $1 - \Pr(T^+)$.

Figure 2.3 shows how the probability of a positive test depends on the prevalence of disease using the example of the Somogyi-Nelson test one hour after the meal in Table 2.3. With a cutoff of 130 mg/dL, the test has a sensitivity of 0.786 and a specificity of 0.906. At low prevalences, the test overestimates the prevalence of diabetes due to imperfect specificity. A high prevalences, it underestimates the prevalence of diabetes due to imperfect sensitivity. The errors cancel out somewhere near a prevalence of 30%.

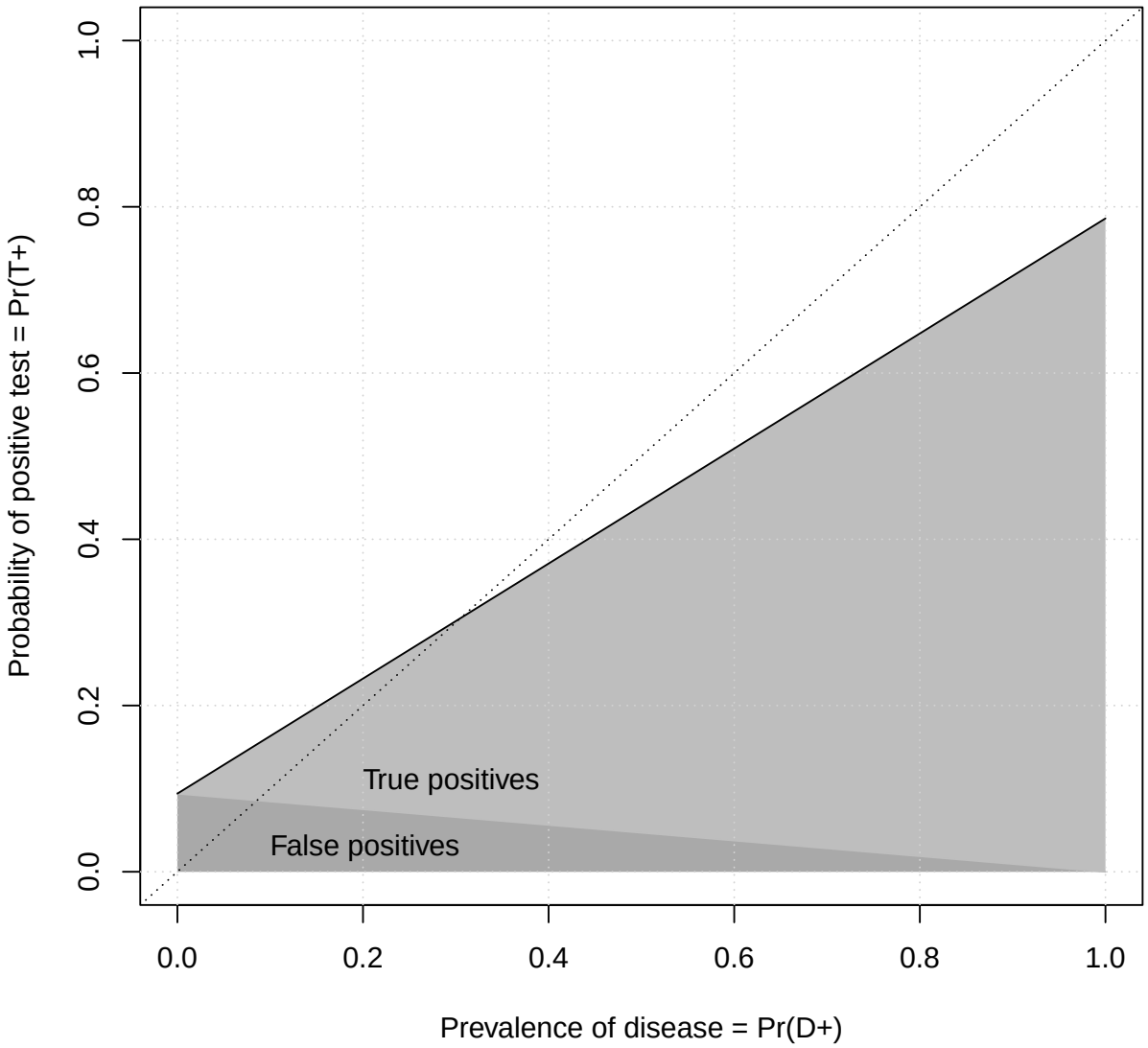

Figure 2.3: The probability of a positive Somogyi-Nelson diabetes test one hour after the meal as a function of the hypothetical prevalence of diabetes. The black dotted line shows the true prevalence of diabetes.

2.7.2 Standardization

In epidemiology, it is often useful to think of our sample space Ω as a population and the outcomes $\omega \in \Omega$ as individuals. The sets $A_1, \ldots, A_n$ into which we partition the sample space are disjoint subpopulations (e.g., age groups). Let $\Pr(D \mid A_i)$ be the prevalence of disease in

subpopulation A_i at a given time point. Then the overall prevalence of disease is

$$\Pr(D) = \sum_{i=1}^{n} \Pr(D \mid A_i) \Pr(A_i). \tag{2.7}$$

This application of the law of total probability is called **standardization**. By changing the $\Pr(A_i)$, we can use the subpopulation prevalences to calculate the prevalence of disease in a population with any desired composition of subpopulations. Equation 2.7 can also be used to calculate population-level risk from the subpopulation-specific risks in any given time interval. In the form of standardization, the law of total probability is one of the most important tools in epidemiology.

2.8 Bayes' rule

Bayes' rule (Bayes 1763) relates the conditional probabilities $\Pr(A \mid B)$ and $\Pr(B \mid A)$:

$$\Pr(A \mid B) = \frac{\Pr(B \cap A)}{\Pr(B)} = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)}. \tag{2.8}$$

In the denominator, the law of total probability is often used to calculate $\Pr(B)$ via partitioning Ω into A and A^{C}. This gives us

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B \mid A) \Pr(A) + \Pr(B \mid A^{\mathsf{C}}) \Pr(A^{\mathsf{C}})}.$$

Bayes' rule is an incredibly useful application of conditional probabilities, and it forms the theoretical foundation for Bayesian statistical inference.

2.8.1 Positive and negative predictive values

Sensitivity and specificity tell us how disease status predicts the result of a test, but they do not tell us how to interpret a test result. If you test positive, it is important to know the conditional probability that you truly have disease given that you tested positive. This is called the **positive predictive value** (PPV):

$$\text{PPV} = \Pr(D^+ \mid T^+).$$

If you test negative, it is important to know the conditional probability that you are truly disease-free given that you tested negative. This is called the **negative predictive value** (NPV):

$$\text{NPV} = \Pr(D^- \mid T^-).$$

These terms were introduced by Vecchio (1966). Table 2.4 shows the PPV and NPV for the Somogyi-Nelson diabetes tests from Table 2.3.

Table 2.4: PPV and NPV of the Somogyi-Nelson blood glucose test for diabetes where T^+ corresponds to a concentration above 130 mg/dL.

	T^+	T^-	PPV and NPV
Before meal			
D^+	31	39	PPV $= 31/36 \approx 0.861$
D^-	5	505	NPV $= 505/544 \approx 0.928$
Total	36	544	
One hour after meal			
D^+	55	15	PPV $= 55/103 \approx 0.534$
D^-	48	462	NPV $= 462/477 \approx 0.969$
Total	103	477	
Two hours after meal			
D^+	45	25	PPV $= 45/61 \approx 0.738$
D^-	16	494	NPV $= 494/519 \approx 0.952$
Total	61	519	
Three hours after meal			
D^+	34	36	PPV $= 34/35 \approx 0.971$
D^-	1	509	NPV $= 509/545 \approx 0.934$
Total	35	545	

Vecchio (1966) showed that the PPV and NPV depend on the prevalence of disease as well as the sensitivity and specificity of the test. To calculate the PPV and NPV, we use Bayes' rule to switch the conditional probabilities from $\Pr(T \mid D)$ to $\Pr(D \mid T)$. From the definition of PPV and Bayes' rule, we get

$$\Pr(D^+ \mid T^+) = \frac{\Pr(T^+ \cap D^+)}{\Pr(T^+)} = \frac{\Pr(T^+ \mid D^+)\Pr(D^+)}{\Pr(T^+)}.$$

The sensitivity of the test and the prevalence of disease are in the numerator, and $\Pr(T+)$ is in Equation 2.6. Putting this all together, we get

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}.$$

The numerator is the probability of a true positive test, and the denominator is the probability of a (true or false) positive test. By a similar argument,

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}}.$$

The numerator is the probability of a true negative test, and the denominator is the probability of a (true or false) negative test.

Figure 2.4 shows how the positive and negative predictive values of a test depend on the prevalence of disease for the Somogyi-Nelson test before the meal and one hour after the meal in Remein and Wilkerson (1961). With a cutoff of 130 mg/dL, the sensitivity and specificity are 0.443 and 0.990 before the meal and 0.786 and 0.906 one hour after the meal. If prevalence equals zero, the PPV is zero and the NPV equals one because no one has disease. As prevalence increases, PPV increases and NPV decreases. If the prevalence equals one, the PPV is one and the NPV is zero because everyone has disease. A perfect test would have PPV and NPV equal to one at all prevalences.

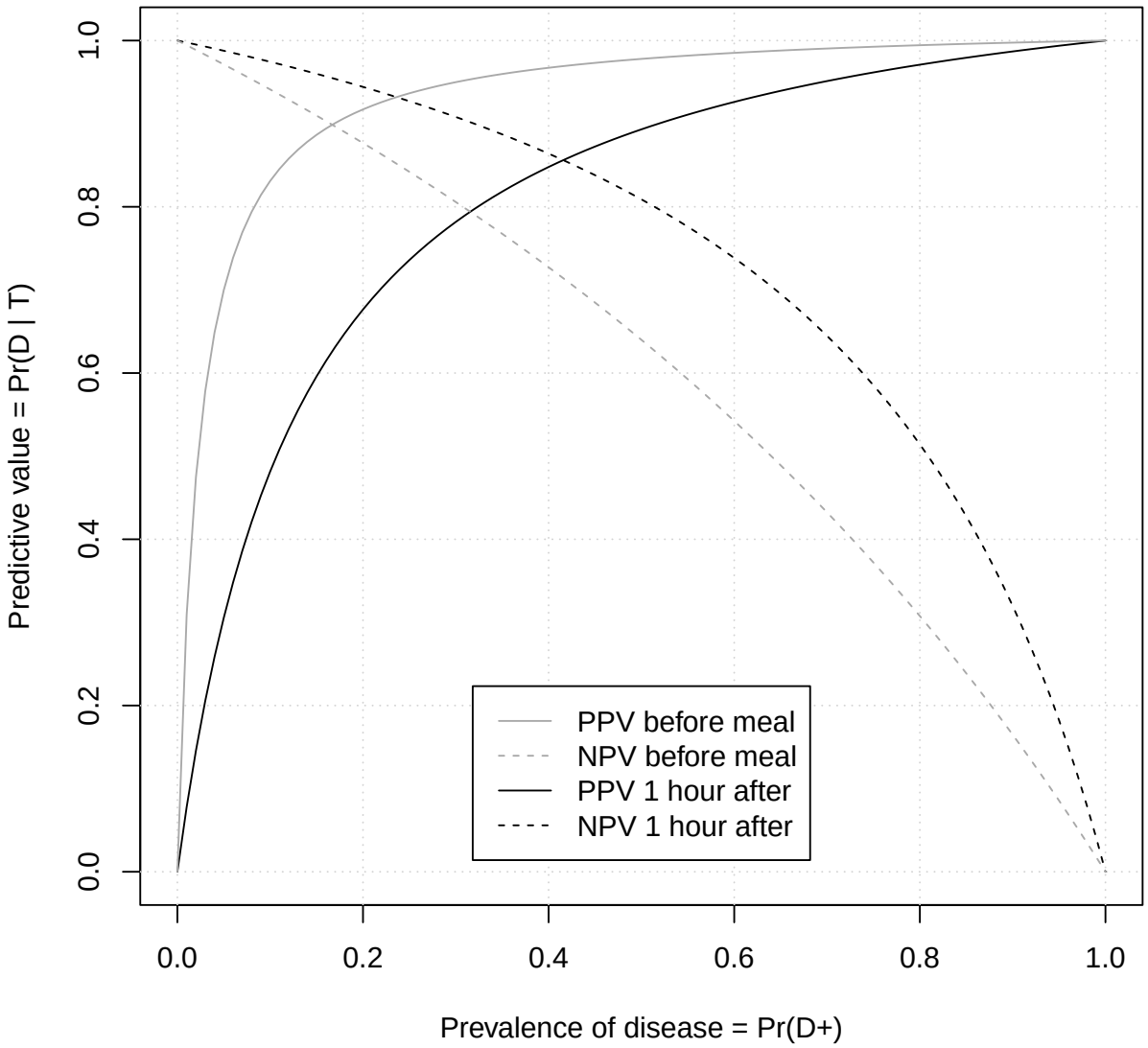

Figure 2.4: Positive and negative predictive values of the Somogyi-Nelson diabetes test before the meal (gray) and one hour after the meal (black) as a function of diabetes prevalence.

