Analytical Epidemiology

Statistical and Causal Inference for Public Health

Eben Kenah

January 30, 2025

Table of contents

Preface

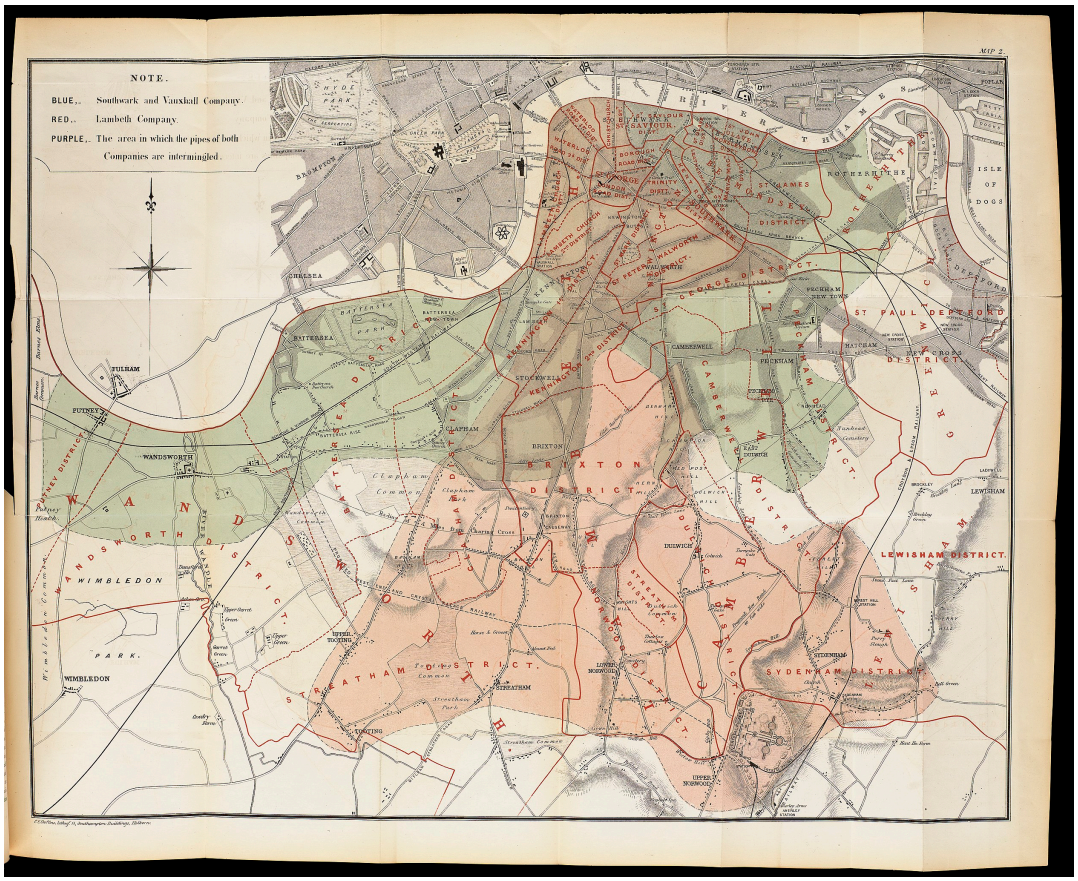


Figure 1: Areas of London supplied by the Southwark & Vauxhall (blue, now green) and Lambeth (red) water companies during the 1849-1854 cholera epidemic in London (Snow 1855). Source: Wellcome Collection via Wikimedia Commons.

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

Defining and Measuring Disease Occurrence

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.

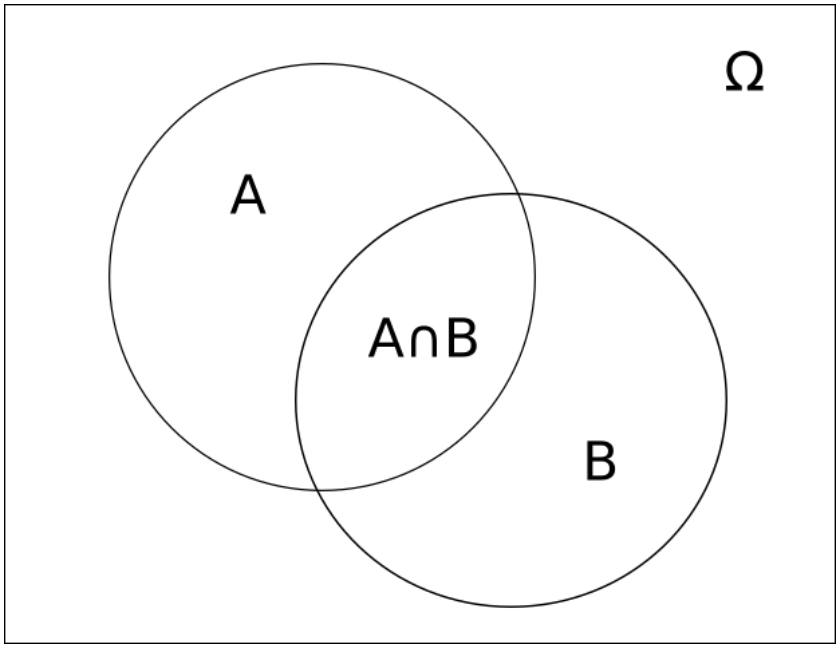


Figure 1.1: Venn diagram showing events A and B. The area contained in both events is their intersection $A \cap B$. The union $A \cup B$ is all area contained in at least one of A and B, including $A \cap B$.

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

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

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 $t^{\text{onset}} \leq t_b$. 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.

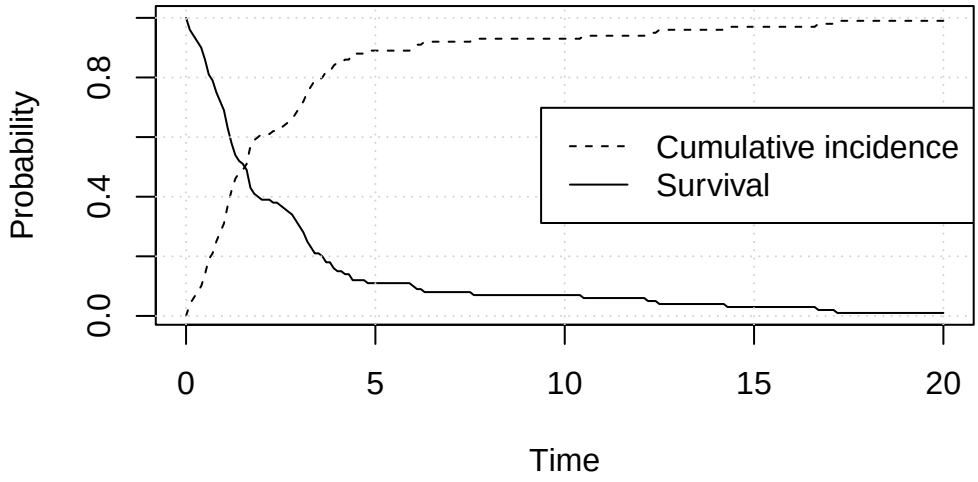


Figure 1.2: Survival and cumulative incidence curves for the data from the prevalence example.

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.

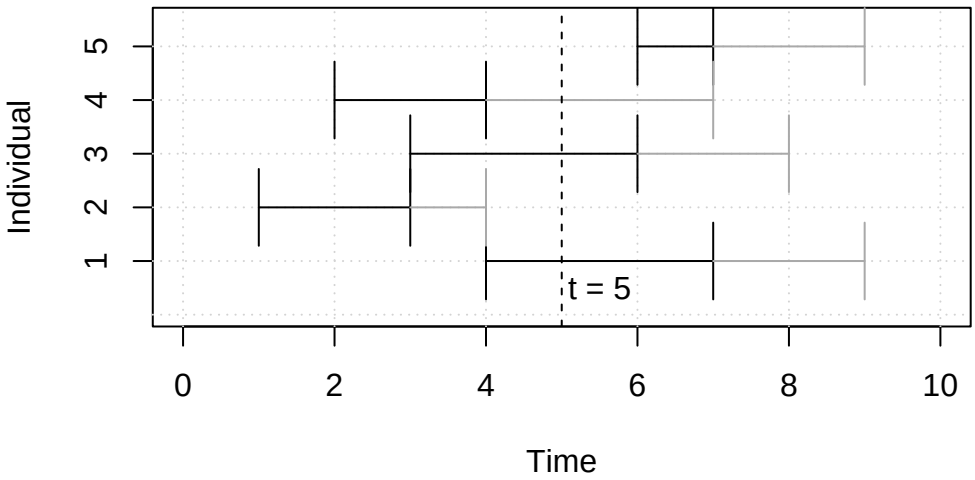


Figure 1.3: Each black horizonal line shows the onset of disease and recovery from disease in a single individual. The gray lines show recoveries from disease if the disease duration increases.

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`

```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.

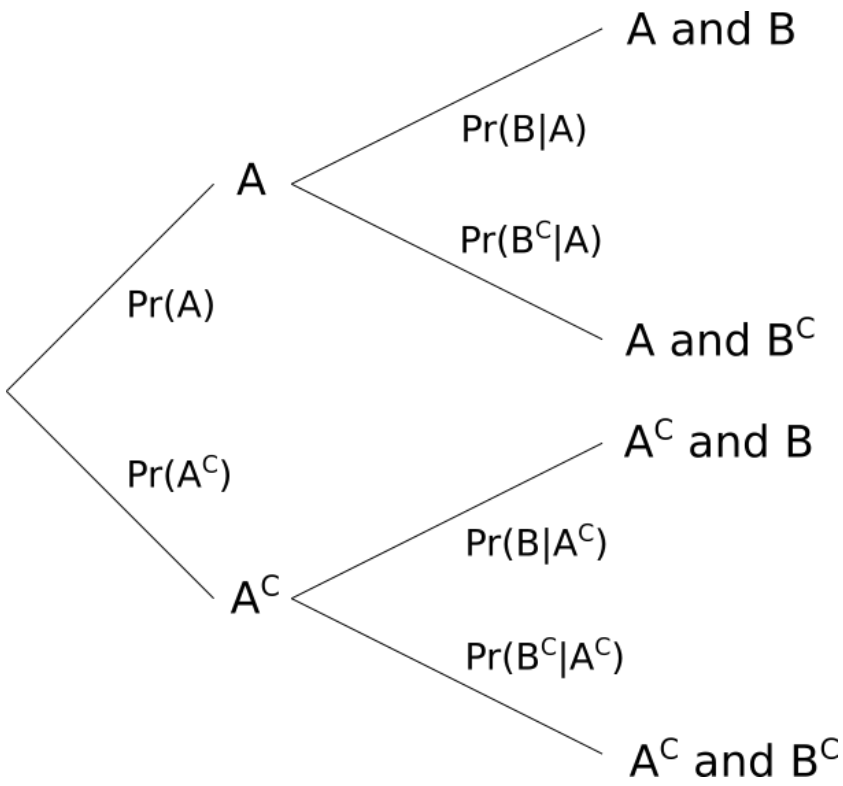


Figure 2.1: A decision tree for events A and B. The probability of each leaf is found by multiplying the probabilities along the branches leading from the leaf back to the root.

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.

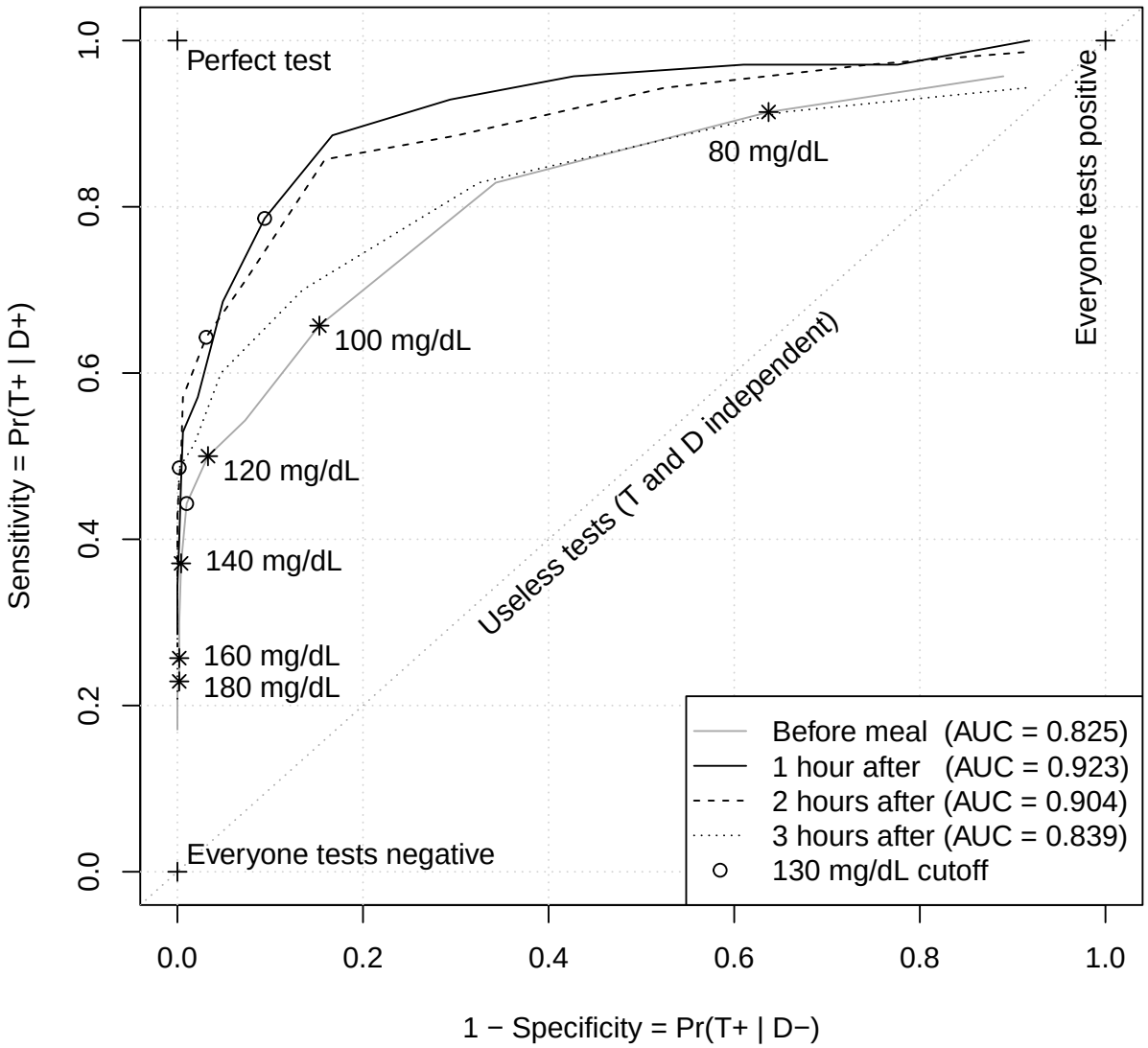


Figure 2.2: ROC curves for Somogyi-Nelson blood tests conducted before the meal and at 1-3 hours after the meal. Cutoff values for the before-meal curve are labeled, and the points corresponding to the 130 mg/dL cutoff along the curve for each blood glucose measurement are circled.

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%.

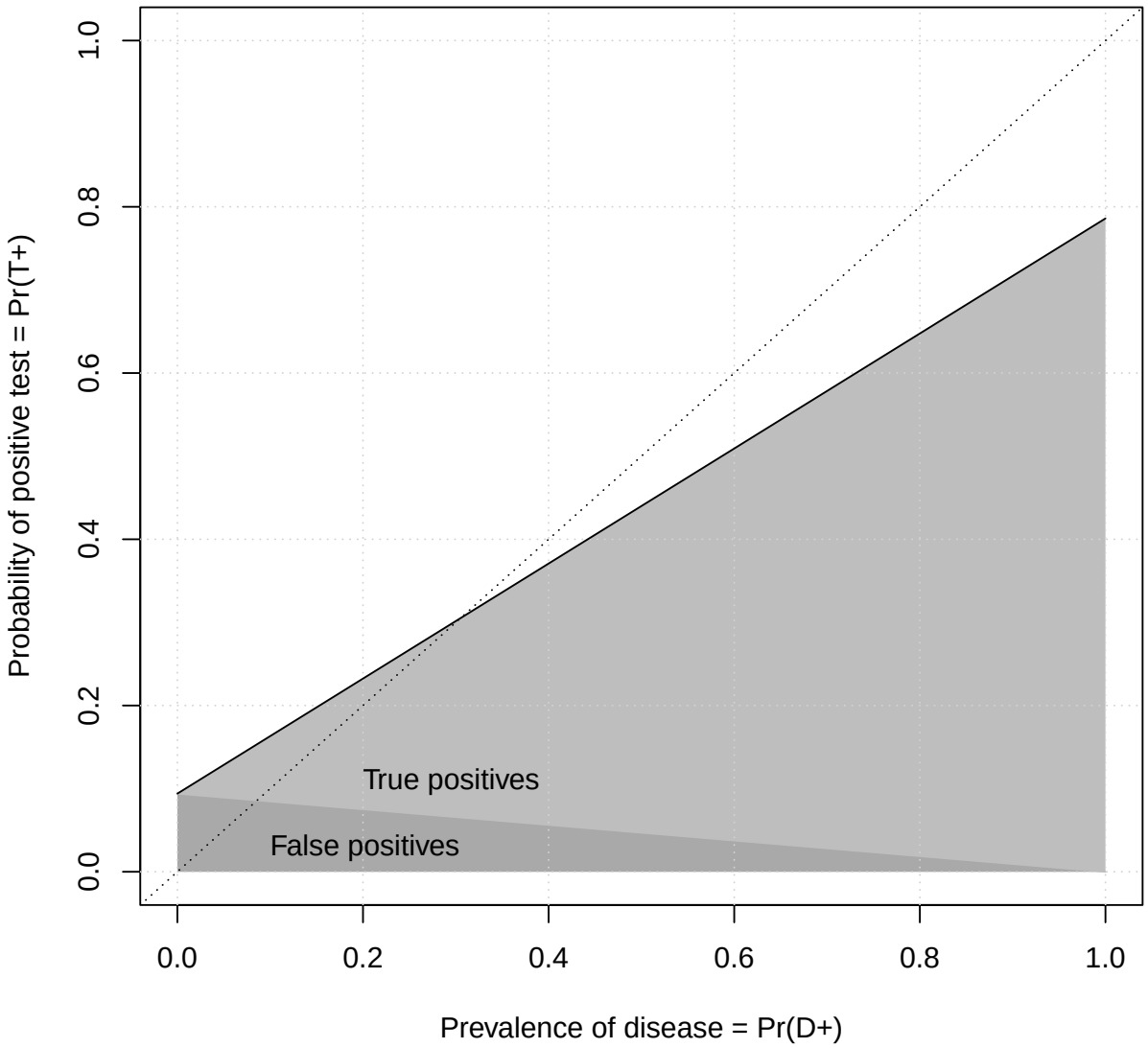


Figure 2.3: The probability of a positive Somogyi-Nelson diabetes test one hour after the meal as a function of the hypothetical prevalence of diabetes. The black dotted line shows the true prevalence of diabetes.

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.

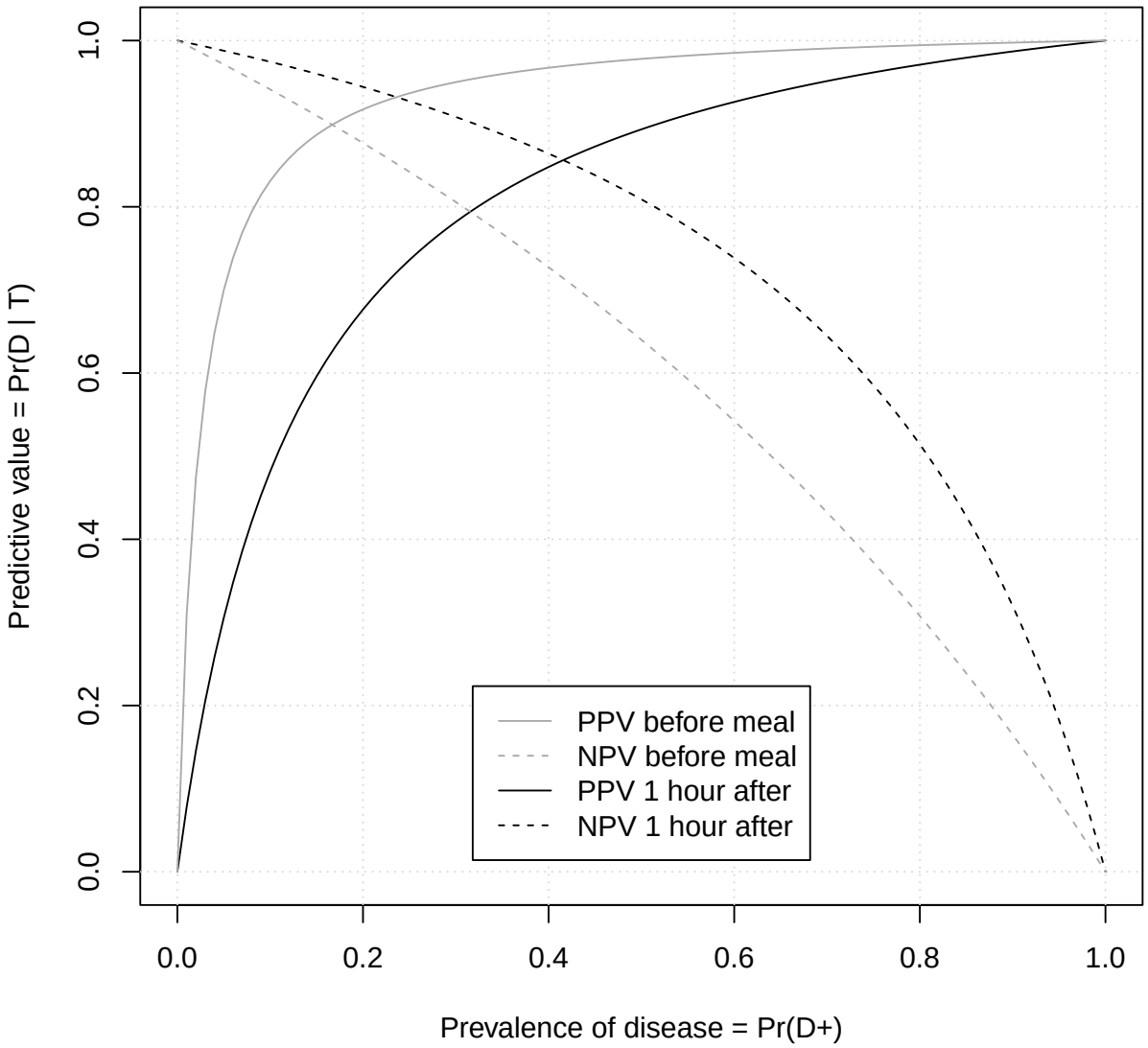


Figure 2.4: Positive and negative predictive values of the Somogyi-Nelson diabetes test before the meal (gray) and one hour after the meal (black) as a function of diabetes prevalence.

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 (I. B. 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** statistical inference, where the probability of an event is interpreted as the limit as $n \to \infty$ of the proportion of n repetitions of an experiment in which the event occurs. Maximum likelihood estimation gives us a way to find point estimates of parameters that are optimal in large samples. It is also the foundation for hypothesis tests and confidence intervals, which are used to quantify uncertainty in frequentist 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

[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.

is to sample n individuals from a population, the entire sample is a single outcome ω and Ω 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 $\Pr(\lim_{n\to\infty} \hat{\mu}_n = \mu) = 1$.

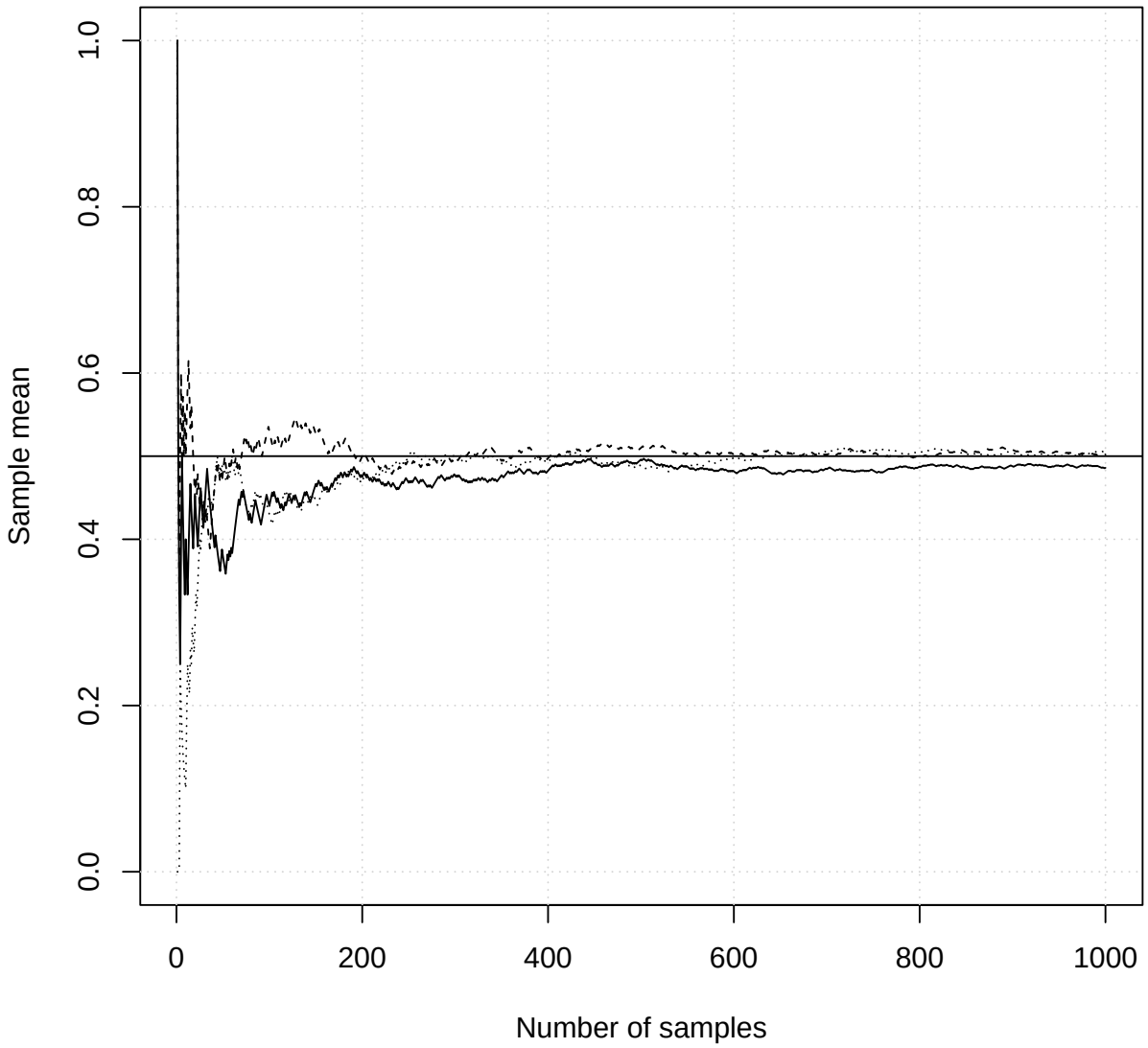


Figure 3.1: The LLN at work. Each line traces the sample means calculated from a sequence of random samples $x_1, x_2, x_3, ...$ from a Bernoulli(0.5) distribution. For each sequence, the y-coordinate above n is the sample mean from the first n random samples in the sequence. The true mean of 0.5 is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

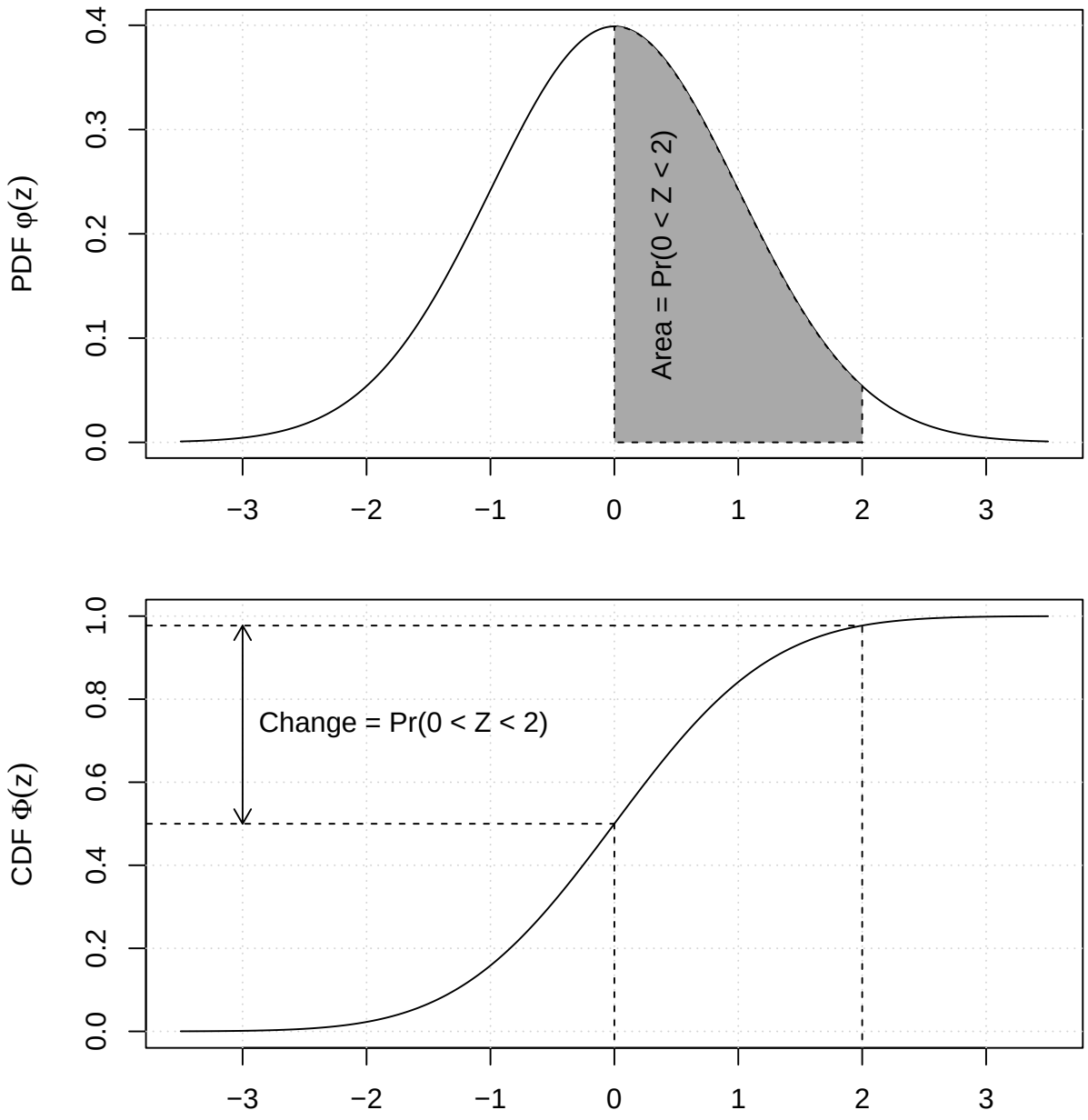


Figure 3.2: The PDF (top) and CDF (bottom) of a standard normal random variable Z. If $X \sim N(0,1)$, then $\Pr(0 < X < 2)$ equals the shaded area under the PDF as well as the change in the CDF from 0 to 2. This same relationship between the CDF and the PDF holds for all continuous random variables 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 $np(1 - p) > 5$.

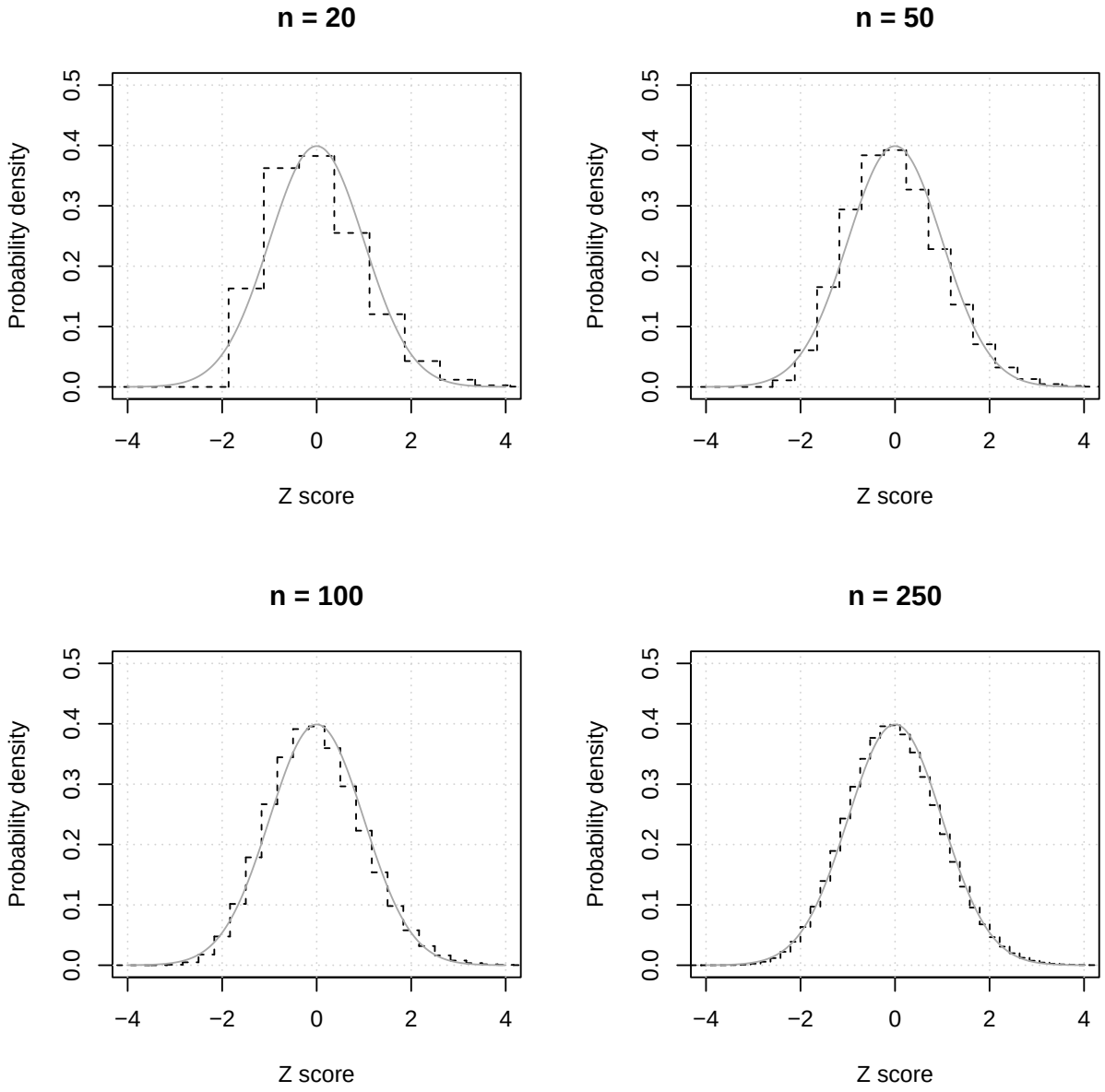


Figure 3.3: The CLT at work. The dashed lines show the PMF of the distribution of the average from a sample of size n from a Bernoulli(0.1) distribution. The solid line is the standard normal PDF.

3.4.1 Efficiency of maximum likelihood estimators*

We have used the LLN and the CLT to show that $\widehat{p}_n$ 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 (D. A. 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.

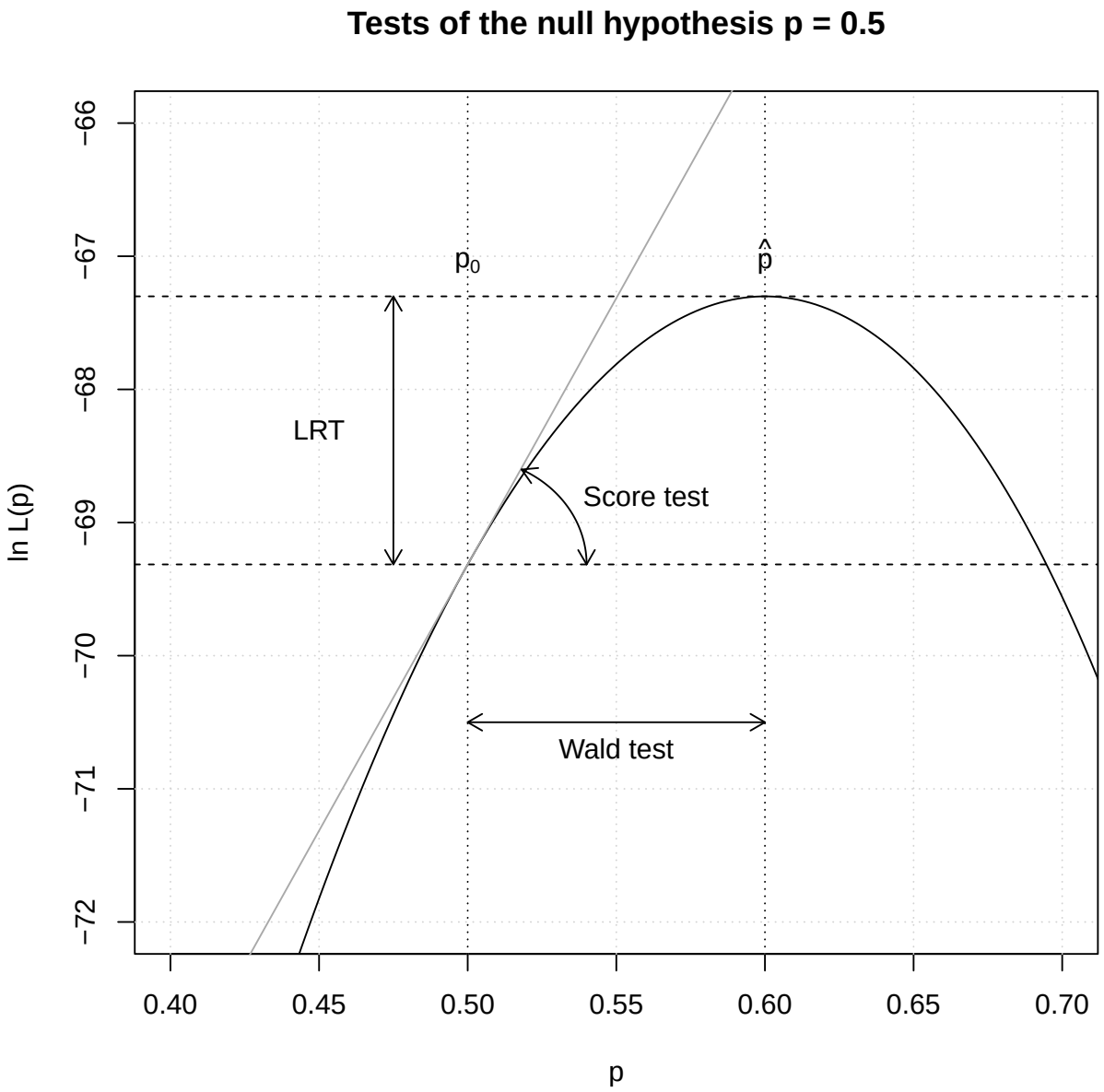


Figure 3.4: Binomial log likelihood function for $x = 60$ and $n = 100$. The null value of p is $p_0 = 0.5$ and the maximum likelihood estimate is $\hat{p} = 0.6$.

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 $g(p_{\text{true}})$.

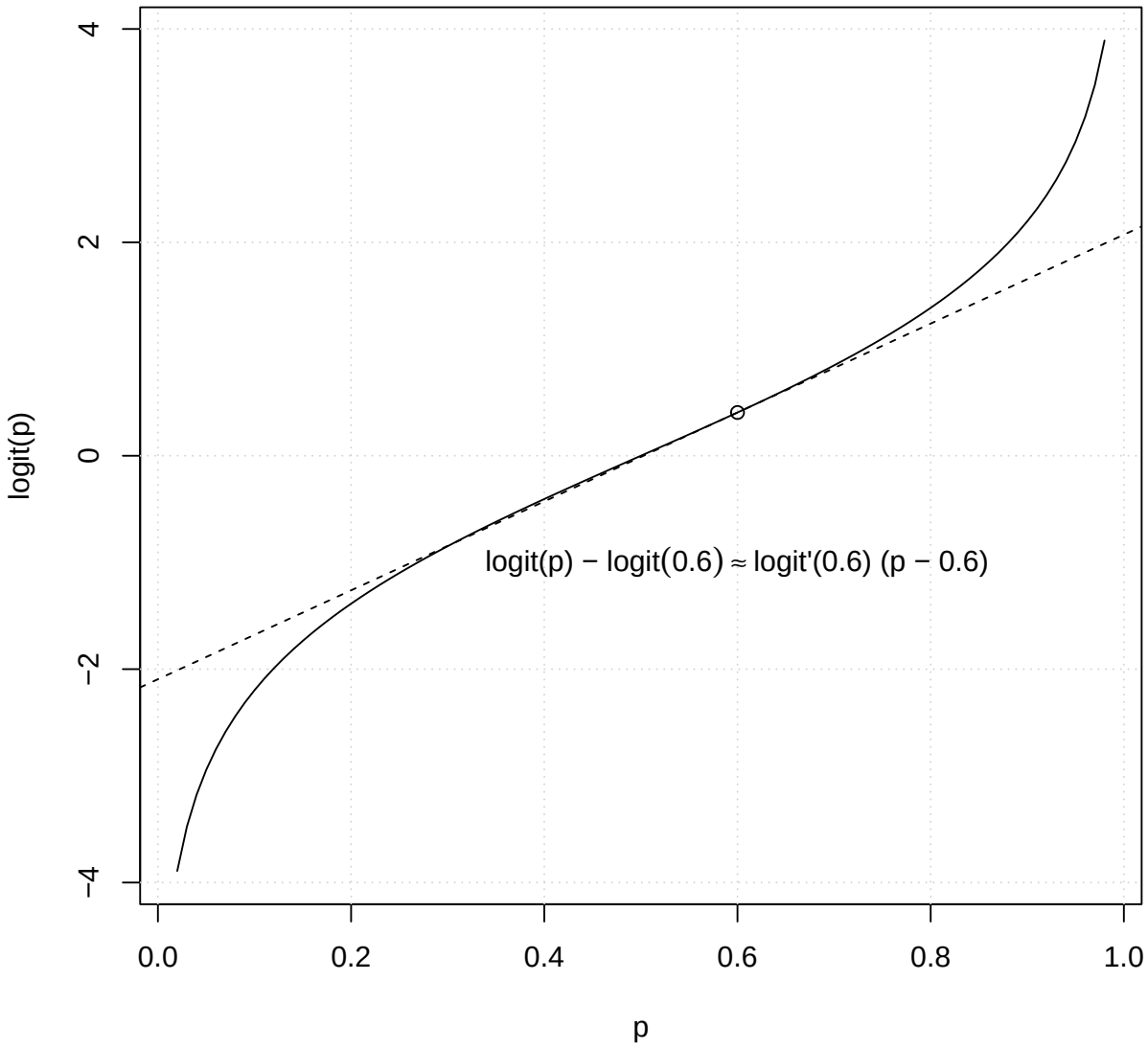


Figure 3.5: The approximation used by the delta method using the logistic transformation for a binomial confidence interval near $\hat{p} = 0.6$. The black curve is $\text{logit}(p)$, and the dashed line shows the tangent line at $p = 0.6$.

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 R

3.8 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).

3.8.1 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 $\mathbb{E}(\hat{p}) = p_{\text{true}}$.

3.8.2 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).

3.9 R

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)")))
```

