Measures of Association for Binary Outcomes

Risks, odds, odds ratios, and the logit and expit functions

Contents

Configuring R

Functions from these packages will be used throughout this document:

```r
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
```

```r
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```r
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

```r
options(digits = 6)
```


Acknowledgements

This content is adapted from:

- Dobson and Barnett (2018), Chapter 7
- Vittinghoff et al. (2012), Chapter 5
- David Rocke[1]'s materials from the 2021 edition of Epi 204[2]
- Nahhas (2024) Chapter 6[3]

1 Introduction

This chapter reviews measures of association for binary outcomes: risks, risk differences, risk ratios, odds, and odds ratios, along with the logit and expit functions that connect probabilities, odds, and log-odds. These concepts are the vocabulary for regression models of binary outcomes; the logistic regression chapter[4] builds on them.

[1] https://dmrocke.ucdavis.edu/
[2] https://dmrocke.ucdavis.edu/Class/EPI204-Spring-2021/EPI204-Spring-2021.html
[3] https://www.bookdown.org/rwnahhas/RMPH/blr.html
[4] logistic-regression.qmd

1.1 Risk estimation and prediction

In Epi 203, you have already seen methods for modeling binary outcomes using one covariate that is also binary (such as exposure/non-exposure). In this section, we review one-covariate analyses, with a special focus on risk ratios and odds ratios, which are important concepts for interpreting logistic regression.

Exm

Example 1.1 (Oral Contraceptive Use and Heart Attack)**.**

- Research question: how does oral contraceptive (OC) use affect the risk of myocardial infarction (MI; a.k.a. heart attack)?

This was a concern when oral contraceptives were first developed, because the original formulations used much higher doses of estrogen than are used today. Cardiovascular risk from combined oral contraceptives increases with estrogen dose, and the dose has been reduced substantially since those early formulations, so modern low-dose formulations carry much lower risk (Roach et al. 2015).

Table 1 contains simulated data for an imaginary follow-up (a.k.a. *prospective*) study in which two groups are identified, one using OCs and another not using OCs, and both groups are tracked for three years to determine how many in each groups have MIs.

```r
library(dplyr)
oc_mi <-
  tribble(
    ~OC, ~MI, ~Total,
    "OC use", 13, 5000,
    "No OC use", 7, 10000
  ) |>
  mutate(`No MI` = Total - MI) |>
  relocate(`No MI`, .after = MI)

totals <-
  oc_mi |>
  summarize(across(c(MI, `No MI`, Total), sum)) |>
  mutate(OC = "Total")

tbl_oc_mi <- bind_rows(oc_mi, totals)