2.8.2 Likelihood ratios*

For a probability p, the **odds** is

$$\theta = \frac{p}{1-p}.$$

While a probability lives in $[0, 1]$, the odds can go from zero (for $p = 0$) to infinity (as p approaches one). There is a one-to-one relationship between probabilities and odds, so we can

calculate the probability of an event if we know the odds. If the odds is θ, the corresponding probability is

$$p = \frac{\theta}{1+\theta}.$$

Odds and odds ratios have an important role in epidemiology and statistical inference. In a Bayesian statistical framework, odds ratios give us a simple way to update our knowledge about the probability of an event given new information.

Suppose we know the prevalence of a disease in a population Ω. We randomly sample an individual $\omega \in \Omega$ and give them a diagnostic test. If we randomly sample an individual ω from a population Ω, the odds that ω has disease is

$$\frac{\Pr(D^+)}{1-\Pr(D^+)} = \frac{\Pr(D^+)}{\Pr(D^-)}.$$

where $\Pr(D^+)$ is the prevalence of disease. This is called the **prior odds** of disease. If ω tests positive for the disease, the conditional odds that they have disease is

$$\frac{PPV}{1-PPV} = \frac{\Pr(D^+ \,|\, T^+)}{\Pr(D^- \,|\, T^+)} = \frac{\Pr(D^+ \cap T^+)}{\Pr(D^- \cap T^+)},$$

where we have cancelled out $\Pr(T^+)$ from the numerator and the denominator in the last expression. This is called the **posterior odds** of disease. The second expression above shows that the probability corresponding to the posterior odds is the PPV.

Using the multiplication rule for conditional probabilities, we get

$$\frac{\Pr(D^+ \cap T^+)}{\Pr(D^- \cap T^+)} = \frac{\Pr(T^+ \,|\, D^+)\Pr(D^+)}{\Pr(T^+ \,|\, D^-)\Pr(D^-)} = \frac{\text{sensitivity}}{1-\text{specificity}} \times \frac{\Pr(D^+)}{\Pr(D^-)}.$$

The term sensitivity$/(1-$specificity$)$ is called the **likelihood ratio**. If our individual ω tests positive for disease,

$$\text{posterior odds of } D^+ = \text{likelihood ratio} \times \text{prior odds of } D^+.$$

The likelihood ratio is a measure of how much we learn from a positive test result, and it does not depend on the prevalence of disease [Lusted (1971b); Swets (1973); Fagan (1975); Albert (1982); Zweig and Campbell (1993)}. Because an ROC curve plots sensitivity on the vertical axis and $1-$specificity on the horizontal axis, the likelihood ratio for a given test is the slope of the line from the point $(0, 0)$ to the point representing the test.

Table 2.5 shows the prior odds, likelihood ratio, posterior odds, and PPV for the Somogyi-Nelson blood glucose tests for diabetes from 580 participants (70 with diabetes and 510 without) in Remein and Wilkerson (1961). Note that the tests with the highest likelihood ratios come from the glucose measurements that had the lowest AUCs in Figure 2.2. These tests have high likelihood ratios despite their low sensitivity because they have specificities near one. The test with the best combination of sensitivity and specificity in Table 2.3 has the lowest likelhood ratio. Like other summaries of diagnostic test performance, the likelihood ratio by itself does not determine the best test for a given purpose.

Table 2.5: Prior odds, likelihood ratios, posterior odds, and PPV for the Somogyi-Nelson blood glucose test for diabetes where T^+ corresponds to a concentration above 130 mg/dL.

Test	Prior odds	Likelihood ratio	Posterior odds	PPV
Before meal	$70/510 \approx 0.137$	45.171	$31/5 = 6.200$	$31/36 \approx 0.861$
1 hour after	$70/510 \approx 0.137$	8.348	$55/48 \approx 1.146$	$55/103 \approx 0.534$
2 hours after	$70/510 \approx 0.137$	20.491	$45/16 \approx 2.813$	$45/61 \approx 0.738$
3 hours after	$70/510 \approx 0.137$	247.714	$34/1 = 34.000$	$34/35 \approx 0.971$

Listing 2.3 `ROCcurve.R`

```
# data from Table 2 in Remein and Wilkerson (Journal of Chronic Disease, 1961)
SNdat <- data.frame(cutoff = seq(70, 200, by = 10))
SNdat$sens_pre <- c(95.7, 91.4, 82.9, 65.7, 54.3, 50.0, 44.3, 37.1, 30.0,
                    25.7, 25.7, 22.9, 21.4, 17.1) / 100
SNdat$spec_pre <- c(11.0, 36.3, 65.7, 84.7, 92.7, 96.7, 99.0, 99.6, 99.8,
                    99.8, 99.8, 99.8, 100.0, 100.0) / 100
SNdat$sens_1hr <- c(100.0, 97.1, 97.1, 95.7, 92.9, 88.6, 78.6, 68.6, 57.1,
                    52.9, 47.1, 40.0, 34.3, 28.6) / 100
SNdat$spec_1hr <- c(8.2, 22.4, 39.0, 57.3, 70.6, 83.3, 90.6, 95.1, 97.8,
                    99.4, 99.6, 99.8, 100.0, 100.0) / 100
SNdat$sens_2hr <- c(98.6, 97.1, 94.3, 88.6, 85.7, 71.4, 64.3, 57.1, 50.0,
                    47.1, 42.9, 38.6, 34.3, 27.1) / 100
SNdat$spec_2hr <- c(8.8, 25.5, 47.6, 69.8, 84.1, 92.5, 96.9, 99.4, 99.6,
                    99.8, 100.0, 100.0, 100.0, 100.0) / 100
SNdat$sens_3hr <- c(94.3, 91.4, 82.9, 70.0, 60.0, 51.4, 48.6, 41.4, 32.9,
                    28.6, 28.6, 28.6, 24.3, 20.0) / 100
SNdat$spec_3hr <- c(8.6, 34.7, 67.5, 86.5, 95.3, 98.2, 99.8,
                    rep(100.0, 7)) / 100
# write.csv(SNdat, "SNdat.csv", row.names = FALSE)

# ROC curves with labels
plot(1 - SNdat$spec_pre, SNdat$sens_pre, type = "n",
     xlim = c(0, 1), ylim = c(0, 1),
     xlab = "1 - Specificity = Pr(T+ | D-)",
     ylab = "Sensitivity = Pr(T+ | D+)")
grid()
lines(1 - SNdat$spec_pre, SNdat$sens_pre, col = "darkgray")
lines(1 - SNdat$spec_1hr, SNdat$sens_1hr, lty = "solid")
lines(1 - SNdat$spec_2hr, SNdat$sens_2hr, lty = "dashed")
lines(1 - SNdat$spec_3hr, SNdat$sens_3hr, lty = "dotted")
points(1 - SNdat[SNdat$cutoff == 130, c(3, 5, 7, 9)],
       SNdat[SNdat$cutoff == 130, c(2, 4, 6, 8)])
points(1 - SNdat$spec_pre[seq(2, 12, by = 2)],
       SNdat$sens_pre[seq(2, 12, by = 2)], pch = 8)
text(1 - SNdat$spec_pre[seq(2, 12, by = 2)] + c(0, .09, .09, .1, .1, .1),
     SNdat$sens_pre[seq(2, 12, by = 2)] + c(-.05, -.02, -.02, 0, 0, -.01),
     labels = c("80 mg/dL", "100 mg/dL", "120 mg/dL", "140 mg/dL",
                "160 mg/dL", "180 mg/dL"))
abline(0, 1, lty = "dotted", col = "darkgray")
text(.51, .49, adj = c(.5, 1), srt = 42,
     label = "Useless tests (T and D independent)")
points(c(0, 0, 1), c(0, 1, 1), pch = 3)
text(.01, .99, adj = c(0, 1), label = "Perfect test")
text(.01, .01, adj = c(0, 0), label = "Everyone tests negative")
text(.99, .99, adj = c(1, 0), srt = 90, label = "Everyone tests positive")
legend("bottomright", bg = "white",
       lty = c("solid", "solid", "dashed", "dotted", NA),
       col = c("darkgray", rep("black", 4)), pch = c(rep(NA, 4), 1),
       legend = c("Before meal  (AUC = 0.825)",
                  "1 hour after   (AUC = 0.923)",
                  "2 hours after (AUC = 0.904)",
```

Listing 2.4 `auc.R`

```r
## areas under the ROC curves

# load Somogyi-Nelson test data generated for Figure 2.2 (if needed)
# The argument can contain a path before the file name.
SNdat <- read.csv("SNdat.csv")

auc <- function(x, y) {
  # x is an increasing list of specificities
  # y is a decreasing list of sensitivities
  roc <- approxfun(c(1, 1 - x, 0), c(1, y, 0), ties = "max")
  area <- integrate(function(x) roc(x), 0, 1)
  return(area)
}
auc(SNdat$spec_pre, SNdat$sens_pre)
auc(SNdat$spec_1hr, SNdat$sens_1hr)
auc(SNdat$spec_2hr, SNdat$sens_2hr)
auc(SNdat$spec_3hr, SNdat$sens_3hr)
```

Listing 2.5 `testpos.R`

```
## probability of testing positive as a function of prevalence

# function to generate testing data
tdat <- function(prev, sens=0.786, spec=0.906) {
  # defaults are sensitivity and sensitivity one hour after the meal
  truepos <- sens * prev
  falsepos <- (1 - spec) * (1 - prev)
  trueneg <- spec * (1 - prev)
  falseneg <- (1 - spec) * prev
  pos <- truepos + falsepos
  neg <- 1 - pos
  ppv <- truepos / pos
  npv <- trueneg / neg
  return(data.frame(prev = prev, sens = sens, spec = spec,
                    truepos = truepos, falsepos = falsepos,
                    trueneg = trueneg, falseneg = falseneg,
                    pos = pos, neg = neg, ppv = ppv, npv = npv))
}
tdat_1hr <- tdat(seq(0, 1, by = .01))
write.csv(tdat_1hr, "R/tdat_1hr.csv", row.names = FALSE)

# plot
plot(tdat_1hr$prev, tdat_1hr$pos, type = "n", xlim = c(0, 1), ylim = c(0, 1),
     xlab = "Prevalence of disease = Pr(D+)",
     ylab = "Probability of positive test = Pr(T+)")
polygon(c(tdat_1hr$prev, 1, 0), c(tdat_1hr$pos, 0, 0),
        border = NA, col = "gray")
polygon(c(tdat_1hr$prev, 1, 0), c(tdat_1hr$falsepos, 0, 0),
        border = NA, col = "darkgray")
grid()
lines(tdat_1hr$prev, tdat_1hr$falsepos, col = "gray")
lines(tdat_1hr$prev, tdat_1hr$pos)
abline(0, 1, lty = "dotted")
text(0.1, 0.02, adj = c(0, 0), label = "False positives")
text(0.2, 0.1, adj = c(0, 0), label = "True positives")
```

Listing 2.6 `predval.R`

```
## Predictive values as a function of prevalence

# uses tdat_1hr data and tdat() function from Figure 2.3 (testpos.R)
# tdat_1hr <- read.csv("tdat_1hr.csv")
# generate data using the sensitivity and specificity of the pre-meal test
tdat_pre <- tdat(seq(0, 1, by = .01), sens = 0.443, spec = 0.990)

# plot of PPV and NPV as a function of diabetes prevalence
plot(tdat_1hr$prev, tdat_1hr$ppv, type = "n", xlim = c(0, 1), ylim = c(0, 1),
     xlab = "Prevalence of disease = Pr(D+)",
     ylab = "Predictive value = Pr(D | T)")
grid()
lines(tdat_1hr$prev, tdat_1hr$ppv)
lines(tdat_1hr$prev, tdat_1hr$npv, lty = "dashed")
lines(tdat_pre$prev, tdat_pre$ppv, col = "darkgray")
lines(tdat_pre$prev, tdat_pre$npv, lty = "dashed", col = "darkgray")
legend("bottom", lty = c("solid", "dashed", "solid", "dashed"),
       col = c("darkgray", "darkgray", "black", "black"),
       bg = "white", inset = 0.05,
       legend = c("PPV before meal", "NPV before meal",
                  "PPV 1 hour after", "NPV 1 hour after"))
```

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]

In probability, we are told the rules of the game and then we predict what it will look like. In statistics, we watch the game and try to figure out the rules. Roughly speaking, statistics (game to rules) is the reverse of probability (rules to game). When done well, statistics helps us learn from observations while accounting honestly for uncertainty. An outstanding early example statistics applied to public health is the work of Florence Nightingale (1820-1910), who collected data and developed statistical graphics to demonstrate the need for public health reforms in the British Army in the 1850s (Cohen 1984; Winkelstein Jr 2009).[2]

Here, we will use estimation of a probability as an example of maximum likelihood estimation, which is used for parameter estimation throughout frequentist statistics. It gives us a way to find point estimates of parameters that are optimal in large samples in a sense that we will explain below. It is also the foundation for hypothesis tests and confidence intervals, which give us an accurate way to account for uncertainty in statistical inference.