Listing 3.8 `binconf.R`

```
## Binomial confidence intervals

# using BinomCI() function from the DescTools package
library(DescTools)
BinomCI(15, 22, method = "wald")            # Wald confidence interval
BinomCI(15, 22, method = "logit")           # logit-transformed Wald CI
BinomCI(15, 22, method = "wilson")          # score CI (default)
BinomCI(15, 22, method = "agresti-coull")   # Agresti-Coull CI
BinomCI(15, 22, method = "lik")             # likelihood ratio CI

# using binconf() function from the Hmisc package
library(Hmisc)
binconf(15, 22, method = "asymptotic")  # Wald CI
binconf(15, 22, method = "wilson")      # score CI (default)

# using prop.test in base R (stats package)
# Wilson confidence interval with continuity correction by default
# The continuity correction is not generally recommended. Like the exact CI,
# it can be too wide and have a coverage probability greater than 1 - \alpha.
prop.test(15, 22)
names(prop.test(15, 22))
prop.test(15, 22, correct = FALSE)      # score CI

# using binom.test (exact confidence interval)
binom.test(15, 22)    # same as binconf with method = "exact"
names(binom.test(15, 22))

# changing the confidence level (1 - alpha) to 80%
# All are score (Wilson) confidence intervals by default.
BinomCI(15, 22, conf.level = 0.8)
binconf(15, 22, alpha = 0.2)
prop.test(15, 22, conf.level = 0.8, correct = FALSE)

# writing a function to get Wald confidence limits
bconf_wald <- function(x, n, level=0.95) {
  # x is number of successes out of n trials
  p_hat <- x / n
  alpha <- 1 - level
  pvar <- p_hat * (1 - p_hat) / n
  p_int <- p_hat + c(-1, 1) * qnorm(1 - alpha / 2) * sqrt(pvar)

  # return named vector (names do not need quotes)
  return(c(point = p_hat, lower = p_int[1], upper = p_int[2]))
}
bconf_wald(15, 22)
bconf_wald(15, 22, level = 0.80)
```

Listing 3.9 `binomial-small.R`

```r
## Small-sample binomial point and interval estimates

# median unbiased estimate
medp_binom <- function(k, n) {
  # k = number of successes, n = number of trials

  # binomial lower tail probability
  lower_tail <- function(p) pbinom(k, n, p) - dbinom(k, n, p) / 2

  # median unbiased estimate
  med <- uniroot(function(p) lower_tail(p) - 1 / 2, interval = c(0, 1))
  med$root
}
medp_binom(15, 22)


# exact (Clopper-Pearson) confidence intervals
binom.test(15, 22)                              # base R (stats)
names(binom.test(15, 22))
binom.test(15, 22, conf.level = 0.8)

library(Hmisc)
binconf(15, 22, method = "exact")
binconf(15, 22, method = "exact", alpha = 0.2)

library(DescTools)
BinomCI(15, 22, method = "clopper-pearson")     # exact CI
BinomCI(15, 22, method = "midp")                # mid-p exact CI
```

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]

In **Bayesian** inference, probability distributions are used to summarize our knowledge about the possible values of an unknown parameter θ_{true}. The distribution of possible values of θ_{true} before we have seen our data is called the **prior distribution**, and the distribution of possible values of θ_{true} after we see the data is called the **posterior distribution**. Most parameters we are interested in estimating (such as probabilities) are continuous, so they have a probability density function (PDF) instead of a probability mass function (PMF). The posterior PDF is proportional to the prior PDF times the likelihood function, and the posterior distribution can be used to get point and interval estimates of an unknown parameter. The interval estimates are called **credible intervals**, and they have important advantages over confidence intervals. The Bayesian approach to statistics is more logically consistent and more intuitive than the frequentist approach, but it can be more computationally complex. While large-sample theory can be useful in Bayesian inference, it does not rely on asymptotic normality to the same degree that maximum likelihood estimation does.

4.1 Prior and posterior distributions

The value of a PDF is not a probability (it can be greater than one), but PDFs can be handled like probabilities in terms of the addition rule and the multiplication of conditional probabilities Boos and Stefanski (2013). Let $\pi(\theta)$ be the **prior** PDF of θ. Before we see our data, we believe that $\theta_{\text{true}} \in [a, b]$ with probability

$$\Pr(a \leq \theta \leq b) = \int_a^b \pi(\theta)\,\mathrm{d}\theta.$$

[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.

This integral is the area under the graph of $\pi(\theta)$ between $\theta = a$ and $\theta = b$. For a given value of θ, the likelihood of our data is $L(\theta) = \pi(\text{data} \mid \theta)$. By Bayes' rule, the **posterior** PDF is

$$\pi(\theta \mid \text{data}) = \frac{L(\theta)\pi(\theta)}{\pi(\text{data})} \propto L(\theta)\pi(\theta).$$

where $\propto$ denotes "proportional to." Calculating $\pi(\text{data})$ is difficult and almost always unnecessary. As long as we can calculate $\pi(\theta \mid \text{data})$ up to a constant of proportionality, we can normalize it to ensure that we have a posterior PDF whose integral over $\mathbb{R}$ equals one. After we see our data, we believe that $\theta_{\text{true}} \in [a, b]$ with probability

$$\Pr(a \leq \theta \leq b \mid \text{data}) = \int_a^b \pi(\theta \mid \text{data})\, \mathrm{d}\theta.$$

Bayesian point and interval estimates of θ_{true} are based on the posterior PDF.

Bayesian methods are often simpler, easier to interpret, and more robust to small sample sizes than frequentist methods like maximum likelihood estimation. In the limit of a large sample size, Bayesian and frequentist methods almost always give equivalent results. The Bayesian approach is also valuable because it emphasizes estimation and the accumulation of knowledge rather than binary decisions (Tukey 1960). However, the adoption of Bayesian methods has been impeded by a historical lack of the computational power needed to use them and by the widespread hesitation to specify prior distributions among epidemiologists, statisticians, and other scientists. The first problem is largely solved, but the latter problem remains with us today.

A common approach to specifying a prior distribution is to use a **noninformative prior** that places few or no restrictions on the value of θ. This lets the data "speak for itself" at the price of ignoring existing knowledge about the underlying scientific question. The ability to incorporate an informative prior distribution in the Bayesian approach to statistical inference should be viewed as a feature, not a bug (Greenland 2006; Greenland and Poole 2013).

4.1.1 Posterior point and interval estimation

The mean, median, or mode of the posterior distribution of θ can be used as a point estimate of θ_{true}. We will use $\bar{\theta}$ to denote the posterior mean and $\tilde{\theta}$ to denote the posterior median. In large samples, the posterior distribution converges to a normal distribution with mean θ_{true} and variance $I(\theta_{\text{true}})^{-1}$ (Le Cam 1953; Gelman et al. 2013),[2] which is the same as the limiting normal distribution of the MLE $\hat{\theta}$. In this limit, the posterior mean, median, and mode are all equal to $\hat{\theta}$.

[2]This convergence follows from the *Laplace approximation* to the posterior distribution (Gelman et al. 2013). It occurs when the likelihood $L(\theta)$ has a continuous second derivative with respect to θ and θ_{true} is not on the boundary of the support of the prior distribution. These are similar to the regularity conditions for maximum likelihood estimation.

A $1-\alpha$ **credible interval** is an interval $[a, b]$ such that

$$\Pr(a \leq \theta \leq b \,|\, \text{data}) = \int_a^b \pi(\theta \,|\, \text{data})\, \mathrm{d}\theta = 1 - \alpha.$$

Given our data, we believe that $\theta_{\text{true}} \in [a, b]$ with probability $1 - \alpha$. There are many different ways that a credible interval can be defined. The one that is conceptually closest to a confidence interval is the **central posterior interval** or **equal-tailed interval**, where

$$\Pr(\theta < a \,|\, \text{data}) = \Pr(\theta > b \,|\, \text{data}) = \frac{\alpha}{2}.$$

Thus, the $1-\alpha$ confidence limits are the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution of θ_{true}. Unlike confidence intervals, credible intervals can accurately be interpreted as containing θ_{true} with probability $1 - \alpha$. Like confidence intervals, credible intervals are only reliable if the likelihood is approximately correct. An equal-tailed credible interval is guaranteed to contain the posterior median $\tilde{\theta}$. It will usually contain the posterior mean $\bar{\theta}$, but this is not guaranteed in small samples.

4.1.2 Bayesian interpretation of confidence intervals

Bayesian credible intervals actually have the properties that people intuitively but naively expect of frequentist confidence intervals. In finite samples, a $1 - \alpha$ equal-tailed credible interval and a $1-\alpha$ confidence interval will be similar when the posterior density $\pi(\theta \,|\, \text{data})$ is proportional to the likelihood $L(\theta) = \pi(\theta \,|\, \text{data})$. This occurs when $\pi(\theta)$ is constant, as in some uninformative priors. In the limit of a large sample, the credible interval and the confidence interval are nearly identical. In general, a frequentist confidence interval can be interpreted as an approximation to a Bayesian credible interval when there is a large sample and a prior distribution that is flat across the range of the confidence interval (Pratt 1965; Greenland and Poole 2013). However, credible intervals do not rely on large-sample approximations, and they are able to incorporate prior knowledge about the possible values of θ_{true}.

4.1.3 Posterior probability of H_0 and p-values

The idea of prior and posterior probabilities from Bayesian inference gives us a useful perspective on the interpretation of p-values, which is a source of much confusion in epidemiology (Diamond and Forrester 1983; Greenland et al. 2016; Baduashvili, Evans, and Cutler 2020). The most common error is to interpret the p-value as the posterior probability that H_0 is true. By Bayes' rule,

$$\Pr(H_0 \text{ true} \,|\, \text{data}) = \frac{\Pr(\text{data} \,|\, H_0 \text{ true}) \Pr(H_0 \text{ true})}{\Pr(\text{data})}.$$

The p-value is analogous to $\Pr(\text{data} \,|\, H_0 \text{ true})$, but the posterior probability $\Pr(H_0 \text{ true} \,|\, \text{data})$ depends on its prior probability $\Pr(H_0)$. It should take more data to convince us of a null

hypothesis that seemed very unlikely than to convince us of a null hypothesis that seemed very likely.

For a null hypothesis of the form $H_0 : \theta_{\text{true}} = \theta_0$ where θ ranges over an interval, it can be difficult to assign a prior probability to H_0. For any continuous distribution of θ, the probability that it takes any particular value is zero. One way around this difficulty is to assign a prior probability mass to the null value θ_0. A more difficult problem is to assign a prior probability to the alternative hypothesis H_1, which is often not clearly specified. This means that the posterior probability of the null hypothesis is often not clearly defined.

However, it is not difficult to calculate a lower bound on the probability that H_0 is true given the data (Edwards, Lindman, and Savage 1963; Berger and Sellke 1987). Let π_0 be the prior probability of H_0, and suppose we have a single alternative hypothesis $H_1 : \theta_{\text{true}} = \theta_1$ with prior probability $1 - \pi_0$. Then the posterior probability of H_0 is

$$\Pr(H_0 \,|\, \text{data}) = \frac{L(\theta_0)\pi_0}{L(\theta_0)\pi_0 + L(\theta_1)(1-\pi_0)} = \left(1 + \frac{1-\pi_0}{\pi_0}\frac{L(\theta_1)}{L(\theta_0)}\right)^{-1}.$$

Given the data, the posterior probability of H_0 is minimized if we happen to get the MLE $\hat{\theta} = \theta_1$, so

$$\Pr(H_0 \text{ true} \,|\, \text{data}) \geq \left(1 + \frac{1-\pi_0}{\pi_0}\frac{L(\hat{\theta})}{L(\theta_0)}\right)^{-1}.$$

From Wilk's theorem (Wilks 1938) for the likelihood ratio test, twice the log likelihood ratio has an approximate χ^2_1 distribution in large samples. To get a p-value of α from the likelihood ratio test, we need

$$2(\ell(\hat{\theta}) - \ell(\theta_0)) = z^2_{1-\frac{\alpha}{2}} \;\Rightarrow\; \frac{L(\hat{\theta})}{L(\theta_0)} = e^{\frac{1}{2}z^2_{1-\frac{\alpha}{2}}}.$$

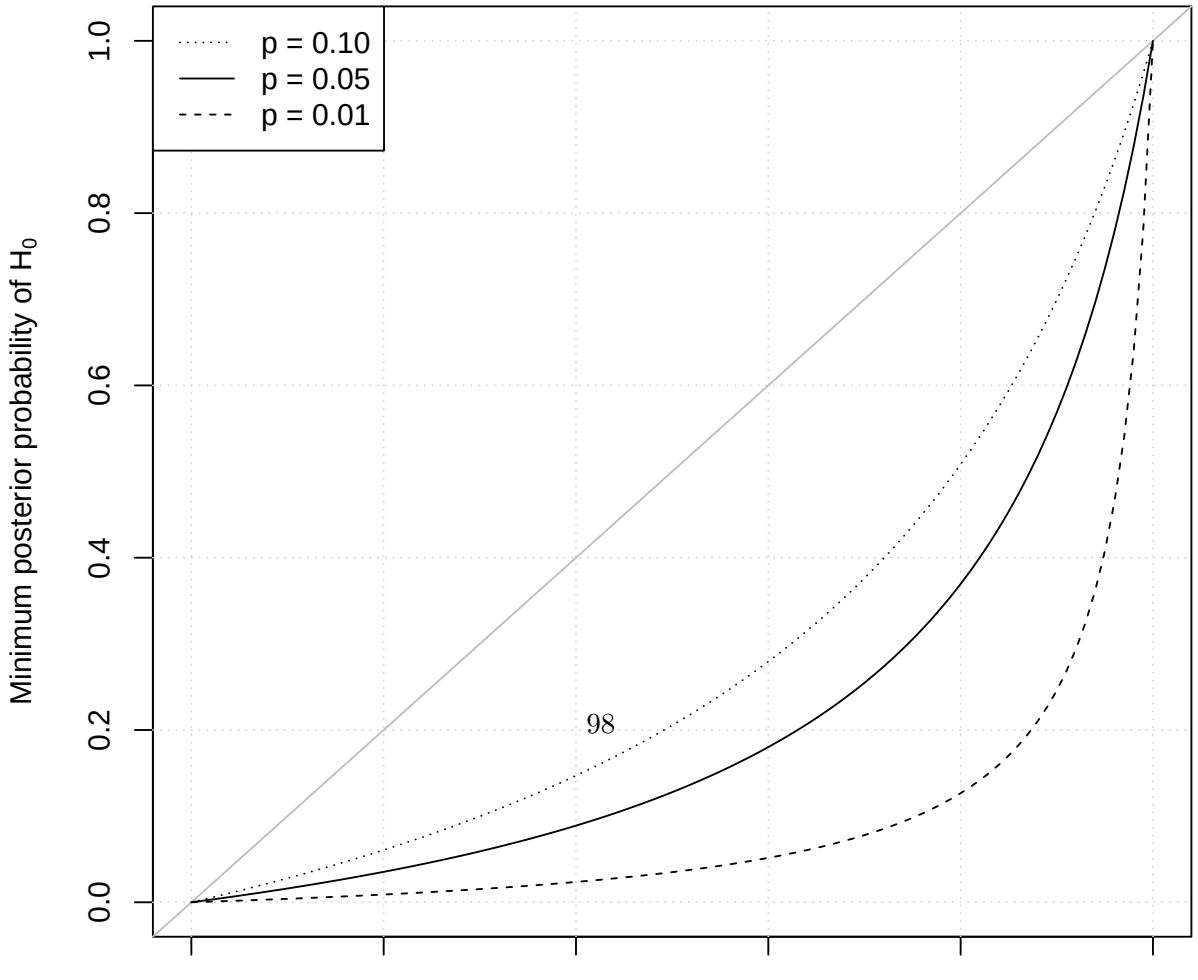
Figure 4.1 shows the minimum posterior probability of H_0 if $\pi_0 = 0.5$ for different p-values. The p-value is almost always much lower than the lower bound of the posterior probability of the null. For $\pi_0 = 0.5$, the lower bounds for $\Pr(H_0 \text{ true} \,|\, \text{data})$ are approximately 0.205 for $p = 0.10$, 0.128 for $p = 0.05$, and 0.035 for $p = 0.01$. In practice, the posterior probability of H_0 can be much larger than its lower bound (Berger and Sellke 1987).

Listing 4.1 `postH0.R`

```r
## Posterior probability of the null hypothesis (H0)

# function to calculate lower bound
lowerb <- function(pi0=0.5, pval=0.05) {
  # args: pi0 = prior probability of H0, pval = p-value
  # return: lower bound on posterior probability of H0
  z <- qnorm(1 - pval / 2)
  1 / (1 + (1 - pi0) / pi0 * exp(0.5 * z^2))
}

# plot of lower bounds for p-value = 0.01, 0.05, and 0.1
x <- seq(0, 1, by = .01)
plot(x, lowerb(x), type = "n", xlim = c(0, 1), ylim = c(0, 1),
     xlab = expression("Prior probability of H"[0]),
     ylab = expression("Minimum posterior probability of H"[0]))
grid()
abline(0, 1, col = "gray")
lines(x, lowerb(x))
lines(x, lowerb(x, pval = .01), lty = "dashed")
lines(x, lowerb(x, pval = .1), lty = "dotted")
legend("topleft", bg = "white", lty = c("dotted", "solid", "dashed"),
       legend = c("p = 0.10", "p = 0.05", "p = 0.01"))
```



4.2 Bayesian estimation of a probability

When estimating a probability p_{true}, our likelihood will be the binomial likelihood

$$L(p) = \binom{n}{k} p^k (1-p)^{n-x}$$

when k out of n samples equal one. Because we only need to calculate the likelihood up to a constant of proportionality, we can safely ignore the $\binom{n}{k}$ because it does not depend on p. The posterior distribution of p will be

$$\pi(p \,|\, \text{data}) \propto p^k (1-p)^{n-k} \pi(p)$$

where $\pi(p)$ is the prior distribution of p. When $\pi(p \,|\, \text{data})$ and $\pi(p)$ are from the same family of distributions, the prior $\pi(p)$ is said to be a **conjugate** distribution for the binomial likelihood. Conjugate distributions exist for many likelihoods used in epidemiology.

4.2.1 Beta distribution

The conjugate distribution for the binomial likelihood is the **beta distribution**. It has is a support on $[0, 1]$ (where p_{true} must live) with the PDF

$$f(x, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \propto x^{\alpha-1}(1-x)^{\beta-1}.$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters and $\Gamma(v)$ is the gamma function.[3] Because we will only need to calculate PDFs up to a mutiplicative constant, the gamma function term can be safely ignored. If $X \sim \text{beta}(\alpha, \beta)$, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

and

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The beta distribution has a number of important special cases:

- When $\alpha = \beta = 1$, it is a uniform$(0, 1)$ distribution.
- The beta distribution with $\alpha = \beta = 1/2$ is the **Jeffreys prior** for the binomial likelihood.[4] A Jeffreys prior is proportional to the square root of the expected information $\mathcal{J}_1(\theta)$ of a single observation(Jeffreys 1946). They are widely used as noninformative priors. For the binomial likelihood, $\mathcal{J}_1(p) = \frac{1}{p(1-p)}$.

[3] The gamma function is $\Gamma(v) = \int_0^\infty y^{v-1} e^{-y} \mathrm{d}y$. It is used to define the gamma distribution, of which the chi-squared distributions (including χ_1^2) are special cases. If v is a positive integer, then $\Gamma(v) = (v-1)!$.

[4] Harold Jeffreys (1891–1989) was an English mathematician, statistician, geophysicist, and astronomer. He helped revive the Bayesian notion of probability as an expression of our knowledge about an unknown quantity, wrote a classic textbook on mathematical physics with his wife Bertha (also a mathematician and physicist), and was a prominent opponent of the theory of plate tectonics.

4.2.2 Posterior point and interval estimates

If the prior distribution of p is a beta(α, β) distribution, then

$$\begin{aligned}\pi(p \,|\, \text{data}) &\propto p^k(1-p)^{n-k} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\end{aligned}$$

so the posterior distribution of p is a beta$(k+\alpha, n-k+\beta)$ distribution. The posterior mean is

$$\bar{p} = \frac{k+\alpha}{(k+\alpha)+(n-k+\beta)} = \frac{k+\alpha}{n+\alpha+\beta},$$

and the posterior variance is

$$\frac{(k+\alpha)(n-k+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)} = \frac{\bar{p}(1-\bar{p})}{n+\alpha+\beta+1}.$$

The endpoints of the $1-\alpha$ central credible interval are the $\alpha/2$ and $1-\alpha/2$ quantiles of the beta$(k+\alpha, n-k+\beta)$ distribution. Figure 4.2 shows prior and posterior distributions for 15 successes out of 22 trials. Although the prior distributions are quite different, the posterior distributions are quite similar. With large samples, the prior distribution disappears into the likelihood.

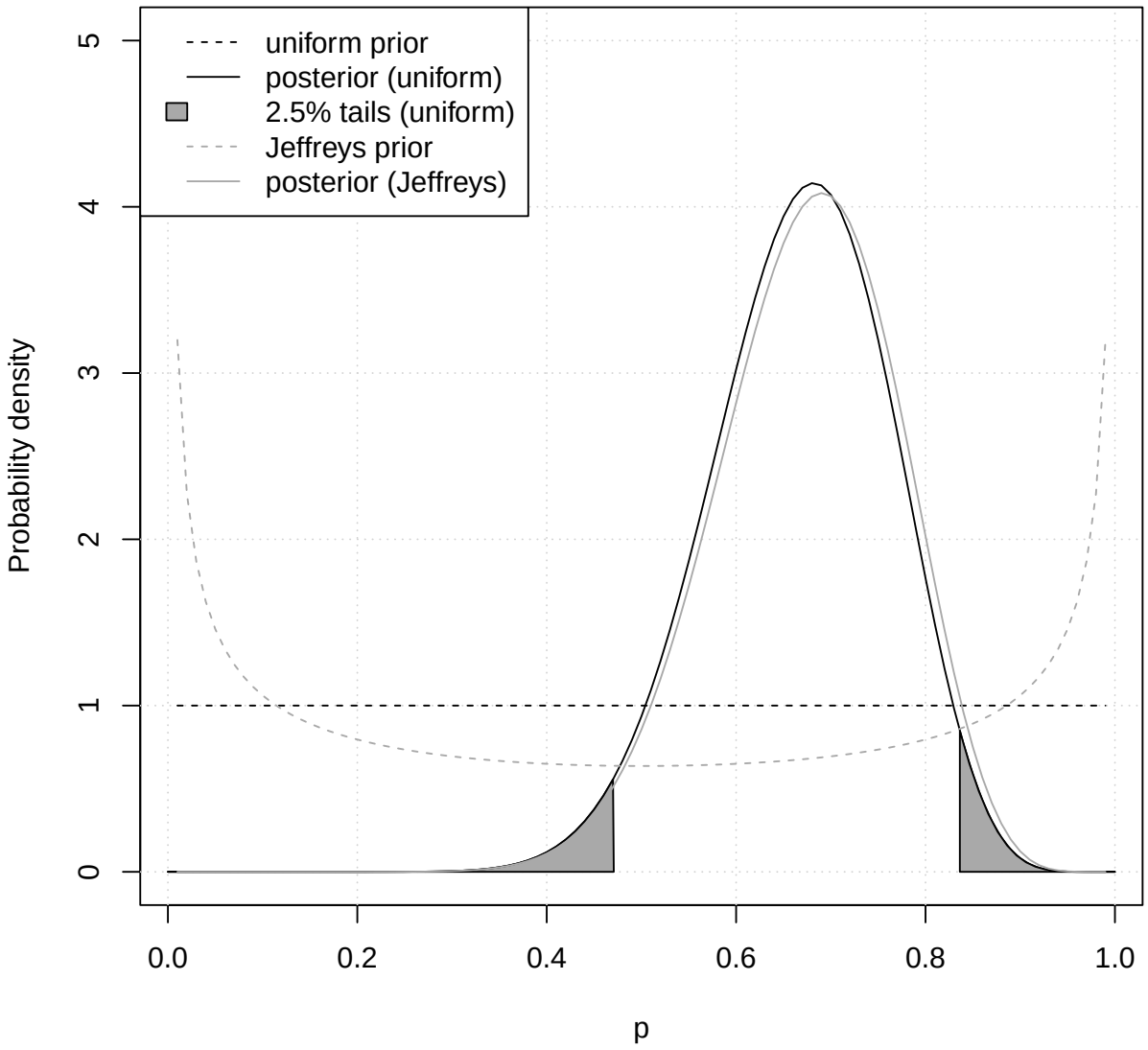


Figure 4.2: Prior and posterior distributions for p after observing 15 successes out of 22 trials. The uniform prior and the resulting posterior distribution are shown in black with the 2.5% tails shaded. The central 95% credible interval includes all values of p between the two tails. The Jeffreys prior and the corresponding posterior distribution are shown in dark gray.

As $n \to \infty$, we have $\bar{p} - \hat{p} \to 0$ and

$$\frac{\bar{p}(1-\bar{p})}{n+\alpha+\beta+1} - \frac{\hat{p}(1-\hat{p})}{n} \to 0. \tag{4.1}$$

Thus, the posterior distribution approaches approximate normal distribution of the maximum

Table 4.1: 95% confidence intervals for the sensitivity, specificity, PPV, and NPV of diabetes test in Remein and Wilkerson (1961).

	Sensitivity	Specificity	PPV	NPV
95% CI type	$55/70 \approx 0.786$	$462/510 \approx 0.906$	$55/103 \approx 0.534$	$462/477 \approx 0.969$
Wald	(0.690, 0.882)	(0.881, 0.931)	(0.438, 0.630)	(0.953, 0.984)
Logit	(0.674, 0.866)	(0.877, 0.928)	(0.438, 0.628)	(0.949, 0.981)
Agresti-Coull	(0.675, 0.867)	(0.877, 0.928)	(0.438, 0.627)	(0.948, 0.981)
Score	(0.676, 0.866)	(0.877, 0.928)	(0.438, 0.627)	(0.949, 0.981)
Likelihood ratio	(0.680, 0.871)	(0.879, 0.929)	(0.438, 0.629)	(0.950, 0.982)
Exact	(0.671, 0.875)	(0.877, 0.930)	(0.433, 0.633)	(0.949, 0.982)
Mid-p	(0.678, 0.870)	(0.878, 0.929)	(0.437, 0.629)	(0.950, 0.982)
Jeffreys	(0.679, 0.869)	(0.878, 0.929)	(0.438, 0.628)	(0.950, 0.982)

likelihood estimate $\hat{p}$ in large samples. However, the Bayesian posterior distribution is valid for any sample size, not just large samples.

4.2.3 Jeffreys confidence interval

The $1-\alpha$ Jeffreys confidence interval is the central credible interval with a beta$(1/2, 1/2)$ prior, which is the Jeffreys prior for a binomial model. When $k > 0$ and $k < n$, its endpoints are the $\alpha/2$ and $1-\alpha/2$ quantiles of the beta$(k+1/2, n-k+1/2)$ posterior distribution that we get if we see k successes in n trials. When $k = 0$, the lower endpoint is 0. When $k = n$, the upper endpoint is 1. For a binomial proportion, the Jeffreys confidence interval has width and coverage probability similar to the score (Wilson) and mid-p exact confidence intervals (Brown, Cai, and DasGupta 2001).

4.3 Comparison of binomial confidence intervals

Table 4.1 shows eight types of confidence intervals for a probability that we have discussed. These are calculated for the sensitivity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) of the diabetes test one hour after the meal, where a positive test was defined as a blood glucose concentration above 130 mg/dL.

Because of their good coverage probabilities and narrow widths, the score (Wilson), likelihood ratio, and Jeffreys intervals are recommended by Agresti and Coull (1998) and Brown, Cai, and DasGupta (2001). The Agresti-Coull intervals are slightly wider than these intervals, but they have slightly higher coverage probabilities and are simpler to calculate. Mid-p exact confidence intervals are similar to the Jeffreys intervals, but they are more difficult to calculate. Exact intervals are too wide, and the Wald intervals have coverage probabilities that are often too

low. The logit-transformed Wald interval has good coverage probabilities, but it can be even wider than the exact interval. In large samples (with larger samples required for probabilities near zero or one), all of the intervals are similar.

4.4 R

Listing 4.2 `normplots.R`

```
## Bayesian prior and posterior distributions for a probability

# beta prior and posterior distributions
# The prior is beta(1, 1), which is the uniform(0, 1) distribution.
# Because k = 15 and n = 22, the posterior is beta(15 + 1, 22 - 15 + 1)
p <- seq(0.01, 0.99, by = 0.01)
plot(p, dbeta(p, 16, 8), type = "n", ylim = c(0, 5),
     xlab = "p", ylab = "Probability density")
grid()
lines(p, dbeta(p, 1, 1), lty = "dashed")  # prior PDF
lines(p, dbeta(p, 16, 8))                 # posterior PDF
postlower <- qbeta(0.025, 16, 8)          # 95% credible interval lower bound
postupper <- qbeta(0.975, 16, 8)          # 95% credible interval upper bound
polygon(c(0, seq(0, postlower, by = 0.01), postlower),
        c(0, dbeta(seq(0, postlower, by = 0.01), 16, 8), 0),
        col = "darkgray")
polygon(c(postupper, seq(postupper, 1, by = 0.01), 1),
        c(0, dbeta(seq(postupper, 1, by = 0.01), 16, 8), 0),
        col = "darkgray")
legend("topleft", bg = "white",
       lty = c("dashed", "solid", NA, "dashed", "solid"),
       col = c("black", "black", NA, "darkgray", "darkgray"),
       fill = c(NA, NA, "darkgray", NA, NA),
       border = c(NA, NA, "black", NA, NA),
       legend = c("uniform prior", "posterior (uniform)",
                  "2.5% tails (uniform)", "Jeffreys prior",
                  "posterior (Jeffreys)"))
lines(p, dbeta(p, 0.5, 0.5), lty = "dashed", col = "darkgray")
lines(p, dbeta(p, 15.5, 7.5), col = "darkgray")
```

Listing 4.3 `binconf-table.R`

```r
## Comparison of binomial confidence limits
library(DescTools)

# Sensitivity (55 / 70)
# Use round() to get rounded numbers shown in the table.
round(BinomCI(55, 70, method = "wald"), 3)
round(BinomCI(55, 70, method = "logit"), 3)
round(BinomCI(55, 70, method = "agresti-coull"), 3)
round(BinomCI(55, 70, method = "wilson"), 3)
round(BinomCI(55, 70, method = "lik"), 3)
round(BinomCI(55, 70, method = "clopper-pearson"), 3)
round(BinomCI(55, 70, method = "midp"), 3)
round(BinomCI(55, 70, method = "jeffreys"), 3)

# Specificity (462 / 510)
round(BinomCI(462, 510, method = "wald"), 3)
round(BinomCI(462, 510, method = "logit"), 3)
round(BinomCI(462, 510, method = "agresti-coull"), 3)
round(BinomCI(462, 510, method = "wilson"), 3)
round(BinomCI(462, 510, method = "lik"), 3)
round(BinomCI(462, 510, method = "clopper-pearson"), 3)
round(BinomCI(462, 510, method = "midp"), 3)
round(BinomCI(462, 510, method = "jeffreys"), 3)

# PPV (55 / 103)
round(BinomCI(55, 103, method = "wald"), 3)
round(BinomCI(55, 103, method = "logit"), 3)
round(BinomCI(55, 103, method = "agresti-coull"), 3)
round(BinomCI(55, 103, method = "wilson"), 3)
round(BinomCI(55, 103, method = "lik"), 3)
round(BinomCI(55, 103, method = "clopper-pearson"), 3)
round(BinomCI(55, 103, method = "midp"), 3)
round(BinomCI(55, 103, method = "jeffreys"), 3)

# NPV (462 / 477)
round(BinomCI(462, 477, method = "wald"), 3)
round(BinomCI(462, 477, method = "logit"), 3)
round(BinomCI(462, 477, method = "agresti-coull"), 3)
round(BinomCI(462, 477, method = "wilson"), 3)
round(BinomCI(462, 477, method = "lik"), 3)
round(BinomCI(462, 477, method = "clopper-pearson"), 3)
round(BinomCI(462, 477, method = "midp"), 3)
round(BinomCI(462, 477, method = "jeffreys"), 3)
```

5 Longitudinal Data, Rates, and Counts

> The *mortality* and *force of mortality* will readily be distinguished, by comparing cholera with consumption; the *mortality* of the latter is 90–100 per cent, but its mean duration is two years, and the *force of mortality* is consequently nearly 0.50; the mortality in cholera is not 50 per cent, while the force of mortality is 2415, for cholera destroys in a week as many as phthisis consumes in a year. Phthisis is more dangerous than cholera; but cholera, probably, excites the greatest terror. (Farr 1838)[1]

The meaning of a risk or cumulative incidence depends on the time interval over which it is observed. William Farr observed that a risk of death slightly less than 50% over one week for cholera patients was widely considered more alarming than a risk of death greater than 90% over two years for tuberculosis patients. In public health, it matters a great deal how quickly things happen. How rapidly an event occurs is measured using a **rate**, which has units of events per unit time (Elandt-Johnson 1975; Morgenstern, Kleinbaum, and Kupper 1980). To estimate a risk or a rate, we use **longitudinal data** where individuals are followed over time. The analysis of longitudinal data is complicated by the fact that individuals can come and go during the study period.