tbl_oc_mi |> pander::pander()
```

Table 1: Simulated data from study of oral contraceptive use and heart attack risk

OC	MI	No MI	Total
OC use	13	4,987	5,000
No OC use	7	9,993	10,000
Total	20	14,980	15,000

Exercise 1.1. Estimate the probabilities of MI for OC users and non-OC users in Example 1.1.

Solution

Solution 1.1.

$$\hat{p}(MI|OC) = \frac{13}{5000} = 0.0026$$

$$\hat{p}(MI|\neg OC) = \frac{7}{10000} = 7 \times 10^{-4}$$

Exercise 1.2. What does the term "controls" mean in the context of study design?

Solution

Solution 1.2.

Def

Definition 1.1 (Two meanings of "controls"). Depending on context, "controls" can mean either:

- individuals who don't experience an *exposure* of interest,
- or individuals who don't experience an *outcome* of interest.

Exercise 1.3. What types of studies do the two definitions of controls correspond to?

Solution

Solution 1.3.

Def

Definition 1.2 (cases and controls in retrospective studies). In *retrospective case-control studies*, participants who experience the outcome of interest are called **cases**, while participants who don't experience that outcome are called **controls** (Rothman et al. 2021, chap. 8).

Def

Definition 1.3 (treatment groups and control groups in prospective studies). In *prospective studies*, the group of participants who experience the treatment or exposure of interest is called the **treatment group**, while the participants who receive the baseline or comparison treatment (for example, clinical trial participants who receive a placebo or a standard-of-care treatment rather than an experimental treatment) are called **controls** (Rothman et al. 2021, chap. 7).

Exm

Example 1.2 (Cases and controls in the OC-MI study)**.** Example 1.1 describes a *prospective* (follow-up) study: participants were enrolled based on their exposure status (OC use versus no OC use) and then followed for three years to observe MI outcomes.
Therefore, Definition 1.3 applies: the 5000 OC users form the treatment (exposed) group, and the 10,000 OC non-users are the **controls**.
Definition 1.2 does not apply to this study design, because participants were not selected based on their outcomes. If we had instead recruited participants by outcome status — enrolling the 20 participants who experienced MIs, along with a sample of participants who did not — the study would have been a *retrospective case-control study*: the 20 participants with MIs would be the **cases**, and the sampled MI-free participants would be the **controls**.

1.2 Comparing probabilities

1.2.1 Risk differences

The simplest comparison of two probabilities, π_1, and π_2, is the difference of their values:

Definition 1.4 (Risk difference)**.** The **risk difference** of two probabilities, π_1, and π_2, is the difference of their values:

$$\delta(\pi_1, \pi_2) \stackrel{\text{def}}{=} \pi_1 - \pi_2$$

Exm

Example 1.3 (Difference in MI risk)**.** Here, we introduce the shorthand notation

$$\pi(OC) \stackrel{\text{def}}{=} \Pr(MI|OC), \quad \pi(\neg OC) \stackrel{\text{def}}{=} \Pr(MI|\neg OC)$$

for the probability of MI given OC use and given OC non-use, respectively.
In Example 1.1, the maximum likelihood estimate of the difference in MI risk between OC users and OC non-users is:

$$\begin{aligned}\hat{\delta}(\pi(OC), \pi(\neg OC)) &= \delta(\hat{\pi}(OC), \hat{\pi}(\neg OC)) \\ &= \hat{\pi}(OC) - \hat{\pi}(\neg OC) \\ &= 0.0026 - 7 \times 10^{-4} \\ &= 0.0019\end{aligned}$$

Exercise 1.4 (interpreting risk differences)**.** How can we interpret the preceding risk difference estimate in prose?

Solution (interpreting risk differences)

Solution 1.4 (interpreting risk differences). "The difference in risk of MI between OC users and non-users was 0.0019."
or
"The difference in risk of MI between OC users and non-users was 0.19 percentage points[a]."

[a]https://en.wikipedia.org/wiki/Percentage_point

See the note about working with percentages in the Appendix[5].

1.2.2 Risk ratios

Exercise 1.5. If π_1 and π_2 are two probabilities, what do we call the following ratio?

$$\rho(\pi_1, \pi_2) = \frac{\pi_1}{\pi_2}$$

Solution

Solution 1.5.

Def

Definition 1.5 (Relative risk ratios). The ratio of two probabilities π_1 and π_2,

$$\rho(\pi_1, \pi_2) = \frac{\pi_1}{\pi_2}$$

is called the:

- **risk ratio**,
- **relative risk ratio**,
- **probability ratio**,
- or **rate ratio**

of π_1 compared to π_2.

Exercise 1.6.
In Solution 1.1, we estimated that:

[5]notation.qmd#percent-sign

$$\hat{\mathrm{p}}(MI|OC) = 0.0026$$

$$\hat{\mathrm{p}}(MI|\neg OC) = 7 \times 10^{-4}$$

Now, estimate the *relative risk* of MI for OC versus non-OC.

Solution

Solution 1.6.
The *relative risk* of MI for OC versus non-OC is:

```
rr <- (13 / 5000) / (7 / 10000)
```

$$\begin{aligned}\hat{\rho}(OC, \neg OC) &= \frac{\hat{\mathrm{p}}(MI|OC)}{\hat{\mathrm{p}}(MI|\neg OC)} \\ &= \frac{0.0026}{7 \times 10^{-4}} \\ &= 3.714286\end{aligned}$$

Exercise 1.7. How can we interpret the preceding relative risk estimate in prose?

Solution

Solution 1.7.
We might summarize this result by saying that:

- "The estimated probability of MI among OC users was 3.714286 times as high as the estimated probability among OC non-users."

or

- "The estimated probability of MI among OC users was 2.714286 times higher than, the estimated probability among OC non-users."

see also Vittinghoff et al. (2012, sec. 8.1.4), which uses similar phrasing.

1.2.3 Relative risk differences

The second approach in Solution 1.7, where we subtract 1 from the risk ratio, is actually reporting a slightly different metric:

Definition 1.6 (Relative risk difference)**.**
Sometimes, we divide the risk difference by the comparison probability; the result is called the **relative risk difference**:

$$\xi(\pi_1, \pi_2) \stackrel{\text{def}}{=} \frac{\delta(\pi_1, \pi_2)}{\pi_2}$$

Theorem 1.1 (Relative risk difference equals risk ratio minus 1)**.**

$$\xi(\pi_1, \pi_2) = \rho(\pi_1, \pi_2) - 1$$

Proof

Proof.

$$
\begin{aligned}
\xi(\pi_1, \pi_2) &\overset{\text{def}}{=} \frac{\delta(\pi_1, \pi_2)}{\pi_2} \\
&= \frac{\pi_1 - \pi_2}{\pi_2} \\
&= \frac{\pi_1}{\pi_2} - 1 \\
&= \rho(\pi_1, \pi_2) - 1
\end{aligned}
$$

□

Exm

Example 1.4 (Relative difference in MI risk)**.** In Example 1.1, the maximum likelihood estimate of the relative risk difference in MI risk between OC users and OC non-users is:

$$
\begin{aligned}
\hat{\xi}(\pi(OC), \pi(\neg OC)) &= \xi(\hat{\pi}(OC), \hat{\pi}(\neg OC)) && \text{(invariance property of MLEs)} \\
&\overset{\text{def}}{=} \frac{\delta(\hat{\pi}(OC), \hat{\pi}(\neg OC))}{\hat{\pi}(\neg OC)} && \text{(definition of relative risk difference)} \\
&= \frac{0.0019}{7 \times 10^{-4}} && \text{(substituting the estimates)} \\
&= 2.714286 && \text{(dividing)}
\end{aligned}
$$

The numerator 0.0019 is the risk difference estimate from Example 1.3, and the denominator 7×10^{-4} is the estimate $\hat{\pi}(\neg OC)$ from Solution 1.1.
We can check this estimate using Theorem 1.1 and the risk ratio estimate from Solution 1.6:

$$
\begin{aligned}
\hat{\xi}(\pi(OC), \pi(\neg OC)) &= \hat{\rho}(OC, \neg OC) - 1 && \text{(relative risk difference as risk ratio minus 1)} \\
&= 3.714286 - 1 && \text{(substituting the risk ratio estimate)} \\
&= 2.714286 && \text{(subtracting)}
\end{aligned}
$$

The two calculations agree.

1.2.4 Changing reference groups in risk comparisons

Risk differences, risk ratios, and relative risk differences are defined by two probabilities, plus a choice of which probability is the **baseline** or **reference** probability (i.e., which probability is the subtrahend of the risk difference or the denominator of the risk ratio).

$$
\delta(\pi_2, \pi_1) = -\delta(\pi_1, \pi_2) \tag{1}
$$

$$
\rho(\pi_2, \pi_1) = (\rho(\pi_1, \pi_2))^{-1} \tag{2}
$$

$$
\xi(\pi_2, \pi_1) = (\xi(\pi_1, \pi_2) + 1)^{-1} - 1 \tag{3}
$$

Exercise 1.8. Prove Equation 1, Equation 2, and Equation 3.

Solution

Solution 1.8. To prove Equation 1:

$$\begin{aligned}
\delta(\pi_2, \pi_1) &\overset{\text{def}}{=} \pi_2 - \pi_1 && \text{(definition of risk difference)} \\