3.1 Binomial likelihood

In Section 3.1.1, we used the prevalence p in our population to figure out the distribution of the number X of diseased individuals in a sample of size n. This is probability. The corresponding statistical problem would be to estimate the prevalence p after seeing $X = x$ infected individuals in a sample of size n.

When our experiment is to sample multiple individuals from a population, the analogy between the outcomes $\omega \in \Omega$ and the individuals in the population breaks down. Recall that when we flip a coin twice, each $\omega \in \Omega$ must specify the outcomes of both flips. When the experiment is to sample n individuals from a population, the entire sample is a single outcome ω and Ω

[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".

[2] She was elected a member of the Royal Statistical Society in 1859, where she was the first woman. In 1860, she founded the world's first modern nursing school at St. Thomas Hospital in London.

contains all possible samples of n individuals from the population. If the population size is N, then the number of possible samples of size n is given by the *binomial coefficient*

$$\binom{N}{n} = \frac{N!}{n!(N-n)!},$$

where $k!$ denotes k factorial. Factorials are defined by $0! = 1$ and $k! = k \cdot (k-1)!$ for any integer $k > 0$. For example, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$, and so on. For $k > 0$, $k!$ is the product of all positive integers up to and including k, which grows extremely fast as k increases.

3.1.1 Binomial distribution

Suppose we sample n individuals from a population Ω and test them for disease. For simplicity, we assume that the diagnostic test has perfect sensitivity and specificity. Let Y_i denote whether person i in the sample has disease, and let X be the total number who have disease. Then

$$X = \sum_{i=1}^{n} Y_i,$$

so it is a linear combination of the Y_i. Each Y_i is a Bernoulli(p) random variable, where p is the prevalence of disease in the population. When N is much larger than n (for which we write $N \gg n$), the test results for each person in the sample are approximately independent.

The distribution of a sum of n independent Bernoulli(p) random variables is called a **binomial(n, p) distribution**.[3] The probability $Y_1 = 1$ is p, and the probability that $Y_1 = 0$ is $(1-p)$, so we can handle both cases by writing

$$\Pr(Y_1 = y_1) = p^{y_1}(1-p)^{1-y_1}.$$

When the Y_i are independent, each Y_i has a Bernoulli(p) distribution (see Section 1.5.2) and

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \Pr(Y_i = y_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

by the multiplication rule for independent events. Substituting $x = \sum_{i=1}^{n} y_i$, we get

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = p^x(1-p)^{n-x}.$$

[3] For finite N, X actually has a *hypergeometric distribution* because the test results are not exactly independent. If the first person in our sample has disease, the probability that the next person we sample has disease is slightly less than p. If the first person in our sample does not have disease, the probability that the next person we sample has disease is slightly greater than p. When $N \gg n$, this hypergeometric distribution is approximately binomial(n, p).

The value of x depends only on the sum of the y_i, and there are $\binom{n}{x}$ different ways to get x cases of disease out of n sampled individuals. By the addition rule for disjoint events, we get

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^x. \tag{3.1}$$

This is the probability mass function (PMF) of the binomial distribution. The set of possible values of a binomial(n, p) random variable X is $\text{supp}(X) = \{0, 1, \ldots, n\}$.

Section 1.5.2 showed that a Bernoulli(p) random variable has expected value p and variance $p(1-p)$. Because a binomial(n, p) random variable is the sum of n independent Bernoulli(p) random variables, its expected value is

$$\mathbb{E}(X) = np.$$

by the rule for expectations of linear combinations in Equation 1.23. Its variance is

$$\text{Var}(X) = np(1-p)$$

by the rule for variances of linear combinations in Equation 1.24. The covariances are all zero because the Y_i are independent.

3.2 R

3.2.1 Likelihood and log likelihood

In probability, we know the prevalence of disease p and we deduce the distribution of the number of diseased individuals X in a sample of size n. In statistics, we observe $X = x$ and use this to estimate p. To do this, we rewrite the binomial PMF Equation 3.1 as a function of p instead of x:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3.2}$$

This is the binomial **likelihood function**. The right-hand sides of Equation 3.1 and Equation 3.2 are identical, and they produce exactly the same value given the same x and p. However, the two equations define different functions. In binomial PMF in Equation 3.1, the prevalence p is fixed and the number of diseased individuals x is the argument of the function. In the binomial likelihood function in Equation 3.2, the number of diseased individuals x is fixed and the prevalence p is the argument of the function. The PMF belongs to probability, and the likelihood belongs to statistics.

Listing 3.1 `binomdist.R`

```
## binomial distribution

# binomial PMF
# The second and third arguments are n ("size") and p ("prob").
dbinom(2, 10, 0.4)
dbinom(0:10, 10, 0.4)
sum(dbinom(0:10, 10, 0.4))

# binomial CDF
pbinom(0:10, 10, 0.4)
cumsum(dbinom(0:10, 10, 0.4))

# binomial quantiles
qbinom(c(0.25, 0.5, 0.75, 1), 10, 0.4)

# random samples
rbinom(20, 10, 0.4)
x <- rbinom(1000, 10, 0.4)
mean(x)
var(x)
```

The **log likelihood** is the natural logarithm (i.e., the logarithm to base $e = 2.718281828\ldots$)[4] of the likelihood function. For binomial log likelihood is

$$\ell(p) = \ln \binom{n}{x} + x \ln p + (n - x) \ln(1 - p).$$

Because the logarithm turns products into sums, it is generally much easier to handle the log likelihood than the likelihood itself. The term $\ln \binom{n}{x}$ does not depend on p, so it can be ignored. Intuitively, this tells us that the total number $x = y_1 + y_2 + \cdots + y_n$ of individuals with disease in our sample contains the same information about the prevalence of disease as the sequence $y_1, y_2, \ldots, y_n$ of disease indicators.

For any given p, we can think of $\ell(p)$ as a random variable whose value is determined by our

[4] Euler's number e is named after Leonhard Euler (1707–1783), a Swiss mathematician who introduced the notation $f(x)$ for mathematical functions and the letter i to denote the imaginary unit $\sqrt{-1}$. He spent most of his life in Berlin and St. Petersburg, and he is widely considered the greatest mathematician of the 18th century. The number e was first discovered in 1683 by Jacob Bernoulli (the namesake of the Bernoulli distribution) when studying compound interest, where $e = \lim_{n\to\infty}(1 + 1/n)^n$. In 1748, Euler proved that $e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots$.

sample of size n. Let p_{true} be the true prevalence of disease. By *Gibb's inequality*,[5]

$$\mathbb{E}[\ell(p_{\text{true}})] > \mathbb{E}[\ell(p)]$$

for all $p \neq p_{\text{true}}$. This inequality is about the expected value of the log likelihood over all possible samples of size n. For any given sample, it is possible that $\ell(p_{\text{true}})$ is not the maximum of the log likelihood. However, this inequality is an important part of the justification for estimating p by maximizing the log likelihood (Boos and Stefanski 2013). Because function $v \mapsto \ln(v)$ is strictly increasing in v, the likelihood $L(p)$ and the log likelihood $\ell(p)$ are maximized at exactly the same value of p.

3.2.2 Score function

To find the maximum of the log likelihood, we find the value of p where its slope is zero. This is the mathematical version of the insight that the ground at the top of a hill is level. The **score function** is the first derivative of the log likelihood

$$U(p) = \frac{\mathrm{d}}{\mathrm{d}p}\ell(p) = \frac{x}{p} - \frac{n-x}{1-p},$$

which is the slope of $\ell(p)$ at p. To find where the slope equals zero, we solve the *score equation*

$$U(\hat{p}) = \frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0 \tag{3.3}$$

where $\hat{p}$ denotes the maximum likelihood estimate (MLE) of p_{true}. When the dust settles, we get

$$\hat{p} = \frac{x}{n}$$

so our MLE of the prevalence is just the proportion of our sample who has disease.

To confirm that this is a maximum instead of a minimum, we need to look at the second derivative of ℓ. When we walk across the top of a hill, we go from walking uphill to walking downhill so the slope is decreasing. If $\ell(p)$ is maximized at $\hat{p}$, then the slope of the slope (i.e., the second derivative) should be negative. The second derivative of $\ell(p)$ at $\hat{p}$ is

$$\frac{\mathrm{d}}{\mathrm{d}p}U(p) = \frac{\mathrm{d}^2}{\mathrm{d}p^2}\ell(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}. \tag{3.4}$$

This is negative for any $p \in (0,1)$. Thus, the log likelihood is maximized at $\hat{p}$ if $x > 0$ and $x < n$.

When $x = 0$ or $x = n$, the log likelihood $\ell(p)$ has no maximum at any $p \in (0,1)$. Instead, the maximum occurs at one of the boundaries of the set of possible p. When $x = 0$, our MLE of p_{true} is $\hat{p} = 0$. When $x = n$, our maximum likelihood estimate is $\hat{p} = 1$.

[5]This is named for Josiah Willard Gibbs (1839–1903), an American scientist who earned the first American doctorate in engineering in 1863 and went on to work on statistical mechanics, thermodynamics, optics, and vector calculus as a professor of physics at Yale. Albert Einstein called him the greatest mind in American history.

3.2.3 Expected and observed information*

For any given p, we can think of the score $U(p)$ as a random variable that has an expected value and a variance. If $p_{\text{true}} = p$, the expected value of the score is always zero:

$$\mathbb{E}_p[U(p)] = \mathbb{E}_p\left[\frac{X}{p} - \frac{n-X}{1-p}\right] = \frac{\mathbb{E}_p(X)}{p} - \frac{\mathbb{E}_p(n-X)}{1-p} = \frac{np}{p} - \frac{n(1-p)}{1-p} = 0$$

where we use the subscript p to indicate that the expected value is calculated assuming that $p_{\text{true}} = p$. Because $\mathbb{E}_p[U(p)] = 0$, the corresponding variance of the score is

$$\mathcal{J}(p) = \text{Var}_p[U(p)] = \mathbb{E}_p[U(p)^2],$$

by Equation 1.22. This is called the **expected Fisher information** or **expected information**.[6] It can be used to calculate confidence limits for p_{true}.

Under *regularity conditions* that are met when $p_{\text{true}} \in (0, 1)$, the Fisher information $\mathcal{J}(p)$ can be calculated using the second derivative of the log likelihood $\ell(p)$ from Equation 3.4.[7] Specifically, $\mathcal{J}(p)$ is the expected value of the negative second derivative of $\ell(p)$:

$$\mathcal{J}(p) = \mathbb{E}_p\left[-\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ell(p)\right] = \mathbb{E}_p\left[\frac{X}{p^2} + \frac{n-X}{(1-p)^2}\right], \tag{3.5}$$

where the subscript p indicates that the expected value is calculated assuming that $p_{\text{true}} = p$. Using Equation 1.23 and the binomial(n, p) distribution for X, this simplifies to

$$\mathcal{J}(p) = \frac{\mathbb{E}(X)}{p^2} + \frac{\mathbb{E}(n-X)}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.$$

Because p_{true} is unknown, the expected information is often evaluated at $\hat{p}$. In some models, the expected information can be difficult to calculate.

The negative second derivative of $\ell(p)$ inside the expectation in Equation 3.5 evaluated is the **observed Fisher information** or **observed information**

$$I(p) = -\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ell(p) = \frac{x}{p^2} + \frac{n-x}{(1-p)^2}. \tag{3.6}$$

[6] Named after Ronald Fisher (1890–1962), who established the foundations of maximum likelihood inference between 1912 and 1922. He was the most important statistician of the 20th century, and he was one of the founders of population genetics. He had poor eyesight for his entire life, which led him to develop a formidable sense of geometry in his head. However, he was also a leading eugenicist and one of the most vocal opponents of the hypothesis that smoking causes lung cancer.

[7] For estimating a parameter θ, the conditions are these: (1) The set of possible values of the observed data X does not depend on θ. (2) Each θ produces a different distribution of X. (3) The true value of θ is in the interior of the set of possible values. (4) The log likelihood $\ell(\theta)$ has continuous first and second derivatives with respect to θ in a neighborhood of θ_{true}. These conditions are met by the binomial likelihood when $p_{\text{true}} \in (0, 1)$.

For the binomial distribution $I(\hat{p}) = \mathcal{J}(\hat{p})$ but this equality does not hold at other values of p. The observed information is an unbiased estimator of the expected information, and it can always be calculated from the data. It often produces more accurate variance estimates than the expected information (Efron and Hinkley 1978; Kenward and Molenberghs 1998; Reid 2003). However, it is generally safe to use whichever is most convenient (Boos and Stefanski 2013).