To estimate the risk of disease onset in a time interval $(t_a, t_b]$, we would follow individuals from time t_a to time t_b to ascertain disease onset. From Section 1.9.1, the risk of disease onset in the time interval $(t_a, t_b]$ is

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

where

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

and t^{onset} denotes the onset time of the disease. In practice, the population Ω is often defined to consist only of individuals who are at risk of disease at time t_a. If we have complete follow-up of the entire population Ω over the entire interval $(t_a, t_b]$, then we know the cumulative incidence exactly.

[1] William Farr (1807–1883) was a British pioneer of epidemiology. As the first statistician in the General Register Office, he was responsible for the collection of medical statistics in England and Wales He set up a system for recording causes of death that allowed comparison of mortality rates. In the quote, "consumption" and "phthisis" both refer to tuberculosis.

More often, we follow a sample of individuals from the population. If selection into the sample is independent of disease onset during $(t_a, t_b]$ and we have complete follow-up over the entire interval, then we can get point and interval estimates of the true cumulative incidence using methods for a binomial proportion from Chapter 3 or Chapter 4. In practice, almost all longitudinal studies have individuals entering or leaving the study during the follow-up period. When this occurs, methods for binomial proportions can produce inefficient or biased estimates of risk.

The analysis of incomplete longitudinal data is called **survival analysis**, and it is the theoretical foundation for many epidemiologic methods. To analyze survival data, it is important to have clear definitions of the following that can be applied equally to all study participants:

- The **time origin**: This is the beginning of the time-to-event, and it defines the time scale used in the analysis. It could be a particular calendar time, a particular age, or the occurrence of a particular event (see Table 5.1).
- The **failure time** or **event time**: This is the end of the time-to-event. Both the outcome and its occurrence time need careful but practical operational definitions that can be applied equally to all participants in the study.
- The **observation process** and **at-risk process**: For their disease onset time to be observed, individual i must be both at risk of the outcome and under observation at t_i^{onset}. The observation and at-risk processes are required to be *predictable*, which means that their value at any time t is determined just before time t.[2] For example, if observation begins at time t, you are under observation just after time t but not at time t itself. If observation ends at time t, then you are under observation up to and including time t.
- The **entry time** t_i^{entry} is the earliest time at which the individual i is both at risk of the outcome and under observation. If the entry time occurs after the origin, we have **delayed entry**, which is also called **left truncation**.
- The **exit time** or **follow-up time** t_i is the last time the individual is both at risk of the outcome and under observation. If the exit time occurs before the failure time, we have **right censoring** (loss to follow-up).

In complex data, an individual can have multiple entry, exit, and failure times and even multiple time origins (e.g., time to heart attack after vigorous exercise or ingestion of cocaine). The time scale is usually defined so that $t = 0$ at the origin time, and we will assume that the origin and population are defined so that all individuals in the population are at risk of the event at $t = 0$.

[2] To be predictable, it is sufficient for a process to be left-continuous. The disease onset process is assumed to be right-continuous with left-hand limits (cadlag) and hence unpredictable.

Table 5.1: Time scales and time origins adapted from Clayton and Hills (1993).

Time scale	Origin
Calendar time	Fixed date
Time since exposure	Exposure time
Time under treatment	Start of treatment
Time since diagnosis	Time of diagnosis
Age	e.g., 65th birthday
Time in hospital	Hospital admission

5.1 Incomplete follow-up

Let t^{event} be the failure time of individual i and t^{cens} be the last time at which they would be under observation if they had no disease onset. We assume all times are defined so that $t = 0$ at the time origin.

5.1.1 Right censoring

When $t_i^{\text{cens}} < t_i^{\text{event}}$, the outcome in person i is **right censored** at time t_i^{cens}. We know that individual i does not have an event at or before t_i^{cens}, but we do not know when or if the event will occur after that. In right-censored data, we see the exit time

$$t_i = \min(t_i^{\text{event}}, t_i^{\text{cens}})$$

and the **event indicator**

$$\delta_i = \begin{cases} 0 & \text{if } t_i^{\text{exit}} = t_i^{\text{cens}}, \\ 1 & \text{if } t_i^{\text{exit}} = t_i^{\text{event}}. \end{cases}$$

Right censoring is often just called *censoring*. There are many reasons that person i might be censored: Perhaps we can no longer find person i (loss to follow-up), person i is no longer at risk of failure (e.g., a woman who has a hysterectomy is no longer at risk of uterine cancer), or observation ends. Different individuals in the same study can be censored for different reasons.

Independent right censoring occurs if those who remain under observation at any given time are a random sample of all of those who would be under observation if there were no right censoring. Censoring is not independent if those who remain under observation have systematically different failure times than those who are censored. There are three canonical types of independent censoring:

- *Type I censoring* occurs when observation of each individual ends at a predetermined time under the control of the investigators.

- *Type II censoring* occurs when observation ends after a predetermined number of events have been observed.
- *Random censoring* occurs when each person i has a random censoring time t_i^{cens} that is independent of his or her failure time t_i^{event} and not under the control of the investigators.

As long as all censoring is independent, different censoring mechanisms can occur within the same study. All of the methods we will discuss assume independent censoring, and they become biased under *dependent* or *informative* censoring.

5.1.2 Delayed entry (left truncation)

Whereas right censoring concerns the observation of failure times, truncation concerns the selection of study participants. **Delayed entry** or **left truncation** occurs when an individual i has an entry time $t_i^{\text{entry}} > 0$ where $t = 0$ denotes the origin. If person i had disease onset before t_i^{entry}, he or she would have been excluded from the study. Person-time prior to entry cannot be included as time at risk; it is called **immortal person-time**. Delayed entry can be handled easily by all of the methods we will discuss. As with independent right censoring, we require that the set of individuals under observation are a random sample of the individuals who would be under observation if we had no right censoring or left truncation. You can think of this as **independent left truncation**

5.1.3 Left censoring and right truncation

Given that we have right censoring and left truncation (delayed entry), it is natural to wonder whether there is also left censoring and right truncation. Unfortunately, the answer is yes:

- *Left censoring* means that we know an individual had an event before a left-censoring time but we do not know when.
- *Right truncation* means that our sample contains only individuals who have already had the event.

Both of these must be handled using strong assumptions or specialized methods. Some areas of epidemiology must contend with left censoring (e.g., we know that a birth defect occurred prior to birth but not exactly when) or right truncation. Usually, they can be prevented or minimized by good study design.

5.2 Failure time distributions

Failure time distributions are descibed in terms of survival functions, cumulative hazard functions, and hazard functions. Any one of these is sufficient to specify the distribution of the

time to an event, but each has a different and useful interpretation. We denote the failure time as a positive random variable T. For simplicity, we will assume that T is continuous.

5.2.1 Survival function

The **survival function** for a failure time T is

$$S(t) = \Pr(T > t),$$

which is the probability that the event does not occur up to and including time t. Several properties follow from this definition:

- Because it is a probability, $S(t) \in [0, 1]$ for all t.
- Because $T > 0$, $S(0) = 1$.
- $S(t)$ is monotonically decreasing, which means that it cannot increase. If $u > t$, you can survive to time u only if you survive to time t, so $S(t) \geq S(u)$.

The survival function $S(t)$ tells us everything there is to know about the distribution of the time-to-event T. The **cumulative incidence function** is

$$F(t) = 1 - S(t) = \Pr(T \leq t),$$

which is also the cumulative distribution function (CDF) for T. If T is a continuous random variable, then

$$-S'(t) = F'(t) = f(t), \tag{5.1}$$

where $f(t)$ is the probability density function (PDF) of T. If T is a discrete or continuous failure time with survival function $S(t)$, it turns out that

$$\mathbb{E}[T] = \int_0^\infty S(t)\,\mathrm{d}t. \tag{5.2}$$

This is often easier to calculate than the integral $\mathbb{E}[T] = \int_0^t tf(t)\mathrm{d}t$ that is normally used to define the mean of a positive random variable.[3] The pth quantile of the failure time distribution is the solution to the equation

$$p = F(t_p) = 1 - S(t_p), \tag{5.3}$$

[3] Using integration by parts, we get that

$$\int_0^\infty tf(t)\,\mathrm{d}t = -tS(t)\Big|_0^\infty + \int_0^\infty S(t)\,\mathrm{d}t = \int_0^\infty S(t)\,\mathrm{d}t$$

when $tS(t) \to 0$ as $t \to \infty$.

so higher quantiles correspond to longer times to events.[4] When $p = 0.5$, we get the median time-to-event. Much of survival analysis is dedicated to calculating and comparing survival functions.

5.2.2 Hazard function

For any $\Delta > 0$, the probability that you have an event in the time interval $(t, t + \Delta]$ is

$$\Pr(\text{event in } (t, t + \Delta]) = S(t) - S(t + \Delta)$$

and the conditional probability that you have an event in the interval given that you survived until time t is

$$\Pr(\text{event in } (t, t + \Delta] \,|\, \text{survival until } t) = \frac{S(t) - S(t + \Delta)}{S(t)}. \tag{5.4}$$

The numerator is the expected number of events in $(t, t + \Delta]$ given that you remain at risk at time t. Dividing by Δ gives us

$$\frac{S(t) - S(t + \Delta)}{S(t)\Delta},$$

which is the expected number of events per unit time in the interval $(t, t + \Delta]$. The **hazard function** is the limit of this expected number of events per unit time as $\Delta \downarrow 0$ (i.e., as Δ decreases to zero):

$$h(t) = \lim_{\Delta \downarrow 0} \frac{S(t) - S(t + \Delta)}{S(t)\Delta}. \tag{5.5}$$

Because the numerator is nonnegative and the denominator is positive, $h(t) \geq 0$. The hazard function $h(t)$ measures the instantaneous expected number of events per unit time like a speedometer measures the instantaneous speed of a vehicle. When $h(t)$ is high, you are likely to have an event soon after time t if you have survived event-free until t. When $h(t)$ is low, you are relatively unlikely to have an event soon after t.

Unlike the survival function, the hazard function has units. Because we divide an expected number of events by a time interval Δ, the hazard has units of events/time. Just like the same speed can be expressed in miles per hour or kilometers per hour, using different measures of time (e.g., month, week, day, hour, minute, or second) changes the numerical value of the hazard but not its meaning.

When Δ is small, then Equation 5.4 and Equation 5.5 give us

$$h(t) \approx \frac{\Pr(\text{event in } (t, t + \Delta] \,|\, \text{survival until } t)}{\Delta}.$$

[4] When T is discrete (or a mixture of continuous and discrete components),

$$t_p = \inf\{t : F(t) \geq p\} = \inf\{t : S(t) \leq 1 - p\}.$$

When T is continuous, this reduces to Equation 5.3.

Rearranging, we get

$$h(t)\Delta \approx \Pr(\text{event in } (t, t+\Delta] \,\big|\, \text{survival until } t).$$

Notice how the units work:

$$h(t)\frac{\text{events}}{\text{time unit}}\Delta \text{ time units} = h(t)\Delta \text{ events}. \tag{5.6}$$

Multiplying the hazard by a time interval gives the expected number of events that would occur in that interval if the hazard remained constant.

When times to failure are continuous, the survival function is differentiable. By definition of the derivative of $S(t)$,

$$\lim_{\Delta \downarrow 0} \frac{S(t) - S(t+\Delta)}{\Delta} = -S'(t).$$

Putting this back into Equation 5.5, we get

$$h(t) = \frac{1}{S(t)} \lim_{\Delta \downarrow 0} \frac{S(t) - S(t+\Delta)}{\Delta} = \frac{-S'(t)}{S(t)}. \tag{5.7}$$

Because $-S'(t) = f(t)$ from Equation 5.1, we can multiply both sides by $S(t)$ to get

$$f(t) = h(t)S(t). \tag{5.8}$$

Thus, the PDF is the product of the hazard and survival functions. This is used to write likelihoods for right-censored and left-truncated data.

5.2.3 Cumulative hazard function

The **cumulative hazard function** for a positive random variable T is

$$H(t) = -\ln S(t). \tag{5.9}$$

Several properties follow from this definition:

- $H(0) = 0$ because $S(0) = 1$.
- $H(t)$ is monotonically increasing, which means that it cannot decrease. If $u > t$, then $H(u) \geq H(t)$.
- When $S(t) > 0$, $H(t) \in [0, \infty)$.
- $S(t) = e^{-H(t)}$.

Taking the derivative with respect to t on both sides of Equation 5.9 (using the chain rule on the right-hand side), we get

$$H'(t) = \frac{-S'(t)}{S(t)} = h(t)$$

where the final equality follows from Equation 5.7. By the fundamental theorem of calculus and the fact that $H(0) = 0$,

$$H(t) = \int_0^t h(u)\mathrm{d}u \tag{5.10}$$

so the cumulative hazard $H(t)$ is the area under the graph of h over $(0, t)$.

Equation 5.10 gives us an interesting way to interpret the cumulative hazard function. If the event is something that can be repeated (e.g., clicks on a Geiger counter), the expected number of events in $(0, t]$ is the sum of the expected numbers of events in a series of intervals $(u_0, u_1], (u_1, u_2], \ldots, (u_{n-1}, u_n]$ where $u_0 = 0$ and $u_n = t$. Taking a limit as the number of subintervals grows larger and each subinterval becomes smaller,

$$\mathbb{E}[\text{number of events in} (0, t]] = \int_0^t h(u)\,\mathrm{d}u = H(t).$$

The units in the integral work in the same way as in Equation 5.6. For an event that cannot be repeated (e.g., death), $H(t)$ can be interpreted as the expected number of events that would occur if the event were made repeatable. After an event at time t, you would be brought back to being at risk at time t to wait for the next event to occur.

5.2.4 Likelihoods for right-censored and left-truncated data

Suppose our survival time distribution has hazard function $h(t, \theta_{\text{true}})$ and survival function $S(t, \theta_{\text{true}})$, where θ_{true} is an unknown parameter (or parameter vector). If individual i has entry time t_i^{entry}, exit time t_i, and event indicator δ_i, then their likelihood contribution is

$$L_i(\theta) = \frac{h(t_i, \theta)^{\delta_i} S(t_i, \theta)}{S(t_i^{\text{entry}}, \theta)}. \tag{5.11}$$

In the numerator, every observation contributes a survival term, but only the observed failure times ($\delta_i = 1$) contribute a hazard term. The survival term in the denominator accounts for the fact that they survived until time t_i^{entry}. When $t_i^{\text{entry}} = 0$, then the denominator equals one. Table 5.2 shows the likelihood contributions for all four possible combinations of right censoring and delayed entry (left truncation).

Taking logarithms on both sides of Equation 5.11 gives us the log likelihood contribution

$$\ell_i(\theta) = \delta_i \ln h(t_i, \theta) - [H(t_i, \theta) - H(t_i^{\text{entry}}, \theta)].$$

Table 5.2: Possible likelihood contributions for $(t_i^{\text{entry}}, t_i, \delta_i)$.

	Right censoring ($\delta_i = 0$)	No right censoring ($\delta_i = 1$)
Delayed entry ($t^{\text{entry}} > 0$)	$\frac{S(t_i,\theta)}{S(t_i^{\text{entry}},\theta)}$	$\frac{h(t_i,\theta)S(t_i,\theta)}{S(t_i^{\text{entry}},\theta)}$
No delayed entry ($t^{\text{entry}} = 0$)	$S(t_i,\theta)$	$h(t_i,\theta)S(t_i,\theta)$

All observations contribute a cumulative hazard term, but only observed event times contribute a hazard term. When $t_i^{\text{entry}} = 0$, then $H(0,\theta) = 0$ for all θ. When we have n independent observations, the log likelihood is the sum of the individual log likelihood contributions:

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

The maximum likelihood estimate $\hat{\theta}$ can be found using the score function $U(\theta) = \ell'(\theta)$, and its approximate variance is $I(\hat{\theta})^{-1} = [-\ell''(\hat{\theta})]^{-1}$. The log likelihood can be used to do Wald, score, and likelihood ratio hypothesis tests and obtain the corresponding confidence intervals. For example, the Wald 95% confidence interval in large samples is

$$\hat{\theta} \pm 1.96\sqrt{I(\hat{\theta})^{-1}}.$$

If necessary, the delta method can be used to get confidence intervals that keep $\hat{\theta}$ within an appropriate range of values. Bayesian estimation can be done using the corresponding likelihood $L(\theta) = \exp(\ell(\theta))$

5.3 Exponential distribution

The exponential distribution is the simplest and most important failure time distribution. It has a constant hazard function

$$h(t) = \lambda$$

where λ is called the **rate parameter** and has units of events/time. Its cumulative hazard function is

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = \int_0^t \lambda \mathrm{d}u = \lambda t,$$

its the survival function is

$$S(t) = e^{-H(t)} = e^{-\lambda t}.$$

A higher rate parameter λ implies shorter survival times. The **scale parameter** is $\sigma = \lambda^{-1}$. A smaller scale parameter corresponds to shorter survival times.

5.3.1 Mean and variance

The mean of an exponential random variable T is found most easily by integrating the survival function:

$$\mathbb{E}(T) = \int_0^\infty e^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}\int_0^\infty \lambda e^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}, \tag{5.12}$$

In that integral, we "creatively multiplied by one'' to turn the integrand into a exponential PDF and then used the fact that the total area under the PDF is one. Integration by parts can be used to show that

$$\mathbb{E}(T^2) = \int_0^\infty t^2 \lambda e^{-\lambda t}\,\mathrm{d}t = \frac{2}{\lambda^2},$$

so

$$\mathrm{Var}(T) = \mathbb{E}(T^2) - \mathbb{E}(T)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

by Equation 1.22. Thus, the standard deviation $\sqrt{\mathrm{Var}(T)} = 1/\lambda$ is equal to the mean.

5.3.2 Incidence rates

Now imagine that we want to estimate an unknown rate parameter λ_{true} for an exponential distribution. Let $(t_1^{\text{entry}}, t_1, \delta_1), \ldots (t_n^{\text{entry}}, t_n, \delta_n)$ denote a set of entry times, exit times, and event indicators in a sample of size n. The likelihood is

$$L(\lambda) = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda(t_i - t_i^{\text{entry}})},$$

and the log likelihood is

$$\ell(\lambda) = \sum_{i=1}^n \left(\delta_i \ln\lambda - \lambda(t_i - t_i^{\text{entry}})\right) = m\ln\lambda - \lambda T, \tag{5.13}$$

where

$$m = \sum_{i=1}^n \delta_i$$

is the total number of observed events and

$$T = \sum_{i=1}^n (t_i - t_i^{\text{entry}})$$

is the total **person-time**. This is the total time under observation added up over all participants in the study. The likelihood assumes that the same rate λ_{true} of events per unit time occurs in all of this person-time.

Maximum likelihood estimation of the rate λ_{true} proceeds in the same way as it did for a probability. The only difference is that we are working with an exponential likelihood for

times to events rather than a binomial likelihood for a binary outcome. Differentiating with $\ell(\lambda)$ with respect to λ, we get the score function

$$U(\lambda) = \frac{m}{\lambda} - T. \tag{5.14}$$

Solving the score equation $U(\hat{\lambda}) = 0$ gives us the point estimate

$$\hat{\lambda} = \frac{m}{T}. \tag{5.15}$$

This is called the **incidence rate** in epidemiology. The incidence rate is the maximum likelihood estimate of an exponential rate parameter.

Differentiating $U(\lambda)$ with respect to λ, we get the observed information function $I(\lambda) = m/\lambda^2$. The estimated variance is

$$I(\hat{\lambda})^{-1} = \left(\frac{m}{(\frac{m}{T})^2} \right)^{-1} = \frac{m}{T^2}, \tag{5.16}$$

so the Wald 95% confidence interval for λ_{true} is

$$\hat{\lambda} \pm 1.96\sqrt{\frac{m}{T^2}}. \tag{5.17}$$

It has relatively poor performance in terms of width and coverage probability. The performance of the Wald confidence interval can be improved using a log transformation, which ensures that the lower bound is greater than zero. Using the delta method, the variance of $\ln \hat{\lambda}$ is approximately

$$\left(\frac{1}{\hat{\lambda}}\right)^2 \frac{m}{T^2} = \frac{1}{m},$$

which depends only on the number of events observed. An approximate 95% confidence interval for $\ln(\lambda_0)$ is

$$\ln\left(\frac{m}{T}\right) \pm 1.96\sqrt{\frac{1}{m}}.$$

Exponentiating, we get the log-transformed Wald 95% confidence interval

$$\frac{m}{T} e^{\pm 1.96\sqrt{\frac{1}{m}}}.$$

Better interval estimates can be obtained by inverting the score or likelihood ratio test or from Bayesian credible intervals. Among the frequentist large-sample confidence intervals, the likelihood ratio interval has the best performance (Brown, Cai, and DasGupta 2003).

5.4 R

5.4.1 Memoryless property

Given that you have reached age a, your **life expectancy at age** a is how many years you are expected to live past age a. According to the Social Security Administration's 2019 life table, life expectancy at birth in the United States in 2019 was 76.22 years for males and 81.28 years for females. Life expectancy at age 40 was 38.74 years for males and 42.76 years for females, which means that the average age at death for those who survive to age 40 was 78.74 years and 82.76 years, respectively. Life expectancy at age 80 was 8.43 years for males and 9.83 years for females, so the average ages at death were 88.43 and 89.83 years, respectively. For humans, remaining life expectancy decreases with age.

Humans do not have exponential lifetimes. If your lifetime has an exponential distribution with rate λ and you survive to age t, the probability that you survive to age $t+u$ is

$$\Pr(\text{lifetime} > t+u \,|\, \text{lifetime} > t) = \frac{e^{-\lambda(t+u)}}{e^{-\lambda t}} = e^{-u}.$$

This does not depend on t, so your life expectancy would be constant with age. This is called the **memoryless property**.

In a population with exponential lifetimes, the old and the young would have equal hazards of death at any given time and equal risks of death over any time interval. This seems to be true of decaying radioactive isotopes and other processes from physics and chemistry, but humans are more complex. The hazard of death is typically high right after birth, drops rapidly to a minimum between the ages of roughly 5 and 30, and then slowly increases. This is called the *bathtub-shaped hazard* or the *Gompertz-Makeham law of mortality* (Gompertz 1825; Makeham 1860). Figure 5.1 shows The bathtub-shaped hazard of death for the United States population in 2019.

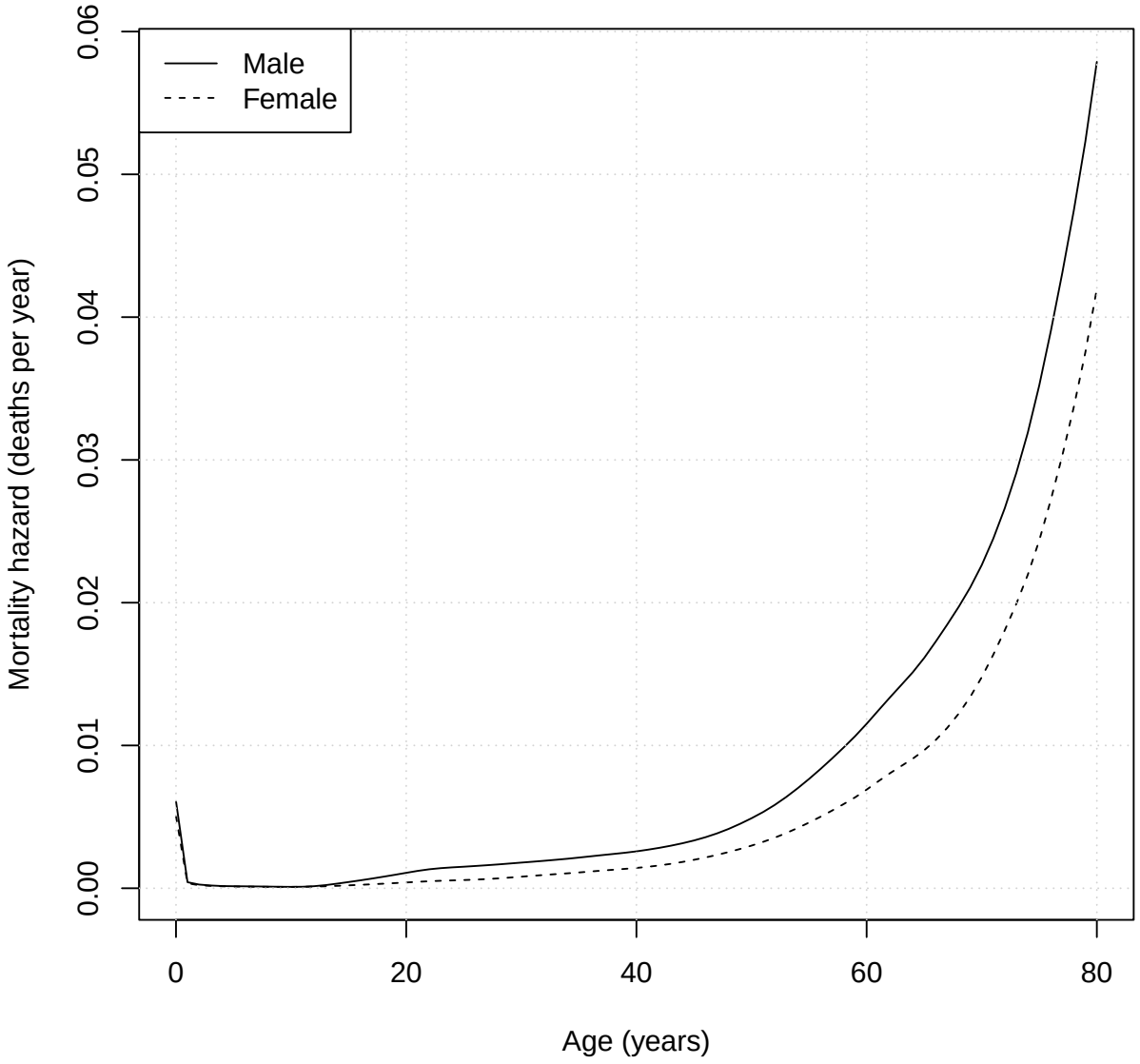


Figure 5.1: Mortality hazard (deaths per year) by age in the United States based on the Social Security Administration 2019 life table.

5.4.2 Prevalence, incidence, and duration of disease*

Suppose the time to disease onset in healthy individuals has an exponential(λ) distribution and the time to recovery in diseased individuals has an exponential(γ) distribution. Then the incidence rate of disease is λ and the mean duration of disease is $1/\gamma$. For simplicity, imagine a closed population where individuals move between the healthy and diseased states.

If the prevalence of disease at time t is p and the population has total size n, then the expected

number of disease onsets in any time interval $(t, t + \mathrm{d}t]$ is $n(1-p)\lambda\,\mathrm{d}t$. The expected number of recoveries in the same interval is $np\gamma\mathrm{d}t$. For the prevalence to remain roughly constant over time (i.e., to randomly fluctuate around an equilibrium), we need the expected number of onsets and recoveries in each time interval to be the same: Thus, we need

$$n(1-p)\lambda = np\gamma$$

The left-hand side is the expected number of disease onsets per time unit, and the right-hand side is the expected number of recoveries per time unit. It is critical to use the same time units (e.g., day, week, month, year) for the incidence rate and the duration of disease. This equation can be rearranged to get

$$\frac{\lambda}{\gamma} = \frac{p}{1-p},$$

so

$$\text{incidence rate} \times \text{mean duration} = \text{prevalence odds}$$

When the prevalence p is low, the prevalence and prevalence odds are roughly equal and we get

$$\text{incidence rate} \times \text{mean duration} \approx \text{prevalence}.$$

Under more realistic conditions, the relationship between prevalence, incidence, and the duration of disease is more complex (Freeman and Hutchison 1980; Preston 1987; Keiding 1991; Alho 1992).

5.5 Poisson distribution

The Poisson distribution[5] is one of the most important distributions in probability and statistics, and it has many applications in epidemiology. Section 5.2.3 showed that the cumulative hazard $H(t)$ is the expected number of events that occur in $(0, t]$ when the event is repeatable. The number of events has a Poisson distribution with mean $H(t)$.

If events occur at a constant rate λ, then the times between events have an exponential(λ) distribution. The number X of events that occur in a time interval of length T will have a Poisson distribution with mean λT. The probability mass function (PMF) of a Poisson(λT) distribution is

$$\Pr(X = x) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \text{ for } x = 0, 1, 2, \ldots. \tag{5.18}$$

and $\text{supp}(X) = \{0, 1, 2, ...\}$. This is a PMF because, by definition,

$$e^{\lambda T} = \sum_{k=0}^{\infty} \frac{(\lambda T)^k}{k!}.$$

[5]Named after Sim'{e}on Denis Poisson (1781-1840), a French mathematician, physicist, and astronomer. He introduced the distribution in an 1837 paper in which he estimated the number of wrongful convictions that would occur over a given time period. His is one of the 72 names inscribed on the Eiffel Tower.

Multiplying $(\lambda T)^k/k!$ by $e^{-\lambda T}$ ensures that the PMF over all possible values of X equals one. The relationship between the exponential and Poisson distributions can be seen easily for $X = 0$:

$$\Pr(X = 0) = \frac{(\lambda T)^0}{0!} e^{-\lambda T} = e^{-\lambda T},$$

which is the probability that no event occurs in $(0, T]$ when the time-to-event has an exponential(λ) distribution.

5.5.1 Mean and variance

The mean of the Poisson(λT) distribution is

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{(\lambda T)^k}{k!} e^{-\lambda T} = \lambda T e^{\lambda T} e^{-\lambda T} = \lambda T.$$

A similar calculation yields

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 \frac{(\lambda T)^k}{k!} e^{-\lambda T} = \lambda T + (\lambda T)^2.$$

Using Equation 1.22, we get

$$\mathrm{Var}(X) = [\lambda T + (\lambda T)^2] - (\lambda T)^2 = \lambda T.$$

Thus, both the mean and the variance are λT. Because it equals the mean, the parameter λT is often written μ. Both λ and μ can be estimated using maximum likelihood or Bayesian methods.

5.5.2 Incidence rates via count data

Suppose we observe n individuals in whom the time to disease onset is exponential(λ_{true}), observing a total person-time of T during which m disease onsets occur. If we string together all of the observations, the memoryless property of the exponential distribution guarantees that we get an interval with total length T in which m events occurred and the times between events were exponential(λ_{true}). Therefore, the total number of events that we see has a Poisson distribution with mean $\lambda_{\text{true}} T$. Using the Poisson PMF in Equation 5.18, we get the likelihood

$$L(\lambda) = \frac{(\lambda T)^m}{m!} e^{-\lambda T}.$$

The corresponding log likelihood is

$$\ell(\lambda) = \ln L(\lambda) = m(\ln \lambda + \ln T) - \ln(m!) - \lambda T.$$

Taking the derivative with respect to λ, we get the score function

$$U(\lambda) = \frac{m}{\lambda} - T,$$

which is exactly the same as the score function for λ that we got using an exponential likelihood in Equation 5.14. Taking another derivative with respect to λ, we also get the same estimated variance

$$I(\hat{\lambda})^{-1} = \frac{m}{T^2}$$

that we got in Equation 5.16. Under the Poisson model for the number of events, the Wald and log-transformed Wald confidence intervals for the unknown incidence rate λ_{true} are exactly the same as those from the exponential model for the times to events. The score and likelihood ratio tests can be inverted to get confidence intervals that perform better in terms of coverage probability and width than the Wald interval (Brown, Cai, and DasGupta 2003).

5.6 R

5.6.1 Small-sample estimation of incidence rates

The Poisson distribution can be used to calculate confidence limits for the incidence rate λ_{true} that do not rely on the approximate normality in large samples that is guaranteed by the central limit theorem (CLT). Exact confidence limits can be calculated in a manner similar to the Clopper-Pearson confidence limits for a binomial probability in Section 3.8.2. Bayesian estimation, which does not require asymptotic normality, can also be used for small samples.

If we observe m events in a total person-time T, the median unbiased estimate of λ_{true} is the value of λ that makes

$$\Pr_\lambda(X \leq m) = \Pr_\lambda(X \geq m)$$

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

$$\sum_{k=0}^{m-1} \frac{(\lambda_{\text{med}} T)^k}{k!} e^{-\lambda_{\text{med}} T} + \frac{1}{2} \frac{(\lambda_{\text{med}} T)^m}{m!} e^{-\lambda_{\text{med}} T} = \frac{1}{2}.$$

It is the value of λ that makes the tail probabilities equal. The median of the distribution of λ_{med} is always λ_{true} (Birnbaum 1964).

The lower exact $1-\alpha$ confidence limit for λ_{true} gives the upper tail of the Poisson distribution a probability of $\alpha/2$ (Garwood 1936). It solves the equation

$$\sum_{k=m}^{\infty} \frac{(\lambda_{\text{lower}} T)^k}{k!} e^{-\lambda_{\text{lower}} T} = 1 - \left[\sum_{k=0}^{m-1} \frac{(\lambda_{\text{lower}} T)^k}{k!} e^{-\lambda_{\text{lower}} T} \right] = \frac{\alpha}{2}.$$

The upper exact $1-\alpha$ confidence limit for λ_{true} gives the lower tail of the Poisson distribution a probability of $\alpha/2$. It solves the equation

$$\sum_{k=0}^{m} \frac{\left(\lambda_{\text{upper}} T\right)^k}{k!} e^{-\lambda_{\text{upper}} T} = \frac{\alpha}{2}.$$

As with the Clopper-Pearson confidence limits for a binomial probability, the exact Poisson confidence limits guarantee a coverage probability (i.e., probability that the confidence interval contains λ_{true}) of at least $1-\alpha$. However, these confidence intervals can be wide and have a coverage probability much higher than $1-\alpha$ (G. R. Cohen and Yang 1994; Swift 2009).

Mid-p confidence limits can produce confidence intervals that are narrower and have a coverage probability closer to $1-\alpha$ (Lancaster 1961). The lower $1-\alpha$ mid-p exact confidence limit for the incidence rate λ_{true} solves the equation

$$1 - \left[\sum_{k=0}^{m-1} \frac{\left(\lambda_{\text{lower}} T\right)^k}{k!} e^{-\lambda_{\text{lower}} T} + \frac{1}{2} \frac{\left(\lambda_{\text{lower}} T\right)^m}{m!} e^{-\lambda_{\text{lower}} T} \right] = \frac{\alpha}{2}.$$

The upper $1-\alpha$ mid-p exact confidence limit solves the equation

$$\sum_{k=0}^{m-1} \frac{\left(\lambda_{\text{upper}} T\right)^k}{k!} e^{-\lambda_{\text{upper}} T} + \frac{1}{2} \frac{\left(\lambda_{\text{upper}} T\right)^m}{m!} e^{-\lambda_{\text{upper}} T} = \frac{\alpha}{2}.$$

The coverage probability of the mid-p confidence interval is usually very close to $1-\alpha$ (G. R. Cohen and Yang 1994; Swift 2009).

5.7 R

5.7.1 Poisson approximation to the binomial for rare events*

Most applications of the Poisson distribution in epidemiology come from its relationship with the exponential distribution, but the Poisson distribution also has a useful relationship with the binomial distribution. When n is large and p is small, the binomial(n, p) distribution can be approximated by a Poisson(np) distribution. More specifically, imagine that n increases and p decreases such that $np = \mu$ stays constant. The binomial(n, p) probability mass function is

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}.$$

When n is much larger than k, we have

$$\frac{n!}{(n-k)!k!} \approx \frac{n^k}{k!}.$$

Because $p = \mu/n$, we also have

$$p^k = \frac{\mu^k}{n^k}$$

for each n and

$$(1-p)^{n-k} = \left(1 - \frac{\mu}{n}\right)^{n-k} \longrightarrow e^{-\mu}$$

as $n \to \infty$. Putting all of this together, we get the following approximation to the binomial PMF

$$\Pr(X = k) \approx \frac{n^k \mu^k}{k! n^k} e^{-\mu} = \frac{\mu^k}{k!} e^{-\mu},$$

which is the Poisson(μ) PMF.

As an approximation to the binomial distribution, the Poisson distribution can be used for rare events in many different contexts. The number of events such as automobile accidents or number of onsets of a rare disease in a given time period or area (or both) will often have a Poisson distribution. Clarke (1946) describes an astonishing application of the Poisson distribution that occurred in London in World War II. The V1 flying bomb was a German cruise missile with an 850 kg warhead that was fired at London (and later at Belgium) in 1944 and 1945. Soon after the attacks began, many people felt that the bomb impacts were clustered in particular areas of London. British investigators used the Poisson distribution to determine whether the V1s were being aimed or fell randomly within the city.

They divided 144 square kilometers of central London into 576 squares of 0.25 square kilometers each. By then, the total number of bombs that had fallen on the entire area was 537. If the bombs were falling randomly, the number of bombs in each square should have a Poisson distribution with mean $537/576 \approx 0.932$. Grouping the squares by the number of bombs that had fallen on them yielded Table 5.3. The close fit to the Poisson distribution suggested that the bombs were falling randomly over the entire 144 square kilometers, not being aimed at particular targets within the city. Near the end of the war, analysis of captured V1s revealed that the guidance system was only accurate to a radius of about 6 kilometers, a circle large enough to encompass almost all of London at the time.

5.8 Bayesian estimation of incidence rates

As with a binomial probability, Bayesian methods can be used to estimate an incidence rate without making any large-sample assumptions. They also allow a prior distribution to be used to incorporate background knowledge about the possible values of λ_{true}.

Table 5.3: Distribution of V1 bomb impacts in London (Clarke 1946).}

Number of impacts	Expected (Poisson) number of squares	Actual number of squares
0	226.74	229
1	211.39	211
2	98.54	93
3	30.62	35
4	7.14	7
5+	1.57	1
Total squares	576.00	576

5.8.1 Gamma conjugate distribution

The conjugate distribution for the exponential(λ) and Poisson(λT) distributions is the **gamma distribution**, which has the PDF

$$f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{5.19}$$

where $\alpha > 0$ is the shape parameter, $\beta > 0$ is the rate parameter, and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{d}x$ is the *gamma function*. If X has a gamma(α, β) distribution, then

$$\mathbb{E}[X] = \frac{\alpha}{\beta}$$

and variance

$$\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$

The exponential(β) distribution is a special case of the gamma distribution with shape $\alpha = 1$.

If the prior PDF of an unknown exponential rate λ_{true} has a gamma(α, β) distribution, then

$$p(\lambda \,|\, \text{data}) \propto (\lambda^m e^{-\lambda T})(\lambda^{\alpha-1} e^{-\beta\lambda}) = \lambda^{m+\alpha-1} e^{-\lambda(T+\beta)}.$$

After normalizing, this is a gamma($m + \alpha, T + \beta$) distribution. The posterior mean is

$$\bar{\lambda} = \frac{m + \alpha}{T + \beta}$$

and the posterior variance is

$$\frac{m + \alpha}{(T + \beta)^2} = \frac{\bar{\lambda}}{T + \beta}$$

The equal-tailed $1 - \alpha$ credible interval has its limits at the $\alpha/2$ and $1 - \alpha/2$ quantiles of this distribution. Figure 5.2 shows an example of an uninformative Bayesian prior and a posterior distribution for the incidence rate.