&= -(\pi_1 - \pi_2) && \text{(factoring out } -1) \\
&= -\delta(\pi_1, \pi_2) && \text{(definition of risk difference)}
\end{aligned}$$

To prove Equation 2:

$$\begin{aligned}
\rho(\pi_2, \pi_1) &\overset{\text{def}}{=} \frac{\pi_2}{\pi_1} && \text{(definition of risk ratio)} \\
&= \left(\frac{\pi_1}{\pi_2}\right)^{-1} && \text{(inverse of the inverted fraction)} \\
&= (\rho(\pi_1, \pi_2))^{-1} && \text{(definition of risk ratio)}
\end{aligned}$$

To prove Equation 3, first rearrange Theorem 1.1 to express the risk ratio in terms of the relative risk difference:

$$\begin{aligned}
\xi(\pi_1, \pi_2) = \rho(\pi_1, \pi_2) - 1 &&& \text{(relative risk difference as risk ratio minus 1)} \\
\xi(\pi_1, \pi_2) + 1 = \rho(\pi_1, \pi_2) &&& \text{(adding 1 to both sides)}
\end{aligned}$$

Then apply Theorem 1.1 to the switched-reference pair (π_2, π_1), substitute Equation 2, and substitute the rearranged form of Theorem 1.1:

$$\begin{aligned}
\xi(\pi_2, \pi_1) &= \rho(\pi_2, \pi_1) - 1 && \text{(relative risk difference as risk ratio minus 1)} \\
&= (\rho(\pi_1, \pi_2))^{-1} - 1 && \text{(risk ratio with switched reference group)} \\
&= (\xi(\pi_1, \pi_2) + 1)^{-1} - 1 && \text{(substituting } \rho(\pi_1, \pi_2) = \xi(\pi_1, \pi_2) + 1)
\end{aligned}$$

Exm

Example 1.5 (Switching the reference group in a risk difference)**.** In Example 1.3, we estimated that the risk difference of MI for OC versus non-OC is:

$$\delta(OC, \neg OC) = 0.0019$$

In comparison, the risk difference for non-OC versus OC is:

$$\begin{aligned}
\delta(\neg OC, OC) &= \hat{\mathrm{p}}(MI|\neg OC) - \hat{\mathrm{p}}(MI|OC) \\
&= 7 \times 10^{-4} - 0.0026 \\
&= -0.0019 \\
&= -\delta(OC, \neg OC)
\end{aligned}$$

as Equation 1 states.

Exm

Example 1.6 (Switching the reference group in a risk ratio)**.** In Solution 1.6, we estimated that the risk ratio of OC versus non-OC is:

$$\rho(OC, \neg OC) = 3.714286$$

In comparison, the risk ratio for non-OC versus OC is:

$$
\begin{aligned}
\rho(\neg OC, OC) &= \frac{\hat{\text{p}}(MI|\neg OC)}{\hat{\text{p}}(MI|OC)} \\
&= \frac{7 \times 10^{-4}}{0.0026} \\
&= 0.269231 \\
&= \frac{1}{\rho(OC, \neg OC)}
\end{aligned}
$$

as Equation 2 states.

Exm

Example 1.7 (Switching the reference group in a relative risk difference)**.** In Example 1.4, we estimated that the relative risk difference of MI for OC versus non-OC is:

$$\xi(OC, \neg OC) = 2.714286$$

In comparison, the relative risk difference for non-OC versus OC is:

$$
\begin{aligned}
\xi(\neg OC, OC) &= \frac{\hat{\text{p}}(MI|\neg OC) - \hat{\text{p}}(MI|OC)}{\hat{\text{p}}(MI|OC)} \\
&= \frac{-0.0019}{0.0026} \\
&= -0.730769
\end{aligned}
$$

We can check this value against the switched-reference identity Equation 3:

$$
\begin{aligned}
\frac{1}{\xi(OC, \neg OC) + 1} - 1 &= \frac{1}{2.714286 + 1} - 1 \\
&= -0.730769
\end{aligned}
$$

The two calculations agree.

1.3 Odds and odds ratios

1.3.1 Odds and probabilities

In logistic regression, we will make use of a mathematically-convenient transformation of probability, called *odds*:

Definition 1.7 (Odds)**.**
The **odds** of an event A, is the probability that the event occurs, divided by the probability that it doesn't occur. We can represent odds with the Greek letter ω ("omega"). [6] Thus, in mathematical notation:

$$\omega \stackrel{\text{def}}{=} \frac{\Pr(A)}{\Pr(\neg A)} \tag{4}$$

This course is about regression models, which are conditional probability models (regression models[7]). Accordingly, we define conditional odds in terms of conditional probabilities:

Definition 1.8 (Conditional odds)**.**
The **conditional odds** of an event A given a condition B, is the (conditional) probability that

[6] The name "omega" is a contraction of "o mega", which means "long o" in Greek, in contrast with "omicron" (o, "short o"). See https://www.etymonline.com/search?q=omega and https://en.wikipedia.org/wiki/Omega for more details.

[7] Intro-to-GLMs.qmd#def-regression-model

event A occurs (given condition B), divided by the (conditional) probability that it doesn't occur (given condition B). In mathematical notation:

$$\omega(A|B) \stackrel{\text{def}}{=} \frac{\Pr(A|B)}{\Pr(\neg A|B)} \tag{5}$$

When the event A is clear from context, we abbreviate $\omega(A|B)$ as $\omega(B)$ or ω_B ("omega bee").

Definition 1.9 (Shorthand for MI risk under OC use)**.** Let $\pi(OC)$ denote the probability of MI given OC use:

$$\pi(OC) \stackrel{\text{def}}{=} \Pr(MI|OC)$$

Exm

Example 1.8 (Computing odds from probabilities)**.** In Exercise 1.1, we estimated that $\pi(OC) = 0.0026$ (Definition 1.9). If this estimate is correct, then the odds of MI, given OC use, is:

$$\begin{aligned}
\omega(OC) &\stackrel{\text{def}}{=} \frac{\Pr(MI|OC)}{\Pr(\neg MI|OC)} \\
&= \frac{\Pr(MI|OC)}{1 - \Pr(MI|OC)} \\
&= \frac{\pi(OC)}{1 - \pi(OC)} \\
&= \frac{0.0026}{1 - 0.0026} \\
&\approx 0.002607
\end{aligned}$$

Exercise 1.9 (Computing odds from probabilities)**.** Estimate the odds of MI, for non-OC users.

Solution

Solution 1.9.

$$\omega(\neg OC) = 7.004903 \times 10^{-4}$$

Exercise 1.10. Find a general formula for converting probabilities into odds.

Solution

Solution 1.10. Using Definition 1.7 and the complementary probability[a]:

$$\begin{aligned}
\omega &\stackrel{\text{def}}{=} \frac{\Pr(A)}{\Pr(\neg A)} \\
&= \frac{\pi}{1 - \pi}
\end{aligned}$$

[a] probability.qmd#cor-p-neg

Theorem 1.2 (Odds as a function of probability)**.** *If π is the probability of an event A and ω is the corresponding odds of A, then:*

$$\omega = \frac{\pi}{1 - \pi} \tag{6}$$

ℹ Proof

Proof. By Solution 1.10. □

The mathematical relationship between odds ω and probabilities π, which is represented in Equation 6, is a core component of logistic regression models, as we will see in the rest of this chapter. Let's give the expression on the righthand side of Equation 6 its own name and symbol, so that we can refer to it concisely:

Definition 1.10 (Odds function)**.** The **odds function** is defined as:

$$\text{odds}\{\pi\} \stackrel{\text{def}}{=} \frac{\pi}{1 - \pi} \tag{7}$$

We can use the odds function (Definition 1.10) to simplify Equation 6 (in Theorem 1.2) into a more concise expression, which is easier to remember and manipulate:

Corollary 1.1 (Odds via the odds function)**.** *If π is the probability of an outcome A and ω is the corresponding odds of A, then:*

$$\omega = \text{odds}\{\pi\} \tag{8}$$

In other words, the odds function rescales probabilities into odds.

ℹ Proof

Proof. By Theorem 1.2 and Definition 1.10. □

Exercise 1.11. Graph the odds function.

Solution

Solution 1.11.
Figure 1 graphs the odds function.