3.3 Large-sample theory

The log likelihood, the score function, and the Fisher and observed information give us all of the pieces we need to calculate point and interval estimates of p_{true}. To put them together, we use two fundamental results from probability theory about the behavior of sample means. The law of large numbers justifies point estimates and the central limit theorem justifies hypothesis tests and interval estimates, which can be obtained in three standard ways.

3.3.1 Sample mean (average)

If $Y_1, Y_2, \ldots, Y_n$ are random variables, then the **sample mean** or **average** is

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

This sample mean can be thought of as a random variable whose value is determined when we observe $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$. If each Y_i has $\mathbb{E}(Y_i) = \mu$, then

$$\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{1}{n} n\mu = \mu \tag{3.7}$$

by Equation 1.23. Thus, the sample mean $\hat{\mu}_n$ is an **unbiased** estimate of μ for any sample size n. When the Y_i are indicator variables, $\hat{\mu}_n$ is just the proportion of the sample with $Y_i = 1$.

3.3.2 Law of large numbers and consistency

If the Y_i are independent and each has $\text{Var}(Y_i) = \sigma^2$, then

$$\text{Var}(\hat{\mu}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{3.8}$$

by Equation 1.24. Thus, the variance of $\hat{\mu}_n$ decreases as the sample size n increases. The standard deviation of $\hat{\mu}_n$ is proportional to $1/\sqrt{n}$. As $n \to \infty$, we should have $\hat{\mu}_n \to \mu$. This is called the **law of large numbers**, and it holds even when $\sigma^2 = \infty$.

Theorem 3.1 (Law of Large Numbers)**.** *If $Y_1, Y_2, \ldots$ is an infinite sequence of independent and identically-distributed (IID) random variables with mean $\mu < \infty$ and variance $\sigma^2 \leq \infty$, then*

$$\hat{\mu}_n \to \mu$$

as $n \to \infty$.[8]

Our maximum likelihood estimate $\hat{p}_n$ is a sample mean:

$$\hat{p}_n = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

where each $Y_i \sim \text{Bernoulli}(p_{\text{true}})$ and the Y_i are independent. Therefore, the LLN implies that

$$\hat{p}_n \to p_{\text{true}}$$

as $n \to \infty$. This convergence is shown in Figure 3.1. An estimate that converges to its true value as $n \to \infty$ is called **consistent**. Intuitively, this means that $\hat{p}_n$ is guaranteed to be close to p_{true} in a large sample. However, the LLN does not specify how close or how large a sample we need.

Listing 3.2 `lln.R`

```
## Law of large numbers

n <- 1000
x <- seq(n)
plot(x, cumsum(rbinom(n, 1, .5)) / x, type = "n", ylim = c(0, 1),
     xlab = "Number of samples", ylab = "Sample mean")
grid()
lines(x, cumsum(rbinom(n, 1, .5)) / x, lty = "solid")
lines(x, cumsum(rbinom(n, 1, .5)) / x, lty = "dashed")
lines(x, cumsum(rbinom(n, 1, .5)) / x, lty = "dotted")
abline(h = .5)
```

[8] For simplicity, we are being vague about what we mean by $\hat{\mu}_n \to \mu$. Probability has several different notions of convergence/. The *weak* LLN guarantees convergence *in probability*, which means that $\lim_{n\to\infty} \Pr(|\hat{\mu}_n - \mu| > \varepsilon) = 0$ for any $\varepsilon > 0$. The *strong* LLN guarantees convergence *almost surely*, which means that $\Pr(\lim_{n\to\infty} \hat{\mu}_n = \mu) = 1$.

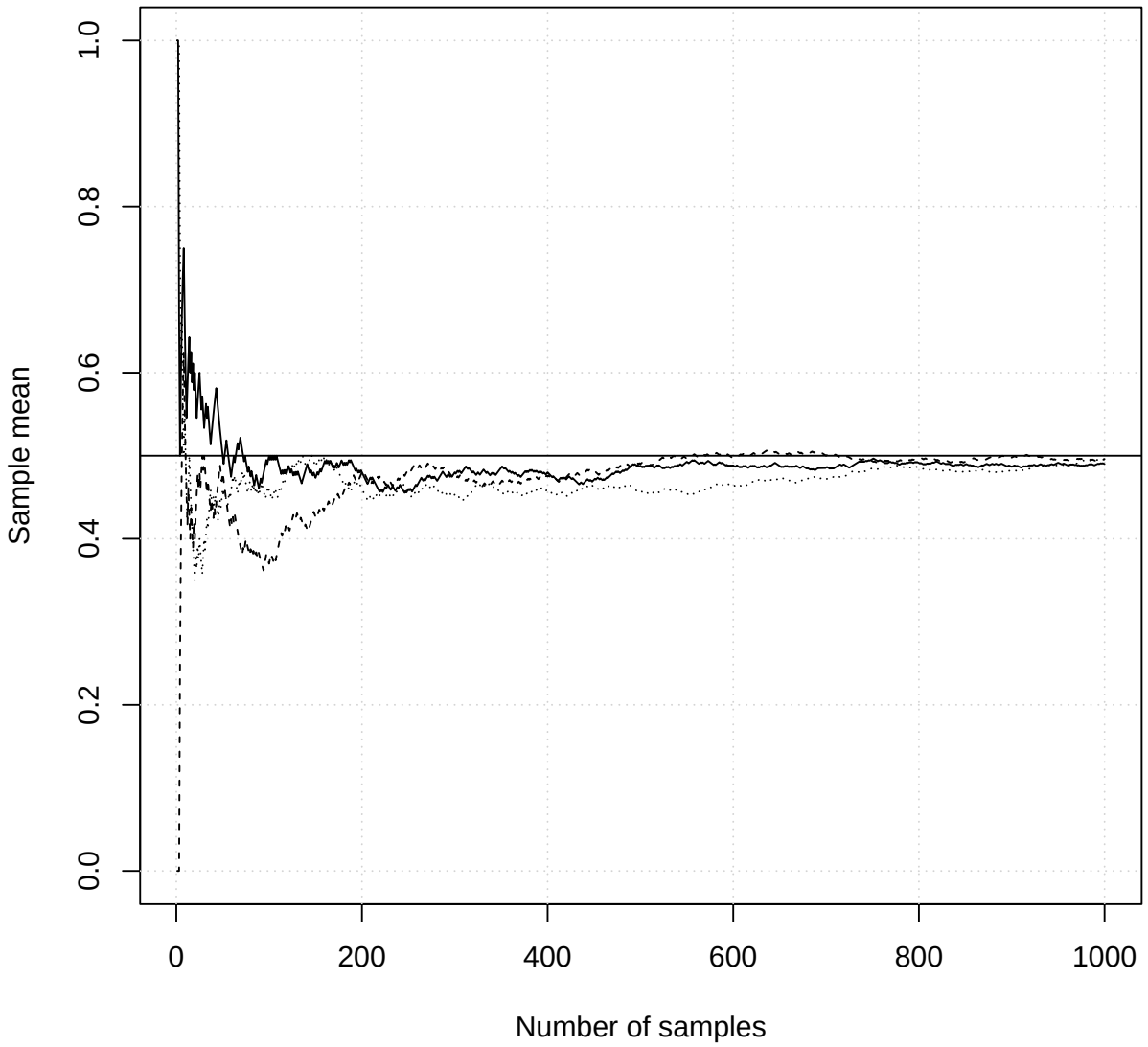

Figure 3.1: The LLN at work. Each line traces the sample means calculated from a sequence of random samples $x_1, x_2, x_3, ...$ from a Bernoulli(0.5) distribution. For each sequence, the y-coordinate above n is the sample mean from the first n random samples in the sequence. The true mean of 0.5 is marked by a solid horizontal line.

3.3.3 Central limit theorem and the normal distribution

When both the mean and variance of the Y_i are finite, the **central limit theorem** (CLT) allows us to say something about how far away our sample mean $\widehat{\mu}_n$ is from the true value μ. It is the most important result in all of probability and statistics.

Theorem 3.2 (Central Limit Theorem)**.** *If $Y_1, Y_2, \ldots$ is an infinite sequence of IID random variables with finite mean μ and variance $\sigma^2 < \infty$, then*

$$Z_n = \frac{\hat{\mu}_n - \mathbb{E}(\hat{\mu_n})}{\sqrt{\mathrm{Var}(\hat{\mu}_n)}} = \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{\sigma^2}}$$

has a distribution that converges to a ***normal distribution*** *or* ***Gaussian distribution*** *with mean zero and variance one as $n \to \infty$.*[9] *Because of this, we say that $\hat{\mu}_n$ is* ***asymptotically normal****.*

The normal distribution is a distribution for a **continuous random variable**, which can take any value on an interval or even on all of $\mathbb{R}$. Instead of a PMF, a continuous random variable Z has a **probability density function** (PDF). If Z is a continuous random variable with PDF $f(z)$ and $[a, b]$ is an interval, then

$$\Pr(Z \in [a, b]) = \int_a^b f(z)\,\mathrm{d}z.$$

The integral on the right-hand side represents the area under $f(z)$ over the interval $[a, b]$. The cumulative distribution function of Z is

$$F(z) = \int_{-\infty}^{z} f(u)\,\mathrm{d}u,$$

where the integral on the right-hand side represents the area under $f(z)$ over the interval $(-\infty, u]$. For the same reason that the values of the PMF for any discrete random variable add up to one, we have

$$\Pr(Z \in \mathbb{R}) = \int_{-\infty}^{\infty} f(z)\,\mathrm{d}z = 1$$

for any continuous random variable Z. Like the PMF and CDF of a discrete random variable, the PDF and CDF of a continuous random variable contain the same information about the distribution of Z.

The PDF of the normal distribution with mean μ and variance σ^2 is

$$f(z, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}.$$

The **standard normal distribution** has $\mu = 0$ and $\sigma^2 = 1$. It is such an important distribution that its PDF and CDF have special notation. The standard normal PDF is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

and its CDF is $\Phi(z)$. These functions and the relationship between them are illustrated in Figure 3.2. A normal distribution is denoted $N(\mu, \sigma^2)$, so the standard normal distribution is written $N(0, 1)$.

[9]Named after Carl Friedrich Gauss (1777-1855), a German mathematician who is widely considered one of the greatest mathematicians of all time. He discovered the normal distribution in 1809, but the CLT itself was first proved by Laplace in 1810 (see Chapter 1).

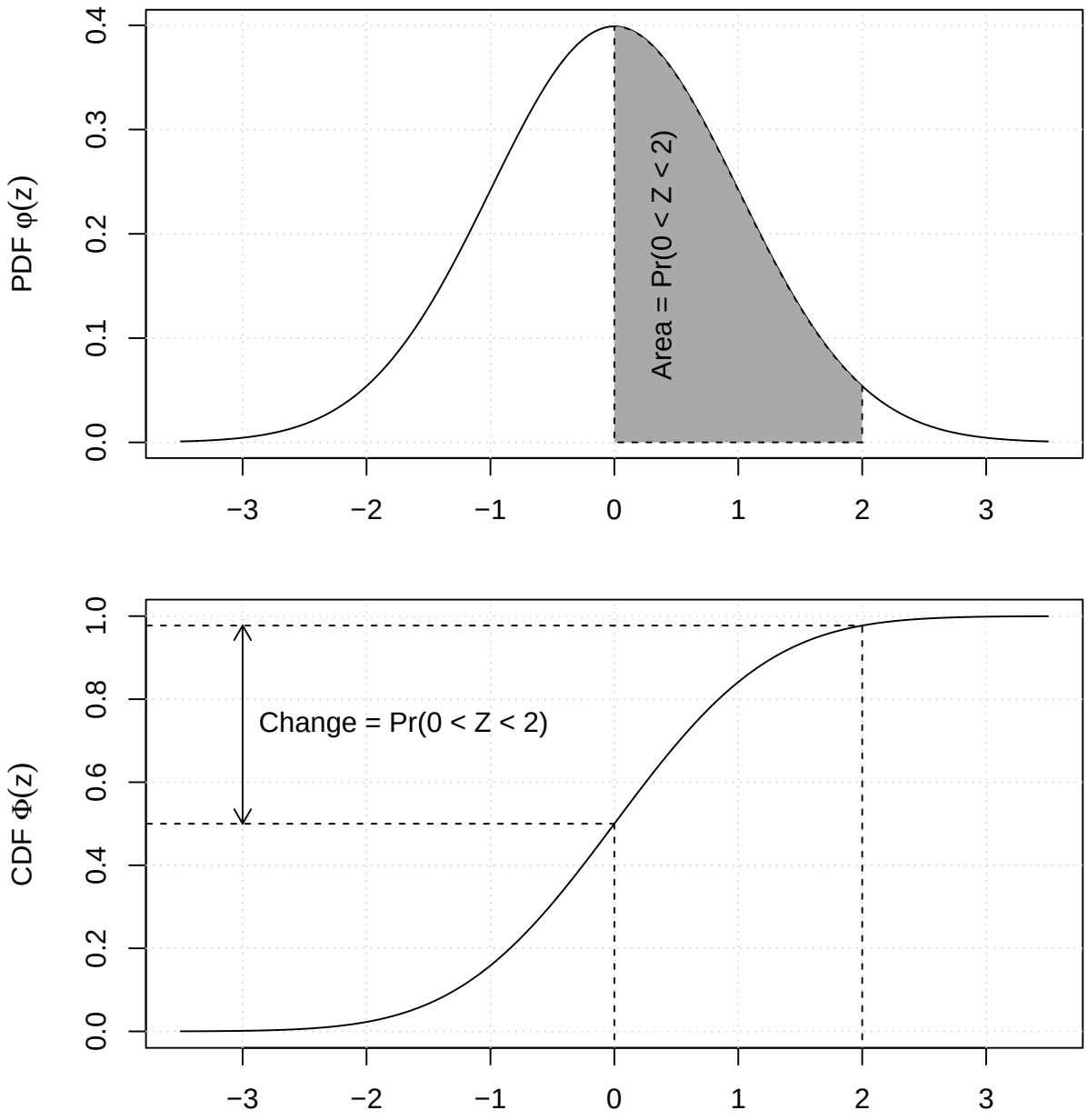

Figure 3.2: The PDF (top) and CDF (bottom) of a standard normal random variable Z. If $X \sim N(0,1)$, then $\Pr(0 < X < 2)$ equals the shaded area under the PDF as well as the change in the CDF from 0 to 2. This same relationship between the CDF and the PDF holds for all continuous random variables and any interval (a, b).