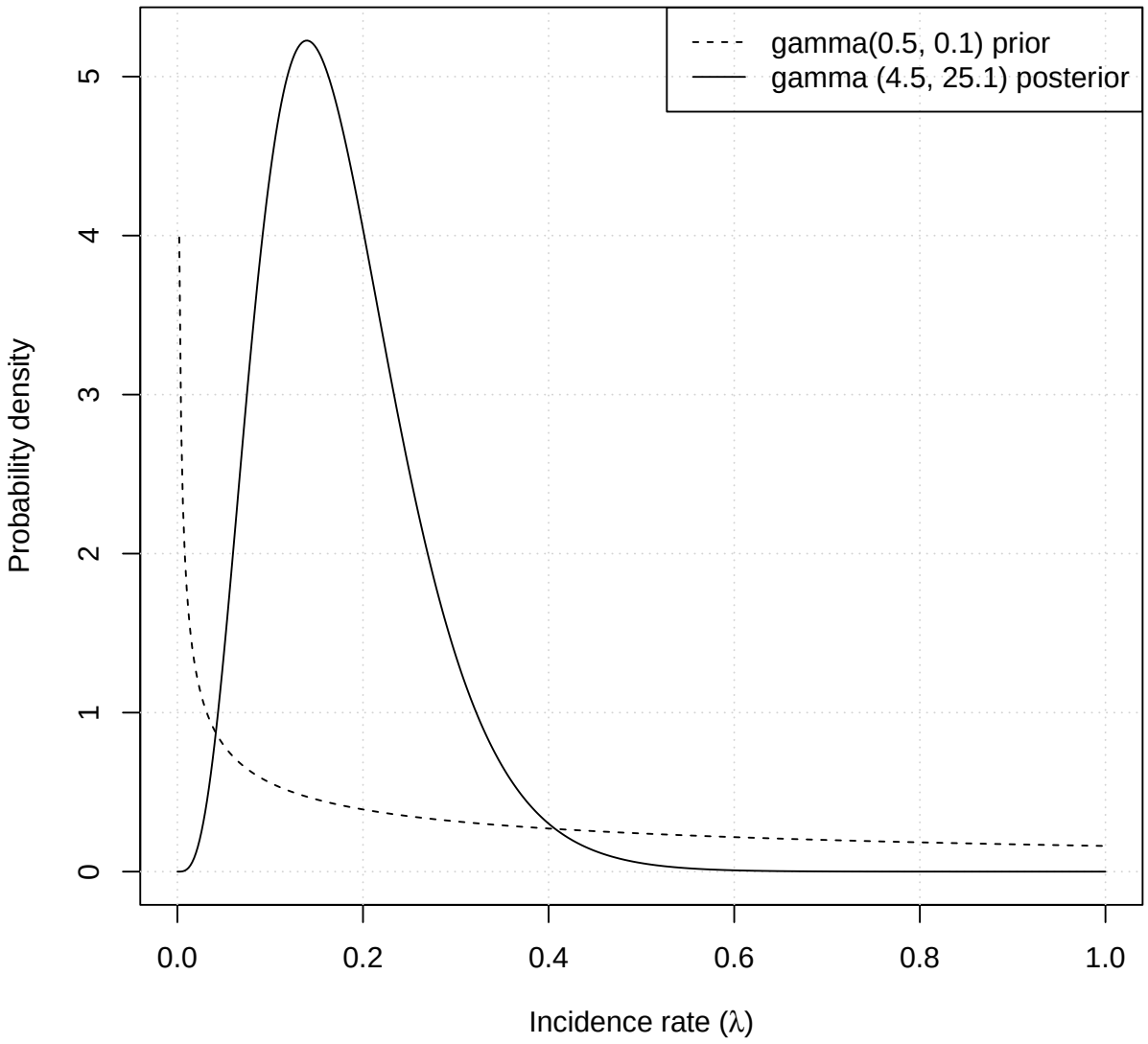


Figure 5.2: The PDFs of a gamma(0.5, 0.1) prior distribution and the gamma(4.5, 25.1) posterior distribution that results from seeing 4 events in 25 person-years of observation. The units of the incidence rate events per year.

5.8.2 Jeffreys confidence interval

If we fix the total person-time T and model the number of events as a Poisson(λT) random variable, the Jeffreys prior has $\alpha = 1/2$ and $\beta = 0$ (Jeffreys 1946). This is an *improper prior* because the PDF has a total area of ∞ under the curve, but we get a proper posterior distribution (i.e., a posterior PDF with a total area of 1) as long as $T > 0$. The Jeffreys confidence interval for λ_{true} is the equal-tailed credible interval from the gamma($m + 1/2, T$)

posterior distribution. If $m = 0$, the lower limit can be set to zero. This confidence interval has good coverage probabilities and narrow widths similar to the likelihood ratio confidence interval (Brown, Cai, and DasGupta 2003; Swift 2009).

5.9 R

Listing 5.1 `exponential.R`

```
## Exponential rate parameter estimation

# generate right-censored exponential distribution
tevent <- rexp(1000, rate = 2)
tcens <- rexp(1000)             # default rate = 1
sdat <- data.frame(texit = pmin(tcens, tevent),
                   event = ifelse(tcens < tevent, 0, 1))

# calculating incidence rate and log-transformed confidence interval
m <- sum(sdat$event)
T <- sum(sdat$texit)
m / T
m / T * exp(c(-1, 1) * qnorm(.975) * sqrt(1 / m))

# fitting intercept-only exponential regression model
# This uses the survreg() function from the survival package.
library(survival)
expfit <- survreg(Surv(texit, event) ~ 1, data = sdat,
                  dist = "exponential")
summary(expfit)
coef(expfit)
confint(expfit)

# log-transformed Wald CI for the exponential rate
# The intercept is ln(scale), which is -ln(rate).
# The rate is exp(-intercept).
exp(-coef(expfit))
exp(-confint(expfit))

# add delayed entry (left truncation) to sdat
sdat2 <- sdat
sdat2$tentry <- rexp(1000, rate = 5)
sdat2 <- subset(sdat2, tentry < texit)

# incidence rate and log-transformed confidence interval
m2 <- sum(sdat2$event)
T2 <- sum(sdat2$texit - sdat2$tentry)
m2 / T2
m2 / T2 * exp(c(-1, 1) * qnorm(.975) * sqrt(1 / m2))

# survreg() does not handle delayed entry, so use flexsurv::flexsurvreg()
library(flexsurv)
expfit2 <- flexsurvreg(Surv(tentry, texit, event) ~ 1, data = sdat2,
                       dist = "exp")
# The summary() function does not work with flexsurvreg objects.
# Type "expfit2" or "expfit2$res" to get point and interval estimates.
# The "se" in expfit2$res is the delta method standard error.
expfit2
expfit2$res       # rate parameter scale
expfit2$res.t     # log rate parameter scale
```

Listing 5.2 `bathtub.R`

```r
# life table for male and female mortality in the United State, 2019
lifetab <- read.csv(file = "R/lifetable-2019.csv")
hdat <- subset(lifetab, age <= 80)
hdat$surv_male <- 1 - hdat$mortality_male
hdat$surv_female <- 1 - hdat$mortality_female

# plot hazard (events per year) for ages 0-80
plot(hdat$age, -log(hdat$surv_male), type = "l",
     xlab = "Age (years)", ylab = "Mortality hazard (deaths per year)")
lines(hdat$age, -log(hdat$surv_female), lty = "dashed")
grid()
legend("topleft", bg = "white", lty = c("solid", "dashed"),
       legend = c("Male", "Female"))
```

Listing 5.3 `Poisson-rate.R`

```
## Poisson regression for incidence rates

# generate right-censored exponential distribution
tevent <- rexp(1000, rate = 2)
tcens <- rexp(1000)             # default rate = 1
sdat <- data.frame(texit = pmin(tcens, tevent),
                   event = ifelse(tcens < tevent, 0, 1))

# Poisson regression model
# Use log(time) offset to get incidence rate from Poisson(time * incidence rate)
poisreg <- glm(event ~ offset(log(texit)), data = sdat, family = poisson())
exp(coef(poisreg))
# GLMs use likelihood ratio confidence intervals in R.
exp(confint(poisreg))

# exponential regression for comparison (log-transformed Wald CI)
library(survival)
expreg <- survreg(Surv(texit, event) ~ 1, data = sdat, dist = "exponential")
exp(-coef(expreg))
exp(-confint(expreg))

# add delayed entry to sdat
sdat2 <- sdat
sdat2$tentry <- rexp(1000, rate = 5)
sdat2 <- subset(sdat2, tentry < texit)

# Poisson regression with delayed entry
poisreg2 <- glm(event ~ offset(log(texit - tentry)), data = sdat2,
                family = poisson())
exp(coef(poisreg2))
exp(confint(poisreg2))

# exponential regression with delayed entry for comparison
library(flexsurv)
expreg2 <- flexsurvreg(Surv(tentry, texit, event) ~ 1, data = sdat2,
                       dist = "exp")
expreg2$res
```

Listing 5.4 `Poisson-small.R`

```r
## Small-sample Poisson point and interval estimation

# median unbiased estimate
medrate_pois <- function(m, T) {
  # m = number of events, T = total person-time

  # Poisson lower tail probability
  lower_tail <- function(rate) {
    mu = rate * T
    ppois(m, mu) - dpois(m, mu) / 2
  }

  # median unbiased estimate
  med <- uniroot(function(rate) lower_tail(rate) - 1 / 2, interval = c(0, 1))
  med$root
}
medrate_pois(7, 22)

# exact confidence limits
# The point estimate is the incidence rate m / T, not the median unbiased rate.
library(DescTools)
PoissonCI(7, 22, method = "exact")
PoissonCI(7, 22, method = "exact", conf.level = 0.8)

# mid-p confidence limits
midp_pois <- function(m, T, level=0.95) {
  # m = number of events, T = total person-time
  # The default confidence level (1 - type I error probability) is 0.95.

  # Poisson mid-p lower tail probability
  lower_tail <- function(rate) {
    mu = rate * T
    ppois(m, mu) - dpois(m, mu) / 2
  }

  # lower confidence limit
  alpha <- 1 - level
  lower <- uniroot(function(rate) lower_tail(rate) - (1 - alpha / 2),
                   interval = c(0, 100), extendInt = "yes")
  # upper confidence limit
  upper <- uniroot(function(rate) lower_tail(rate) - alpha / 2,
                   interval = c(0, 100), extendInt = "yes")

  # names for confidence limits
  lower_perc <- paste(round(alpha / 2 * 100, 3), "%", sep = "")
  upper_perc <- paste(round((1 - alpha / 2) * 100, 3), "%", sep = "")

  # return named vector of confidence limits
  conflimits <- c(lower$root, upper$root)
  names(conflimits) <- c(lower_perc, upper_perc)
  conflimits
```

Listing 5.5 `incrate-Bayes-plot.R`

```
## Bayesian estimation of incidence rates

x <- seq(0, 1, by = 0.002)
m <- 4
T <- 25

# plot of prior and posterior distributions
plot(x, dgamma(x, shape = 0.5 + m, rate = 0.1 + T), type = "n",
     xlab = expression(paste("Incidence rate (", lambda, ")")),
     ylab = "Probability density")
grid()
lines(x, dgamma(x, shape = 0.5, rate = 0.1), lty = "dashed")
lines(x, dgamma(x, shape = 0.5 + m, rate = 0.1 + T))
legend("topright", lty = c("dashed", "solid"),
       legend = c("gamma(0.5, 0.1) prior", "gamma (4.5, 25.1) posterior"))
```

Listing 5.6 `incrate-Bayes.R`

```
## Bayesian estimation of incidence rates with gamma conjugate distribution

# incidence rate posterior mean, median, and equal-tailed credible limits
incrate_bayes <- function(m, T, level=0.95, priora=0.5, priorb=0) {
  # default arguments are for the Jeffreys confidence interval
  alpha <- 1 - level
  posta  <- priora + m
  postb <- priorb + T
  if (m == 0) {
    lower <- 0
  } else {
    lower <- qgamma(alpha / 2, shape = posta, rate = postb)
  }
  upper <- qgamma(1 - alpha / 2, shape = posta, rate = postb)
  postmean <- posta / postb
  postmedian <- qgamma(0.5, shape = posta, rate = postb)
  return(c(postmean = postmean, postmedian = postmedian,
           lower = lower, upper = upper,
           priora = priora, priorb = priorb, level = level))
}

# 7 events in 22 units of person-time
incrate_bayes(7, 22)                          # Jeffreys 95% confidence interval
incrate_bayes(7, 22, level = 0.8)             # Jeffreys 80% confidence interval
incrate_bayes(7, 22, priora = 1, priorb = 1)  # uniform prior
```

6 Survival Analysis

> In many estimation problems it is inconvenient or impossible to make complete measurements on all members of a random sample. For example, in medical follow-up studies to determine the distribution of survival times after an operation, contact with some individuals will be lost before their death, and others will die from causes it is desired to exclude from consideration. Similarly, observation of the life of a vacuum tube may be ended by breakage of the tube, or a need to use the test facilities for other purposes. In both examples, incomplete observations may also result from a need to get out a report within a reasonable time. (Kaplan and Meier 1958)

In **nonparametric** survival analysis, we do not assume that the failure time distributions are defined by a small number of parameters, such as the rate parameter in an exponential model for times to events or a Poisson model for the number of events. Whenever possible, it is good to incorporate existing knowledge into the estimation of unknown parameters, and the use of parametric models and Bayesian methods accomplishes this. When such knowledge is not available, nonparametric methods allow us to avoid making assumptions we cannot defend.

However, this flexibility comes at a price. For example, suppose we know that our time-to-event has an exponential distribution. If we use a nonparametric model anyway, then:

- The nonparametric estimates of the survival, cumulative hazard, or hazard functions will be less precise than the estimates from an exponential model.
- We might be unable to estimate mean or median survival times or other quantities that require extrapolation beyond the data used for estimation.

Parametric and nonparametric methods are at opposite ends of the bias-variance tradeoff. The assumptions of parametric models can induce bias, but they produce estimates with low variance when the assumptions are approximately correct. The flexibility of nonparametric models avoids bias, but they produce estimates with higher variance than an equivalent parametric method based on sound assumptions. Survival analysis has nonparametric estimators of the survival and cumulative hazard functions that can be used with relatively little loss of efficiency. Because of this combination of flexibility and efficiency, they are widely used in epidemiologic research.

6.1 Empirical cumulative distribution function

The cumulative distribution function (CDF) of a random variable X is the function

$$F(x) = \Pr(X \leq x). \tag{6.1}$$

For each value of x, $F(x)$ is a probability that we can estimate using methods for a binomial proportion. However, we can get a more complete picture of the distribution of X by linking the estimates for different x together to estimate the whole function $F(x)$.

If $x_1, \ldots, x_n$ are observations of a random variable X, the **empirical CDF** is the function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x} \tag{6.2}$$

where $\mathbb{1}_{x_i \leq x} = 1$ if $x_i \leq x$ is true and zero otherwise. For a fixed value of x, $\hat{F}_n(x)$ is just the proportion of the observations that are $\leq x$. At each x, the number of observations with $X_i \leq x$ is a binomial$(n, F(x))$ random variable with expected value $nF(x)$ and variance $nF(x)(1 - F(x))$. Thus,

$$\mathbb{E}(\hat{F}_n(x)) = F(x)$$

and

$$\mathrm{Var}(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)). \tag{6.3}$$

As $n \to \infty$, we have $\hat{F}_n(x) \to F(x)$ by the law of large numbers (LLN) and By the central limit theorem (CLT),

$$\frac{\sqrt{n}(\hat{F}_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \overset{\text{approx}}{\sim} N(0, 1) \tag{6.4}$$

by the central limit theorem (CLT).[1] Unlike a single proportion estimate $\hat{p}$, the empirical CDF links all of these estimated probabilities together—like beads on a necklace—through the variable x.

At any given x, interval estimates for $F(x)$ can be obtained using any of the methods we have discussed for probabilities, including the Wald, score (Wilson), likelihood ratio, exact (Clopper-Pearson), mid-p, or Jeffreys confidence intervals as well as Bayesian credible intervals. For example,

$$\hat{F}_n(x) \pm 1.96 \sqrt{\frac{1}{n} \hat{F}_n(x)(1 - \hat{F}_n(x))} \tag{6.5}$$

is the 95% Wald confidence interval. To force this confidence interval to stay inside $(0, 1)$, we can use the delta method with a logit or log-log transformation. These are called **pointwise** confidence intervals because we have a separate confidence interval for $F(x)$ at each x.

[1] The *Glivenko-Cantelli theorem* guarantees that $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0$ as $n \to \infty$, so the convergence happens simultaneously for all x (Van der Vaart 2000).

6.2 Kaplan-Meier estimator

In data with right censoring and left truncation, we cannot calculate the empirical CDF directly. The **Kaplan-Meier estimator** (Kaplan and Meier 1958) uses conditional probabilities to estimate the survival function $S(t) = 1 - F(t)$ for failure time data.[2] The basic idea behind the Kaplan-Meier estimator is to solve the problems of right censoring and left truncation (delayed entry) by breaking analysis time into periods where no one enters or leaves the study. In each such interval $(t_a, t_b]$, we can estimate the conditional probability of surviving to time t_b given that you were at risk of disease at time t_a.

If there are n individuals at risk throughout the interval $(t_a, t_b]$, then the number of events at time t_b can be treated like a binomial(n, p) random variable where p is the risk of the event in $(t_a, t_b]$. Our point estimate of this conditional probability is

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

where d is the number of failures at time t_b. Its variance is approximately

$$\text{Var}(\hat{p}) = \frac{1}{n}\hat{p}(1-\hat{p}).$$

Given that you were at risk of disease at time t_a,

$$\hat{q} = 1 - \hat{p}$$

is the conditional probability of surviving past time t_b.

6.2.1 At-risk process and risk sets

To estimate the risk in an interval $(t_a, t_b]$, it is critical to define who is at risk of failure at time t_a. We assume that all times are defined relative to a time origin that can differ between individuals. The **at-risk process** for individual i is

$$Y_i(t) = \begin{cases} 1 & \text{if } i \text{ is at risk and under observation at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The at-risk process is assumed to be predictable (i.e., its value at t is determined just before time t), so $Y_i(t) = 1$ even if i fails or is censored at time t. Note that a person is only at risk

[2] Named after American statisticians Edward L. Kaplan (1920–2006) and Paul Meier (1924–2011). Kaplan worked at Bell Telephone Laboratories, the Lawrence Radiation Laboratory (now the Lawrence Berkeley National Laboratory), and Oregon State University. Meier worked at Johns Hopkins and the University of Chicago and was an early advocate for the use of randomization in clinical trials. Both were doctoral students of John Tukey at Princeton. Their 1958 paper is one of the most-cited papers in statistics, with 66,740 citations as of 22 January 2025.

if they are both at risk of failure and under observation. If person i is not under observation until an entry time $t_i^{\text{entry}} > 0$, then $Y_i(t) = 0$ when $t \leq t_i^{\text{entry}}$.

The set of individuals under observation and at risk of failure at time t is called the **risk set** at time t and written

$$\mathcal{R}(t) = \{i : Y_i(t) = 1\}$$

The risk set $\mathcal{R}(t)$ includes everyone under observation who fails at time t, who is censored at time t, or who survives past time t.

6.2.2 Survival via multiplication of conditional probabilities

Let T denote the random failure time in the analysis time scale (i.e., with the origin as time zero), and suppose we have times $0 < t_1 < t_2$. To have $T > t_2$, we must also have $T > t_1$, so

$$\begin{aligned}\Pr(T > t_2) &= \Pr(T > t_2 \text{ and } T > t_1) \\ &= \Pr(T > t_2 \,|\, T > t_1) \Pr(T > t_1).\end{aligned}$$

In other words, the probability of survival in $(0, t_2]$ is the product of the survival probabilities in the intervals $(0, t_1]$ and $(t_1, t_2]$. If $t_3 > t_2$, then

$$\begin{aligned}\Pr(T > t_3) &= \Pr(T > t_3 \,|\, T > t_2) \Pr(T > t_2) \\ &= \Pr(T > t_3 \,|\, T > t_2) \Pr(T > t_2 \,|\, T > t_1) \Pr(T > t_1)\end{aligned}$$

which is the product of the survival probabilities in the intervals $(0, t_1]$, $(t_1, t_2]$, and $(t_2, t_3]$. This logic extends to any number of intervals. If we have distinct times $0 = t_0 < t_1 < \cdots < t_m$, then

$$\Pr(T > t_m) = \prod_{i=1}^{m} \Pr\left(T > t_i \,|\, T > t_{i-1}\right).$$

This uses the multiplication rule for conditional probabilities, so it does not assume that failures in different intervals are independent. In single-failure data, an individual who fails in one interval cannot fail in a later interval, so failures in different intervals cannot be independent.

Let $0 = t_0 < t_1 < t_2 < ... < t_m$ be the endpoints of intervals $(t_{i-1}, t_i]$ within which there are no entries or exits from the study. Let $n_j = \sum_{i=1}^{n} Y_i(t_j)$ be the number of people in the risk set $\mathcal{R}(t_j)$ and let $d_j \geq 0$ be the number of failures that occur at time t_j. The estimated conditional probability q_j of surviving through the interval $(t_{j-1}, t_j]$ given survival to t_{j-1} is

$$\hat{q}_j = 1 - \frac{d_j}{n_j},$$

and the Kaplan-Meier estimator of $S(t)$ is

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \hat{q}_j = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{6.6}$$

This is the product of the conditional survival probabilities in $(t_{j-1}, t_j]$ for all intervals such that $t_j \leq t$. To survive to time t, you need to survive through all intervals up to and including time t. This makes it easier to calculate the survival function than to calculate cumulative incidence directly. The Kaplan-Meier estimator is a consistent and asymptotically normal estimator of the true survival function (Fleming and Harrington 2005; Aalen, Borgan, and Gjessing 2008). Figure 6.1 shows an example based on a right-censored sample of size 500 from a log-logistic distribution with shape $\alpha = 1$ and rate $\lambda = 2$.

6.3 R

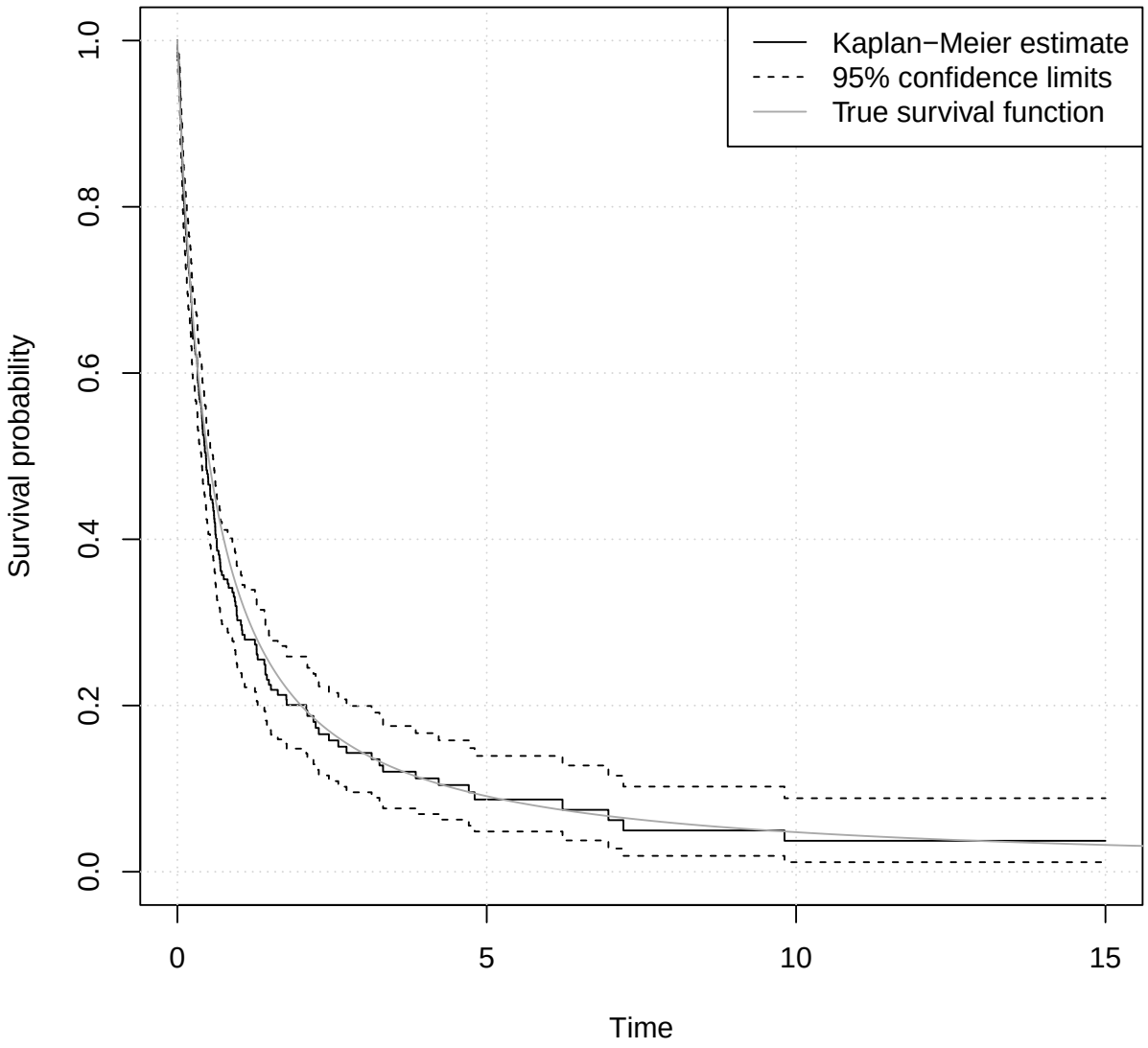


Figure 6.1: True log-logistic survival function and Kaplan-Meier estimate with complementary log-log 95% confidence limits.

6.3.1 Greenwood formula and confidence intervals

Calculating the variance of a product is difficult and tedious, but calculating the variance of a sum is easy. Taking logarithms in Equation 6.6, we get

$$\ln \hat{S}(t) = \sum_{j:t_j \leq t} \ln \hat{q}_j.$$

For each j, the estimated variance of $\hat{q}_j$ is

$$\operatorname{Var}(\hat{q}_j) = \frac{1}{n_j}\hat{q}_j(1-\hat{q}_j).$$

Since $\ln x$ has the derivative $\frac{1}{x}$,

$$\operatorname{Var}(\ln \hat{q}_j) \approx \frac{1}{\hat{q}_j^2}\operatorname{Var}(\hat{q}_j) = \frac{d_j}{n_j(n_j-d_j)}.$$

by the delta method from Section 3.6.1.

The estimated survival probabilities in each time interval are conditionally independent, so

$$\operatorname{Var}(\ln \hat{S}(t)) = \sum_{t_j \le t} \operatorname{Var}(\ln \hat{q}_j) = \sum_{t_j \le t} \frac{d_j}{n_j(n_j-d_j)}.$$

by Equation 1.24. Since $\hat{S}(t) = \exp(\ln \hat{S}(t))$, we can use the delta method again to get an estimated variance for $\hat{S}(t)$. The function $\exp(x) = e^x$ is its own derivative, so we get

$$\operatorname{Var}(\hat{S}(t)) = \hat{S}(t)^2 \operatorname{Var}(\ln \hat{S}(t)) = \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j-d_j)}. \tag{6.7}$$

This is called the *Greenwood formula* (Greenwood 1926).[3] It was developed originally for life tables and applied later to the Kaplan-Meier estimator.

For each t, a pointwise Wald 95% confidence interval for $S(t)$ is

$$\hat{S}(t) \pm 1.96\sqrt{\operatorname{Var}(\hat{S}(t))}.$$

This confidence interval can have a lower bound less than zero or an upper bound greater than one, outside the possible values of $S(t)$. Better confidence intervals can be obtained using a *complementary log-log* transformation, which is

$$\ln(-\ln S(t)) = \ln H(t)$$

where $H(t)$ is the cumulative hazard. The logit (log odds) transformation can also be used.

[3] Major Greenwood (1880–1949) was an English epidemiologist and statistician. He worked at the Lister Institute (now part of the University of London) and joined the newly-created Ministry of Health after serving in the Royal Army Medical Corps in World War I. He studied the health effects of factory work and developed early models of infectious disease transmission. In 1928, he became the first professor of epidemiology at the London School of Hygiene and Tropical Medicine. In an obituary, Austin Bradford Hill wrote that one of Greenwood's greatest contributions "lay merely in his outlook, in his statistical approach to medicine, then a new approach and one long regarded with suspicion. And he fought this fight continuously and honestly."

6.3.2 Cumulative incidence and cumulative hazard

The Kaplan-Meier estimator of the survival function can also be used to estimate the cumulative hazard function $H(t) = -\ln S(t)$ and the cumulative incidence function $F(t) = 1 - S(t)$, which is the CDF of the time-to-event distribution. The estimated cumulative hazard function is

$$\hat{H}_{\text{KM}}(t) = -\ln \hat{S}(t), \tag{6.8}$$

which is defined whenever $\hat{S}(t) > 0$. The estimated cumulative incidence function is

$$\hat{F}_{\text{KM}}(t) = 1 - \hat{S}(t).$$

When there is no right censoring or left truncation (delayed entry), $\hat{F}(t)$ equals the empirical CDF of the times to events as in Equation 6.2 and the Greenwood variance equals the corresponding variance in Equation 6.3. Confidence limits for $F(t)$ or $H(t)$ can be obtained from the corresponding confidence limits for $S(t)$.

6.4 Nelson-Aalen estimator

The Kaplan-Meier estimator is based on estimating conditional survival probabilities in intervals within which there are no entries or exits. The **Nelson-Aalen** esimator uses the same time intervals, but it estimates an expected number of events in each interval (Nelson 1969, 1972; Altshuler 1970; Aalen 1978). It is based on the interpretation of the cumulative hazard $H(t)$ as an expected number of events in $(0, t]$ if the event could be made repeatable, and it uses the fact that the expected number of events in different intervals can be added together by Equation 1.23. The at-risk process and risk sets are defined exactly as in Section 6.2.1.

6.4.1 Cumulative hazard via addition of expected values

As above, let $0 = t_0 < t_1 < t_2 < ... < t_m$ be the endpoints of intervals $(t_{i-1}, t_i]$ within which there are no entries or exits from the study. Let $n_j = \sum_{i=1}^{n} Y_i(t_j)$ be the number of people in the risk set $\mathcal{R}(t_j)$ and let $d_j \geq 0$ be the number of failures that occur at time t_j. The estimated expected number of events per individual under observation in this time interval is

$$\frac{1}{n_j} \sum_{i \in \mathcal{R}_j} \mathbb{1}_{t_i^{\text{event}} \in (t_{j-1}, t_j]} = \frac{d_j}{n_j}.$$

Adding these up over all time intervals with endpoints at or before time t, we get the Nelson-Aalen estimator

$$\hat{H}(t) = \sum_{j: t_j \leq t} \frac{d_j}{n_j}. \tag{6.9}$$

The Nelson-Aalen estimator is an unbiased, consistent, and asymptotically normal estimator of the true cumulative hazard (Fleming and Harrington 2005; Aalen, Borgan, and Gjessing 2008). Figure 6.2 shows an example based on a right-censored sample of size 500 from a log-logistic distribution with shape $\alpha = 1$ and rate $\lambda = 2$.

The **Fleming-Harrington correction for ties** (Fleming and Harrington 1984) replaces each d_j/n_j in Equation 6.9 with

$$\frac{1}{n_j} + \frac{1}{n_j - 1} + \cdots + \frac{1}{n_j - (d_j - 1)} > \frac{d_j}{n_j}. \tag{6.10}$$

The resulting estimator of the cumulative hazard is sometimes called the *Fleming-Harrington estimator*. It accounts for the fact that the d_j events did not really happen at the same time. They appear to be tied because the times were defined and measured with the limited precision possible in a real study. The example in Figure 6.2 uses this correction.

6.5 R

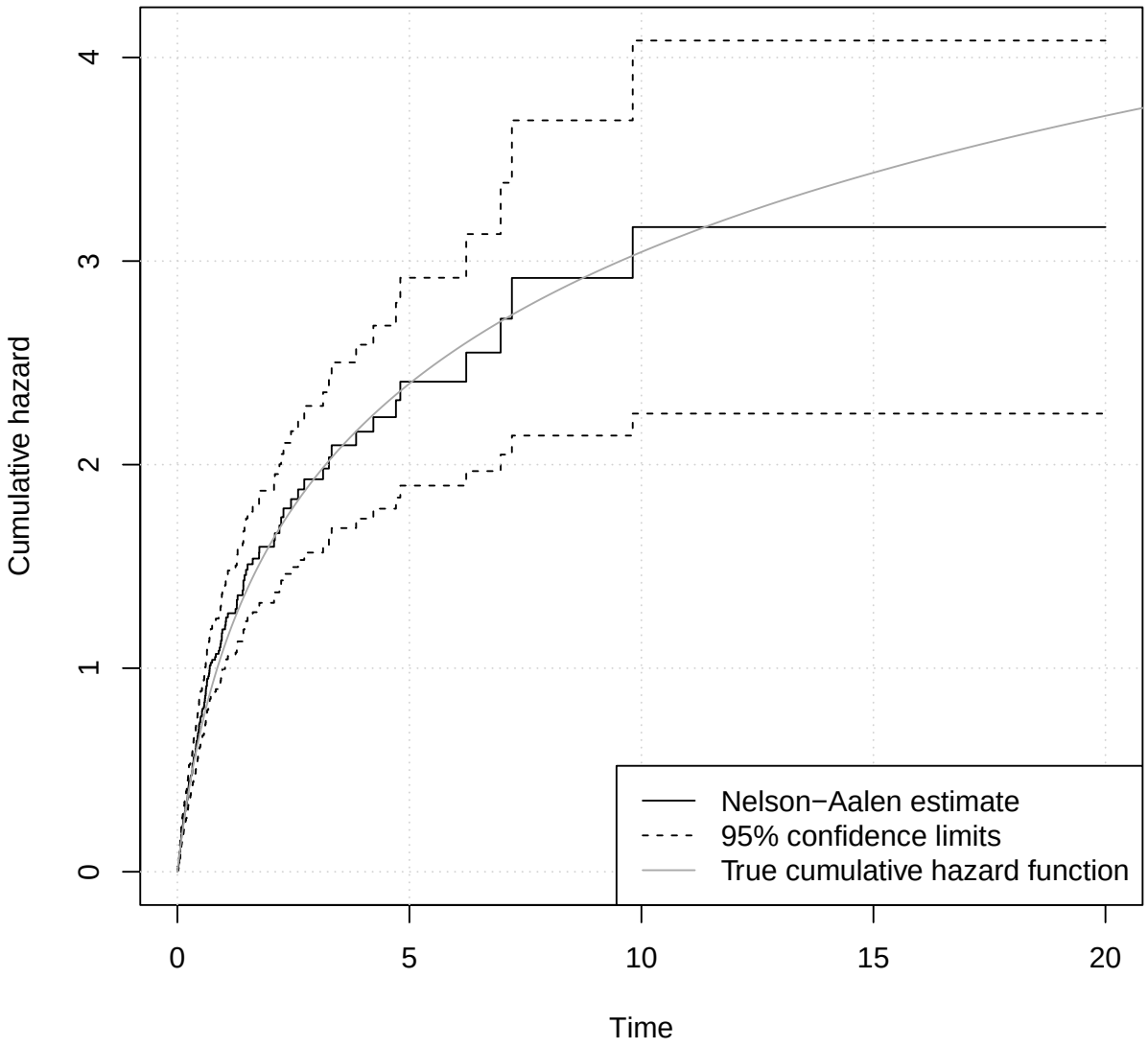


Figure 6.2: True log-logistic cumulative hazard function and Nelson-Aalen estimate with log-transformed 95% confidence limits.

6.5.1 Variance and confidence intervals

Let D_j be the random number of events in the interval $(t_{j-1}, t_j]$ whose observed value is d_j. If an event is repeatable, the number of events per individual in any interval $(t_{j-1}, t_j]$ has a Poisson distribution with mean

$$\Delta H_j = H(t_j) - H(t_{j-1}).$$

When an event cannot be repeated, this is true in an interval sufficiently small that ΔH_j is much smaller than one (i.e., $\Delta H_j \ll 1$). Because our time intervals are defined to be small enough that no one leaves the study by having an event, $\Delta H_j \ll 1$ in each interval where n_j is large.

Because we have n_j individuals at risk in this interval,

$$D_j \sim \text{Poisson}(n_j \Delta H_j).$$

because D_j is the sum of n_j Poisson(ΔH_j) random variables. Because $\text{Var}(D_j) = n_j \Delta H_j$,

$$\text{Var}\left(\frac{D_j}{n_j}\right) = \frac{1}{n_j^2}\text{Var}(D_j) = \frac{\Delta H_j}{n_j}.$$

Replacing the unknown ΔH_j with

$$\Delta\hat{H}_j = \hat{H}(t_j) - \hat{H}(t_{j-1}) = \frac{d_j}{n_j},$$

we get the estimated variance

$$\text{Var}(\Delta\hat{H}_j) = \frac{d_j}{n_j^2}.$$

The variance of $\hat{H}(t)$ is the sum of these variances over all time intervals with endpoints on or before time t, so

$$\text{Var}(\hat{H}(t)) = \sum_{j:t_j \le t} \text{Var}(\Delta\hat{H}_j) = \sum_{j:t_j \le t} \frac{d_j}{n_j^2}.$$

With the Fleming-Harrington correction for ties from Equation 6.10, each d_j/n_j^2 is replaced by

$$\frac{1}{n_j^2} + \frac{1}{(n_j-1)^2} + \cdots + \frac{1}{(n_j-(d_j-1))^2} > \frac{d_j}{n_j^2}.$$

This estimator of the variance (with or without the Fleming-Harrington correction) can be justified more rigorously using the theory of *counting processes* and *martingales* (Fleming and Harrington 2005; Aalen, Borgan, and Gjessing 2008). For our purposes, it is enough to highlight its connection to the Poisson distribution.