3.4 R

For our estimated probability $\hat{p}_n$ is a sample mean of IID Y_i with $\mathbb{E}(Y_i) = p_{\text{true}}$ and $\text{Var}(Y_i) = p_{\text{true}}(1 - p_{\text{true}})$. When n is large,

$$Z_n = \frac{\sqrt{n}(\hat{p}_n - p_{\text{true}})}{\sqrt{p_{\text{true}}(1 - p_{\text{true}})}} = \frac{\hat{p}_n - p_{\text{true}}}{\sqrt{\mathcal{J}(p_{\text{true}})^{-1}}} \tag{3.9}$$

has a distribution that is close to a standard normal distribution. Figure 3.3 shows this convergence is shown for sample means where $Y_i \sim \text{Bernoulli}(0.1)$. The CLT does not guarantee that the distribution of Z_n is approximately normal in any given sample. It only guarantees that the normal approximation holds eventually as n increases. When the $Y_i \sim \text{Bernoulli}(p)$, the normal approximation is typically good when $np(1 - p) > 5$.

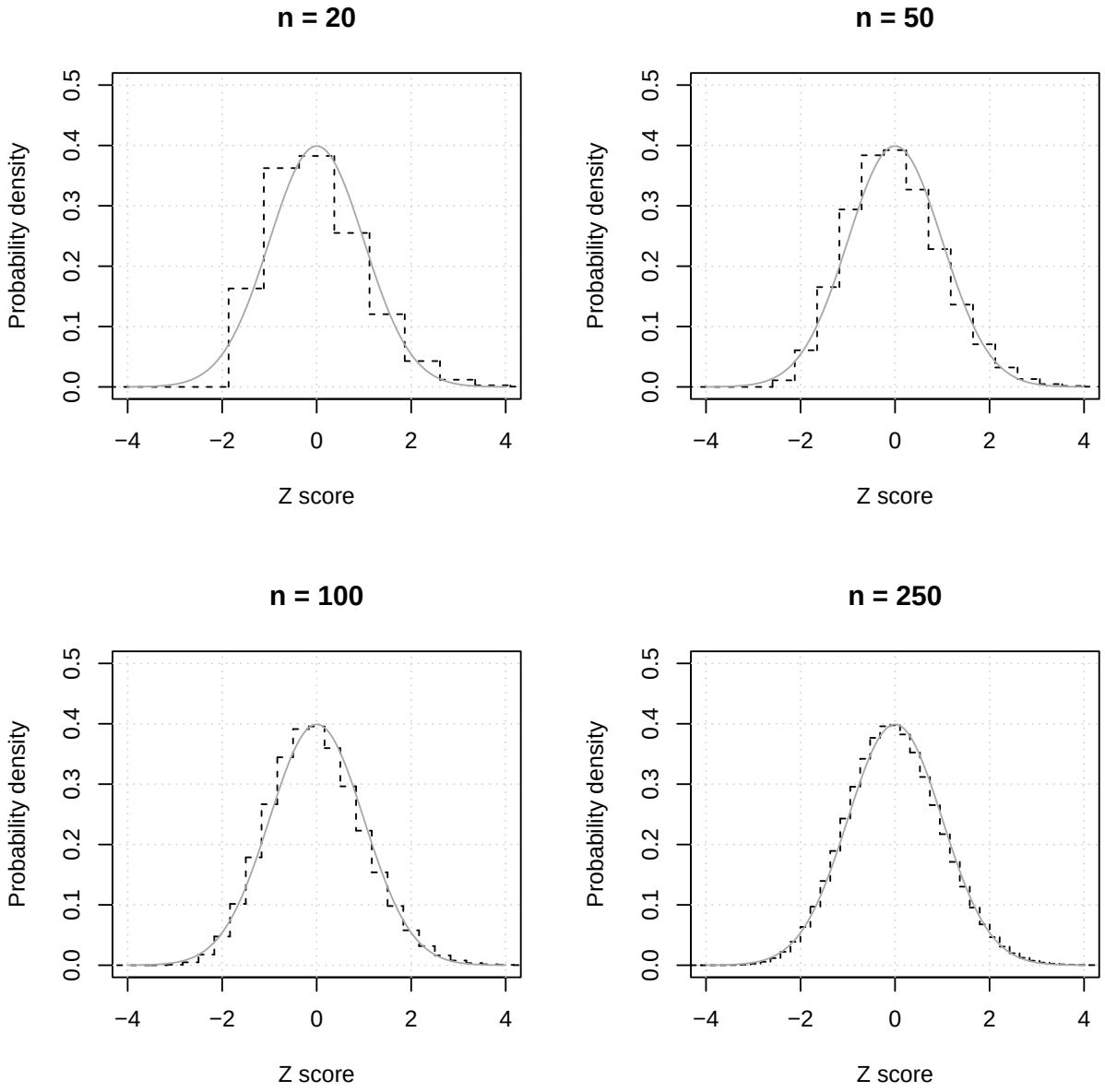

Figure 3.3: The CLT at work. The dashed lines show the PMF of the distribution of the average from a sample of size n from a Bernoulli(0.1) distribution. The solid line is the standard normal PDF.

3.4.1 Efficiency of maximum likelihood estimators*

We have used the LLN and the CLT to show that $\hat{p}_n$ is consistent and asymptotically normal, which are both wonderful properties for an estimator to have. However, they do not prove

that $\hat{p}_n$ is the best estimator of p_{true} in any particular sense. In Equation 3.9, the variance of $\hat{p}_n$ was

$$\mathcal{I}(p_{\text{true}})^{-1} = \frac{p_{\text{true}}(1 - p_{\text{true}})}{n},$$

which is the inverse of the Fisher information. It turns out that no other unbiased estimator of p_{true} can have lower variance, so $\hat{p}_n$ is the minimum-variance unbiased estimator of p_{true}.

Suppose θ is a parameter for a family of PMFs or PDFs $f(y, \theta)$ such that the true PMF or PDF is $f(y, \theta_{\text{true}})$. When we observe $Y_1 = y_1, Y_2 = y_2, \ldots Y_n = y_n$, the likelihood is

$$L(\theta) = \prod_{i=1} f(y_i, \theta),$$

and the log likelihood is

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i, \theta).$$

The score function is

$$U(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta),$$

and the MLE is the solution of the score equation $U(\hat{\theta}) = 0$. The Fisher information is

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta) \right],$$

and $\text{Var}(\hat{\theta}) = \mathcal{I}(\theta)^{-1}$. If $\bar{\theta}$ is any unbiased estimator of the true value θ_{true}, then

$$\text{Var}(\bar{\theta}) \geq \mathcal{I}(\theta_{\text{true}})^{-1}.$$

This result is called the *Cramér-Rao lower bound* (Rao 1945; Cramér 1946),[10] No unbiased estimator of θ_{true} can have smaller variance than the MLE $\hat{\theta}$. Maximum likelihood estimates are consistent, asymptotically normal, and asymptotically efficient when the likelihood is correct (Boos and Stefanski 2013).

3.5 Hypothesis testing

In a **hypothesis test**, we specify a **null hypothesis** and then decide whether to reject it based on the value of a **test statistic**. A null hypothesis often takes the form

$$H_0 : \theta_{\text{true}} = \theta_0. \tag{3.10}$$

We reject H_0 if the test statistic appears inconsistent with its distribution under H_0. Otherwise, we *fail to reject* H_0. It is traditional to avoid saying that H_0 was accepted.

[10] Named after Swedish statistician Harald Cramér (1893–1985), who was a professor at Stockholm University, and Indian-American statistician Calyampudi Radhakrishna (C. R.) Rao (1920–2023), who was a professor at the Indian Statistical Institute, the University of Cambridge, the University of Pittsburgh, and Pennsylvania State University.

Table 3.1: Truth of H_0 and hypothesis test results.

	Reject H_0 (T^+)	Fail to reject H_0 (T^-)
H_0 false (D^+)	True positive	False negative = type II error
H_0 true (D^-)	False positive = type I error	True negative

Table 3.2: Truth of H_0 and hypothesis test results

3.5.1 Hypothesis tests and diagnostic tests

If we think of H_0 as not having the disease and rejecting H_0 as testing positive for the disease, a hypothesis test is analogous to a diagnostic test. Table 3.1 shows the possible outcomes of a hypothesis test, and its margins show the correspondence to diagnostic testing (Diamond and Forrester 1983). A false positive occurs when we reject H_0 when it is true, which is called a **type I error**. A false negative occurs when we fail to reject H_0 when it is false, which is called **type II error**.

A hypothesis test has analogs of sensitivity and specificity. The equivalent of specificity is $1-\alpha$ where

$$\alpha = \Pr(\text{reject } H_0 \,|\, H_0 \text{ true})$$

is the probability of a type I error. This is also called the **significance level** of the test. The equivalent of sensitivity is the **power** of the test, which is $1-\beta$ where

$$\beta = \Pr(\text{fail to reject } H_0 \,|\, H_0 \text{ false})$$

is the probability of a type II error.

A hypothesis test also has analogs of positive and negative predictive values (PPV and NPV). Just like the PPV and NPV of a dignostic test depend on the prevalence of disease, the PPV and NPV of a hypothesis test depend on the **prior probability** that H_0 is true, which is the probability that H_0 is true based on what we know before we see the test result. For a hypothesis test, the PPV is

$$\Pr(H_0 \text{ false} \,|\, H_0 \text{ rejected}) = \frac{(1-\beta)\Pr(H_0 \text{ false})}{(1-\beta)\Pr(H_0 \text{ false}) + \alpha \Pr(H_0 \text{ true})} \tag{3.11}$$

by Bayes' rule. Similarly, the NPV of the hypothesis test is

$$\Pr(H_0 \text{ true} \,|\, H_0 \text{ not rejected}) = \frac{(1-\alpha)\Pr(H_0 \text{ true})}{(1-\alpha)\Pr(H_0 \text{ true}) + \beta \Pr(H_0 \text{ false})}. \tag{3.12}$$

The conditional probability that H_0 is true given the result of the hypothesis test is called the **posterior probability** of H_0.

3.5.2 Wald, score, and likelihood ratio tests

In a maximum likelihood framework, there are three classical tests for a null hypothesis of the form

$$H_0 : p_{\text{true}} = p_0.$$

These tests are asymptotically equivalent, which means that they produce similar results in large samples. The best way to visualize the different tests is to look at a graph of the log likelihood function. Figure 3.4 shows the log likelihood function for a binary outcome with $x = 60$ events out of $n = 100$ trials and a null hypothesis $H_0 : p_{\text{true}} = 0.5$. All three tests generalize to null hypotheses involving multiple parameters (Boos and Stefanski 2013).

The **Wald test** (Wald 1943) of H_0 looks at the distance between the MLE $\hat{p}$ and the hypothesized value p_0(Wald 1943), rejecting H_0 when this distance is sufficiently large.[11] An example is shown in Figure 3.4. The Wald test statistic is

$$W = \frac{(\hat{p} - p_0)^2}{I(\hat{p})} = \frac{n(\hat{p} - p_0)^2}{\hat{p}(1 - \hat{p})} \overset{\text{approx}}{\sim} \chi^2_1 \tag{3.13}$$

under H_0, where $I(\hat{p})$ is the observed information from Equation 3.6. The χ^2_1 distribution is the distribution of Z^2 if $Z \sim N(0, 1)$.

The **score test** looks at the slope of the log likelihood at p_0, rejecting H_0 if this slope is sufficiently far from zero (Rao 1948; Aitchison and Silvey 1958). An example is shown in Figure 3.4. It score test statistic is

$$S = \frac{U(p_0)^2}{\mathcal{I}(p_0)} = \frac{n(\hat{p} - p_0)^2}{p_0(1 - p_0)} \overset{\text{approx}}{\sim} \chi^2_1 \tag{3.14}$$

under H_0, where $\mathcal{I}(p_0)$ is the expected information from Equation 3.5. The numerator of the score statistic is the same as for the Wald statistic in Equation 3.13, but the denominator uses the expected information at p_0 instead of the observed information at $\hat{p}$. In score tests, it generally better to use the expected information than the observed information (Freedman 2007). The most important advantage of the score test is that it only needs the hypothesized null value p_0, so it can be done without finding the maximum likelihood estimate $\hat{p}$.

The **likelihood ratio test** looks at the vertical distance between $\ell(\hat{p})$ (which is the maximum) and $\ell(p_0)$, rejecting H_0 if this distance is sufficiently large Wilks (1938).[12] An example is shown in Figure 3.4. The likelihood ratio test statistic is

$$L = 2(\ell(\hat{p}) - \ell(p_0)) \overset{\text{approx}}{\sim} \chi^2_1 \tag{3.15}$$

[11] Named after Abraham Wald (1902–1950), a Jewish Hungarian mathematician who was invited to move from Vienna to the United States in 1938 after Nazi Germany annexed Austria. He worked at the Statistical Research Group at Columbia University during World War II. In 1950, he and his wife were killed in a plane crash in India, where he was visiting the Indian Statistical Institute.

[12] Samuel S. Wilks (1906–1964) was an American mathematician and statistician who grew up on a farm in Texas, got a Ph.D. at the University of Iowa, and went on to be a professor at Princeton University.

under H_0. The *Neyman-Pearson lemma* (Neyman and Pearson 1933) shows that the likelihood ratio test is the most powerful of all hypothesis test for comparing two hypotheses $H_0 : p_{\text{true}} = p_0$ and $H_1 : p_{\text{true}} = p_1$ at a fixed significance level.

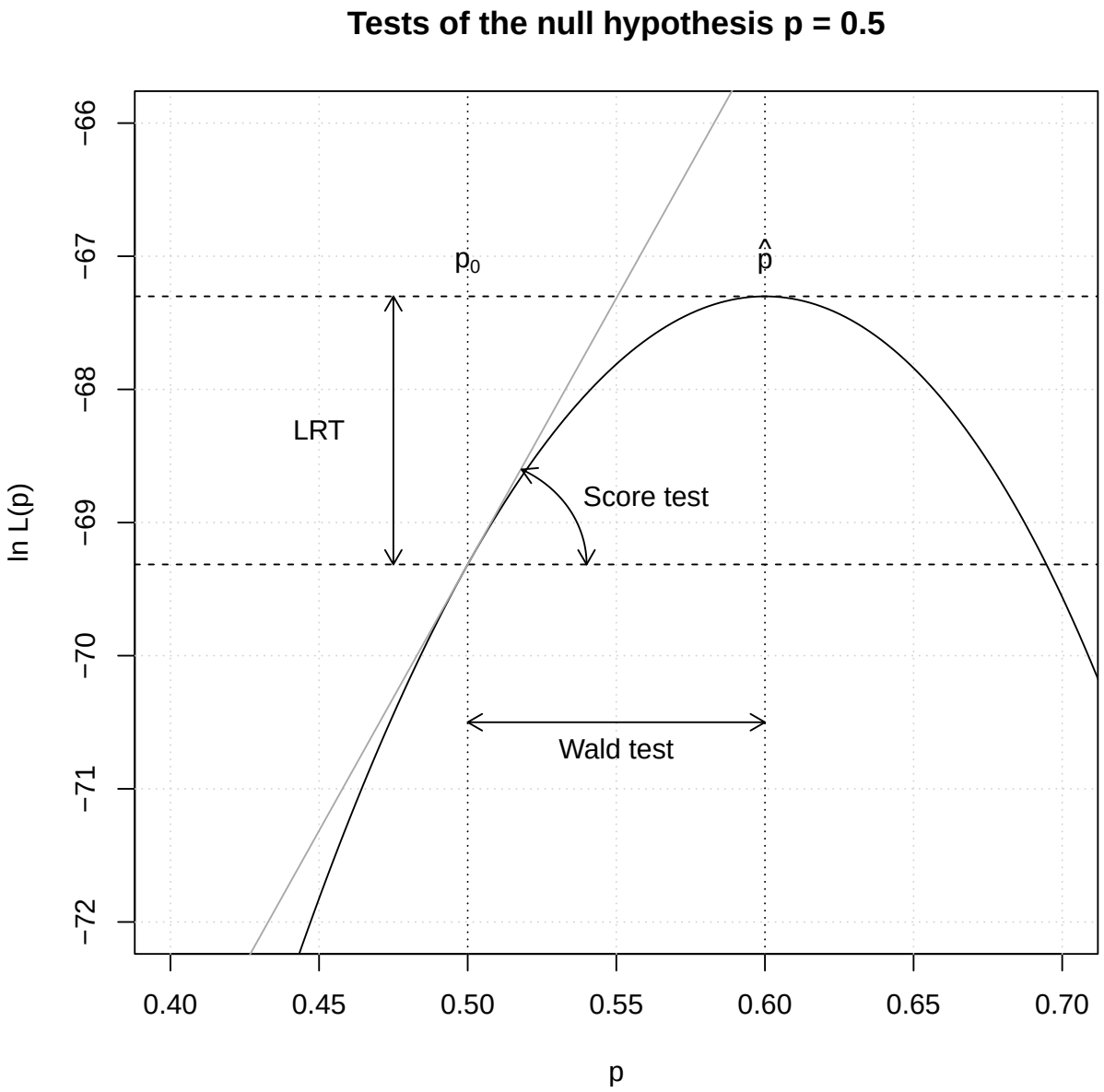

Figure 3.4: Binomial log likelihood function for $x = 60$ and $n = 100$. The null value of p is $p_0 = 0.5$ and the maximum likelihood estimate is $\hat{p} = 0.6$.

3.5.3 Critical values and p-values

The Neyman-Pearson approach to hypothesis testing fixes the significance level α before calculating the test statistic and deciding whether to reject H_0.[13] The decision to reject the null hypothesis depends on the value of the test statistic, which is compared to a **critical value** calculated based on the distribution of the test statistic under H_0. If $Z \sim N(0,1)$ under H_0

$$\Pr\left(|Z| \geq z_{1-\frac{\alpha}{2}} \,|\, H_0 \text{ true}\right) = 1 - \alpha.$$

Because $Z^2 \sim \chi^2_1$ when $Z \sim N(0,1)$, this is equivalent to

$$\Pr\left(Z^2 \geq z^2_{1-\frac{\alpha}{2}} \,|\, H_0 \text{ true}\right) = 1 - \alpha.$$

In the Wald, score, and likelihood ratio tests above, H_0 is rejected if the test statistic is larger than the critical value $z^2_{1-\frac{\alpha}{2}}$. For $\alpha = 0.05$, we have $z_{0.975} \approx 1.96$ so critical value for the χ^2_1 distribution is $1.96^2 \approx 3.84$. The test statistic and critical value in a hypothesis test are analogous to the clinical measurement and cutoff in a diagnostic test.

Instead of making a binary decision, it is more informative to calculate a measure of the evidence against H_0. The **p-value** for a given test statistic is the lowest value of α at which the test would still fail to reject H_0. A hypothesis test with significance level α rejects H_0 if the p-value is $\leq \alpha$. For the Wald, score, or likelihood ratio tests above,

$$\text{p-value} = 1 - F_{\chi^2_1}(\text{test statistic})$$

where $F_{\chi^2_1}$ is the CDF of the χ^2_1 distribution If we think of the test statistic as the clinical measurement underlying a diagnostic test, the p-value equals $1 - \text{spec}_{\text{max}}$ where spec_{max} is the highest specificity under which we would still get a positive test (i.e., reject H_0).

3.6 Confidence intervals

A p-value is more informative than a binary decision whether to reject H_0, but it is still more useful to know what values of p are plausibly consistent with the data we observed (Rothman 1978). The $1-\alpha$ **confidence interval** for p_{true} is the set of all possible null values p_0 such that we would fail to reject $H_0 : p_{\text{true}} = p_0$ in a hypothesis test with significance level α. The endpoints of the confidence interval are called are called *confidence limits*. Just as different clinical measurements lead to different diagnostic tests, different hypothesis tests lead to different confidence intervals.

[13] This approach to hypothesis testing was pioneered in the 1929s by Jerzy Neyman (1894–1981), a Polish mathematician and statistician who founded the first department of statistics in the United States at the University of California, Berkeley in 1938, and Egon Pearson (1895–1980), a British statistician who was a professor at University College London like his father Karl Pearson.

If we calculate a confidence interval many times with independent data sets, the $1-\alpha$ confidence interval should contain p_{true} with probability $1-\alpha$. The actual probability that the confidence interval contains p_{true} is called the **coverage probability**. A good confidence interval should have a coverage probability close to $1-\alpha$ while being as narrow as possible. The Wald, score, and likelihood ratio tests from Section 3.5.2 are large-sample tests because they rely on consistency and asymptotic normality of the maximum likelihood estimate $\hat{p}$. All three tests can be inverted to produce confidence intervals that perform well in large samples. In smaller samples, the score and likelihood ratio confidence intervals often have better coverage probability and width than the Wald confidence interval (Agresti and Coull 1998; Brown, Cai, and DasGupta 2001).

3.6.1 Wald confidence intervals and the delta method

The Wald confidence limits come from solving the equation

$$\frac{(\hat{p}-p)^2}{\hat{p}(1-\hat{p})/n} = z^2_{1-\frac{\alpha}{2}}. \tag{3.16}$$

for p, which gives us

$$\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{3.17}$$

The coverage probabilities of Wald confidence intervals can be much lower than $1-\alpha$, especially when p_{true} is close to zero or one (Agresti and Coull 1998; Brown, Cai, and DasGupta 2001).

Another problem with the Wald confidence interval for p_{true} is that it can have bounds outside $[0,1]$. One way to avoid this is to calculate confidence limits for a transformation of $\hat{p}$ using the **delta method**. A good transformation $g(p)$ should have continuous first derivatives and be strictly increasing or decreasing, so each value of $g(p)$ corresponds to a single value of p (i.e., g is *one-to-one*). The delta method derives the approximate normal distribution $g(\hat{p})$ using the approximation

$$g(\hat{p}) \approx g(p_{\text{true}}) + g'(p_{\text{true}})(\hat{p} - p_{\text{true}}).$$

where $g'(p_{\text{true}})$ is the slope of g at p_{true}. An example of this approximation is shown in Figure 3.5. The key insight is that

$$\text{Var}[g(\hat{p})] \approx g'(p_{\text{true}})^2 \,\text{Var}(\hat{p}),$$

which is a generalization of the fact that $\text{Var}(c\hat{p}) = c^2 \,\text{Var}(\hat{p})$ for any constant c. If $\hat{p}$ has an approximate $N(p_{\text{true}}, \text{Var}(\hat{p}))$ distribution in large samples, then

$$g(\hat{p}) \overset{\text{approx}}{\sim} N(g(p_{\text{true}}), g'(p_{\text{true}})^2 \,\text{Var}(\hat{p})).$$

in large samples. Because our estimator $\hat{p}$ is consistent, we can replace the unknown p_{true} with $\hat{p}$. Because g is one-to-one, we can calculate confidence limits for p_{true} using the confidence limits for $g(p_{\text{true}})$.