For each t, a pointwise Wald 95% confidence interval for $H(t)$ is

$$\hat{H}(t) \pm 1.96\sqrt{\text{Var}(\hat{H}(t))}.$$

This can produce confidence intervals with negative lower bounds, outside the possible values of $H(t)$. A better confidence interval is produced using a log transformation. By the delta method,

$$\text{Var}(\ln\hat{H}(t)) = \frac{1}{\hat{H}(t)^2}\text{Var}(\hat{H}(t)).$$

The corresponding confidence interval for $\hat{H}(t)$ has the endpoints

$$\hat{H}(t)e^{\pm 1.96\sqrt{\mathrm{Var}\left(\ln \hat{H}(t)\right)}}. \tag{6.11}$$

Both endpoints of this confidence interval are nonnegative, and they are strictly positive for all t such that $\hat{H}(t) > 0$. Because $H(t) = -\ln S(t)$, the log transformation for the cumulative hazard function $H(t)$ corresponds to the log-log transformation for the survival function $S(t)$.

6.5.2 Survival and cumulative incidence functions

The Nelson-Aalen estimate $\hat{H}(t)$ can be used to estimate the survival function

$$S(t) = e^{-H(t)},$$

and the cumulative incidence function

$$F(t) = 1 - S(t) = 1 - e^{-H(t)}.$$

The estimated survival function is

$$\hat{S}_{\mathrm{NA}}(t) = e^{-\hat{H}(t)},$$

and the estimated cumulative incidence function is

$$\hat{F}_{\mathrm{NA}}(t) = 1 - \hat{S}_{\mathrm{NA}}(t) = 1 - e^{-\hat{H}(t)}.$$

In both of these estimators, $\hat{H}(t)$ can incorporate the Fleming-Harrington correction for ties from Equation 6.10. Confidence limits for $S(t)$ and $F(t)$ can be obtained from the corresponding confidence limits for $H(t)$.

If there is any t_j where all individuals at risk have an event, the Kaplan-Meier estimator $\hat{S}(t) = 0$ for all $t > t_j$. Once you multiply by zero, there is no going back. This never happens for the estimate of $S(t)$ based on the Nelson-Aalen estimator. Because $\hat{H}(t) < \infty$, we always have

$$\hat{S}_{\mathrm{NA}}(t) = \exp(-\hat{H}(t)) > 0.$$

This is a practical advantage of the Nelson-Aalen estimator over the Kaplan-Meier estimator.

More generally, the Kaplan-Meier estimator produces smaller estimates of $S(t)$ than the Nelson-Aalen estimator does. Similar inequalities exist for the estimated $H(t)$ and $F(t)$:

$$\begin{aligned} \hat{S}(t) &\leq \hat{S}_{\mathrm{NA}}(t) \\ \hat{H}_{\mathrm{KM}}(t) &\geq \hat{H}(t) \\ \hat{F}_{KM}(t) &\geq \hat{F}_{\mathrm{NA}}(t). \end{aligned}$$

Each inequality implies the others, but the cumulative hazard inequality is the simplest. As shown in Figure 6.3,

$$-\ln\left(1-\frac{d_j}{n_j}\right) \geq \frac{d_j}{n_j}$$

with equality only if $d_j > 0$. By Equation 6.6 and Equation 6.8,

$$\hat{H}_{KM}(t) = -\sum_{j:t_j \leq t} \ln\left(1-\frac{d_j}{n_j}\right) \geq \sum_{j:t_j \leq t} \frac{d_j}{n_j} = \hat{H}(t).$$

with equality only when all $d_j = 0$ in the sums. The Nelson-Aalen estimate with the Fleming-Harrington correction for ties is greater than the uncorrected $\hat{H}(t)$ and less than $\hat{H}_{\text{KM}}(t)$. Although not equal, the Nelson-Aalen estimator and Kaplan-Meier estimators of the survival $S(t)$, cumulative hazard $H(t)$, and the cumulative incidence $F(t)$ produce similar results in large samples, where d_j/n_j is small for each j.

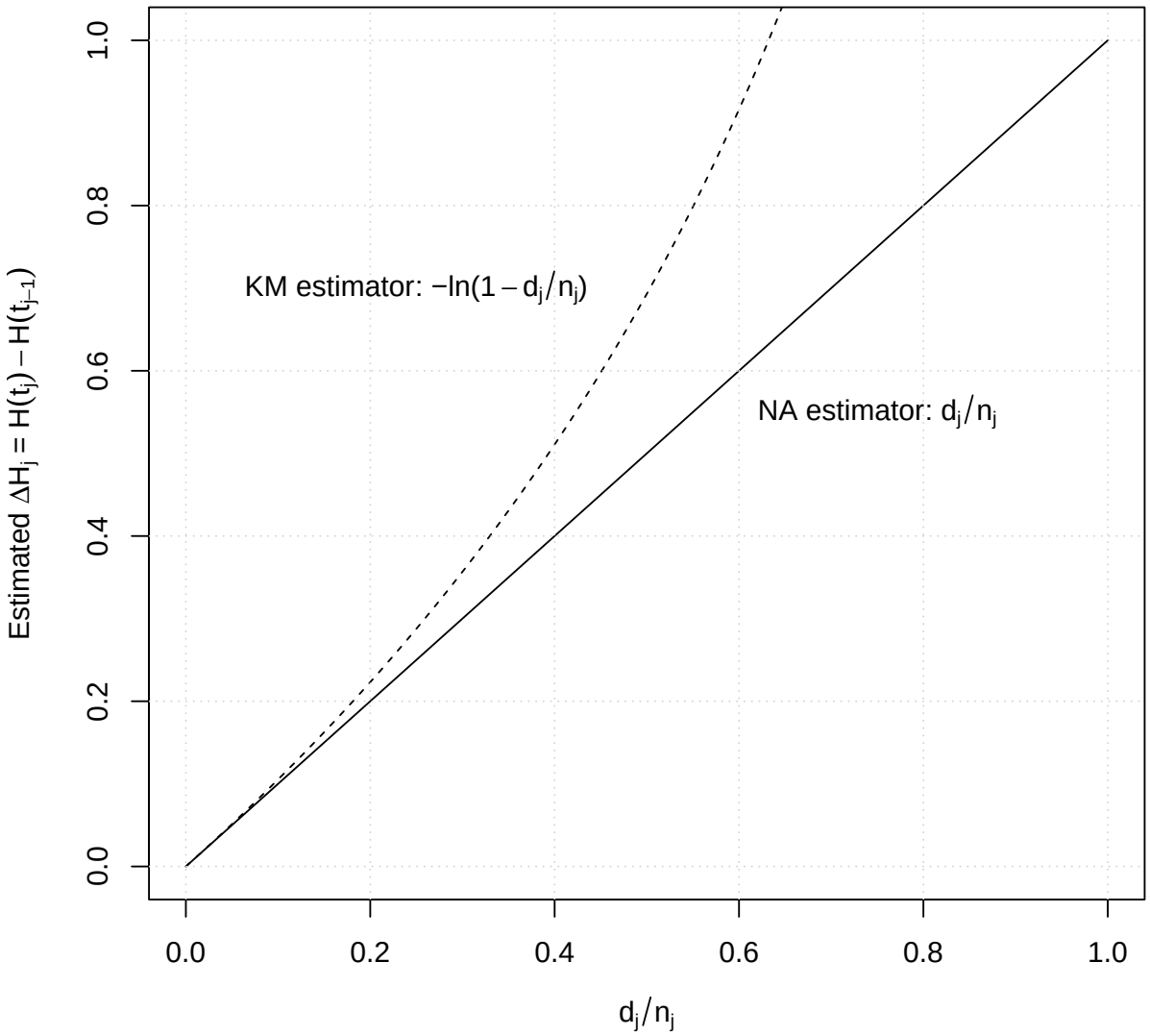


Figure 6.3: Increments in the estimated cumulative hazard in the interval $(t_{j-1}, t_j]$ based on the Kaplan-Meier and Nelson-Aalen estimators.

6.6 Parametric failure time distributions

Many times to events that are important in epidemiology cannot be accurately described using an exponential distribution. In particular, it is important to allow the hazard function $h(t)$ to change over time. Here, we introduce two simple failure time distributions where $h(t)$ is not constant. They each have a rate parameter λ and a shape parameter α. As with the exponential distribution, the scale parameter is $\sigma = 1/\lambda$. The gamma distribution from

equation Equation 5.19 is also used as a failure time distribution, but its survival and hazard functions do not have a simple closed form.

For all parametric failure time models, likelihoods are constructed as in Section 5.2.4. The rate parameter λ and shape parameter α can be estimated using frequentist methods such as maximum likelihood or Bayesian methods. When using parametric methods, it is important to evaluate goodness-of-fit to check whether the underlying assumptions are reasonable.

6.6.1 Weibull distribution

The Weibull distribution (Weibull et al. 1951) is a two-parameter generalization of the exponential distribution.[4] It has the survival function

$$S(t, \alpha, \lambda) = \exp\big(-(\lambda t)^{\alpha}\big),$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the rate parameter. The Weibull cumulative hazard function is

$$H(t, \alpha, \lambda) = -\ln S(t) = (\lambda t)^{\alpha}, \tag{6.12}$$

and its hazard function is

$$h(t, \alpha, \lambda) = \frac{\partial}{\partial t} H(t, \alpha, \lambda) = \alpha \lambda^{\alpha} t^{\alpha - 1}. \tag{6.13}$$

The notation $\partial / \partial t$ means that we take a derivative with respect to t while holding the other arguments, α and λ, constant. Figure 6.4 shows examples of these hazard functions. The exponential distribution is a special case of the Weibull distribution with shape $\alpha = 1$.

[4] Named after Waloddi Weibull (1887-1979), a Swedish engineer and applied mathematician who was a member of the Swedish Coast Guard and invented a technique of using explosives to determine the type and thickness of sediments beneath the sea floor.

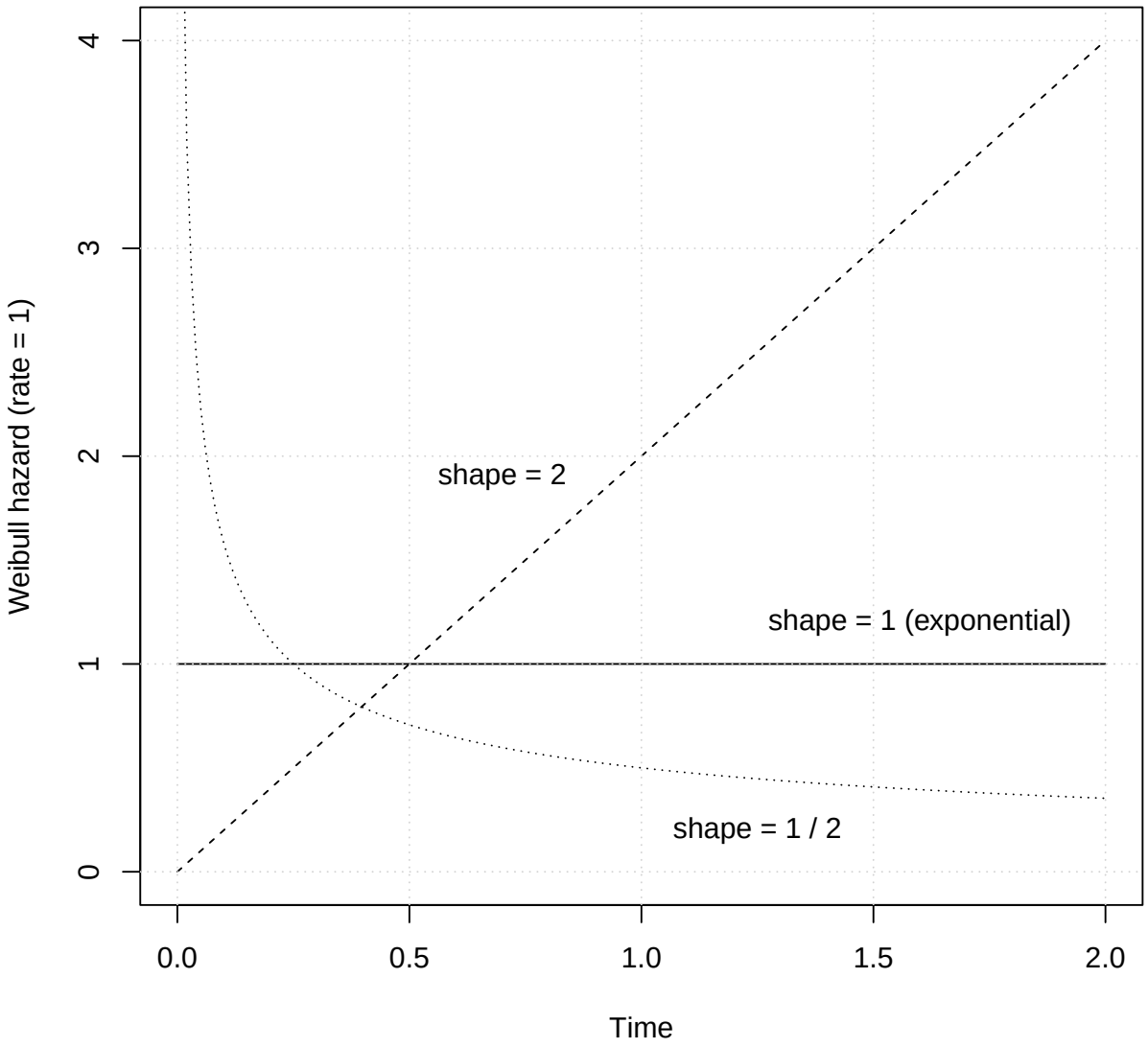


Figure 6.4: Hazard functions for the Weibull distribution with different shape parameters α. All have rate $\lambda = 1$.

6.7 R

6.7.1 Log-logistic distribution

The exponential distribution has a constant hazard, and the Weibull hazard function is increasing, decreasing, or constant. The log-logistic distribution has a more flexible hazard function.

Its survival function is

$$S(t) = \frac{1}{1 + (\lambda t)^{\alpha}}.$$

where $\lambda > 0$ is the rate pameter and $\alpha > 0$ is the shape parameter. Its cumulative hazard function is

$$H(t, \alpha, \lambda) = -\ln S(t) = \ln\left(1 + (\lambda t)^{\alpha}\right),$$

and its the hazard function is

$$h(t, \alpha, \lambda) = \frac{\partial}{\partial t} H(t, \alpha, \lambda) = \frac{\lambda\alpha(\lambda t)^{\alpha-1}}{1 + (\lambda t)^{\alpha}}.$$

As before, we differentiate $H(t, \alpha, \lambda)$ with respect to t while holding α and λ constant. There are three general shapes that the hazard function can take depending on the shape parameter α:

$$h(t) \begin{cases} \text{decreases from } \infty & \text{if } \alpha < 1, \\ \text{decreases from } \lambda & \text{if } \alpha = 1, \\ \text{increases then decreases} & \text{if } \alpha > 1. \end{cases}$$

Figure 6.5 shows examples of these hazard functions. The exponential distribution is not a special case of the log-logistic distribution for any shape α.

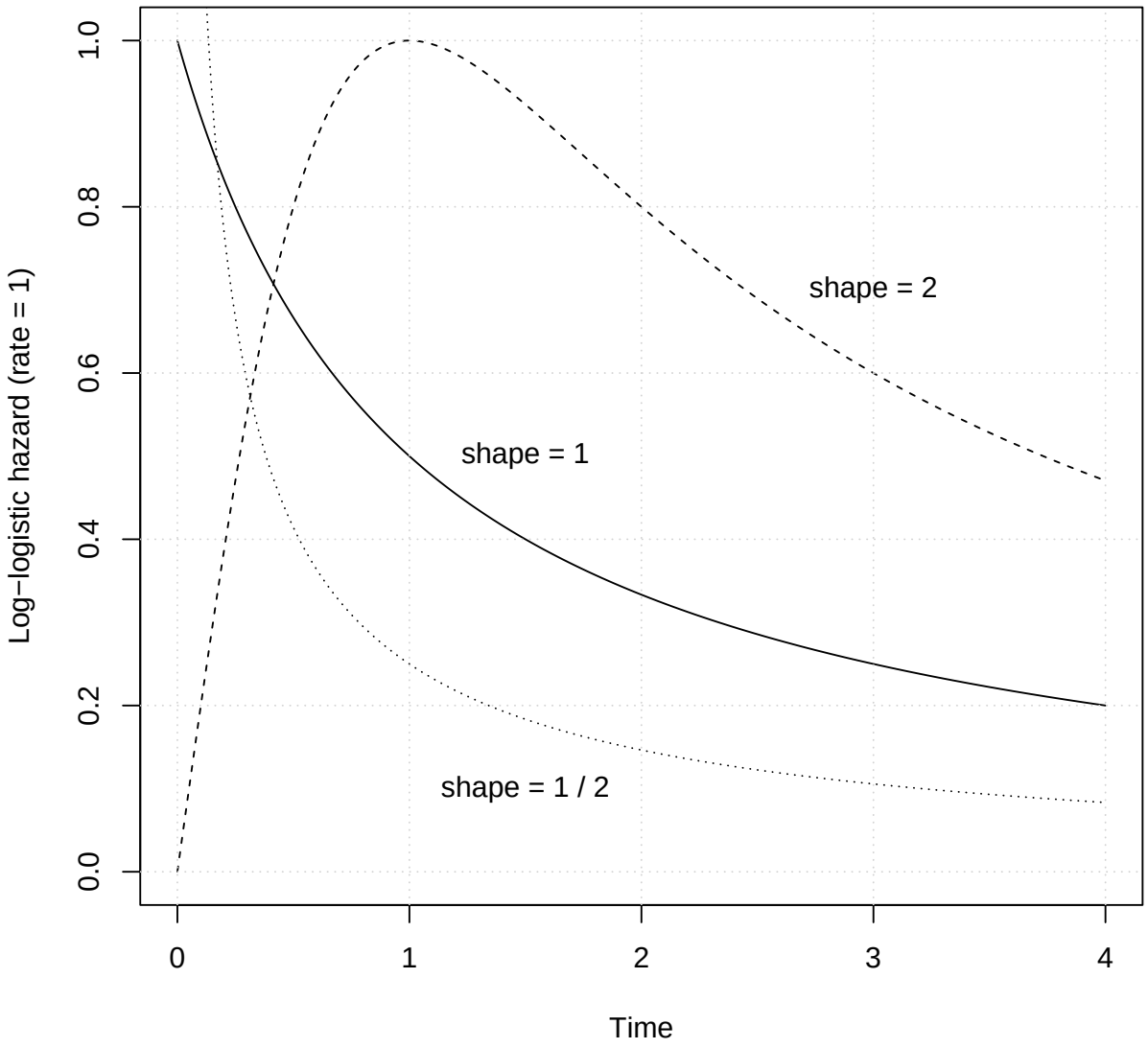


Figure 6.5: Hazard functions for the log-logistic distribution with different shape parameters α. All have rate $\lambda = 1$.

The name of the log-logistic distribution comes from the fact that the logarithm of a log-logistic random variable has a *logistic distribution* (just like the logarithm of a log-normal random variable has a normal distribution). The log-logistic distribution is used in economics to model the distribution of wealth or income (Fisk 1961), where it is known as the *Fisk distribution*.

6.8 R

6.8.1 Cox-Snell residuals

One way to check goodness-of-fit for a parametric failure time model is to use **Cox-Snell residuals** (Cox and Snell 1968). For an observation $(t_i^{\text{entry}}, t_i^{\text{exit}}, \delta_i)$, the Cox-Snell residual is

$$\left(\hat{H}(t_i^{\text{entry}}), \hat{H}(t_i^{\text{exit}}), \delta_i\right),$$

where the estimated cumulative hazards are playing the role of entry and exit times. When the parametric model is correct, the Cox-Snell residuals are a right-censored and left-truncated sample from an exponential distribution with rate $\lambda = 1$. To check this, we can plot the Nelson-Aalen estimate of the cumulative hazard for the Cox-Snell residuals and compare it to the exponential(1) cumulative hazard, which is $H(t) = t$.

Listing 6.1 `loglogistic.R`

```
## Fitting a log-logistic distribution
# To sample a log-logistic random variable in R, you sample a logistic random
# variable with location = -log(rate) and scale = 1 / shape.
# The exponential of the logistic variable has a log-logistic distribution
# with the correct rate and shape parameters.
library(survival)

# Log-logistic distribution and regression (rate = 2, shape = 3)
llsample <- exp(rlogis(1000, location = -log(2), scale = 1 / 3))
llogdat <- data.frame(time = llsample, event = 1)
llogreg <- survreg(Surv(time, event) ~ 1, data = llogdat,
                   dist = "loglogistic")
exp(-coef(llogreg))       # point estimate of rate
exp(-confint(llogreg))    # 95% confidence interval for rate
1 / llogreg$scale         # point estimate of shape

# log-transformed 95% confidence interval for the shape parameter
exp(-log(llogreg$scale) + c(-1, 1)
    * qnorm(.975) * sqrt(vcov(llogreg)["Log(scale)", "Log(scale)"]))

# plot of true and estimated log-logistic hazard functions
llrate_est <- exp(-coef(llogreg))
llshape_est <- 1 / llogreg$scale
llrate_true <- 2
llshape_true <- 3
h_llog <- function(time, rate, shape) {
  # returns last expression if there is no return() statement
  (rate * shape * (time * rate)^(shape - 1) /
   (1 + (time* rate)^shape))
}
t <- seq(0, 4, by = 0.01)
plot(t, h_llog(t, llrate_true, llshape_true), type = "n",
     xlab = "Time", ylab = "Hazard (Log-logistic)")
grid()
lines(t, h_llog(t, llrate_est, llshape_est))
lines(t, h_llog(t, llrate_true, llshape_true), lty = "dashed")
legend("topright", lty = c("solid", "dashed"), bg = "white",
       legend = c("Estimated hazard function", "True hazard function"))
```

Listing 6.2 `KMcurve.R`

```
## Kaplan-Meier survival curve
library(survival)

# right-censored sample from log-logistic dist with rate = 2 and shape = 1
# Uses samples from logistic distribution with location = -log(rate) and
# scale = 1 / shape.
set.seed(42)
llog_f <- exp(rlogis(500, location = -log(2), scale = 1))
llog_c <- exp(rlogis(500, location = -log(2), scale = 2))
t <- pmin(llog_f, llog_c)
d <- ifelse(llog_c < llog_f, 0, 1)
llogdat <- data.frame(time = t, delta = d)

# Kaplan-Meier estimate with complementary log-log confidence intervals
llog_km <- survfit(Surv(time, delta) ~ 1, data = llogdat,
                   conf.type = "log-log")

# Log-logistic survival function
llog_surv <- function(t, lambda=1, gamma=1) 1 / (1 + (lambda * t)^gamma)

# Kaplan-Meier curve and log-logistic survival curve
t <- seq(0, max(llogdat$time), .01)
plot(llog_km, xlim = c(0, 15),
     xlab = "Time", ylab = "Survival probability")
grid()
lines(t, llog_surv(t, 2, 1), col = "darkgray")
legend("topright", bg = "white",
       lty = c("solid", "dashed", "solid"),
       col = c("black", "black", "darkgray"),
       legend = c("Kaplan-Meier estimate", "95% confidence limits",
                  "True survival function"))
```

Listing 6.3 `NelsonAalen.R`

```
## Nelson-Aalen estimator
# The Nelson-Aalen estimator is calculated using survival::survfit().
# Use the argument "stype = 2" to get the survival function estimated from the
# Nelson-Aalen estimate of the cumulative hazard.
# Use the argument "ctype = 2" to get the Fleming-Harrington correction.

library(survival)
?survfit          # get general help about survfit
?survfit.formula  # get help with the specific version we use below

# right-censored sample from log-logistic dist with rate = 2 and shape = 1
# Uses samples from logistic distribution with location = -log(rate) and
# scale = 1 / shape.
set.seed(42)
llog_f <- exp(rlogis(500, location = -log(2), scale = 1))
llog_c <- exp(rlogis(500, location = -log(2), scale = 2))
t <- pmin(llog_f, llog_c)
d <- ifelse(llog_c < llog_f, 0, 1)
llogdat <- data.frame(time = t, delta = d)

# Nelson-Aalen estimator with log-transformed confidence intervals
# The log transformation of H is the log-log transformation of S, so we use
# the argument conf.type = "log-log".
llog_na <- survfit(Surv(time, delta) ~ 1, data = llogdat,
                   conf.type = "log-log", stype = 2, ctype = 2)

# point and interval estimates of the survival function
summary(llog_na, times = 1:15)

# calculate point and interval estimates of the cumulative hazard function
names(summary(llog_na, times = 1:15))
summary(llog_na, times = 1:15)$cumhaz
-log(summary(llog_na, times = 1:15)$surv)
-log(summary(llog_na, times = 1:15)$lower)
-log(summary(llog_na, times = 1:15)$upper)
```

Listing 6.4 `NAcurve.R`

```
## Nelson-Aalen cumulative hazard curve
library(survival)

# right-censored sample from log-logistic dist with rate = 2 and shape = 1
# Uses samples from logistic distribution with location = -log(rate) and
# scale = 1 / shape.
set.seed(42)
llog_f <- exp(rlogis(500, location = -log(2), scale = 1))
llog_c <- exp(rlogis(500, location = -log(2), scale = 2))
t <- pmin(llog_f, llog_c)
d <- ifelse(llog_c < llog_f, 0, 1)
llogdat <- data.frame(time = t, delta = d)

# Nelson-Aalen estimate of the survival function with FH correction
llog_na <- survfit(Surv(time, delta) ~ 1, data = llogdat,
                   conf.type = "log-log", stype = 2, ctype = 2)

# log-logistic cumulative hazard function
llog_cumhaz <- function(t, lambda, gamma) log(1 + (lambda * t)^gamma)

# Nelson-Aalen curve and log-logistic cumulative hazard curve
t <- seq(0, max(llogdat$time), .01)
plot(llog_na, fun = "cumhaz", xlim = c(0, 20),
     xlab = "Time", ylab = "Cumulative hazard")
grid()
lines(t, llog_cumhaz(t, 2, 1), col = "darkgray")
legend("bottomright", bg = "white", lty = c("solid", "dashed", "solid"),
       col = c("black", "black", "darkgray"),
       legend = c("Nelson-Aalen estimate", "95% confidence limits",
                  "True cumulative hazard function"))
```

Listing 6.5 `estHineq.R`

```
## Kaplan-Meier H(t) >= Nelson-Aalen H(t)

# plot of estimated cumulative hazard function increments
x <- seq(0, 1, by = 0.01)
plot(x, x, type = "l", ylim = c(0, 1),
     xlab = expression(d[j] / n[j]),
     ylab = expression(paste("Estimated ", Delta, H[j], " = ",
                             H(t[j]) - H(t[j - 1]))))
lines(x, -log(1 - x), lty = "dashed")
grid()
text(0.75, 0.55, expression(paste("NA estimator: ", d[j] / n[j])))
text(0.25, 0.7, expression(paste("KM estimator: -ln(", 1 - d[j] / n[j], ")",
                                 sep = "")))
```

Listing 6.6 `Weibull-haz.R`

```
## Weibull hazard functions

# hazard function
hweib <- function(t, shape=1, rate=1) shape * rate^shape * t^(shape - 1)

# hazard plots for shapes 2, 1, and 1 / 2
t <- seq(0, 2, by = 0.01)
plot(t, hweib(t, 2), type = "l", lty = "dashed",
     xlab = "Time", ylab = "Weibull hazard (rate = 1)")
lines(t, hweib(t))
lines(t, hweib(t, 0.5), lty = "dotted")
grid()
text(1.6, 1.2, "shape = 1 (exponential)")
text(1.25, 0.2, "shape = 1 / 2")
text(0.7, 1.9, "shape = 2")
```

Listing 6.7 `Weibull.R`

```
## Fitting a Weibull distribution
# In R, the shape is 1 / the "scale" parameter.
# In standard terminology, the scale is 1 / rate.
library(survival)

# Weibull distribution and regression (rate = 2, shape = 3)
# Weibull is the default distribution for survival::survreg().
wsample <- rweibull(1000, shape = 3, scale = 1 / 2)
weibdat <- data.frame(time = wsample, event = 1)
weibreg <- survreg(Surv(time, event) ~ 1, data = weibdat)
summary(weibreg)
exp(-coef(weibreg))       # point estimate of rate
exp(-confint(weibreg))    # 95% confidence interval for rate
1 / weibreg$scale         # point estimate of shape

# log-transformed Wald confidence interval for the shape parameter
# vcov() returns the estimated covariance matrix from the model
exp(-log(weibreg$scale) + c(-1, 1)
    * qnorm(.975) * sqrt(vcov(weibreg)["Log(scale)", "Log(scale)"]))

# plot of true and estimated Weibull hazard functions
wrate_est <- exp(-coef(weibreg))
wshape_est <- 1 / weibreg$scale
wrate_true <- 2
wshape_true <- 3
h_weib <- function(time, rate, shape) rate * shape * (time * rate)^(shape - 1)
t <- seq(0, 4, by = 0.01)
plot(t, h_weib(t, wrate_true, wshape_true), type = "n",
     xlab = "Time", ylab = "Hazard (Weibull)")
grid()
lines(t, h_weib(t, wrate_est, wshape_est))
lines(t, h_weib(t, wrate_true, wshape_true), lty = "dashed")
legend("topleft", lty = c("solid", "dashed"), bg = "white",
       legend = c("Estimated hazard function", "True hazard function"))
```

Listing 6.8 `loglogistic-haz.R`

```r
## Log-logistic hazard function plot

# hazard function
hllog <- function(t, shape=1, rate=1) {
  shape * rate^shape * t^(shape - 1) / (1 + (rate * t)^shape)
}

# hazard plots for shape = 2, 1, 1 / 2
t <- seq(0, 4, by = 0.01)
plot(t, hllog(t, 2), type = "l", lty = "dashed",
     xlab = "Time", ylab = "Log-logistic hazard (rate = 1)")
lines(t, hllog(t))
lines(t, hllog(t, 0.5), lty = "dotted")
grid()
text(1.5, 0.5, "shape = 1")
text(3, 0.7, "shape = 2")
text(1.5, 0.1, "shape = 1 / 2")
```

Listing 6.9 `loglogistic.R`

```
## Fitting a log-logistic distribution
# To sample a log-logistic random variable in R, you sample a logistic random
# variable with location = -log(rate) and scale = 1 / shape.
# The exponential of the logistic variable has a log-logistic distribution
# with the correct rate and shape parameters.
library(survival)

# Log-logistic distribution and regression (rate = 2, shape = 3)
llsample <- exp(rlogis(1000, location = -log(2), scale = 1 / 3))
llogdat <- data.frame(time = llsample, event = 1)
llogreg <- survreg(Surv(time, event) ~ 1, data = llogdat,
                   dist = "loglogistic")
exp(-coef(llogreg))       # point estimate of rate
exp(-confint(llogreg))    # 95% confidence interval for rate
1 / llogreg$scale         # point estimate of shape

# log-transformed 95% confidence interval for the shape parameter
exp(-log(llogreg$scale) + c(-1, 1)
    * qnorm(.975) * sqrt(vcov(llogreg)["Log(scale)", "Log(scale)"]))

# plot of true and estimated log-logistic hazard functions
llrate_est <- exp(-coef(llogreg))
llshape_est <- 1 / llogreg$scale
llrate_true <- 2
llshape_true <- 3
h_llog <- function(time, rate, shape) {
  # returns last expression if there is no return() statement
  (rate * shape * (time * rate)^(shape - 1) /
   (1 + (time* rate)^shape))
}
t <- seq(0, 4, by = 0.01)
plot(t, h_llog(t, llrate_true, llshape_true), type = "n",
     xlab = "Time", ylab = "Hazard (Log-logistic)")
grid()
lines(t, h_llog(t, llrate_est, llshape_est))
lines(t, h_llog(t, llrate_true, llshape_true), lty = "dashed")
legend("topright", lty = c("solid", "dashed"), bg = "white",
       legend = c("Estimated hazard function", "True hazard function"))
```

Part II

Study Design and Measures of Association

7 Cohort and Case-Control Studies

> Like fire, the χ^2 test is an excellent servant and a bad master. (Hill 1965)

Some the most important questions in public health involve the association between a disease and a possible predictor or cause, which we call an exposure. Here, we consider testing the null hypothesis that exposure and disease are independent in a population based on a sample from that population. If exposure and disease are independent, an individual's exposure status contains no information about their risk of disease and vice versa. For simplicity, we focus on a binary exposure and a binary disease outcome and we focus on association, not causation. It turns out this null hypothesis can be tested most efficiently when we sample study participants according to exposure or according to disease (but not both). Sampling by exposure leads to the **cohort study** design, and sampling by disease leads to the **case-control** study design.

7.1 Sampling from a population

Suppose we take a random sample of size n from a population of size $N \gg n$ (i.e., N is much greater than n) and classify each individual in the sample by exposure and disease in a contingency table. We assume that each possible sample of size n is equally likely. Each of the cell counts in the resulting 2x2 table is a random variable in the sample space Ω_n that consists of all possible samples of size n from the population Ω. In Table 7.1, these random variables are $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$. The random row totals are $\mathcal{R}_1 = \mathcal{A} + \mathcal{B}$ and $\mathcal{R}_0 = \mathcal{C} + \mathcal{D}$, and the random column totals are $\mathcal{K}_1 = \mathcal{A} + \mathcal{C}$ and $\mathcal{K}_0 = \mathcal{B} + \mathcal{D}$. The total sample size n is fixed, which means that it is the same for every sample $\omega_n \in \Omega_n$.

Table 7.2 shows the observed values of these random variables from a single sample. These are the values available to us for statistical inference about the independence of exposure and disease.

Table 7.1: Random 2x2 table of exposure (X) and disease (D).

	$D = 1$	$D = 0$	Total
$X = 1$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{R}_1$
$X = 0$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{R}_0$
Total	$\mathcal{K}_1$	$\mathcal{K}_0$	n

Table 7.2: Observed 2x2 table of exposure (X) and disease (D).

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

7.1.1 Hypergeometric distribution*

Over all possible samples from the population Ω, the joint distribution of the cell counts $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ in Table 7.1 is a **multivariate hypergeometric** distribution. Its probability mass function (PMF) is

$$\Pr(\mathcal{A}=a,\mathcal{B}=b,\mathcal{C}=c,D=d)=\frac{\binom{Np_{\mathcal{A}}}{a}\binom{Np_{\mathcal{B}}}{b}\binom{Np_{\mathcal{C}}}{c}\binom{Np_{\mathcal{D}}}{d}}{\binom{N}{n}} \tag{7.1}$$

for all $a,b,c,d\geq 0$ such that $a+b+c+d=n$, where

$$\begin{aligned}p_{\mathcal{A}}&=\Pr(X=1 \text{ and } D=1)\\ p_{\mathcal{B}}&=\Pr(X=1 \text{ and } D=0)\\ p_{\mathcal{C}}&=\Pr(X=0 \text{ and } D=1)\\ p_{\mathcal{D}}&=\Pr(X=0 \text{ and } D=0)\end{aligned}$$

in the underlying population (i.e., where Ω is the population and we sample a single individual ω at random). The numerator in Equation 7.1 is the number of ways of getting cell counts $\mathcal{A}=a$, $\mathcal{B}=b$, $\mathcal{C}=c$, and $\mathcal{D}=d$ in a sample of size n, and the denominator is the number of samples of size n that can be chosen from our population Ω of size $N\geq n$.

The marginal distribution of each cell count is a **hypergeometric distribution**. The PMF of $\mathcal{A}$, which is the number of individuals who are exposed and have disease (or disease onset), is

$$\Pr(\mathcal{A}=a)=\frac{\binom{Np_{\mathcal{A}}}{a}\binom{N(1-p_{\mathcal{A}})}{n-a}}{\binom{N}{n}}.$$

where $a\geq 0$ and $a\leq n$. Its mean is

$$\mathbb{E}(\mathcal{A})=np_{\mathcal{A}},$$

which is identical to the binomial(n, $p_{\mathcal{A}}$) mean. Its variance is

$$\mathrm{Var}(\mathcal{A})=np_{\mathcal{A}}(1-p_{\mathcal{A}})\frac{N-n}{N-1},$$

which is smaller than the binomial(n, $p_{\mathcal{A}}$) variance for all $n>1$. The factor $(N-n)/(N-1)$ is called the *finite population correction*.

The row totals $\mathcal{R}_1$ and $\mathcal{R}_0$ and the column totals $\mathcal{K}_1$ and $\mathcal{K}_0$ from Table 7.1 also have hypergeometric distributions. For $\mathcal{R}_1$, we have

$$\Pr(\mathcal{R}_1 = r_1) = \frac{\binom{N\pi}{r_1}\binom{N(1-\pi)}{n-r_1}}{\binom{N}{n}}$$

where $\pi = \Pr(X = 1)$ is the marginal probability of exposure in the population. For $\mathcal{K}_1$, we have

$$\Pr(\mathcal{K}_1 = k_1) = \frac{\binom{Np}{k_1}\binom{N(1-p)}{n-k_1}}{\binom{N}{n}}$$

where $p = \Pr(D = 1)$ is the marginal prevalence or risk of disease in the population.

As the population size $N \to \infty$, the distribution of $\mathcal{A}$ converges to a binomial$(n,\ p_{\mathcal{A}})$ distribution. If $N \to \infty$ and $n \to \infty$ such that $n^2/N \to 0$, the distribution of

$$\frac{\mathcal{A} - \mathbb{E}(\mathcal{A})}{\sqrt{\mathrm{Var}(\mathcal{A})}}$$

converges to the standard normal distribution $N(0, 1)$. The hypergeometric distributions of the other cell counts and marginal totals also converge to binomial or normal distributions.

7.1.2 Multinomial distribution

If we fix the sample size n and let the population size $N \to \infty$, the multivariate hypergeometric distribution converges to the **multinomial distribution**. Its PMF is

$$\Pr(\mathcal{A} = a, \mathcal{B} = b, \mathcal{C} = c, \mathcal{D} = d) = \frac{n!}{a!b!c!d!} p_{\mathcal{A}}^a p_{\mathcal{B}}^b p_{\mathcal{C}}^c p_{\mathcal{D}}^d.$$

for $a, b, c, d \geq 0$ such that $a+b+c+d = n$. This PMF is written in terms of four probabilities, but there are only three degrees of freedom because $p_{\mathcal{A}} + p_{\mathcal{B}} + p_{\mathcal{C}} + p_{\mathcal{D}} = 1$. In the multinomial distribution, the covariance of $\mathcal{A}$ and $\mathcal{B}$ is

$$\mathrm{Cov}(\mathcal{A}, \mathcal{B}) = -np_{\mathcal{A}}p_{\mathcal{B}},$$

and the covariances for the other five pairs of cell counts follow the same pattern. The multinomial approximation to the multivariate hypergeometric distribution and the binomial approximation to the hypergeometric distribution can be used when N is much larger than n (i.e., $N \gg n$), which is a common situation in epidemiologic studies.[1]

When the *joint* distribution of the cell counts is multinomial, the *marginal* distribution of each cell count is binomial. For example, the distribution of $\mathcal{A}$ is binomial$(n,\ p_{\mathcal{A}})$, so its

[1] The distribution of the cell counts in Table 7.1 is exactly multinomial (and each cell count and row or column total is exactly binomial) if we sample with replacement.

mean is $np_{\mathcal{A}}$ and its variance is $np_{\mathcal{A}}(1-p_{\mathcal{A}})$. The row and column sums also have binomial distributions. The distribution of $\mathcal{R}_1$ is binomial(n, π) where

$$\pi = \Pr(X = 1)$$

is the marginal prevalence of exposure. The distribution of $\mathcal{K}_1$ is binomial(n, p) where

$$p = \Pr(D = 1)$$

is the marginal prevalence or risk of disease.