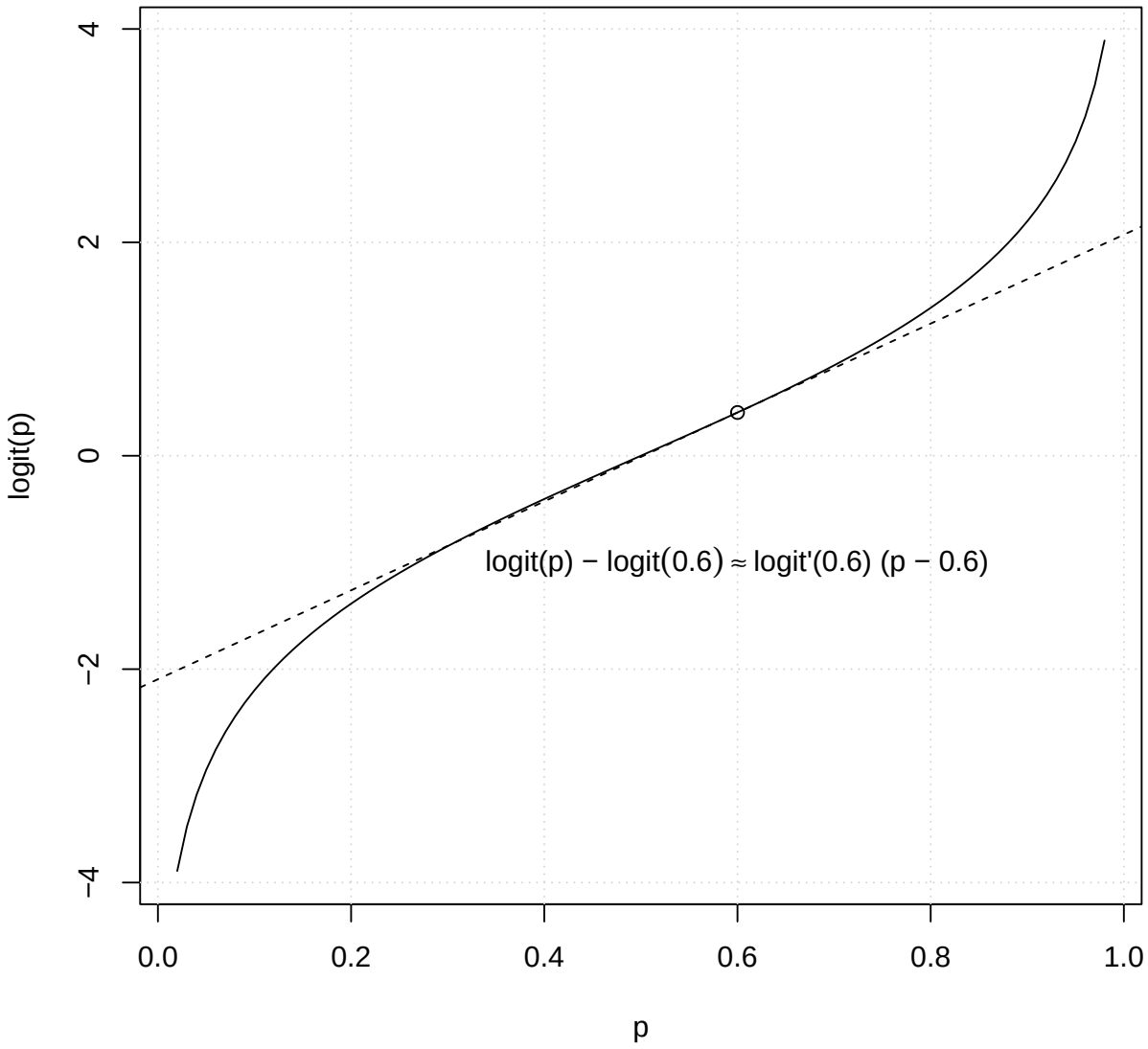

Figure 3.5: The approximation used by the delta method using the logistic transformation for a binomial confidence interval near $\hat{p} = 0.6$. The black curve is $\text{logit}(p)$, and the dashed line shows the tangent line at $p = 0.6$.

A widely used transformation for probabilities is the **logit transformation**

$$\text{logit}(p) = \ln\Big(\frac{p}{1-p}\Big).$$

The **odds** corresponding to the probability p is $\frac{p}{1-p}$, so the logit is the natural logarithm of the odds. The logit transformation maps the interval $(0, 1)$ onto all of $\mathbb{R}$:

- As $p \to 0$, the odds $p/(1-p) \to 0$ and $\text{logit}(p) \to -\infty$.

- When $p = 1/2$, the odds $p/(1-p) = 1$ and $\text{logit}(p) = 0$.
- As $p \to 1$, the odds $p/(1-p) \to \infty$ and $\text{logit}(p) \to \infty$.

To use the delta method, we need to calculate the derivative of $\text{logit}(p)$. By the chain rule,

$$\text{logit}'(p) = \frac{1-p}{p}\frac{1}{(1-p)^2} = \frac{1}{p(1-p)},$$

which is continuous and strictly positive for all $p \in (0,1)$. By the delta method, the variance of $\text{logit}(\hat{p})$ is approximately

$$\text{logit}'(p_{\text{true}})^2 \frac{p_{\text{true}}(1-p_{\text{true}})}{n} = \frac{1}{p_{\text{true}}^2(1-p_{\text{true}})^2}\frac{p_{\text{true}}(1-p_{\text{true}})}{n} = \frac{1}{np_{\text{true}}(1-p_{\text{true}})}.$$

When we replace the unknown p_{true} with our MLE $\hat{p}$, we get the following confidence limits for $\text{logit}(p_{\text{true}})$:

$$\text{logit}(\hat{p}) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}.$$

To get confidence limits for p_{true}, we use the inverse function for the logit, which is

$$\text{expit}(v) = \frac{e^v}{1+e^v} = \frac{1}{1+e^{-v}}.$$

This is called the *logistic function*. If the confidence limits for $\text{logit}(p_{\text{true}})$ are a and b, then the confidence limits for p_{true} are $\text{expit}(a)$ and $\text{expit}(b)$. These are guaranteed to be in $(0,1)$ because $\text{expit}(v) \in (0,1)$ for any $v \in \mathbb{R}$. The logit-transformed confidence interval can have narrower width and a coverage probability closer to $1-\alpha$ than the untransformed Wald confidence interval (Agresti 2013).

3.6.2 Score (Wilson) confidence intervals

The **score** or **Wilson** confidence limits come from solving the equation

$$\frac{(\hat{p}-p)^2}{p(1-p)/n} = z_{1-\frac{\alpha}{2}}^2. \tag{3.18}$$

for p (Wilson 1927). This differs from Equation 3.16 for the Wald confidence interval because it uses p instead of $\hat{p}$ in the denominator. It is a quadratic equation in p, so it has two solutions. The center of the resulting confidence interval is

$$\tilde{p} = \hat{p}\left(\frac{n}{n+z_{1-\frac{\alpha}{2}}^2}\right) + \frac{1}{2}\left(\frac{z_{1-\frac{\alpha}{2}}^2}{n+z_{1-\frac{\alpha}{2}}^2}\right) = \frac{x+\frac{1}{2}z_{1-\frac{\alpha}{2}}^2}{n+z_{1-\frac{\alpha}{2}}^2}, \tag{3.19}$$

where x is the number of diseased individuals in our sample. This is a weighted average of $\hat{p}$ and $1/2$ with weights proportional to n and $z^2_{1-\frac{\alpha}{2}}$, respectively. The resulting confidence interval is

$$\tilde{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\tilde{V}}$$

where

$$\tilde{V} = \frac{\hat{p}(1-\hat{p})}{n + z^2_{1-\frac{\alpha}{2}}} \left(\frac{n}{n + z^2_{1-\frac{\alpha}{2}}} \right) + \frac{\left(\frac{1}{2}\right)^2}{n + z^2_{1-\frac{\alpha}{2}}} \left(\frac{z^2_{1-\frac{\alpha}{2}}}{n + z^2_{1-\frac{\alpha}{2}}} \right).$$

This variance is a weighted average of the variances of sample proportions equal to $\hat{p}$ and $1/2$ with the same weights as in $\tilde{p}$ and with $n + z^2_{1-\frac{\alpha}{2}}$ instead of n in the denominator. Wilson confidence intervals are narrower than the corresponding Wald intervals, and they have coverage probabilities much closer to $1 - \alpha$ (Agresti and Coull 1998; Brown, Cai, and DasGupta 2001).

The *Agresti-Coull confidence interval* is a simplification of the Wilson confidence interval that replaces $\hat{p}$ with $\tilde{p}$ in the Wald confidence interval to get the confidence limits

$$\tilde{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}.$$

Because $z_{0.975} \approx 1.96$, we have $\tilde{p} \approx \frac{k+2}{n+4}$ for a 95% confidence interval. In this case, the Agresti-Coull interval is often implemented as follows: "Add two successes and two failures and then use the Wald formula'' (Agresti and Coull 1998). This interval is only slightly wider than the score confidence interval, and the two intervals are nearly identical for $n > 40$ (Brown, Cai, and DasGupta 2001).

The likelihood ratio test can also be inverted to get confidence intervals, but these can only be calculated numerically. For the binomial model, the likelihood ratio and score confidence intervals are nearly identical (Agresti and Coull 1998; Brown, Cai, and DasGupta 2001). The score intervals are more common in practice because they are easier to calculate.

3.7 Small-sample estimation*

Maximum likelihood estimates are consistent, asymptotically normal, and asymptotically efficient. However, they are not guaranteed to perform well in any finite sample. For a sample of n independent Bernoulli(p) random variables, the sum has a binomial(n, p) distribution and this can be used to find the finite-sample distribution of the sample mean. This distribution can be used directly to calculate point estimates, p-values, and confidence limits.

Confidence limits calculated using the finite-sample distribution of a test statistic under H_0 are called **exact confidence limits**. They can often be constructed to have a coverage probability of at least $1-\alpha$. However, their coverage probabilities are often higher than $1-\alpha$, and they can

be much wider than approximate $1-\alpha$ confidence intervals for the same parameter (Agresti and Coull 1998).

If the finite-sample distribution of the test statistic is not known exactly, it is possible to calculate point estimates, p-values, or confidence limits using simulations. This is the basic idea behind the *bootstrap* (Efron and Tibshirani 1994) and *Monte Carlo methods* (Robert and Casella 2004).

3.7.1 Median unbiased estimate

The **median unbiased estimate** of p_{true} is the value of p that makes

$$\Pr_p(X < x) = \Pr_p(X > x)$$

where we use the subscript p to indicate that these probabilities are calculated assuming $p_{\text{true}} = p$. If p_{med} is the median unbiased estimate, then

$$\sum_{k=0}^{x-1} \binom{n}{k} p_{\text{med}}^k (1-p_{\text{med}})^{n-k} + \frac{1}{2}\binom{n}{x} p_{\text{med}}^k (1-p_{\text{med}})^{n-x} = \frac{1}{2},$$

and

$$\frac{1}{2}\binom{n}{x} p_{\text{med}}^x (1-p_{\text{med}})^{n-x} + \sum_{k=x+1}^{n} \binom{n}{k} p_{\text{med}}^k (1-p_{\text{med}})^{n-k} = \frac{1}{2}.$$

The median of the distribution of p_{med} is always p_{true} (Birnbaum 1964), which is a slightly different notion of unbiasedness than the unbiasedness of $\hat{p}$ where $\mathbb{E}(\hat{p}) = p_{\text{true}}$.

3.7.2 Exact (Clopper-Pearson) and mid-p confidence intervals

The **exact** or **Clopper-Pearson** confidence limits for p_{true} use the finite-sample distribution of the sample mean $\hat{p}$ Clopper and Pearson (1934). When $x > 0$, the lower $1-\alpha$ confidence limit is the solution to

$$\sum_{k=x}^{n} \binom{n}{k} p_{\text{lower}}^k (1-p_{\text{lower}})^{n-k} = \frac{\alpha}{2}, \tag{3.20}$$

so the *upper tail* of the binomial(n, p_{lower}) distribution has probability $\alpha/2$. When $x = 0$, we set $p_{\text{lower}} = 0$. When $x < n$, the upper confidence limit is the solution to

$$\sum_{k=0}^{x} \binom{n}{k} p_{\text{upper}}^k (1-p_{\text{upper}})^{n-k} = \frac{\alpha}{2}, \tag{3.21}$$

so the *lower tail* of the binomial(n, p_{upper}) distribution has probability $\alpha/2$. When $x = n$, we set $p_{\text{upper}} = 1$. This interval is guaranteed to have a coverage probability of at least $1-\alpha$, but the price for this is that it is always wider than the Wald and Wilson confidence intervals

(Agresti and Coull 1998; Brown, Cai, and DasGupta 2001). In general, the score or likelihood ratio confidence intervals have better combinations of coverage probability and width.

To make exact confidence limits less conservative, we can include only $\frac{1}{2}\Pr(X = x)$ instead of $\Pr(X = x)$ in the calculation of the tail probabilities in Equation 3.21 and Equation 3.20. The resulting confidence intervals are called **mid-p exact confidence intervals** (Lancaster 1961, berry1995mid). The lower $1 - \alpha$ mid-p exact confidence limit is the solution to

$$\frac{1}{2}\binom{n}{x}p_{\text{lower}}^{x}(1-p_{\text{lower}})^{n-x} + \sum_{k=x+1}^{n}\binom{n}{k}p_{\text{lower}}^{k}(1-p_{\text{lower}})^{n-k} = \frac{\alpha}{2}.$$

and the upper limit is the solution to

$$\sum_{k=0}^{x-1}\binom{n}{k}p_{\text{upper}}^{k}(1-p_{\text{upper}})^{n-k} + \frac{1}{2}\binom{n}{x}p_{\text{upper}}^{x}(1-p_{\text{upper}})^{n-x} = \frac{\alpha}{2}.$$

The mid-p exact confidence limits are have good combinations of coverage probability and width as well as good perfomance in small samples (Brown, Cai, and DasGupta 2001).