7.2 Hypothesis tests for independence in a 2x2 table

When exposure and disease are independent, the multiplication rule for independent events implies that

$$\Pr(X = x \text{ and } D = d) = \Pr(X = x)\Pr(D = d)$$

for all possible values x of X and d of D. There are two equivalent ways to express this null hypothesis that will prove useful in thinking about epidemiologic study design: one in terms of conditional risks of disease given exposure and one in terms of conditional prevalences of exposure given disease.

7.2.1 Equality of conditional probabilities

Independence of exposure and disease can be expressed in terms of equality of conditional probabilities of disease (or disease onset) given exposure. Let

$$p_1 = \Pr(D = 1 \,|\, X = 1)$$

be the risk of disease among the exposed and

$$p_0 = \Pr(D = 1 \,|\, X = 0)$$

be the prevalence or risk of disease among the unexposed. If exposure and disease are independent, then

$$\Pr(D = 1 \,|\, X = x) = \frac{\Pr(D = 1)\Pr(X = x)}{\Pr(X = x)} = \Pr(D = 1)$$

for $x = 1$ and $x = 0$. Therefore, $p_1 = p_0$ if exposure and disease are independent. Conversely, suppose $p_1 = p_0$. By definition of p_1 and p_0,

$$\Pr(D = 1 \,|\, X = 1) = \Pr(D = 1 \,|\, X = 0).$$

Expanding the conditional probabilities, we get

$$\frac{\Pr(D = 1 \text{ and } X = 1)}{\Pr(X = 1)} = \frac{\Pr(D = 1 \text{ and } X = 0)}{\Pr(X = 0)}.$$

This can be rewritten as

$$\frac{\Pr(D = 1 \text{ and } X = 1)}{\Pr(X = 1)} = \frac{\Pr(D = 1) - \Pr(D = 1 \text{ and } X = 1)}{1 - \Pr(X = 1)}.$$

Cross-multiplying the numerators and denominators shows that this equality holds if and only if

$$\Pr(D = 1 \text{ and } X = 1) = \Pr(D = 1) \Pr(X = 1).$$

Because D and X are binary, this is establishes that D and X are independent random variables. Therefore, $p_1 = p_0$ implies that exposure and disease are independent. Combining both results shows that $H_0 : p_1 = p_0$ is equivalent to the null hypothesis that exposure and disease are independent.

A similar argument applies to the conditional prevalence of exposure given disease status. Let

$$\pi_1 = \Pr(X = 1 \,|\, D = 1)$$

be the prevalence of exposure among cases and

$$\pi_0 = \Pr(X = 1 \,|\, D = 0)$$

be the prevalence of exposure among controls. The null hypothesis $H_0 : \pi_1 = \pi_0$ is equivalent to the null hypothesis that exposure and disease are independent.

7.2.2 Hypergeometric chi-squared test

Under the null hypothesis that exposed and disease are independent, we have

$$\begin{aligned}
p_{\mathcal{A}} &= \Pr(X = 1) \Pr(D = 1) = \pi p, \\
p_{\mathcal{B}} &= \Pr(X = 1) \Pr(D = 0) = \pi(1 - p), \\
p_{\mathcal{C}} &= \Pr(X = 0) \Pr(D = 1) = (1 - \pi)p, \\
p_{\mathcal{D}} &= \Pr(X = 0) \Pr(D = 0) = (1 - \pi)p.
\end{aligned}$$

The marginal prevalence of exposure π and the marginal risk of disease p are both unknown. In a score test of the null hypothesis, these are *nuisance parameters* that can be replaced by maximum likelihood estimates (Rao 1948; Boos and Stefanski 2013). Because $\mathcal{R}_1$ has an approximate binomial(n, π) distribution when $N \gg n$,

$$\hat{\pi} = \frac{r_1}{n}. \tag{7.2}$$

Table 7.3: 2x2 table determined by $\mathcal{A}$ and the margins.

	$D=1$	$D=0$	Total
$X=1$	$\mathcal{A}$	$r_1-\mathcal{A}$	r_1
$X=0$	$k_1-\mathcal{A}$	$\mathcal{A}-(a-d)$	r_0
Total	k_1	k_0	n

is the maximum likelihood estimate of π based on Table 7.2. Because $\mathcal{K}_1$ has an approximate binomial$(n,\, p)$ distribution when $N \gg n$,

$$\hat{p} = \frac{k_1}{n} \tag{7.3}$$

is the maximum likelihood estimate of p based on Table 7.2. When we use these maximum likelihood estimates of π and p to test independence, we are conditioning on the row and column totals in the observed 2x2 table in Table 7.2.

Given the margins of a 2x2 table, the entire table is determined by any one of the four cell counts. Table 7.3 shows how the cell counts in Table 7.2 are determined by $\mathcal{A}$ and the margins. Because all cell counts must be nonnegative, we must have $\mathcal{A} \geq 0$, $\mathcal{A} \leq r_1$, and $\mathcal{A} \leq k_1$. In the bottom right cell of the 2×2 table, we must have

$$\mathcal{A} - (a-d) \geq 0.$$

Therefore,

$$a_{\min} = \max(0, a-d) \leq \mathcal{A} \leq \min(r_1, k_1) = a_{\max}.$$

Note that the cells along the diagonal of the 2×2 table (the $\mathcal{A}$ and $\mathcal{D}$ cells) both increase with $\mathcal{A}$, while the cells off the diagonal (the $\mathcal{B}$ and $\mathcal{C}$ cells) both decrease with $\mathcal{A}$. Any of the other cells could also determine the entire table given the margins, and constraints on the possible values of $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ given the margins could be found in a similar way.

The conditional distribution of the cell count $\mathcal{A}$ given the margins of Table 7.2 is hypergeometric. Imagine our sample as a bowl of n marbles, r_1 of which are exposed and r_0 of which are unexposed. If we randomly choose k_1 marbles without replacement to represent the individuals with disease, then $\mathcal{A}$ is the number of exposed marbles in our sample. The probability that we get a exposed marbles and $k_1 - a$ unexposed marbles is

$$\Pr(\mathcal{A} = a \,|\, \text{margins}) = \frac{\binom{r_1}{a}\binom{r_0}{k_1-a}}{\binom{n}{k_1}} = \frac{\binom{r_1}{a}\binom{r_0}{c}}{\binom{n}{k_1}} = \frac{r_1!\, r_0!\, k_1!\, k_0!}{a!\, b!\, c!\, d!\, n!}$$

We could also view our sample as a bowl of n marbles of which k_1 have disease (or disease onset) and k_0 do not. In that case, A is the number of diseased marbles in a sample of r_1 marbles that represent exposed individuals and we get exactly the same hypergeometric distribution

of $\mathcal{A}$. The cell counts $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ also have hypergeometric distributions given the margins of the table.

Under the null hypothesis that exposure and disease are independent, the conditional mean of $\mathcal{A}$ given the margins of Table 7.2 is

$$\mathbb{E}(\mathcal{A} \,|\, \text{margins}) = n\hat{\pi}\hat{p} = \frac{r_1 k_1}{n}$$

and its conditional variance is

$$\text{Var}(\mathcal{A} \,|\, \text{margins}) = \frac{r_1 r_0 k_1 k_0}{n^2(n-1)}.$$

For large n, the hypergeometric distribution is approximately normal so the hypergeometric chi-squared statistic is

$$\chi^2_{\text{H}} = \frac{\big(a - \mathbb{E}(\mathcal{A} \,|\, \text{margins})\big)^2}{\text{Var}(\mathcal{A} \,|\, \text{margins})} = \frac{(n-1)(ad-bc)^2}{r_1 r_0 k_1 k_0} \tag{7.4}$$

Under the null hypothesis, $\chi^2_{\text{H}} \overset{\text{approx}}{\sim} \chi^2_1$ (i.e., the chi-squared distribution with one degree of freedom). The p-value is $1 - F(\chi^2_{\text{H}})$ where F is the cumulative distribution function (CDF) of the χ^2_1 distribution. We reject the null hypothesis at significance level α when χ^2_{H} is sufficiently large that the p-value is less than α. We get exactly the same hypothesis test using $\mathcal{B}$, $\mathcal{C}$, or $\mathcal{D}$ instead of $\mathcal{A}$.

7.2.3 Pearson's chi-squared test

A more general approach to testing independence of the rows and columns in a contingency table is **Pearson's chi-squared test** (Pearson 1900, 1922).[2] Like the hypergeometric test, Pearson's chi-squared test conditions on the margins of the table. In a contingency table with I rows and J columns, let O_{ij} be the observed cell count in row i and column j. Let r_i be the total for row i and k_j be the total for column j. Under independence, the expected cell count is

$$E_{ij} = \frac{r_i k_j}{n}.$$

Pearson's chi-squared statistic is

$$\chi^2_{\text{P}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{7.5}$$

Under the null hypothesis that the variables defining the rows and the colums are independent, χ^2_{P} has a chi-squared distribution with $(I-1)(J-1)$ degrees of freedom (Fisher 1922; Boos

[2] Named after Karl Pearson (1857–1936), an English statistician who appeared in the context of the Pearson correlation coefficient in Section 1.7.2.

and Stefanski 2013). In any contingency table, Pearson's chi-squared test is the score test of the null hypothesis that the rows and columns are independent based on a multinomial model (see Section 7.1.2) for the joint distribution of the cell counts (Boos and Stefanski 2013).

In Table 7.2, we have $I = J = 2$ with $O_{11} = a$, $O_{12} = b$, $O_{21} = c$, and $O_{22} = d$. Using the multivariate hypergeometric distribution or its multinomial approximation, we have the following estimated expected cell counts under the null hypothesis that exposure and disease are independent:

$$\begin{aligned}
\mathbb{E}_{11} &= \mathbb{E}(\mathcal{A} \,|\, \text{margins}) = n\hat{\pi}\hat{p} = \frac{r_1 k_1}{n} \\
\mathbb{E}_{12} &= \mathbb{E}(\mathcal{B} \,|\, \text{margins}) = n\hat{\pi}(1-\hat{p}) = \frac{r_1 k_0}{n} \\
\mathbb{E}_{21} &= \mathbb{E}(\mathcal{C} \,|\, \text{margins}) = n(1-\hat{\pi})\hat{p} = \frac{r_0 k_1}{n} \\
\mathbb{E}_{22} &= \mathbb{E}(\mathcal{D} \,|\, \text{margins}) = n(1-\hat{\pi})(1-\hat{p}) = \frac{r_0 k_0}{n}.
\end{aligned} \tag{7.6}$$

As in the hypergeometric chi-squared test, we are conditioning on the margins of the table because we are using the maximum likelihood estimates of π (the prevalence of exposure) and p (the risk of disease). When the dust settles in Equation 7.5, we get

$$\chi^2_{\text{P}} = \frac{n(ad-bc)^2}{r_1 r_0 k_1 k_0} = \frac{n}{n-1}\chi^2_{\text{H}}. \tag{7.7}$$

When exposure and disease are independent, χ^2_{P} has a chi-squared distribution with $(2-1)(2-1) = 1$ degrees of freedom. The p-value is $1 - F(\chi^2_{\text{P}})$ where F is the CDF of the χ^2_1 distribution, and we reject the null hypothesis at significance level α when χ^2_{P} is sufficiently large that the p-value is less than α.

The chi-squared approximation to the distribution of χ^2_{P} is generally considered acceptable if the minimum expected cell count is greater than or equal to five, and it is likely to be accurate whenever the average expected cell count is greater than or equal to 7.5 (Roscoe and Byars 1971), which is equivalent to $n \geq 30$ for a 2x2 table.[3] Because $\chi^2_{\text{H}} < \chi^2_{\text{P}}$, the hypergeometric chi-squared test is slightly more conservative than Pearson's chi-squared test in the sense that it is less likely to reject the null hypothesis of independence. For large n, there is no practical difference.

7.2.4 Small samples and exact tests*

In small samples, the hypergeometric distribution can be used to calculate "exact" p-values. For two-sided alternative hypotheses, this leads to **Fisher's exact test** (Fisher 1935; Irwin

[3] These rules of thumb are for chi-squared tests with one degree of freedom and significance level $\alpha = 0.05$. Smaller α require larger average cell counts to estimate smaller p-values accurately, and chi-squared tests with more than one degree of freedom are more robust to small expected cell counts (Roscoe and Byars 1971).

et al. 1935) or **Blaker's exact test** (Blaker 2000). These use the hypergeometric PMF to calculate a p-value for the null hypothesis of independent rows and columns. These tests differ slightly in the way that they define the tails of the distribution of $\mathcal{A}$, and there are two versions of Fisher's exact test.

The *minimum likelihood* Fisher's exact test defines the p-value as the sum of the probabilities of all possible a such that $\Pr(A = a \,|\, \text{margins}) \leq \Pr(\mathcal{A} = a \,|\, \text{margins})$. The simpler but slightly less powerful *central* Fisher's exact test defines the p-value as twice the twice the minimum of the tail probabilities $\Pr(\mathcal{A} \leq a \,|\, \text{margins})$ and $\Pr(A \geq a \,|\, \text{margins})$.[4]

Blaker's exact test defines the p-value as the minumum tail probability plus the probability of an opposite tail defined so that its probability is less than or equal to that of the smaller tail. For example: If the smaller tail is $\mathcal{A} \leq a$, then the p-value is

$$\Pr(A \leq a \,|\, \text{margins}) + \sum_{a'=a_{\text{opp}}}^{a_{\max}} \Pr(A = a' \,|\, \text{margins})$$

where a_{opp} is chosen so that the sum in the second term is less than or equal to $\Pr(A \leq a \,|\, \text{margins})$. Blaker's test is sometimes more powerful and never less powerful than both versions of Fisher's exact test (Blaker 2000; Fay 2010).

These tests are "exact' ' in the sense that they reject a true null hypothesis with probability less than or equal to the nominal significance level α. However, they are often overly conservative in that the true significance level (i.e., the actual probability of rejecting the null hypothesis when it is true) can be substantially less than α. Using mid-p values mitigates this problem, ensuring that the true significance level stays closer to α. The price of this is that the true significance level of the test can be slightly greater than α, so the mid-p tests are no longer "exact" (Lancaster 1961; Routledge 1992; Agresti 2013).

7.3 Cohort studies

Random sampling from the population is not the most efficient way to detect a departure from independence of exposure and disease. By rearranging the Pearson chi-squared statistic χ^2_{P} from equation Equation 7.7, we can identify two strategies for generating a more powerful test. One is to select participants by exposure, which leads to the **cohort study** design. The other is to select participants by disease, which leads to the **case-control** study design. In both cases, a balanced study design is optimal (or near-optimal) and Pearson's chi-squared test is the score test of the null hypothesis that exposure and disease are independent. If participation in the study involves any cost, risk, or inconvenience, then maximizing the power of the study

[4] Inversion of the minimum likelihood Fisher's or Blaker's exact tests can produce confidence regions for the odds ratio that consist of two disjoint intervals, but inversion of the central Fisher's exact test always produces a confidence region that consists of a single interval (Fay 2010).

for a given number of participants is an important ethical consideration because an inefficient study will place an unnecessary burden on some participants.

7.3.1 Selection by exposure

The Pearson chi-squared statistic χ^2_P from Equation 7.5 can be rewritten in terms of the risks of disease in exposed and unexposed individuals. As above, let p_1 be the risk of disease in the exposed and p_0 be the risk of disease in the unexposed. In Table 7.2, their maximum likelihood estimates are $\hat{p}_1 = a/r_1$ and $\hat{p}_0 = c/r_0$. The maximum likelihood estimate of $p_1 - p_0$ is

$$\hat{p}_1 - \hat{p}_0 = \frac{a}{a+b} - \frac{c}{c+d} = \frac{ad-bc}{(a+b)(c+d)} = \frac{ad-bc}{r_1 r_0}. \tag{7.8}$$

Section 7.2.1 showed that the null hypothesis that exposure and disease are independent is equivalent to $H_0 : p_1 = p_0 = p$ where p is the marginal risk of disease.

When $n \ll N$ and the null hypothesis is true, $\mathcal{A}$ has an approximate binomial(r_1, p) conditional distribution, $\mathcal{C}$ has an approximate binomial$(r_0,\ p)$ conditional distribution, and they are conditionally independent given the row sums r_1 and r_0. Thus, the large-sample variance of $\hat{p}_1 - \hat{p}_0$ under the null is approximately

$$\mathrm{Var}_0(\hat{p}_1 - \hat{p}_0) = p(1-p)\left(\frac{1}{r_1} + \frac{1}{r_0}\right) = p(1-p)\frac{n}{r_1 r_0} \tag{7.9}$$

where we used $n = r_1 + r_0$. Replacing the unknown p with its maximum likelihood estimate $\hat{p} = k_1/n$, we get the estimated null variance

$$\hat{\mathrm{Var}}_0(\hat{p}_1 - \hat{p}_0) = \hat{p}(1-\hat{p})\frac{n}{r_1 r_0} = \frac{k_1 k_0}{r_1 r_0 n} \tag{7.10}$$

where we used $1 - \hat{p} = k_0/n$. Combining Equation 7.8 and Equation 7.9, we get

$$\frac{(\hat{p}_1 - \hat{p}_0)^2}{\hat{p}(1-\hat{p})\left(\frac{1}{r_1} + \frac{1}{r_0}\right)} = \frac{n(ad-bc)^2}{r_1 r_0 k_1 k_0} = \chi^2_P$$

(see Equation 7.7). Let φ be the proportion of our sample that is exposed, so $r_1 = \varphi n$ and $r_0 = (1-\varphi)n$. As $n \to \infty$, we have $r_1 \to \infty$ and $r_0 \to \infty$. The law of large numbers (LLN) guarantees that $\hat{p}_1 \to p_1$, $\hat{p}_0 \to p_0$, and

$$\hat{p} \to p_\varphi = \varphi p_1 + (1-\varphi)p_0.$$

In large samples,

$$\chi^2_P \approx \frac{(p_1 - p_0)^2}{p_\varphi(1-p_\varphi)\left(\frac{1}{r_1} + \frac{1}{r_0}\right)} = \frac{(p_1 - p_0)^2}{p_\varphi(1-p_\varphi)\frac{1}{\phi(1-\phi)n}} \tag{7.11}$$

The numerator of Equation 7.11 is fixed, but the denominator depends on r_1, r_0, and $\varphi = r_1/n$. By sampling according to exposure, we can choose r_1 and r_0 to increase the power of the Pearson chi-squared test for a fixed total number of participants.

7.3.2 Score test for independence in a cohort study*

We need to make sure that sampling by exposure does not change the score test of the null hypothesis that exposure and disease are independent. Using a binomial(r_1, p_1) distribution for the number of individuals with disease the exposed group and a binomial(r_0, p_0) distribution for the number of individuals with disease in the unexposed group, the log likelihood is

$$\ell(p_1, p_0) = \mathcal{A} \ln p_1 + \mathcal{B} \ln(1 - p_1) + \mathcal{C} \ln p_0 + \mathcal{D} \ln(1 - p_0),$$

where we have dropped terms that do not depend on p_1 or p_0. In order to calculate the expected information for the score test, we view the log likelihood as a random variable whose value will be determined by the realized values a, b, c, and d of the random variables $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$. The score function and the information function will also be treated as random variables.

Because $\ell(p_1, p_0)$ depends on two parameters, the score function is a column vector of length two:

$$U(p_1, p_0) = \begin{pmatrix} \frac{\partial}{\partial p_1}\ell(p_1, p_0) \\ \frac{\partial}{\partial p_0}\ell(p_1, p_0) \end{pmatrix} = \begin{pmatrix} \frac{\mathcal{A}}{p_1} - \frac{\mathcal{B}}{1-p_1} \\ \frac{\mathcal{C}}{p_0} - \frac{\mathcal{D}}{1-p_0} \end{pmatrix}.$$

The information $I(p_1, p_0)$ is a 2x2 matrix

$$\begin{bmatrix} \frac{\partial^2}{\partial p_1^2}\ell(p_1, p_0) & \frac{\partial^2}{\partial p_1 \partial p_0}\ell(p_1, p_0) \\ \frac{\partial^2}{\partial p_0 \partial p_1}\ell(p_1, p_0) & \frac{\partial^2}{\partial p_0^2}\ell(p_1, p_0) \end{bmatrix} = \begin{bmatrix} \frac{\mathcal{A}}{p_1^2} + \frac{\mathcal{B}}{(1-p_1)^2} & 0 \\ 0 & \frac{\mathcal{C}}{p_0^2} + \frac{\mathcal{D}}{(1-p_0)^2} \end{bmatrix}.$$

The realized value of $U(p_1, p_0)$ and the observed information $I(p_1, p_0)$ are obtained by replacing the random variables $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ with their realized values a, b, c, and d.

The score statistic is calculated under the null hypothesis $H_0 : p_1 = p_0 = p$, and we use the expected information (D. A. Freedman 2007). Let $\mathbb{E}_0(Y)$ be the expected value of a random variable Y calculated under H_0. Then $\mathbb{E}_0(\mathcal{A}) = n_1 p$, $\mathbb{E}_0(\mathcal{B}) = n_1(1 - p)$, $\mathbb{E}_0(\mathcal{C}) = n_0 p$, and $\mathbb{E}_0(\mathcal{D}) = n_0(1 - p)$, so the expected information under H_0 is

$$\mathcal{I}(p, p) = \mathbb{E}_0[I(p, p)] = \begin{bmatrix} \frac{n_1}{p} + \frac{n_1}{1-p} & 0 \\ 0 & \frac{n_0}{p} + \frac{n_0}{1-p} \end{bmatrix}.$$

Both $U(p, p)$ and $\mathcal{I}(p, p)$ depend on the unknown p, which we replace with its maximum likelihood estimate $\hat{p} = k_1/n$. This gives us the score

$$U(\hat{p}, \hat{p}) = \begin{pmatrix} \frac{a}{\hat{p}} - \frac{b}{1-\hat{p}} \\ \frac{c}{\hat{p}} - \frac{d}{1-\hat{p}} \end{pmatrix} = \begin{pmatrix} \frac{na}{k_1} - \frac{nb}{k_0} \\ \frac{nc}{k_1} - \frac{nd}{k_0} \end{pmatrix} = \begin{pmatrix} \frac{n(ad-bc)}{k_1 k_0} \\ -\frac{n(ad-bc)}{k_1 k_0} \end{pmatrix}.$$

where we used $k_1 = a + c$ and $k_0 = b + d$. The expected information at $p = \hat{p}$ is

$$\mathcal{I}(\hat{p}, \hat{p}) = \begin{bmatrix} \frac{r_1 n^2}{k_1 k_0} & 0 \\ 0 & \frac{r_0 n^2}{k_1 k_0} \end{bmatrix} \Rightarrow \mathcal{I}^{-1}(\hat{p}, \hat{p}) = \begin{bmatrix} \frac{k_1 k_0}{r_1 n^2} & 0 \\ 0 & \frac{k_1 k_0}{r_0 n^2} \end{bmatrix}$$

where we used $n_1 = r_1$ and $n_0 = r_0$. The score statistic is

$$U(\hat{p}, \hat{p})^\mathsf{T} \mathcal{J}(\hat{p}, \hat{p})^{-1} U(\hat{p}, \hat{p}) = \frac{n(ad - bc)^2}{r_1 r_0 k_1 k_0} = \chi^2_\text{P},$$

from Equation 7.7. Because H_0 reduces the degrees of freedom from two (p_1 and p_0) to one ($p_1 = p_0 = p$), χ^2_P has an asymptotic χ^2 distribution with $2 - 1 = 1$ degree of freedom under the null. Therefore, Pearson's chi-squared test is the score test of independence of exposure and disease in a cohort study. The row sums r_1 and r_0 are fixed by design, and we condition on the column sums k_1 because we use the maximum likelihood estimate $\hat{p} = k_1/n$ of the risk of disease under H_0.

When it uses the expected information, the score test does not depend on the parameterization of the model for p_1 and p_0 (Boos and Stefanski 2013). We get the same score statistic χ^2_P and the same χ^2_1 distribution under the null even if the model uses transformations of p_1 and p_0 (e.g., log or logit) or if it is parameterized in terms of the *risk difference* $\text{RD} = p_1 - p_0$, the *risk ratio* $\text{RR} = p_1/p_0$, or the *odds ratio* $\text{OR} = \text{odds}(p_1)/\text{odds}(p_0)$ where $\text{odds}(p) = p/(1-p)$. All roads lead to the same score test of the null hypothesis that exposure and disease are independent, which corresponds to $\text{RD} = 0$ and $\text{RR} = \text{OR} = 1$.

7.3.3 Optimal sampling by exposure

Having established that χ^2_P is the score statistic for testing the independence of exposure and disease in a cohort study, we can choose r_1 and r_0 to maximize the power of the test for a given number of participants $n = r_1 + r_0$. The value of the chi-squared statistic in Equation 7.11 depends on r_1 and r_0 only in the denominator, so we can maximize the statistic by minimizing its denominator. Writing the denominator of Equation 7.10 in terms of p_1, p_0, and $\varphi = r_1/n$ of the sample that is exposed and simplifying gives us

$$\frac{np(1-p)}{r_1 r_0} = \frac{\varphi}{1-\varphi} p_1(1-p_1) + \frac{1-\varphi}{\varphi} p_0(1-p_0) + C(p_1, p_0) \tag{7.12}$$

where $C(p_1, p_0) = p_1(1-p_0) + p_0(1-p_1)$ does not depend on φ. The derivative of this with respect to φ is

$$\frac{\mathrm{d}}{\mathrm{d}\varphi} \frac{np(1-p)}{r_1 r_0} = \frac{p_1(1-p_1)}{(1-\varphi)^2} - \frac{p_0(1-p_0)}{\varphi^2}, \tag{7.13}$$

which equals zero when

$$\frac{\varphi}{1-\varphi} = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}. \tag{7.14}$$

To see that this is a minimum and not a maximum, notice that the function in equation Equation 7.12 takes large values for φ near one when $p_1(1-p_1) > 0$ and for φ near zero when $p_0(1-p_0) > 0$. It also has a positive second derivative with respect to φ.

Solving for φ in Equation 7.14 shows that a proportion exposed of

$$\varphi^* = \frac{1}{1 + \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}}. \tag{7.15}$$

maximizes the expected value of the Pearson chi-squared statistic χ^2_{P} for a given n (Walter 1977). The expression inside the square root is the variance of a Bernoulli(p_1) random variable divided by the variance of a Bernoulli(p_0) random variable. Figure 7.1 shows how φ^* depends on this variance ratio. When $p_1 \approx p_0$, the Bernoulli variance ratio is approximately one and $\varphi^* \approx 0.5$.

Listing 7.1 `optim-phi.R`

```
## Optimal proportion exposed in a cohort study

# plot of optimal phi as a function of the Bernoulli variance ratio
logvratio <- seq(-3, 3, by = 0.01)
phi <- function(v) 1 / (1 + sqrt(v))
plot(logvratio, phi(exp(logvratio)), type = "l", xaxt = "n", ylim = c(0, 1),
     xlab = "Bernoulli variance ratio (log scale)",
     ylab = expression(paste("Optimal proportion exposed or cases (", phi, "*)")))
axis(1, at = log(c(1 / c(16, 8, 4, 2), 1, c(2, 4, 8, 16))),
     labels = c("1/16", "1/8", "1/4", "1/2", 1, 2, 4, 8, 16))
grid()
abline(h = 0.5, col = "darkgray")
```

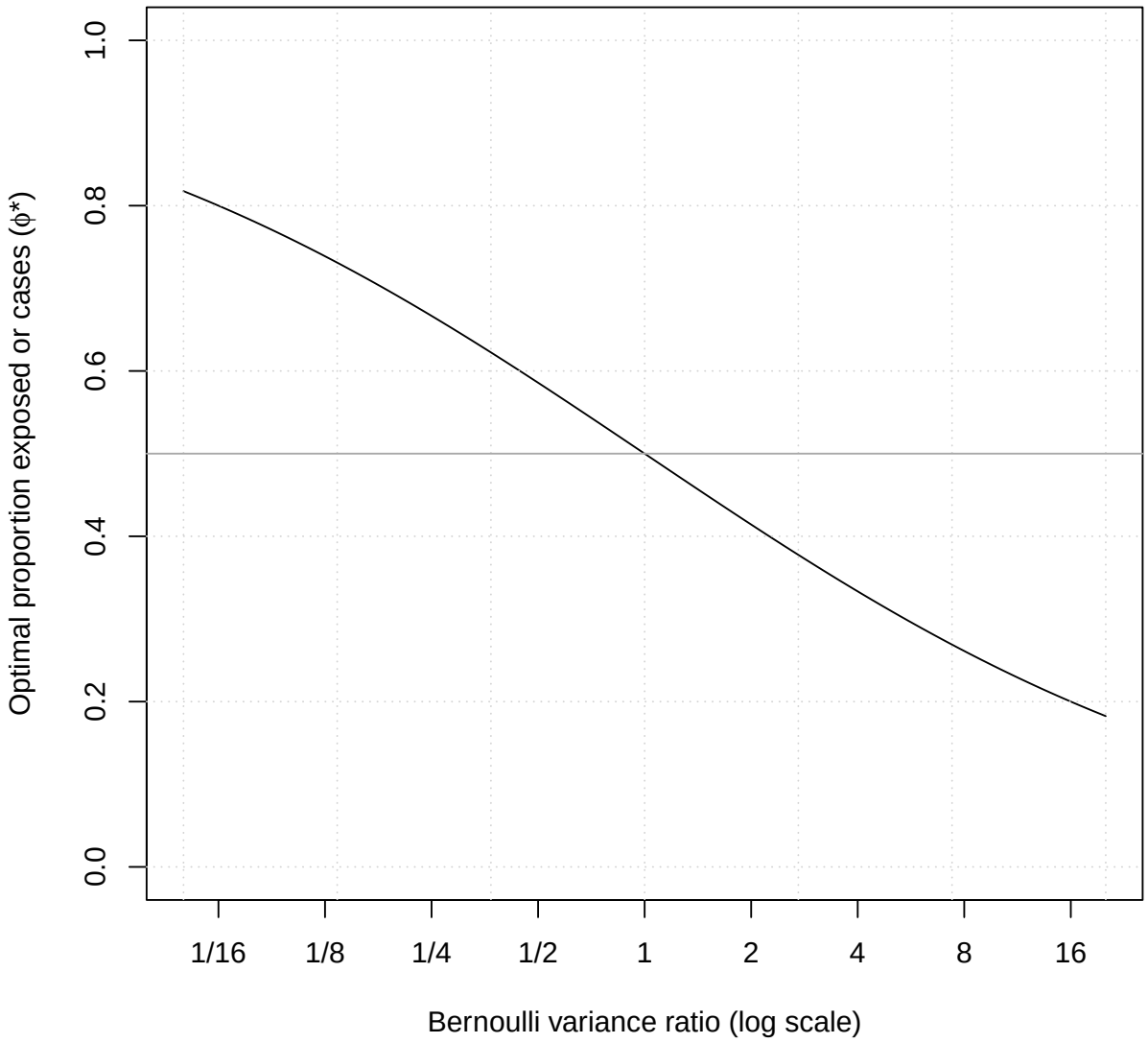


Figure 7.1: The optimal proportion exposed φ^* in a cohort study as a function of the Bernoulli variance ratio $p_1(1-p_1)/(p_0(1-p_0))$. In a case-control study, φ^* represents the optimal proportion of the sample who are cases and the Bernoulli variance ratio is $\pi_1(1-\pi_1)/(\pi_0(1-\pi_0))$. There is a dark gray horizontal line at $\varphi = 0.5$, which represents a balanced study.

The "optimal" proportion exposed φ^* from Equation 7.15 is based on maximizing the value of χ^2_P in large samples. For a given sample size, the power of the test is actually determined by the distribution of possible values of χ^2_P, so the maximum power can occur at a value of φ slightly different from φ^*. Figure 7.2 shows the power achieved by Pearson's chi-squared test at several combinations of p_1, p_0, and n. In all cases, the power at $\varphi = 0.5$ is close to that at φ^*. In several cases, the power at $\varphi = 0.5$ exceeds that at φ^*. If we have strong enough

prior information about p_1 and p_0 to justify an imbalanced study design, the value of testing the null hypothesis that $p_1 = p_0$ is questionable. Without such prior information, a balanced study is a safe bet to be optimal or near-optimal in terms of the power to detect an association between exposure and disease (Walter 1977).

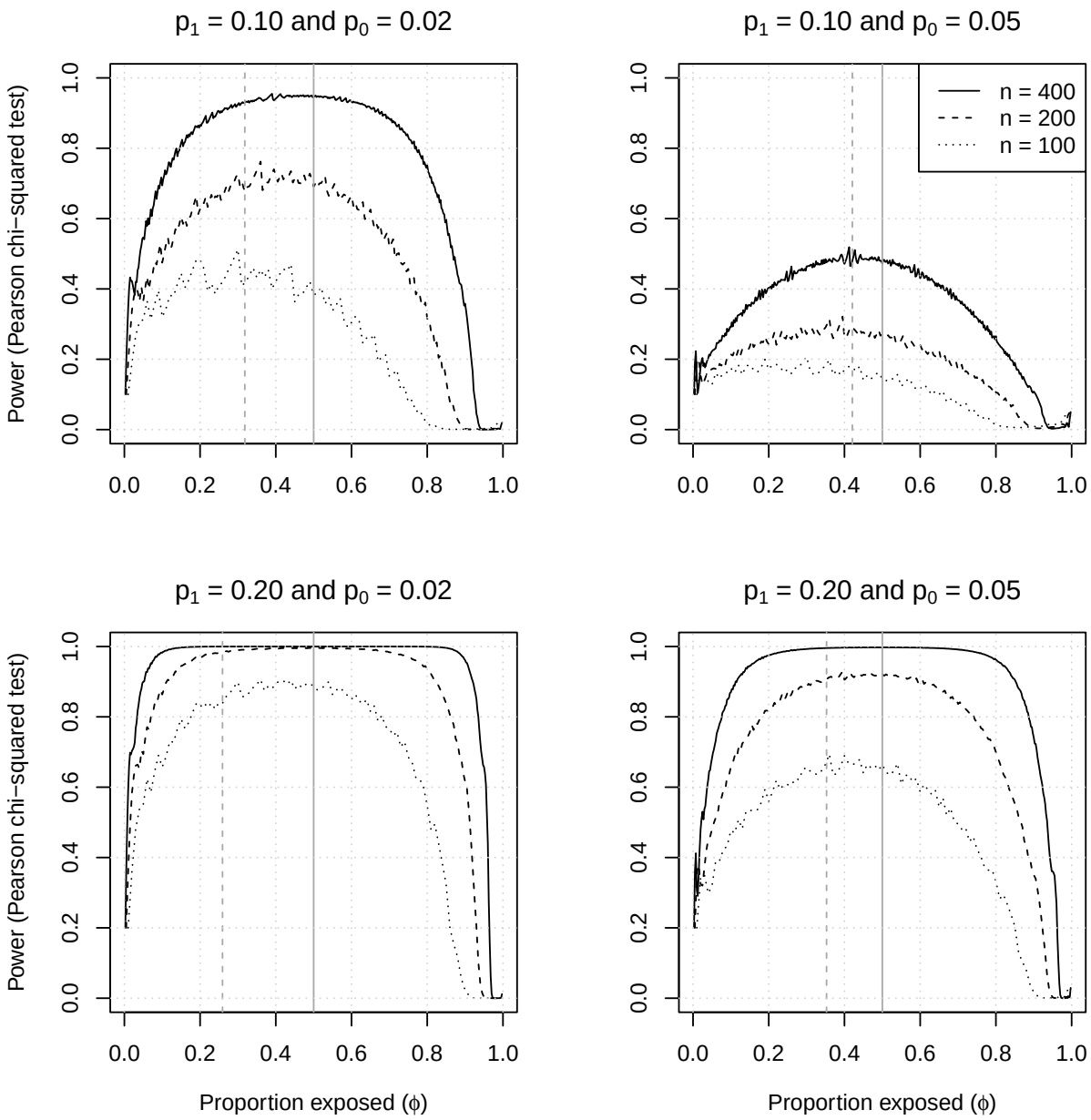


Figure 7.2: The power of the Pearson chi-squared test from a cohort study as a function of the proportion of the sample exposed (φ) at several combinations of p_1 and p_0 for $n = 400$ (solid), $n = 200$ (dashed), and $n = 100$ (dotted). There is a dark gray solid line at $\varphi = 0.5$, representing a balanced study, and a dark gray dashed line at φ^* from Equation 7.15. The same power is achieved by a case control study where π_1 replaces p_1, π_0 replaces p_0, and φ is the proportion of the sample who are cases.

7.4 Case-control studies

The Pearson chi-squared statistic χ^2_P from Equation 7.7 can also be rewritten in terms of the prevalence of exposure among **cases** (participants who have disease or disease onset) and **controls** (participants who do not have disease or disease onset). This leads to the **case-control** study design.

7.4.1 Selection by disease

As above, let π_1 be the exposure prevalence in cases and π_0 be the exposure prevalence in controls. Their maximum likelihood estimates are $\hat{\pi}_1 = a/k_1$ and $\hat{\pi}_0 = c/k_0$, so the maximum likelihood estimate of $\pi_1 - \pi_0$ is

$$\hat{\pi}_1 - \hat{\pi}_0 = \frac{a}{a+c} - \frac{b}{b+d} = \frac{ad-bc}{(a+c)(b+d)} = \frac{ad-bc}{k_1 k_0}. \tag{7.16}$$

Section 7.2.1 showed that null hypothesis that exposure and disease are independent is equivalent to $H_0 : \pi_1 = \pi_0 = \pi$ where π is the marginal prevalence of exposure.

In large samples under the null, $\mathcal{A}$ has a binomial(k_1, π) conditional distribution, $\mathcal{B}$ has a binomial(k_0, π) conditional distribution, and they are conditionally independent given the column sums k_1 and k_0. Thus, the large-sample variance of $\hat{\pi}_1 - \hat{\pi}_0$ under the null is

$$\mathrm{Var}_0(\hat{\pi}_1 - \hat{\pi}_0) = \pi(1-\pi)\left(\frac{1}{k_1} + \frac{1}{k_0}\right) = \pi(1-\pi)\frac{n}{k_1 k_0} \tag{7.17}$$

where we used $k_1 + k_0 = n$. Replacing the unknown π with its maximum likelihood estimate $\hat{\pi} = r_1/n$, we get the estimated null variance

$$\hat{\mathrm{Var}}_0(\hat{\pi}_1 - \hat{\pi}_0) = \hat{\pi}(1-\hat{\pi})\frac{n}{k_1 k_0} = \frac{r_1 r_0}{k_1 k_0 n} \tag{7.18}$$

where we used $1 - \hat{\pi} = r_0/n$. Combining the results in Equation 7.16} and Equation 7.18, we get

$$\frac{(\hat{\pi}_1 - \hat{\pi}_0)^2}{\hat{\mathrm{Var}}_0(\hat{\pi}_1 - \hat{\pi}_0)} = \frac{n(ad-bc)^2}{r_1 r_0 k_1 k_0} = \chi^2_P$$

(see Equation 7.7). The LLN guarantees that $\hat{\pi}_1 \to \pi_1$ as $k_1 \to \infty$ and that $\hat{\pi}_0 \to \pi_0$ as $k_0 \to \infty$. In large samples,

$$\chi^2_P \approx \frac{(\pi_1 - \pi_0)^2}{\pi(1-\pi)\left(\frac{1}{k_1} + \frac{1}{k_0}\right)} \tag{7.19}$$

because the sample average

$$\frac{k_1\pi_1 + k_0\pi_0}{n} \to \pi$$

as $n \to \infty$ by the LLN. The numerator of Equation 7.19 is fixed, but the denominator depends on k_1 and k_0. By sampling according to disease status, we can choose k_1 and k_0 to increase the power of the Pearson chi-squared test for a fixed total number of participants.

7.4.2 Score test for independence in a case-control study*

As with sampling by exposure in a cohort study, sampling by disease in a case-control study does not affect the score test of the null hypothesis that exposure and disease are independent. Using a binomial(k_1, π_1) distribution for the number of exposed cases and a binomial(k_0, π_0) distribution for the number of exposed controls, we get the log likelihood

$$\ell(\pi_1, \pi_0) = \mathcal{A} \ln \pi_1 + \mathcal{C} \ln(1 - \pi_1) + \mathcal{B} \ln \pi_0 + \mathcal{D} \ln(1 - \pi_0)$$

as a random variable whose value will be determined by the data. Calculating the score $U(\pi, \pi)$ and the expected information $\mathcal{I}(\pi, \pi)$ under the null hypothesis $H_0 : \pi_1 = \pi_0 = \pi$ and evaluating them at $\hat{\pi} = r_1/n$, we get

$$U(\hat{\pi}, \hat{\pi}) = \begin{pmatrix} \frac{a}{\hat{\pi}} + \frac{c}{1-\hat{\pi}} \\ \frac{b}{\hat{\pi}} + \frac{d}{1-\hat{\pi}} \end{pmatrix} = \begin{pmatrix} \frac{n(ad-bc)}{r_1 r_0} \\ -\frac{n(ad-bc)}{r_1 r_0} \end{pmatrix}$$

and

$$\mathcal{I}(\hat{\pi}, \hat{\pi}) = \begin{bmatrix} \frac{k_1 n^2}{r_1 r_0} & 0 \\ 0 & \frac{k_0 n^2}{r_1 r_0} \end{bmatrix} \Rightarrow \mathcal{I}^{-1}(\hat{\pi}, \hat{\pi}) = \begin{bmatrix} \frac{r_1 r_0}{k_1 n^2} & 0 \\ 0 & \frac{r_1 r_0}{k_0 n^2} \end{bmatrix}$$

The score statistic is

$$U(\hat{\pi}, \hat{\pi})^{\mathsf{T}} \mathcal{I}(\hat{\pi}, \hat{\pi})^{-1} U(\hat{\pi}, \hat{\pi}) = \frac{n(ad - bc)^2}{r_1 r_0 k_1 k_0} = \chi^2_{\mathrm{P}},$$

which is the Pearson chi-squared statistic from Equation 7.7. The null hypothesis reduces the degrees of freedom from two (π_1 and π_0) to one ($\pi_1 = \pi_0 = \pi$), so the score statistic has a χ^2_1 distribution under H_0. Therefore, Pearson's chi-squared test is the score test of the null hypothesis $H_0 : \pi_0 = \pi_1$ in a case-control study. The column sums k_1 and k_0 are fixed by design, and we condition on the row sums r_1 and r_0 because we use the maximum likelihood estimate $\hat{\pi} = r_1/n$ for the prevalence of exposure under H_0. Because of the invariance of the score test when it uses the expected information, any parameterization of the model for the exposure prevalences π_1 and π_0 leads to the same test of the null hypothesis that exposure and disease are independent.

7.4.3 Optimal sampling by disease

Having established that χ^2_{P} is the score statistic for testing the independence of exposure and disease in a case-control study, we can choose k_1 and k_0 to maximize the power of the test for a given number of participants $n = k_1 + k_0$. Let φ be the proportion of the sample who are cases. Then

$$\begin{aligned} k_1 &= \varphi n \\ k_0 &= (1 - \varphi) n \\ \pi &= \varphi \pi_1 + (1 - \varphi) \pi_0. \end{aligned}$$

Substituting these into equation Equation 7.17 and simplifying gives us the denominator as a function of φ:

$$\frac{n\pi(1-\pi)}{k_1 k_0} = \frac{\varphi}{1-\varphi}\pi_1(1-\pi_1) + \frac{1-\varphi}{\varphi}\pi_0(1-\pi_0) + C(\pi_1, \pi_0)$$

where $C(\pi_1, \pi_0) = \pi_1(1-\pi_0) + \pi_0(1-\pi_1)$ does not depend on φ. This is identical to The derivative with respect to φ is

$$\frac{\mathrm{d}}{\mathrm{d}\varphi}\frac{n\pi(1-\pi)}{r_1 r_0} = \frac{\pi_1(1-\pi_1)}{(1-\varphi)^2} - \frac{\pi_0(1-\pi_0)}{\varphi^2}.$$

This is identical to Equation 7.13 if we replace p_1 with π_1 and p_0 with π_0, so the same argument used in Section 7.3.3 tells us that the Pearson chi-squared statistic χ^2_{P} from a case-control study is maximized when the proportion of the sample comprised of cases is

$$\varphi^* = \frac{1}{1 + \sqrt{\frac{\pi_1(1-\pi_1)}{\pi_0(1-\pi_0)}}}. \tag{7.20}$$

Here, the expression inside the square root is the variance of a Bernoulli(π_1) random variable divided by the variance of a Bernoulli(π_0) random variable. Figure 7.1 shows how ϕ^* depends on this variance ratio. When $\pi_1 \approx \pi_0$, the Bernoulli variance ratio is approximately one $\phi^* \approx 0.5$.

The power functions shown in Figure 7.2 apply to a case-control study if we replace p_1 with π_1 and p_0 with π_0. The justification for recruiting equal numbers of cases and controls in a case-control study is exactly the same as that for recruiting equal numbers of exposed and unexposed in a cohort study: When testing the null hypothesis can be justified, a balanced study is almost always optimal or near-optimal in terms of its power to detect an association between exposure and disease (Walter 1977).

7.5 Choice of study design

We have shown that the power of the Pearson and hypergeometric chi-squared tests can be increased by sampling participants according to exposure (in a cohort study) or disease (in a case-control study) instead of taking a random sample from the population. It remains to see how to choose between a cohort study and a case-control study.

7.5.1 Odds ratio

To choose between the cohort and case-control study designs, it is extremely helpful that the estimated odds ratio is the same for all three study designs. In Table 7.2, the estimated

odds ratio comparing the risks of disease in the exposed (numerator) and the unexposed (denominator) is

$$\frac{\text{odds}(\hat{p}_1)}{\text{odds}(\hat{p}_0)} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

where r_1 canceled out of the numerator and r_0 canceled out of the denominator in the middle expression. The estimated odds ratio comparing the prevalence exposure in cases (numerator) and controls (denominator) is

$$\frac{\text{odds}(\hat{\pi}_1)}{\text{odds}(\hat{\pi}_0)} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

where k_1 canceled out of the numerator and k_0 canceled out of the denominator in the middle expression. The Pearson chi-squared statistic can be rewritten in terms of the odds ratio:

$$\chi^2_{\text{P}} = \frac{n(\frac{ad}{bc} - 1)^2 b^2 c^2}{r_1 r_0 k_1 k_0} = \frac{n(\widehat{\text{OR}} - 1)^2 b^2 c^2}{r_1 r_0 k_1 k_0}.$$

Let

$$\Delta_n = n(\widehat{\text{OR}} - 1),$$

which does not depend on which study design we use.

A random sample from the population has

$$\chi^2_{\text{P}} = \Delta_n \hat{p}_0(1 - \hat{p}_1)\hat{\pi}_0(1 - \hat{\pi}_1).$$

because $b = r_1(1 - \hat{p}_1) = k_0\hat{\pi}_0$ and $c = r_0\hat{p}_0 = k_1(1 - \hat{\pi}_1)$. Close to the null hypothesis, $\hat{p}_1 \approx \hat{p}_0 \approx \hat{p}$ and $\hat{\pi}_0 \approx \hat{\pi}_1 \approx \hat{\pi}$. In large samples close to the null hypothesis,

$$\chi^2_{\text{P}} \approx \Delta_n p(1 - p)\pi(1 - \pi).$$

because $\hat{p} \to p$ and $\hat{\pi} \to \pi$ as $n \to \infty$ by the LLN. A balanced cohort study has $r_0 = r_1 = n/2$ and

$$\chi^2_{\text{P}} = \frac{\Delta_n(1 - \hat{p}_1)^2\hat{p}_0^2}{4\hat{p}(1 - \hat{p})}.$$

because $bc = n^2(1-\hat{p}_1)\hat{p}_0/4$ and $k_1k_0 = n^2\hat{p}(1-\hat{p})$. In a large sample close to (but not under) the null hypothesis,

$$\chi^2_{\text{P}} \approx \frac{\Delta_n p^2(1 - p)^2}{4p(1 - p)} = \frac{\Delta_n}{4} p(1 - p).$$

Following similar logic for a case-control study, we get

$$\chi^2_{\text{P}} \approx \frac{\Delta_n}{4}\pi(1 - \pi)$$

in large samples near the null hypothesis. Because $v(1-v) \leq 1/4$ for $v \in [0, 1]$, the χ^2_{P} statistics from the cohort and case-control studies are both upper bounds for the χ^2_{P} statistic from a random sample of the population.

Close to the null, a cohort study will be more powerful than a case-control study when

$$p(1-p) > \pi(1-\pi)$$

and a case-control study will be more powerful than a cohort study when

$$p(1-p) < \pi(1-\pi).$$

The advantage of a cohort study will be greatest for a rare exposure and a risk of disease close to 1/2, and the advantage of a case-control study will be greatest for rare disease and a prevalence of exposure close to 1/2. Both study designs are always more powerful than a random sample from the population.

7.5.2 Imbalance and efficiency on a fixed budget

Even when testing the null hypothesis is defensible, an imbalanced study design can be justified when one exposure or disease group is substantially more difficult or expensive to recruit than the other. In a cohort study with a rare exposure, exposed individuals might be harder to recruit than unexposed individuals. In a case-control study with a rare disease, cases might be harder to recruit than controls. Even when the greatest power for a given number of participants is achieved with a balanced study, the greatest power for a given study's resources may occur with imbalanced groups.

Deliberately imbalanced designs are used most often in case-control studies, but the principle is the same in cohort studies. Let C be the ratio of the cost of recruiting a case to that of recruiting a control, and B be the budget of the study (expressed as the total number of controls that could be enrolled if no cases were enrolled). As in Table 7.2, k_1 is the number of cases and k_0 is the number of controls. We need to minimize the variance of $\hat{\pi}_1 - \hat{\pi}_0$ from Equation 7.17 given that

$$k_1 C + k_0 = B.$$

For simplicity, we will assume that the prevalences of exposure π_1 (in cases) and π_0 (in controls) are approximately equal, so we can ignore the fact that $\hat{\pi}$ depends on $\varphi = k_1/n$.[5] To maximize the value of χ^2_{P} close to (but not under) the null, we need to minimize

$$\frac{1}{k_1} + \frac{1}{k_0} = \frac{1}{k_1} + \frac{1}{B - k_1 C}$$

over k_1 The derivative with respect to k_1 is

$$\frac{\mathrm{d}}{\mathrm{d}k_1}\left(\frac{1}{k_1} + \frac{1}{B - k_1 C}\right) = -\frac{1}{k_1^2} + \frac{C}{(B - k_1 C)^2},$$

[5]Without this assumption, it is difficult or impossible to derive an explicit expression for the optimal ratio φ^* because the total sample size n depends on φ, which complicates the derivatives. An optimal ratio can be calculated numerically.

which equals zero when

$$k_0^2 = k_1^2 C.$$

This corresponds to $k_0 = k_1\sqrt{C}$ or recruiting $\sqrt{C}$ controls per case (Miettinen 1969; Nam 1973; Gail et al. 1976). The optimal proportion of the sample who are cases is

$$\varphi_C^* = \frac{1}{1+\sqrt{C}}.$$

A nearly identical argument based on Equation 7.9 shows that this φ_C^* is also the optimal proportion exposed in a cohort study where the cost of recruiting an exposed individual is C times that of recruiting an unexposed individual. This $\sqrt{C}$ rule is a good approximation to more accurate and complicated optimal sampling rules (Meydrech and Kupper 1978; Pike and Casagrande 1979; Morgenstern and Winn 1983).

With a total sampling budget of B, the optimal numbers of cases is

$$k_1^* = \frac{B}{\sqrt{C}+C}$$

and the optimal number of controls is

$$k_0^* = k_1^*\sqrt{C} = \frac{B}{1+\sqrt{C}}.$$

The minimum variance of the risk difference that we can achieve near the null is proportional to

$$V^* = \frac{1}{k_1^*} + \frac{1}{k_0^*} = \frac{\left(1+\sqrt{C}\right)^2}{B}.$$

If we use a balanced study design, $k_1 = k_0 = B/(1+C)$ and the variance of the risk difference is proportional to

$$V^{\text{bal}} = \frac{2}{k_1} = \frac{2(1+C)}{B}.$$

For any given budget B, the asymptotic relative efficiency of the optimal study compared to a balanced study is

$$\frac{V^{\text{bal}}}{V^*} = \frac{2(1+C)}{\left(1+\sqrt{C}\right)^2}.$$

It is plotted as a function of C in Figure 7.3. The difference is small for moderate values of C, with relative efficiencies of approximately 1.029 for $C = 2$ and 1.146 for $C = 5$. In extreme scenarios (i.e., as $C \to 0$ or $C \to \infty$), the optimal study is twice as efficient as a balanced study with the same budget (Nam 1973).

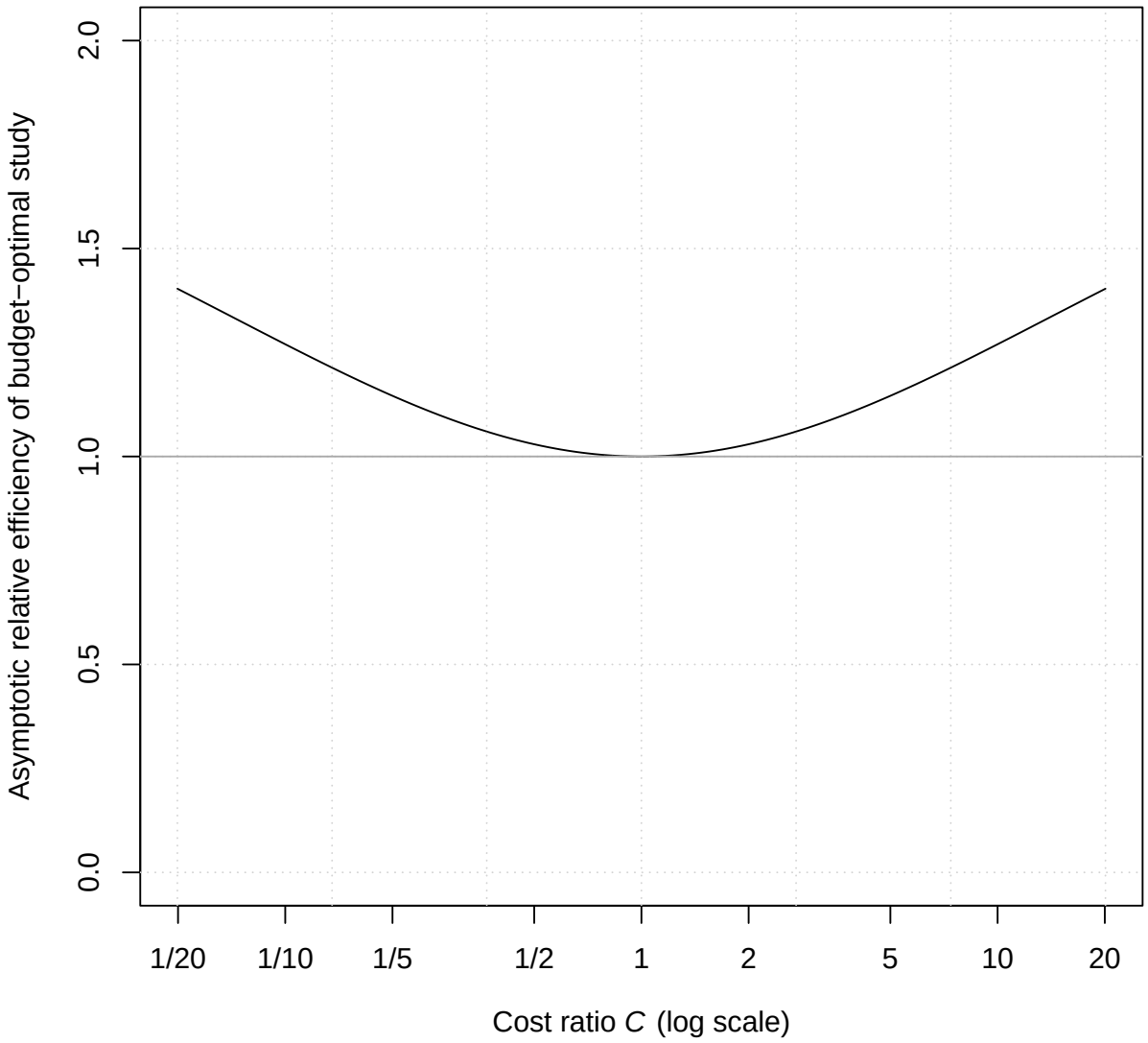


Figure 7.3: The asymptotic relative efficiency of an optimal case-control study compared to a balanced study with the same budget when recruiting a case costs C times as much as recruiting a control. There is a dark gray line at a relative efficiency of one. The same relative efficiency applies to cohort studies when recruiting an exposed individual costs C times as much as recruiting an unexposed individual.

The relative efficiency can thought of as the ratio of the sampling budgets of a balanced study and an optimal study that achieve the same power (Nam 1973; Gail et al. 1976). Thus, a balanced study requires at most twice the budget of an optimal study to achieve the same power. Brittain, Schlesselman, and Stadel (1981) found that sampling costs were approximately 33-66% of total costs in five case-control studies funded by the National Institute of Child Health

and Human Development in the 1970s. Compared to a balanced study design, they found that optimal sampling of cases and controls would reduce total study costs by at most 8.5% for $C \leq 5$ and at most 4.5% for $C \leq 3$. As usual, balanced study designs are close to optimal.

Listing 7.2 `chisq-power.R`

```
## Actual power of a Pearson chi-squared test

# calculate Pearson chi-squared test power
# This can take a few minutes to run with large n.
powers <- function(p1, p0, n, level = 0.95) {
  chisq_alpha <- qchisq(level, df = 1)
  htest <- function(r1) {
    r0 <- n - r1
    joint_dbinom <- outer(0:r1, 0:r0,
                          function(a, c) dbinom(a, r1, p1) * dbinom(c, r0, p0))
    joint_include <- outer(0:r1, 0:r0,
                           function(a, c) max(a, c) > 0 & a + c < n)
    acpower <- Vectorize(function(a, c) {
      if (max(a, c) > 0 & a + c < n) {
        b <- r1 - a
        d <- r0 - c
        k1 <- a + c
        k0 <- b + d
        chisqP <- n * (a * d - b * c)^2 / (r1 * r0 * k1 * k0)
        return(chisqP > chisq_alpha)
      } else {
        return(0)
      }
    })
    joint_power <- outer(0:r1, 0:r0, acpower)
    return(sum(joint_dbinom * joint_power) / sum(joint_dbinom * joint_include))
  }
  r1s <- 1:(n - 1)
  powers <- sapply(r1s, htest)
  return(data.frame(r1 = r1s, power = powers, n = n))
}

# optimal value proportion exposed (or proportion cases)
optimphi <- function(p1, p0) 1 / (1 + sqrt(p1 * (1 - p1) / (p0 * (1 - p0))))

# save values of graphical parameter "mar" before changing them
orig_mar <- par("mar")
orig_mfrow <- par("mfrow")
par(mar = c(4, 4, 3, 2))
par(mfrow = c(2, 2))

# Pearson chi-squared test power for p1 = 0.1 and p0 = 0.02
power_10_02_400 <- powers(0.10, 0.02, 400)
power_10_02_200 <- powers(0.10, 0.02, 200)
power_10_02_100 <- powers(0.10, 0.02, 100)
plot(power_10_02_400$r1 / 400, power_10_02_400$power,
     type = "l", ylim = c(0, 1),
     main = expression(paste(p[1], " = 0.10", " and ", p[0], " = 0.02")),
     xlab = "",
     ylab = "Power (Pearson chi-squared test)")
lines(power_10_02_200$r1 / 200, power_10_02_200$power, lty = "dashed")
```

Listing 7.3 `optimal-budget.R`

```
## Relative efficiency of imbalanced study design on fixed budget

# variance ratio comparing balanced study to budget-optimal study
logC <- seq(-3, 3, by = 0.01)
releff <- function(C) 2 * (1 + C) / (1 + sqrt(C))^2
plot(logC, releff(exp(logC)), type = "l", ylim = c(0, 2), xaxt = "n",
     xlab = expression(paste("Cost ratio ", italic("C"), " (log scale)")),
     ylab = "Asymptotic relative efficiency of budget-optimal study")
axis(1, at = log(c(1 / c(20, 10, 5, 2), 1, c(2, 5, 10, 20))),
     labels = c("1/20", "1/10", "1/5", "1/2", 1, 2, 5, 10, 20))
grid()
abline(h = 1, col = "darkgray")
```

8 Internal and External Validity

> Validity will be evaluated in terms of two major criteria. First, and as a basic minimum, is what can be called *internal validity:* did in fact the experimental stimulus make some significant difference in this specific instance? The second criterion is that of *external validity, representativeness,* or *generalizability:* to what populations, settings, and variables can this effect be generalized? Both criteria are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness. (Campbell 1957)

In statistics, we make inferences about a population based on a sample. A study is said to have **internal validity** if it makes accurate measurements or inferences within the sample itself, and it is said to have **external validity** if these inferences accurately describe the population up to random sampling error (Campbell 1957). Both internal and external validity are best thought of as continuous, not binary. High internal validity is a prerequisite for high external validity, but there is often a tradeoff between them in practice. For simplicity, we focus on internal and external validity for descriptive epidemiology (i.e., for association and not necessarily causation).

So far, our discussion of 2x2 tables has assumed that the classification of exposure and disease is completely accurate and that the participants are a random sample from the population. Table 8.1 shows our 2x2 table based on true exposure and disease classifications. In reality, **misclassification** and **selection bias** threaten the validity of almost all epidemiologic studies. It is critical to understand where they come from and what they do.

Table 8.1: 2x2 table of true disease and exposure

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

8.1 Misclassification

Misclassification of exposure and disease threatens both the internal and external validity of an epidemiologic study. In a cohort study, we compare the exposed and unexposed groups and we have to classify disease outcomes in each group. In a case-control study, we compare cases and controls and we have to classify exposure in each group.

Nondifferential misclassification occurs when the same classification errors affect both populations being compared. Under nondifferental misclassification, a test of the null hypothesis is still has the correct significance level but the power of the test is reduced—much like a reduction in the effective sample size (Bross 1954; Rubin, Rosenbaum, and Cobb 1956). Nondifferential classification almost always causes bias toward the null, making the expected value of a given measure of association closer to the null than its true value. However, a measure of association under nondifferential misclassification can be farther away from the null than its true value due to random variation (Gullen, Bearman, and Johnson 1968; Sorahan and Gilthorpe 1994; Wacholder et al. 1995; Yland et al. 2022).

Differential misclassification occurs when classification errors differ between the two populations being compared. Differental misclassification can distort both the significance level and power of a hypothesis test, and it can cause bias toward the null, away from the null, or across the null. The unpredictability of the size and direction of the bias makes differential misclassification fundamentally more dangerous than nondifferential misclassification.

8.1.1 Nondifferential misclassification of disease

In our discussion of diagnostic tests, we let D^+ indicate $D = 1$, D^- indicate $D = 0$, T^+ indicate testing positive for disease, and T^- indicate testing negative. Let D^{obs} be the measured disease outcome of individuals in a cohort study where disease is detected using a diagnostic test with sensitivity

$$\text{sens}_D = \Pr(\mathrm{T}^+ \,|\, \mathrm{D}^+) = \Pr(D^{\text{obs}} = 1 \,\big|\, D = 1)$$

and specificity

$$\text{spec}_D = \Pr(\mathrm{T}^- \,|\, \mathrm{D}^-) = \Pr(D^{\text{obs}} = 0 \,\big|\, D = 0).$$

We assume that

$$\text{sens}_D = \Pr(\mathrm{T}^+ \,|\, \mathrm{D}^+) > \Pr(\mathrm{T}^+ \,\big|\, \mathrm{D}^-) = 1 - \text{spec}_D,$$

so individuals with disease are more likely to test positive than individuals without disease.[1] This is equivalent to assuming that $\text{sens}_D + \text{spec}_D > 1$. We also assume that the misclassification of each participant is independent of the misclassification of all other participants.

[1] These tests are in the top left half of a receiver operating characteristic (ROC) plot from Section 2.5.1, where $1 - \text{spec}_D$ is the horizontal axis and sens_D is the vertical axis. A test with $\text{sens}_D = 1 - \text{spec}_D$ is a useless test (on the diagonal of an ROC plot). A test with $\text{sens}_D < 1 - \text{spec}_D$ (in the bottom right half of an ROC plot) needs to have the definitions of T^+ and T^- reversed.

Table 8.2: 2x2 table with misclassified disease status

	$D^{\text{obs}} = 1$	$D^{\text{obs}} = 0$	Total
$X = 1$	a^{obs}	b^{obs}	r_1
$X = 0$	c^{obs}	d^{obs}	r_0
Total	k_1^{obs}	k_0^{obs}	n

Table 8.2 shows a 2x2 table with misclassification of disease. The row sums r_1 and r_0 are the same as in Table 8.1 because there is no misclassification of exposure.

Misclassification of disease is **nondifferential** when the sensitivity and specificity of the test are the same in all exposure groups. In other words, we have nondifferential misclassification of disease if and only if

$$\Pr(\mathrm{T}^+ \,|\, \mathrm{D}^+, X = x) = \Pr(\mathrm{T}^+ \,|\, \mathrm{D}^+) = \mathrm{sens}_D$$

and

$$\Pr(\mathrm{T}^- \,|\, \mathrm{D}^-, X = x) = \Pr(\mathrm{T}^- \,|\, \mathrm{D}^-) = \mathrm{spec}_D$$

for all possible values x of exposure X. It is critical that nondifferential misclassification is defined in terms of the sensitivity and specificity of the test, not its positive predictive value (PPV) or negative predictive value (NPV). When there is an association between exposure and disease, nondifferential misclassification of disease may produce different PPVs and NPVs in the exposed and unexposed because these predictive values depend on the prevalence of disease in addition to the sensitivity and specificity of the test (D. J. Newell 1962; Buell and Dunn Jr 1964).

Under nondifferential misclassification, the probability that an exposed person tests positive for disease is

$$\begin{aligned} p_1^{\text{obs}} &= p_1 \mathrm{sens}_D + (1 - p_1)(1 - \mathrm{spec}_D) \\ &= (1 - \mathrm{spec}_D) + (\mathrm{sens}_D + \mathrm{spec}_D - 1)p_1 \end{aligned}$$

where p_1 is the true risk of disease in the exposed. Similarly, the probability that an unexposed person tests positive for disease is

$$p_0^{\text{obs}} = (1 - \mathrm{spec}_D) + (\mathrm{sens}_D + \mathrm{spec}_D - 1)p_0$$

where p_0 is the true risk of disease in the unexposed. The misclassified risk difference is

$$\mathrm{RD}^{\text{obs}} = p_1^{\text{obs}} - p_0^{\text{obs}} = (\mathrm{sens}_D + \mathrm{spec}_D - 1)(p_1 - p_0). \tag{8.1}$$

Given the margins of Table 8.2, the number $\mathcal{A}^{\text{obs}}$ of exposed individuals who test positive for disease has a hypergeometric distribution with mean $r_1 p_1^{\text{obs}}$ and the number $\mathcal{C}^{\text{obs}}$ of unexposed

people who test positive for disease has a hypergeometric distribution with mean $r_0 p_0^{\text{obs}}$. The estimated risk difference based on the misclassified data is

$$\widehat{\text{RD}}^{\text{obs}} = \hat{p}_1^{\text{obs}} - \hat{p}_0^{\text{obs}},$$

where $\hat{p}_1^{\text{obs}} = a^{\text{obs}}/r_1$ and $\hat{p}_0^{\text{obs}} = c^{\text{obs}}/r_0$. It is an unbiased estimate of RD^{obs}.

When $\text{sens}_D < 1$ or $\text{spec}_D < 1$, the misclassified risk difference RD^{obs} in Equation 8.1 is closer to zero than the true risk difference (Bross 1954; Rubin, Rosenbaum, and Cobb 1956; D. Newell 1963). The risk ratio and odds ratio are also biased toward the null under nondifferential misclassification (Goldberg 1975; Copeland et al. 1977). This bias operates on average, not for every single estimate based on misclassified data. Even under nondifferential misclassification of disease, random variation can produce an estimate of the risk difference, risk ratio, or odds ratio that is farther from the null than the true value (Gullen, Bearman, and Johnson 1968; Sorahan and Gilthorpe 1994; Wacholder et al. 1995; Yland et al. 2022).

When $\text{sens}_D + \text{spec}_D > 1$ (as is true of any useful diagnostic or screening test), Equation 8.1 implies that the null hypothesis that $p_1^{\text{obs}} = p_0^{\text{obs}}$ is equivalent to the null hypothesis that $p_1 = p_0$. Both null hypotheses are equivalent to the null hypothesis that exposure and disease are independent (see Section 7.2.1). Therefore, a test of the null hypothesis that X and D^{obs} are independent is also a valid test of the independence of X and D (Bross 1954; Rubin, Rosenbaum, and Cobb 1956). The Pearson chi-squared statistic for Table 8.2 is

$$\chi^2_{\text{Pobs}} = \frac{n(a^{\text{obs}}d^{\text{obs}} - b^{\text{obs}}c^{\text{obs}})^2}{r_1 r_0 k_1^{\text{obs}} k_0^{\text{obs}}}, \tag{8.2}$$

and it has a χ^2_1 distribution under the null hypothesis that $p_1 = p_0$. If we set the critical value at the $1-\alpha$ quantile of the χ^2_1 distribution, the test will reject the null with probability α when $p_1 = p_0$ even under nondifferential misclassification of disease. A similar result holds for the hypergeometric chi-squared test, Fisher's exact test, and other tests of independence for 2x2 tables from Section 7.2.

Although nondifferential misclassification does not affect the significance level of a hypothesis test of the null hypothesis that $p_1 = p_0$, it reduces the power of the test away from the null (Bross 1954; Rubin, Rosenbaum, and Cobb 1956; Rogot 1961). The Pearson chi-squared statistic based on the misclassified data in Table 8.2 can be rewritten

$$\chi^2_{\text{Pobs}} = \frac{(\hat{p}_1^{\text{obs}} - \hat{p}_0^{\text{obs}})^2}{\hat{p}^{\text{obs}}(1-\hat{p}^{\text{obs}})\left(\frac{1}{r_1} + \frac{1}{r_0}\right)}.$$

where $\hat{p}^{\text{obs}} = k_1^{\text{obs}}/n$ is the misclassified estimate of the marginal risk of disease among the study participants. Let $\varphi \in (0,1)$ be the proportion of the sample that is exposed, which we assume to be (approximately) constant as $n \to \infty$. Let

$$K_D = \text{sens}_D + \text{spec}_D - 1,$$

so $K_D \in (0,1)$ whenever we have a useful but imperfect test for disease. When both $r_1 = \varphi n$ and $r_0 = (1-\varphi)n$ are large, the numerator on the right-hand side of Equation 8.2 is approximately

$$\mathbb{E}(\hat{p}_1^{\text{obs}} - \hat{p}_0^{\text{obs}})^2 = K_D^2(p_1 - p_0)^2$$

by the law of large numbers (LLN) and the *continuous mapping theorem*.[2] Similarly, $\hat{p}^{\text{obs}}$ is approximately

$$\begin{aligned} p^{\text{obs}} &= p\,\text{sens}_D + (1-p)(1-\text{spec}_D) \\ &= (1-\text{spec}_D) + K_D p, \end{aligned}$$

where

$$p = \varphi p_1 + (1-\varphi)p_0$$

is the marginal risk of disease among the study participants. Thus,

$$\hat{p}^{\text{obs}}(1-\hat{p}^{\text{obs}}) \approx (1-\text{spec}_D + K_D p)(1-\text{sens}_D + K_D(1-p)).$$

in large samples. It follows that

$$\chi^2_{\text{Pobs}} \approx \frac{K_D^2 p(1-p)}{(1-\text{spec}_D + K_D p)(1-\text{sens}_D + K_D(1-p))}\chi^2_{\text{P}}$$

where χ^2_{P} is the Pearson chi-squared statistic based on the correctly classified data in Table 8.1. Therefore, nondifferential misclassification of disease has approximately the same effect as multiplying the sample size n by the effective sample size ratio

$$\text{ESSR}_D = \frac{K_D^2 p(1-p)}{(1-\text{spec}_D + K_D p)(1-\text{sens}_D + K_D(1-p))} \leq 1$$

with equality if and only if $\text{sens}_D = \text{spec}_D = 1$. Thinking about nondifferential misclassification as a reduction in the effective sample size is a good way to remember both that it preserves the correct significance level under the null and that it reduces power away from the null (Bross 1954; Rubin, Rosenbaum, and Cobb 1956).

When the prevalence or risk of disease is low, the effective sample size depends much more on the specificity of the test than on the sensitivity of the test (Rubin, Rosenbaum, and Cobb 1956). Figure 8.1 shows the effective sample size ratio ESSR_D for three values of the marginal risk of disease in the sample (p) under two different scenarios: one where $\text{spec}_D = 1$ while sensitivity varies from zero to one and one where $\text{sens}_D = 1$ while specificity varies from zero to one. For all three values of p, ESSR_D is substantially lower in the scenario with varying specificity. This difference is greatest for $p = 0.02$ and smallest for $p = 0.40$. If $p = 0.5$, the two scenarios produce identical curves. If $p > 0.5$, then ESSR_D depends more on the sensitivity than the specificity of the test for disease.

[2]The *continuous mapping theorem* says that if a statistic $\hat{\theta}_n \to \theta$ in probability and f is a continuous function in a neighborhood of θ, then $f(\hat{\theta}_n) \to f(\theta)$ in probability. Convergence in probability can be replaced with convergence in distribution or convergence almost surely. See Chung (2000). Here, the statistic is $\hat{p}_1 - \hat{p}_0$, which converges to $p_1 - p_0$, and the function is $f(v) = v^2$.

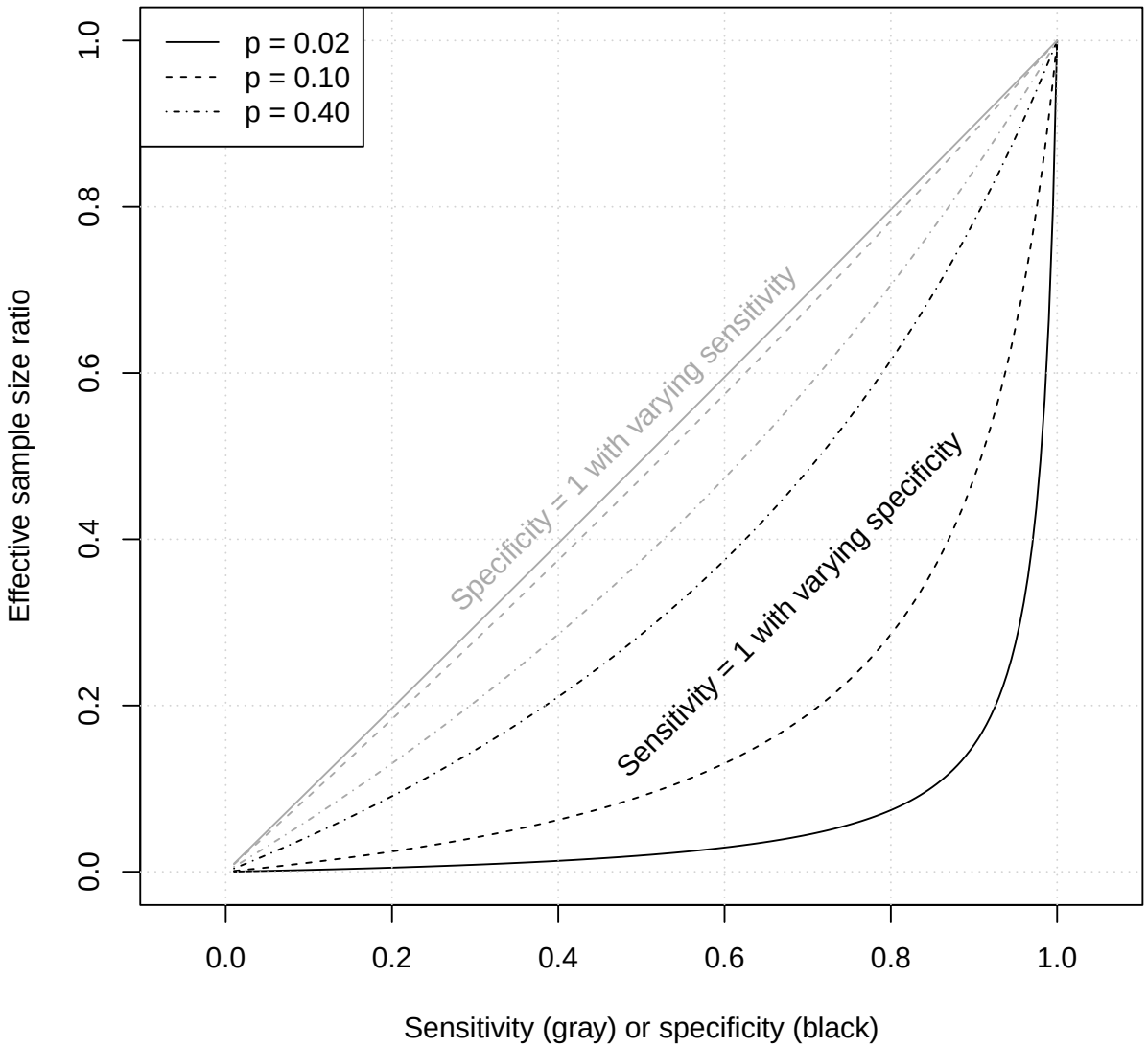


Figure 8.1: The effective sample size ratio ESSR_D as a function of sens_D when $\text{spec}_D = 1$ (gray lines) and as a function of spec_D when $\text{sens}_D = 1$ (black lines).