Listing 3.3 `normplots.R`

```
## Normal distribution PDF and CDF

# set grid of plots
par(mfrow = c(2, 1), mar = c(2, 5, 2, 2) + 0.1)

# define variables
x <- seq(-3.5, 3.5, by = 0.01)
a <- 0
b <- 2

# plot of PDF
plot(x, dnorm(x), type = "n",
    ylab = expression(paste("PDF ", phi1(z))))
grid()
lines(x, dnorm(x))
polygon(x = c(b, a, seq(a, b, by = 0.01)),
        y = c(0, 0, dnorm(seq(a, b, by = 0.01))),
        lty = "dashed", col = "darkgray")
text(0.4, 0.18, labels = "Area = Pr(0 < Z < 2)", srt = 90)

# plot of CDF
plot(x, pnorm(x), type = "n",
     ylab = expression(paste("CDF ", Phi(z))))
grid()
lines(x, pnorm(x))
segments(c(-4, -4), pnorm(c(a, b)), c(a, b), pnorm(c(a, b)),
         lty = "dashed")
segments(c(a, b), c(-1, -1), c(a, b), pnorm(c(a, b)), lty = "dashed")
arrows(-3, pnorm(a), -3, pnorm(b), code = 3, length = 0.1)
text(-1.7, sum(pnorm(c(a, b))) / 2, labels = "Change = Pr(0 < Z < 2)")
```

Listing 3.4 `normdist.R`

```
## normal (Gaussian) distribution

# normal PDF
# Second and third arguments are mean and SD (not variance).
# The defaults are mean = 0 and SD = 1.
dnorm(2, 1.2, 5)

# normal CDF (using default mean and variance)
pnorm(1.96)
pnorm(1.96) - pnorm(-1.96)

# normal quantiles
qnorm(0.975)
pnorm(qnorm(0.975))

# random samples (using named arguments)
rnorm(25, mean = 2.3, sd = 3)
```

Listing 3.5 `clt.R`

```
## Central limit theorem

# probability mass function for sample mean
dbline <- function(n, p=.5, ...) {
  x <- (seq(-.5, n + .5) / n - p) * sqrt(n / (p * (1 - p)))
  y <- c(0, dbinom(0:n, n, p), 0) * sqrt(p * (1 - p) * n)
  lines(stepfun(x, y), pch = NA, ...)
}

# define grid of plots
par(mfrow = c(2, 2))
x <- seq(-4, 4, by = .01)

# n = 20
plot(x, dnorm(x), type = "n", ylim = c(0, .5),
     main = "n = 20", xlab = "Z score", ylab = "Probability density")
grid()
dbline(20, p = .1, lty = "dashed")
lines(x, dnorm(x), col = "darkgray")

# n = 50
plot(x, dnorm(x), type = "n", ylim = c(0, .5),
     main = "n = 50", xlab = "Z score", ylab = "Probability density")
grid()
dbline(50, p = .1, lty = "dashed")
lines(x, dnorm(x), col = "darkgray")

# n = 100
plot(x, dnorm(x), type = "n", ylim = c(0, .5),
     main = "n = 100", xlab = "Z score", ylab = "Probability density")
grid()
dbline(100, p = .1, lty = "dashed")
lines(x, dnorm(x), col = "darkgray")

# n = 250
plot(x, dnorm(x), type = "n", ylim = c(0, .5),
     main = "n = 250", xlab = "Z score", ylab = "Probability density")
grid()
dbline(250, p = .1, lty = "dashed")
lines(x, dnorm(x), col = "darkgray")
```

Listing 3.6 `htests.R`

```
## Hypothesis tests based on the log likelihood

# binomial log likelihood, score, and information functions
bin_loglik <- function(p, k=60, n=100) {
  k * log(p) + (n - k) * log(1 - p)
}
bin_score <- function(p, k=60, n=100) {
  k / p - (n - k) / (1 - p)
}
bin_information <- function(p, k=60, n=100) {
  k / p^2 + (n - k) / (1 - p)^2
}

# plot showing Wald, score, and likelihood ratio tests
p <- seq(0.4, 0.8, length.out = 200)
plot(p, bin_loglik(p), type = "n",
     xlim = c(0.40, 0.70), ylim = c(-72, -66),
     main = "Tests of the null hypothesis p = 0.5",
     xlab = "p", ylab = "ln L(p)")
grid()
lines(p, bin_loglik(p))
abline(v = c(0.5, 0.6), lty = "dotted")
abline(h = c(bin_loglik(0.5), bin_loglik(0.6)), lty = "dashed")
abline(a = bin_loglik(0.5) - bin_score(0.5) * 0.5, b = bin_score(0.5),
       col = "darkgray")
text(c(0.5, 0.6), c(-67.05, -67),
     labels = c(expression(p[0]), expression(hat(p))))
text(0.55, -70.7, labels = "Wald test")
arrows(0.5, -70.5, 0.6, code = 3, length = 0.1)
arrows(0.475, bin_loglik(0.5), y1 = bin_loglik(0.6),
       code = 3, length = 0.1)
text(0.45, -68.3, labels = "LRT")

# The slope is the tangent of the angle to the x-axis.
# We also must account for the different scales on the x- and y-axes.
# 0.3 / 6 is xdist / ydist (see xlim and ylim above)
score_angle <- atan(bin_score(0.5) * 0.3 / 6)
angles <- seq(0, score_angle, by = 0.01)
score_x <- 0.5 + 0.04 * cos(angles)
score_y <- bin_loglik(0.5) + 0.04 * (6 / 0.3) * sin(angles)
lines(score_x, score_y)
text(0.56, -68.8, "Score test")
arrows(score_x[2], score_y[2], score_x[1], score_y[1], length = 0.1)
arrows(rev(score_x)[2], rev(score_y)[2], rev(score_x)[1], rev(score_y)[1],
       length = 0.1)
```

Listing 3.7 `delta.R`

```
## Approximation used by the delta method

p <- seq(0.02, 0.98, by = 0.01)
logit <- function(p) log(p) - log(1 - p)

# plot
plot(p, logit(p), type = "n",
     xlab = "p", ylab = "logit(p)")
grid()
lines(p, logit(p))
points(0.6, logit(0.6))
abline(logit(0.6) - 2.5, 1 / 0.24, lty = "dashed")
text(0.6, -1,
     labels = expression(paste("logit(p) - ", logit(0.6) %~~% logit,
                               "'(0.6) (p - 0.6)")))
```

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". In the late 1950s and the 1960s, he argued that scientific evidence did not establish that smoking causes lung cancer.

5 Longitudinal Data and Rates

6 Survival Analysis

Part II

Two-Sample Inference and Study Design

Part III

Principles of Causal Inference

Part IV

Epidemologic and Statistical Methods for Causal Inference

References

Agresti, Alan. 2013. *Categorical Data Analysis*. Third. Vol. 792. John Wiley & Sons.

Agresti, Alan, and Brent A Coull. 1998. "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician* 52 (2): 119–26.

Aitchison, John, and SD Silvey. 1958. "Maximum-Likelihood Estimation of Parameters Subject to Restraints." *The Annals of Mathematical Statistics* 29: 813–28.

Albert, Adelin. 1982. "On the Use and Computation of Likelihood Ratios in Clinical Chemistry." *Clinical Chemistry* 28 (5): 1113–19.

Alho, Juha M. 1992. "On Prevalence, Incidence, and Duration in General Stable Populations." *Biometrics* 48 (2): 587–92.

Bamber, Donald. 1975. "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph." *Journal of Mathematical Psychology* 12 (4): 387–415.

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.

Bengtsson, Ewert, and Patrik Malm. 2014. "Screening for Cervical Cancer Using Automated Analysis of PAP-Smears." *Computational and Mathematical Methods in Medicine* 2014: 842037.

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

Birnbaum, Allan. 1964. "Median-Unbiased Estimators." *Bulletin of Mathematical Statistics* 11: 25–34.

Blumberg, Mark S. 1957. "Evaluating Health Screening Procedures." *Operations Research* 5 (3): 351–60.

Boos, Dennis D, and Leonard A Stefanski. 2013. *Essential Statistical Inference*. Springer.

Bostrom, RC, HS Sawyer, and WE Tolles. 1959. "Instrumentation for Automatically Prescreening Cytological Smears." *Proceedings of the IRE* 47 (11): 1895–1900.

Brown, Lawrence D, T Tony Cai, and Anirban DasGupta. 2001. "Interval Estimation for a Binomial Proportion." *Statistical Science* 16 (2): 101–17.

Clopper, Charles J, and Egon S Pearson. 1934. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." *Biometrika* 26 (4): 404–13.

Cohen, I Bernard. 1984. "Florence Nightingale." *Scientific American* 250 (3): 128–37.

Cramér, Harald. 1946. *Mathematical Methods of Statistics*. Princeton University Press.

Diamond, George A, and James S Forrester. 1983. "Clinical Trials and Statistical Verdicts: Probable Grounds for Appeal." *Annals of Internal Medicine* 98 (3): 385–94.

Dunn Jr, John E. 1962. "The Use of Incidence and Prevalence in the Study of Disease Development in a Population." *American Journal of Public Health* 52 (7): 1107–18.

Efron, Bradley, and David V Hinkley. 1978. "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information." *Biometrika* 65 (3): 457–83.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Fagan, Terrence J. 1975. "Nomogram for Bayes's Theorem." *New England Journal of Medicine* 293 (5): 257.

Freedman, David A. 2007. "How Can the Score Test Be Inconsistent?" *The American Statistician* 61 (4): 291–95.

Freeman, Jonathan, and George B Hutchison. 1980. "Prevalence, Incidence and Duration." *American Journal of Epidemiology* 112 (5): 707–23.

Hanley, James A, and Barbara J McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology* 143 (1): 29–36.

Keiding, Niels. 1991. "Age-Specific Incidence and Prevalence: A Statistical Perspective." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154 (3): 371–96.

Kenward, Michael G, and Geert Molenberghs. 1998. "Likelihood Based Frequentist Inference When Data Are Missing at Random." *Statistical Science* 13 (3): 236–47.

Kessel, Elton. 1962. "Diabetes Detection: An Improved Approach." *Journal of Chronic Diseases* 15 (12): 1109–21.

Lancaster, H Oliver. 1961. "Significance Tests in Discrete Distributions." *Journal of the American Statistical Association* 56 (294): 223–34.

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

Lusted, Lee B. 1971a. "Decision-Making Studies in Patient Management." *New England Journal of Medicine* 284 (8): 416–24.

———. 1971b. "Signal Detectability and Medical Decision-Making." *Science* 171 (3977): 1217–19.

———. 1984. "ROC Recollected." *Medical Decision Making* 4: 131–35.

MacMahon, Brian, and William D Terry. 1958. "Application of Cohort Analysis to the Study of Time Trends in Neoplastic Disease." *Journal of Chronic Diseases* 7 (1): 24–35.

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

Neyman, Jerzy, and Egon Sharpe Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (694-706): 289–337.

Preston, Samuel H. 1987. "Relations Among Standard Epidemiologic Measures in a Population." *American Journal of Epidemiology* 126 (2): 336–45.

Rao, C Radhakrishna. 1945. "Information and Accuracy Attainable in the Estimation of Statistical Parameters." *Bulletin of the Calcutta Mathematical Society* 37 (3): 81–91.

———. 1948. "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation." In *Mathematical Proceedings of the*

Cambridge Philosophical Society, 44:50–57. Cambridge University Press.

Reid, Nancy. 2003. "Asymptotics and the Theory of Inference." *The Annals of Statistics* 31 (6): 1695–2095.

Remein, Quentin R, and Hugh LC Wilkerson. 1961. "The Efficiency of Screening Tests for Diabetes." *Journal of Chronic Diseases* 13 (1): 6–21.

Robert, Christian P, and George Casella. 2004. *Monte Carlo Statistical Methods*. Second edition. Springer.

Rothman, Kenneth J. 1978. "A Show of Confidence." *New England Journal of Medicine* 299 (24): 1362–63.

———. 1981. "Induction and Latent Periods." *American Journal of Epidemiology* 114 (2): 253–59.

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

Swets, John A. 1973. "The Relative Operating Characteristic in Psychology: A Technique for Isolating Effects of Response Bias Finds Wide Use in the Study of Perception and Cognition." *Science* 182 (4116): 990–1000.

———. 1988. "Measuring the Accuracy of Diagnostic Systems." *Science* 240 (4857): 1285–93.

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

Vecchio, Thomas J. 1966. "Predictive Value of a Single Diagnostic Test in Unselected Populations." *New England Journal of Medicine* 274 (21): 1171–73.

Wald, Abraham. 1943. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large." *Transactions of the American Mathematical Society* 54 (3): 426–82.

Wilks, Samuel S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics* 9 (1): 60–62.

Wilson, Edwin B. 1927. "Probable Inference, the Law of Succession, and Statistical Inference." *Journal of the American Statistical Association* 22 (158): 209–12.

Winkelstein Jr, Warren. 2009. "Florence Nightingale: Founder of Modern Nursing and Hospital Epidemiology." *Epidemiology* 20 (2): 311.

Yerushalmy, Jacob. 1947. "Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques." *Public Health Reports (1896-1970)* 62 (40): 1432–49.

Zweig, Mark H, and Gregory Campbell. 1993. "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine." *Clinical Chemistry* 39 (4): 561–77.

A Calculus