8.1.2 Nondifferential misclassification of exposure

Nondifferential misclassification of exposure has effects similar to those of nondifferential misclassification of disease. Although we usually discuss sensitivity and specificity in the context of a test for disease, the same ideas can be applied to a test or measurement used to determine exposure status. For simplicity, we will focus on a binary exposure X with X^+ indicating $X = 1$ and X^- indicating $X = 0$.

Table 8.3: 2x2 table with misclassified exposure

	$D=1$	$D=0$	Total
$X^{\text{obs}}=1$	a^{obs}	b^{obs}	r_1^{obs}
$X^{\text{obs}}=0$	c^{obs}	d^{obs}	r_0^{obs}
Total	k_1	k_0	n

Let X^{obs} be the measured disease outcome of individuals in a case-control study when we classify exposure using a test T_X that has sensitivity

$$\text{sens}_X = \Pr(\text{T}_X^+ \,|\, \text{X}^+) = \Pr(\text{X}^{\text{obs}} = 1 \,|\, X = 1)$$

and specificity

$$\text{spec}_X = \Pr(\text{T}_X^- \,|\, \text{X}^-) = \Pr(\text{X}^{\text{obs}} = 0 \,|\, X = 0).$$

We assume that $\text{sens}_X > 1 - \text{spec}_X$, so exposed individuals are more likely to test positive for exposure than unexposed individuals. This is equivalent to assuming $\text{sens}_X + \text{spec}_X > 1$.[3] We also assume that the misclassification of each participant is independent of the misclassification of all other participants. Table 8.3 shows a 2x2 table with exposure misclassification. The column sums k_1 and k_0 are the same in both tables because there is no misclassification of disease status.

The misclassification of exposure is nondifferential when the sensitivity and specificity of the exposure test are the same in cases and controls. In other words, we need

$$\Pr(\text{T}_X^+ \,|\, \text{X}^+, D = d) = \Pr(\text{T}_X^+ \,|\, \text{X}^+)$$

and

$$\Pr(\text{T}_X^- \,|\, \text{X}^-, D = d) = \Pr(\text{T}_X^- \,|\, \text{X}^-)$$

for all possible values d of disease status D. As with disease, it is critical that nondifferential misclassification of exposure is defined through the sensitivity and specificity of the test for exposure. When there is an association between exposure and disease, nondifferential misclassification can produce a PPV and NPV that differ between cases and controls because these predictive values depend on the prevalence of exposure in addition to the sensitivity and specificity of the test (D. J. Newell 1962; Buell and Dunn Jr 1964).

Under nondifferential misclassification, the probability that a case tests positive for exposure is

$$\pi_1^{\text{obs}} = (1 - \text{spec}_X) + (\text{sens}_X + \text{spec}_X - 1)\pi_1$$

[3] As with a test for disease, a test for exposure with $\text{sens}_X + \text{spec}_X = 1$ would be a useless test (on the diagonal of an ROC curve) and a test with $\text{sens}_X + \text{spec}_X < 1$ would need to have the definitions of T_X^+ and T_X^- reversed.

where π_1 is the true prevalence of exposure among cases. Similarly, the probability that a control tests positive for exposure is

$$\pi_0^{\text{obs}} = (1 - \text{spec}_X) + (\text{sens}_X + \text{spec}_X - 1)p_0$$

where π_0 is the true prevalence of exposure in controls. The misclassified difference in exposure prevalences is

$$\pi_1^{\text{obs}} - \pi_0^{\text{obs}} = (\text{sens}_X + \text{spec}_X - 1)(\pi_1 - \pi_0).$$

When $\text{sens}_X + \text{spec}_X > 1$ (as is true of any useful test for exposure), the null hypothesis that $\pi_1^{\text{obs}} = \pi_0^{\text{obs}}$ is equivalent to the null hypothesis that $\pi_1 = \pi_0$, which is equivalent to the null hypothesis that exposure and disease are independent (see Section 7.2.1). Therefore, any test of the independence of X^{obs} and D has the correct significance level under the null hypothesis that X and D are independent (Bross 1954).[4]

Like nondifferential misclassification of disease, nondifferential misclassification of exposure reduces the power of the hypothesis test that exposure and disease are independent (Bross 1954; Rubin, Rosenbaum, and Cobb 1956; Rogot 1961). The maximum likelihood estimates $\hat{\pi}_1^{\text{obs}} = a^{\text{obs}}/k_1$ and $\hat{\pi}_0^{\text{obs}} = b^{\text{obs}}/k_0$ are unbiased. The Pearson chi-squared statistic based on the misclassified data in Table~**??** can be rewritten

$$\chi^2_{\text{Pobs}} = \frac{(\hat{\pi}_1^{\text{obs}} - \hat{\pi}_0^{\text{obs}})^2}{\hat{\pi}^{\text{obs}}(1 - \hat{\pi}^{\text{obs}})\left(\frac{1}{k_1} + \frac{1}{k_0}\right)}. \tag{8.3}$$

where $\hat{\pi}^{\text{obs}} = k_1^{\text{obs}}/n$ is the misclassified estimate of the marginal prevalence of exposure among the study participants. Let φ_X be the proportion of the sample that consists of cases, and let $K_X = \text{sens}_X + \text{spec}_X - 1$, so $K_X \in (0, 1)$ whenever we have a useful but imperfect test for exposure. When both $k_1 = \varphi_X n$ and $k_0 = (1 - \varphi_X)n$ are large, the numerator on the right-hand side of Equation 8.3 is approximately

$$\mathbb{E}(\hat{\pi}_1^{\text{obs}} - \hat{\pi}_0^{\text{obs}})^2 = K_X^2(\pi_1 - \pi_0)^2,$$

and

$$\hat{\pi}^{\text{obs}} \approx \mathbb{E}(\hat{\pi}^{\text{obs}}) = (1 - \text{spec}_X) + K_X\pi$$

where

$$\pi = \varphi_X\pi_1 + (1 - \varphi_X)\pi_0$$

is the marginal prevalence of exposure among the study participants, In large samples,

$$\hat{\pi}^{\text{obs}}(1 - \hat{\pi}^{\text{obs}}) \approx (1 - \text{spec}_X + K_X\pi)(1 - \text{sens}_X + K_X(1 - \pi)).$$

[4] Case control studies typically use the odds ratio $\text{odds}(\pi_1)/\text{odds}(\pi_0)$, not the difference $\pi_1 - \pi_0$, to compare the exposure prevalences in cases and controls. The difference between the prevalences is being used here only to establish that a hypothesis test of the independence of X and D based on misclassified data has the correct significance level.

For hypothesis testing, nondifferential misclassification of exposure has approximately the same effect as multiplying the sample size n by the effective sample size ratio

$$\text{ESSR}_X = \frac{K_X^2 \pi(1-\pi)}{(1-\text{spec}_X + K_X\pi)(1-\text{sens}_X + K_X(1-\pi))} \leq 1$$

with equality if and only if $\text{sens}_X = \text{spec}_X = 1$. Just like nondifferential misclassification of disease, nondifferential misclassification of exposure acts like reduction in the effective sample size. It preserves the significance level under the null, but it reduces the power of the test away from the null. A similar reduction in power occurs when more complex exposures (such as dietary intakes) are measured with error, requiring larger sample sizes to achieve a given power (L. S. Freedman, Schatzkin, and Wax 1990).

The curves in Figure 8.1 are the same if we replace the marginal risk of disease p with the marginal prevalence of exposure π. The effective sample size ratio ESSR_X depends on the specificity more than the sensitivity when $\pi < 0.5$, and it depends on the sensitivity more than the specificity when $\pi > 0.5$. While diseases typically (and fortunately) have low risks, exposures can have both low and high prevalences.

When there are more than two levels of exposure, nondifferential misclassification does not always bias a measure of association toward the null for all exposure categories (Walker, Velema, and Robins 1988; Dosemeci, Wacholder, and Lubin 1990; Verkerk and Buitendijk 1992; Correa-Villaseñor et al. 1995). Misclassification causes the risks of disease in different exposure categories to get closer to each other on average. Without loss of generality, suppose that higher exposure is associated with a higher risk of disease. Misclassification of high-exposure individuals to lower-exposure categories can increase the apparent risk of disease in these categories, and misclassification of low-exposure individuals into higher-exposure categories can decrease the apparent risk of disease in these categories. The risk in the highest-exposure category can only go down on average due to misclassification, and the risk in the lowest-exposure category can only go up on average. In both cases, this results in bias toward the null—and these are the only possible cases for a binary exposure. The risk in an intermediate-exposure category can go up or down on average, so some measures of association can be biased away from the null. Despite this exception, bias toward the null remains the most likely outcome of nondifferential misclassification of exposure (Dosemeci, Wacholder, and Lubin 1990; Correa-Villaseñor et al. 1995). However, random variation can produce point estimates closer to, farther from, or across the null compared to the true value of a measure of association.

8.1.3 Simultaneous nondifferential misclassification

Although we discussed nondifferential misclassification of disease in the context of a cohort study and nondifferential misclassification of exposure in the context of a case-control study, simultaneously misclassification of X and D can occur in any epidemiologic study. Table 8.4

Table 8.4: 2x2 table for a study with misclassified exposure and disease

	$D^{\text{obs}} = 1$	$D^{\text{obs}} = 0$	Total
$X^{\text{obs}} = 1$	a^{obs}	b^{obs}	r_1^{obs}
$X^{\text{obs}} = 0$	c^{obs}	d^{obs}	r_0^{obs}
Total	k_1^{obs}	k_0^{obs}	n

shows a 2x2 table with X and D both misclassified. The row totals are affected by misclassification of X, and the column totals are affected by misclassification of D. Only the total sample size n is unaffected.

The effects of nondifferential misclassification of both X and D can be derived by imagining that we misclassify one first and then the other. Here, we will consider misclassifying X and then D. When we misclassify X, we have the equivalent null hypotheses

$$X \perp\!\!\!\perp D \iff X^{\text{obs}} \perp\!\!\!\perp D.$$

where the symbol $\perp\!\!\!\perp$ indicates independence (Dawid 1979). When we misclassify D in addition to X, we have the equivalent null hypotheses

$$X^{\text{obs}} \perp\!\!\!\perp D \iff X^{\text{obs}} \perp\!\!\!\perp D^{\text{obs}}.$$

Therefore,

$$X \perp\!\!\!\perp D \iff X^{\text{obs}} \perp\!\!\!\perp D^{\text{obs}}$$

so a test of the null hypothesis that X and D are independent based on the misclassified data in Table 8.4 has the correct significance level (as long as $\text{sens}_X + \text{spec}_X > 1$ and $\text{sens}_D + \text{spec}_D > 1$). However, the power of the test is reduced when we misclassify X and reduced again when we misclassify D. Simultaneous misclassification of X and D has approximately the same effect as multiplying the sample size by

$$\text{ESSR}_X \text{ESSR}_D \leq 1$$

with equality if and only if $\text{sens}_X = \text{spec}_X = 1$ and $\text{sens}_D = \text{spec}_D = 1$.

8.2 Selection bias

Participants in an epidemiologic study can be selected according to exposure (in a cohort study) or according to disease (in a case-control study), but they cannot be selected according to both. In a cohort study, selection must be conditionally independent of disease given exposure so that risks of disease are estimated accurately in all exposure groups. In a case-control study, selection must be conditionally independent of exposure given disease, so the prevalence

of exposure is measured accurately among both cases and controls.. Selection according to exposure and disease simultaneously leads to **selection bias**, which is a threat to the external validity of a study.[5] A study with uncontrolled selection bias is neither generalizable nor transportable.

8.2.1 Selection bias in cohort studies

In a cohort study, selection bias leads to biased estimates of the risks of disease in exposure groups. Let S indicate selection into the study, so $S_i = 1$ if individual i is selected into the study and $S_i = 0$ otherwise. When only X and D are measured, there is no selection bias in a cohort study if and only if the conditional probability of disease given exposure in sampled individuals equals that in the population:

$$\Pr(D = 1 \,|\, X = x, S = 1) = \Pr(D = 1 \,|\, X = x) \tag{8.4}$$

for both $x = 1$ and $x = 0$. By the definition of conditional probability,

$$\Pr(D = 1 \,|\, X = x, S = 1) = \frac{\Pr(D = 1, S = 1 \,|\, X = x)}{\Pr(S = 1 \,|\, X = x)}.$$

Multiplying both sides of Equation 8.4 by $\Pr(S = 1 \,|\, X = x)$ show that there is no selection bias in a cohort study if and only if

$$\Pr(D = 1, S = 1 \,|\, X = x) = \Pr(D = 1 \,|\, X = x) \Pr(S = 1 \,|\, X = x), \tag{8.5}$$

which means that D and S are conditionally independent given X. This condition can be relaxed somewhat if additional covariates are measured. In that case, we only need D and S to be conditionally independent given the measured covariates.

8.2.2 Selection bias in case-control studies

In a case-control study, selection bias leads to biased estimates of exposure prevalences in cases and controls. When only X and D are measured, there is no selection bias in a case-control study if and only if the conditional probability of exposure given disease in sampled individuals equals that in the underlying population:

$$\Pr(X = 1 \,|\, D = d) = \Pr(X = 1 \,|\, D = d, S = 1) \tag{8.6}$$

for $d = 1$ and $d = 0$. By the same argument used for the cohort study, there is no selection bias in a case-control study if and only if

$$\Pr(X = 1, S = 1 \,|\, D = d) = \Pr(X = 1 \,|\, D = d) \Pr(S = 1 \,|\, D = d),$$

which means that X and S are conditionally independent given D. As with selection bias in a cohort study, this condition can be relaxed if additional covariates are measured.

[5] In causal inference, selection bias can threaten both the internal and external validity of a study. It can cause an apparent association between X and D within the study sample that does not represent a causal effect (Hernán, Hernández-Díaz, and Robins 2004).

8.2.3 Prospective and retrospective studies

A **prospective study** is one in which exposure information is collected and recorded prior to disease onset. A **retrospective study** is one in which exposure information is collected after the onset of disease. Traditionally, cohort studies were called "prospective studies" and case-control studies were called "retrospective studies". While this classification is often accurate, it is possible for either design to be prospective or retrospective (Rothman, Greenland, and Lash 2008).

Because selection into a retrospective study occurs after disease onset and relevant exposures occur prior to disease onset, retrospective studies are more susceptible to selection bias than prospective studies. Because knowledge of disease occurrence can affect recall or measurement of exposure, retrospective studies are also more susceptible to differential misclassification.

8.2.4 Generalizability and transportability

To discuss sources of bias in epidemiologic studies, we will use terminology adapted from Dahabreh and Hernán (2019) and illustrated in Figure 8.2. These terms are meant to be consistent with the Consolidated Standards of Reporting Trials (CONSORT) statement (Moher et al. 2001; Altman et al. 2001) as well as the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (Elm et al. 2007; Vandenbroucke et al. 2007). The **eligible population** is the population of individuals who meet the eligibility criteria for a study—whether or not they are invited to participate or willing to participate. Individuals within the eligible population who are invited to participate are the **invited population**, and those within the invited population who enroll in the study are the **study sample**.[6] Members of the study sample are called **participants**. Inferences based on data from the study sample are applied to a **target population** that could be the eligible population or a population that includes individuals outside the eligible population.

A study that makes valid inferences for the eligible population has **generalizability**, and a study that makes valid inferences for a larger or different target population has **transportability** to that population. Generalizability and transportability live along a spectrum of external validity. Generalizability is typically a prerequisite for transportability, and a generalizable study can be transportable to some target populations but not to others. Qualitative insights (e.g., smoking causes lung cancer) might be generalizable or transportable even when the estimated risks of disease or prevalences of exposure are not. The best way to ensure that the results of a study are widely applicable is to make the eligible population as inclusive as possible within ethical and logistical constraints (Bibbins-Domingo and Helman 2022).

[6] We avoid use of the term "study population" because it sometimes refers to the study sample and sometimes to the eligible population.

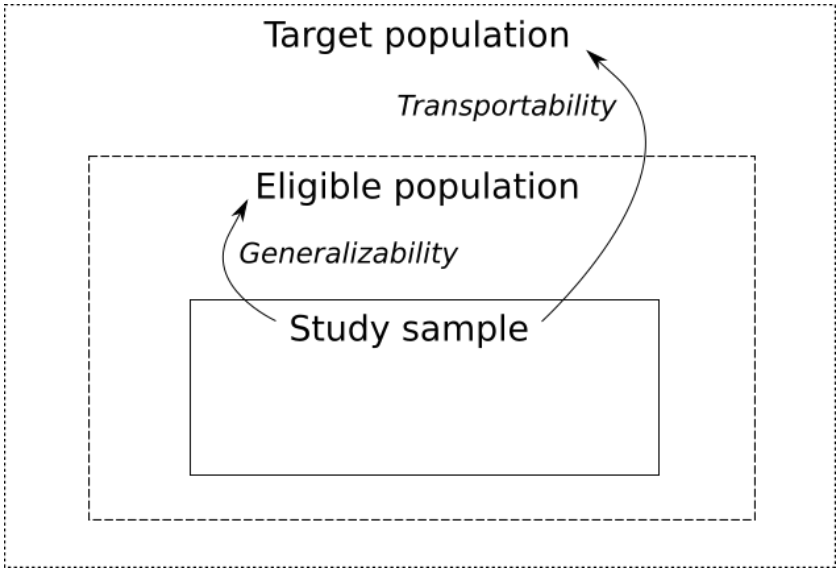


Figure 8.2: Schematic illustration of the relationship between the study sample, the eligible population, and a target population. A target population can contain all, part, or none of the eligible population.

8.2.5 Example: Berkson's bias

Berkson (1946) discussed a hypothetical case-control study to test whether cholecystitis (i.e., gallbladder inflammation) is associated with diabetes. The cases are patients who come to the clinic for diabetes treatment. The controls are nondiabetic patients who come to the clinic to get eyeglasses to correct refractive errors, a diagnosis considered to be independent of cholecystitis based on existing knowledge of human biology. The overall population has 10,000,000 individuals. The prevalence of diabetes is 1%, the prevalence of refractive errors is 10%, and the prevalence of cholecystitis is 3%. All three conditions are assumed to occur independently.

Table 8.5 shows the distribution of cholecystitis among individuals with diabetes and individuals with refractive errors. Eligible cases are individuals with diabetes, and eligible controls are individuals with refractive errors but no diabetes. The prevalence of cholecystitis is 3% in both groups, so $\pi_1 = \pi_0 = 0.03$. It follows that the Pearson chi-squared statistic $\chi^2_{\mathrm{P}} = 0$ in the 2x2 table for the eligible population.

The study sample consists of patients who visit the clinic for diabetes or refractive errors. Each diagnosis comes with a probability of visiting the clinic in the relevant time period. In one version of the example, diabetes patients visit the clinic with probability 0.05, refractive error patients visit with probability 0.20, and cholecystitis patients visit with probability 0.15. For patients with multiple conditions, "we shall say that these selective probabilities operate independently, as though a person who had two diseases were like [conjoined] twins, each one

Table 8.5: Cholecystitis among eligible cases and controls in Berkson (1946)

	Eligible cases	Eligible controls	Total
Cholecystitis	3,000	29,700	32,700
No cholecystitis	97,000	960,300	1,057,300
Total	100,000	990,000	1,090,000

Table 8.6: Combinations of cholecystitis (C), diabetes (D), and refractive errors (R) in Berkson (1946).

Condition	Population	Selection probability	Clinic visitors
None	8,642,700	0	0
C only	267,300	0.15	40,095
D only	87,300	0.05	4,365
R only	960,300	0.20	192,060
C and D	2,700	0.1925	520
C and R	29,700	0.32	9,504
D and R	9,700	0.24	2,328
C, D, and R	300	0.354	106
Total	10,000,000	0.0249	248,978

of whom had one disease, so that the probability of the twins' coming to the hospital is the probability of either one getting there, but the presence of one disease does not affect the other in any way" (Berkson 1946). For example, a patient with both diabetes and refractive error visits the clinic with probability $1-(1-0.05)(1-0.20)=0.24$. Table 8.6 shows the numbers of individuals in the population and among clinic visitors at each combination of cholecystitis, diabetes, and refractive errors.

Table 8.7 shows the 2x2 table for the study sample. The cases with cholecystitis include the 520 individuals with diabetes and cholecystitis only and the 106 individuals with all three conditions. The cases without cholecystitis include the 4,365 individuals with diabetes only and the 2,328 individuals with diabetes and refractive errors only. In this table, we have

$$\chi^2_{\mathrm{P}} = \frac{208{,}883(626 \times 192{,}060 - 6{,}693 \times 192{,}060)^2}{10{,}130 \times 198{,}753 \times 7{,}319 \times 201{,}564} \approx 225.447.$$

The p-value is 5.9×10^{-51}.

If we ignored selection bias, we would conclude that there is almost certainly an association between cholecystitis and diabetes in the eligible population. The example is constructed with no such association. Because the study sample included only clinic visitors and individuals

Table 8.7: Cholecystitis among cases and controls in Berkson (1946)

	Cases	Controls	Total
Cholecystitis	520 + 106 = 626	9,504	10,130
No cholecystitis	4,365 + 2,328 = 6,693	192,060	198,753
Total	7,319	201,564	208,883

with multiple conditions (including cholecystitis) were more likely to visit the clinic, selection and exposure (cholecystitis) are not conditionally independent given disease in this example.

8.3 R

Listing 8.1 ESSratio.R

```
## Effective sample size under nondifferential misclassification

# function that returns the effective sample size ratio
ess_ratio <- function(sens, spec, p) {
  # returns the multiplier of the sample size to get the effective sample size
  # under nondifferential misclassification with marginal risk (or prevalence)
  # p in the sample
  K <- sens + spec - 1
  Kden <- (1 - spec + K * p) * (1 - sens + K * (1 - p))
  return(K^2 * p * (1 - p) / Kden)
}

# data frame for ESSR with specificity = 1 and varying sensitivity
s <- seq(0.01, 1, by = 0.005)
p <- c(0.02, 0.05, 0.1, 0.2, 0.4)
pnames <- c("p02", "p05", "p10", "p20", "p40")
essr_sens <- outer(s, p, function(s, p) ess_ratio(sens = s, spec = 1, p = p))
colnames(essr_sens) <- pnames
essr_sens <- as.data.frame(essr_sens)

# data frame for ESSR with sensitivity = 1 and varying specificity
essr_spec <- outer(s, p, function(s, p) ess_ratio(sens = 1, spec = s, p = p))
colnames(essr_spec) <- pnames
essr_spec <- as.data.frame(essr_spec)

# plot
plot(s, essr_spec$p02, type = "l", asp = 1, xlim = c(0, 1), ylim = c(0, 1),
     xlab = "Sensitivity (gray) or specificity (black)",
     ylab = "Effective sample size ratio")
lines(s, essr_spec$p10, lty = "dashed")
lines(s, essr_spec$p40, lty = "dotdash")
lines(s, essr_sens$p02, col = "darkgray")
lines(s, essr_sens$p10, col = "darkgray", lty = "dashed")
lines(s, essr_sens$p40, col = "darkgray", lty = "dotdash")
grid()
legend("topleft", bg = "white",
       lty = c("solid", "dashed", "dotdash"),
       legend = c("p = 0.02", "p = 0.10", "p = 0.40"))
text(0.48, 0.52, srt = 45, col = "darkgray",
     "Specificity = 1 with varying sensitivity")
text(0.68, 0.32, srt = 45, "Sensitivity = 1 with varying specificity")
```

Listing 8.2 `Berkson.R`

```r
## Berkson (1946) example of selection bias

# Pearson chi-squared test for the eligible population (X and D independent)
poptab <- matrix(c(3000, 97000, 29700, 960300), nrow = 2)
chisq.test(poptab, correct = FALSE)
# Fisher's exact test (with confidence limits for odds ratio)
fisher.test(poptab)

# Pearson chi-squared test for the study sample (X and D not independent)
sampletab <- matrix(c(626, 6693, 9504, 192060), nrow = 2)
chisq.test(sampletab, correct = FALSE)
# Fisher's exact test (with confidence limits for odds ratio)
fisher.test(sampletab)
```

Part III

Principles of Causal Inference

Part IV

Epidemologic and Statistical Methods for Causal Inference

References

Aalen, Odd. 1978. "Nonparametric Inference for a Family of Counting Processes." *The Annals of Statistics* 6: 701–26.

Aalen, Odd, Ørnulf Borgan, and Håkon Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media.

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.

Altman, Douglas G, Kenneth F Schulz, David Moher, Matthias Egger, Frank Davidoff, Diana Elbourne, Peter C Gøtzsche, Thomas Lang, and CONSORT Group. 2001. "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134 (8): 663–94.

Altshuler, Bernard. 1970. "Theory for the Measurement of Competing Risks in Animal Experiments." *Mathematical Biosciences* 6: 1–11.

Baduashvili, Amiran, Arthur T Evans, and Todd Cutler. 2020. "How to Understand and Teach p-Values: A Diagnostic Test Framework." *Journal of Clinical Epidemiology* 122: 49–55.

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.

Berger, James O, and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence." *Journal of the American Statistical Association* 82 (397): 112–22.

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

———. 1946. "Limitations of the Application of Fourfold Table Analysis to Hospital Data." *Biometrics Bulletin* 2 (3): 47–53.

Bibbins-Domingo, Kirsten, and Alex Helman, eds. 2022. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. National Academies of Sciences, Engineering,; Medicine (National Academies Press).

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

Blaker, Helge. 2000. "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions." *Canadian Journal of Statistics* 28 (4): 783–98.

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.

Brittain, Erica, James J Schlesselman, and Bruce V Stadel. 1981. "Cost of Case-Control Studies." *American Journal of Epidemiology* 114 (2): 234–43.

Bross, Irwin. 1954. "Misclassification in 2 x 2 Tables." *Biometrics* 10 (4): 478–86.

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

———. 2003. "Interval Estimation in Exponential Families." *Statistica Sinica* 13: 19–49.

Buell, Philip, and John E Dunn Jr. 1964. "The Dilution Effect of Misclassification." *American Journal of Public Health* 54 (4): 598–602.

Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54 (4): 297.

Chung, Kai Lai. 2000. *A Course in Probability Theory*. Third. Elsevier.

Clarke, RD. 1946. "An Application of the Poisson Distribution." *Journal of the Institute of Actuaries* 72 (3): 481–81.

Clayton, David, and Michael Hills. 1993. *Statistical Models in Epidemiology*. Oxford University Press.

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, Geoffrey R, and Shu-Ying Yang. 1994. "Mid-p Confidence Intervals for the Poisson Expectation." *Statistics in Medicine* 13 (21): 2189–2203.

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

Copeland, Karen T, Harvey Checkoway, Anthony J McMichael, and Robert H Holbrook. 1977. "Bias Due to Misclassification in the Estimation of Relative Risk." *American Journal of Epidemiology* 105 (5): 488–95.

Correa-Villaseñor, Adolfo, Walter F Stewart, Francisco Franco-Marina, and Hui Seacat. 1995. "Bias from Nondifferential Misclassification in Case-Control Studies with Three Exposure Levels." *Epidemiology* 6 (3): 276–81.

Cox, David R, and E Joyce Snell. 1968. "A General Definition of Residuals." *Journal of the*

Royal Statistical Society: Series B (Methodological) 30 (2): 248–65.
Cramér, Harald. 1946. *Mathematical Methods of Statistics*. Princeton University Press.
Dahabreh, Issa J, and Miguel A Hernán. 2019. "Extending Inferences from a Randomized Trial to a Target Population." *European Journal of Epidemiology* 34 (8): 719–22.
Dawid, A Philip. 1979. "Conditional Independence in Statistical Theory." *Journal of the Royal Statistical Society: Series B (Methodological)* 41 (1): 1–15.
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.
Dosemeci, Mustafa, Sholom Wacholder, and Jay H Lubin. 1990. "Does Nondifferential Misclassification of Exposure Always Bias a True Effect Toward the Null Value?" *American Journal of Epidemiology* 132 (4): 746–48.
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.
Edwards, Ward, Harold Lindman, and Leonard J Savage. 1963. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70 (3): 193–242.
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.
Elandt-Johnson, Regina C. 1975. "Definition of Rates: Some Remarks on Their Use and Misuse." *American Journal of Epidemiology* 102 (4): 267–71.
Elm, Erik von, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, and Jan P Vandenbroucke. 2007. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *The Lancet* 370 (9596): 1453–57.
Fagan, Terrence J. 1975. "Nomogram for Bayes's Theorem." *New England Journal of Medicine* 293 (5): 257.
Farr, William. 1838. "On Prognosis." *British Medical Almanack* Supplement: 199–216.
Fay, Michael P. 2010. "Confidence Intervals That Match Fisher's Exact or Blaker's Exact Tests." *Biostatistics* 11 (2): 373–74.
Fisher, Ronald A. 1922. "On the Interpretation of χ 2 from Contingency Tables, and the Calculation of p." *Journal of the Royal Statistical Society* 85 (1): 87–94.
———. 1935. "The Logic of Inductive Inference." *Journal of the Royal Statistical Society* 98 (1): 39–82.
Fisk, Peter R. 1961. "The Graduation of Income Distributions." *Econometrica: Journal of the Econometric Society* 29 (2): 171–85.
Fleming, Thomas R, and David P Harrington. 1984. "Nonparametric Estimation of the Survival Distribution in Censored Data." *Communications in Statistics-Theory and Methods* 13 (20): 2469–86.
———. 2005. *Counting Processes and Survival Analysis*. Vol. 625. John Wiley & Sons.
Freedman, David A. 2007. "How Can the Score Test Be Inconsistent?" *The American Statistician* 61 (4): 291–95.

Freedman, Laurence S, Arthur Schatzkin, and Yohanan Wax. 1990. "The Impact of Dietary Measurement Error on Planning Sample Size Required in a Cohort Study." *American Journal of Epidemiology* 132 (6): 1185–95.

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

Gail, Mitchell, Roger Williams, David P Byar, Charles Brown, et al. 1976. "How Many Controls?" *Journal of Chronic Diseases* 29 (11): 723–31.

Garwood, F. 1936. "Fiducial Limits for the Poisson Distribution." *Biometrika* 28 (3/4): 437–42.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC press.

Goldberg, Judith D. 1975. "The Effects of Misclassification on the Bias in the Difference Between Two Proportions and the Relative Odds in the Fourfold Table." *Journal of the American Statistical Association* 70 (351): 561–67.

Gompertz, Benjamin. 1825. "XXIV. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." *Philosophical Transactions of the Royal Society of London* 115: 513–83.

Greenland, Sander. 2006. "Bayesian Perspectives for Epidemiological Research: I. Foundations and Basic Methods." *International Journal of Epidemiology* 35 (3): 765–75.

Greenland, Sander, and Charles Poole. 2013. "Living with p Values: Resurrecting a Bayesian Perspective on Frequentisi Statistics." *Epidemiology* 24 (1): 62–68.

Greenland, Sander, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. "Statistical Tests, p Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (4): 337–50.

Greenwood, Major. 1926. "The Natural Duration of Cancer." *Reports on Public Health and Medical Subjects* 33: 1–26.

Gullen, Warren H, Jacob E Bearman, and Eugene A Johnson. 1968. "Effects of Misclassification in Epidemiologic Studies." *Public Health Reports* 83 (11): 914–18.

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.

Hernán, Miguel Á, Sonia Hernández-Díaz, and James M Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15 (5): 615–25.

Hill, Sir Austin Bradford. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58: 295–300.

Irwin, JO et al. 1935. "Tests of Significance for Differences Between Percentages Based on Small Numbers." *Metron* 12 (2): 84–94.

Jeffreys, Harold. 1946. "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186 (1007): 453–61.

Kaplan, Edward L, and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81.

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.

Le Cam, Lucien. 1953. "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates." *University of California Publications in Statistics* 1 (11): 277–330.

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.

Makeham, William Matthew. 1860. "On the Law of Mortality and the Construction of Annuity Tables." *The Assurance Magazine, and Journal of the Institute of Actuaries* 8 (6): 301–10.

Meydrech, Edward F, and Lawrence L Kupper. 1978. "Cost Considerations and Sample Size Requirements in Cohort and Case-Control Studies." *American Journal of Epidemiology* 107 (3): 201–5.

Miettinen, Olli S. 1969. "Individual Matching with Multiple Controls in the Case of All-or-None Responses." *Biometrics* 25 (2): 339–55.

Moher, David, Kenneth F Schulz, Douglas G Altman, Matthias Egger, Frank Davidoff, Diana Elbourne, Peter C. Gøtzsche, Thomas Lang, and CONSORT Group. 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials." *Annals of Internal Medicine* 134 (8): 657–62.

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

Morgenstern, Hal, David G Kleinbaum, and Lawrence L Kupper. 1980. "Measures of Disease Incidence Used in Epidemiologic Research." *International Journal of Epidemiology* 9 (1): 97–104.

Morgenstern, Hal, and Deborah M Winn. 1983. "A Method for Determining the Sampling Ratio in Epidemiologic Studies." *Statistics in Medicine* 2 (3): 387–96.

Nam, Jun-Mo. 1973. "Optimum Sample Sizes for the Comparison of the Control and Treatment." *Biometrics* 29: 101–8.

Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *Journal of Quality Technology* 1: 27–52.

———. 1972. "Theory and Applications of Hazard Plotting for Censored Failure Data." *Technometrics* 14 (4): 945–66.

Newell, David J. 1962. "Errors in the Interpretation of Errors in Epidemiology." *American*

Journal of Public Health 52 (11): 1925–28.

Newell, DJ. 1963. "Note: Misclassification in 2 x 2 Tables." *Biometrics* 19 (1): 187–88.

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.

Pearson, Karl. 1900. "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (302): 157–75.

———. 1922. "On the χ 2 Test of Goodness of Fit." *Biometrika* 14 (1/2): 186–91.

Pike, MC, and JT Casagrande. 1979. "Re:'cost Considerations and Sample Size Requirements in Cohort and Case-Control Studies'." *American Journal of Epidemiology* 110 (1): 100–102.

Pratt, John W. 1965. "Bayesian Interpretation of Standard Inference Statements." *Journal of the Royal Statistical Society: Series B (Methodological)* 27 (2): 169–203.

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.

Rogot, Eugene. 1961. "A Note on Measurement Errors and Detecting Real Differences." *Journal of the American Statistical Association* 56 (294): 314–19.

Roscoe, John T, and Jackson A Byars. 1971. "An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic." *Journal of the American Statistical Association* 66 (336): 755–59.

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.

Rothman, Kenneth J, Sander Greenland, and Timothy L Lash. 2008. *Modern Epidemiology*. Lippincott Williams & Wilkins.

Routledge, RD. 1992. "Resolving the Conflict over Fisher's Exact Test." *Canadian Journal of Statistics* 20 (2): 201–9.

Rubin, Theodore, Joseph Rosenbaum, and Sidney Cobb. 1956. "The Use of Interview Data for the Detection of Associations in Field Studies." *Journal of Chronic Diseases* 4 (3): 253–66.

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

Sorahan, Tom, and Mark S Gilthorpe. 1994. "Non-Differential Misclassification of Exposure Always Leads to an Underestimate of Risk: An Incorrect Conclusion." *Occupational and Environmental Medicine* 51 (12): 839–40.

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.

Swift, Michael Bruce. 2009. "Comparison of Confidence Intervals for a Poisson Mean–Further Considerations." *Communications in Statistics–Theory and Methods* 38 (5): 748–59.

Tukey, John W. 1960. "Conclusions Vs Decisions." *Technometrics* 2 (4): 423–33.

———. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1): 1–67.

Van der Vaart, Aad W. 2000. *Asymptotic Statistics*. Cambridge University Press.

Vandenbroucke, Jan P, Erik von Elm, Douglas G Altman, Peter C Gøtzsche, Cynthia D Mulrow, Stuart J Pocock, Charles Poole, James J Schlesselman, Matthias Egger, and Strobe Initiative. 2007. "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration." *Annals of Internal Medicine* 147 (8): W–163.

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

Verkerk, PH, and SE Buitendijk. 1992. "Non-Differential Underestimation May Cause a Threshold Effect of Exposure to Appear as a Dose-Response Relationship." *Journal of Clinical Epidemiology* 45 (5): 543–45.

Wacholder, Sholom, Patricia Hartge, Jay H Lubin, and Mustafa Dosemeci. 1995. "Non-Differential Misclassification and Bias Towards the Null: A Clarification." *Occupational and Environmental Medicine* 52 (8): 557.

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.

Walker, Alexander M, Johan P Velema, and James M Robins. 1988. "Analysis of Case-Control Data Derived in Part from Proxy Respondents." *American Journal of Epidemiology* 127 (5): 905–14.

Walter, Samuel D. 1977. "Determination of Significant Relative Risks and Optimal Sampling Procedures in Prospective and Retrospective Comparative Studies of Various Sizes." *American Journal of Epidemiology* 105 (4): 387–97.

Weibull, Waloddi et al. 1951. "A Statistical Distribution Function of Wide Applicability." *Journal of Applied Mechanics* 18 (3): 293–97.

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 Hospi-

tal 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.

Yland, Jennifer J, Amelia K Wesselink, Timothy L Lash, and Matthew P Fox. 2022. "Misconceptions about the Direction of Bias from Nondifferential Misclassification." *American Journal of Epidemiology* 191 (8): 1485–95.

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