```r
odds <- function(pi) pi / (1 - pi)
library(ggplot2)
ggplot() +
  geom_function(
    fun = odds,
    arrow = arrow(ends = "last"),
    mapping = aes(col = "odds function")
  ) +
  xlim(0, .99) +
  xlab("Probability") +
  ylab("Odds") +
  geom_abline(aes(
    intercept = 0,
    slope = 1,
    col = "y=x"
  )) +
  theme_bw() +
  labs(colour = "") +
  theme(legend.position = "bottom")
```

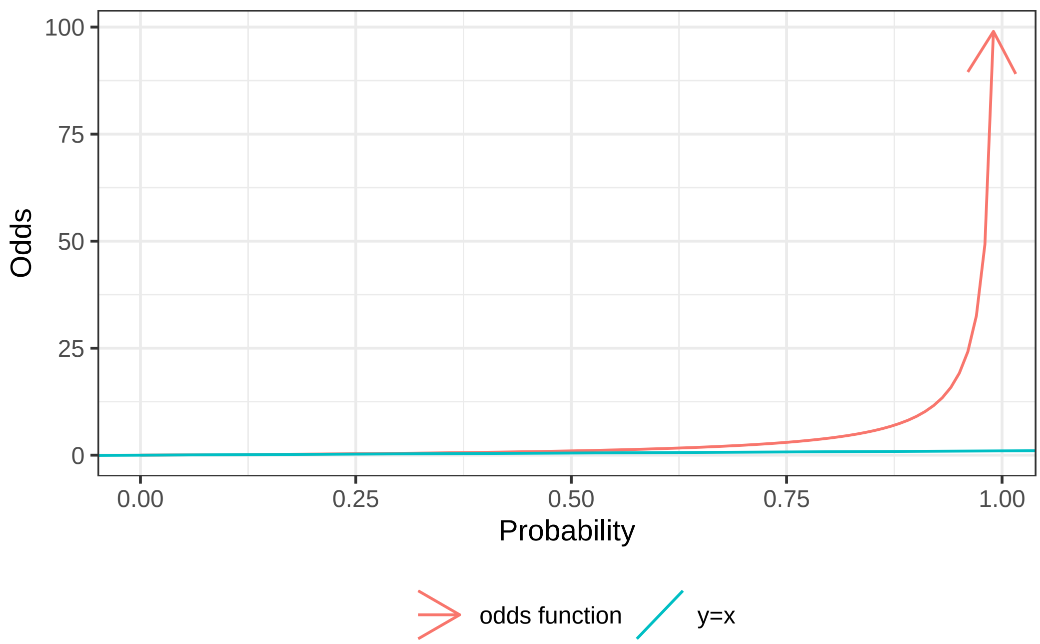

Figure 1: Odds versus probability

Theorem 1.3 (One-sample MLE for odds)**.** *Let $X_1, ...X_n$ be a set of n iid Bernoulli trials, and let $X = \sum_{i=1}^{n} X_i$ be their sum.*
Then the maximum likelihood estimate of the common per-trial success odds ω (that is, the odds of $X_i = 1$) is:

$$\hat{\omega} = \frac{x}{n - x}$$

Proof

Proof.

$$
\begin{aligned}
1-\hat{\pi} &= 1-\frac{x}{n} \\
&= \frac{n}{n}-\frac{x}{n} \\
&= \frac{n-x}{n}
\end{aligned}
$$

Thus, the estimated odds is:

$$
\begin{aligned}
\frac{\hat{\pi}}{1-\hat{\pi}} &= \frac{\left(\frac{x}{n}\right)}{\left(\frac{n-x}{n}\right)} \\
&= \frac{x}{n-x}
\end{aligned}
\tag{9}
$$

That is, the odds estimate can be computed directly as "# events" divided by "# nonevents", without needing to compute $\hat{\pi}$ and $1-\hat{\pi}$ first.

□

Exm

Example 1.9 (Calculating odds using the shortcut)**.** In Example 1.8, we calculated

$$\omega(OC) = 0.002607$$

Let's recalculate this result using our shortcut.

Solution

Solution 1.12.

$$
\begin{aligned}
\omega(OC) &= \frac{13}{5000-13} \\
&= 0.002607
\end{aligned}
$$

Same answer as in Example 1.8!

Theorem 1.4 (Simplified expressions for odds function)**.**
Two equivalent expressions for the odds function are:

$$
\begin{aligned}
\text{odds}\{\pi\} &= \frac{1}{\pi^{-1}-1} \\
&= \left(\pi^{-1}-1\right)^{-1}
\end{aligned}
\tag{10}
$$

Exercise 1.12. Prove Theorem 1.4.

Solution

Solution 1.13. Starting from Definition 1.10:

$$
\begin{aligned}
\text{odds}\{\pi\} &= \frac{\pi}{1-\pi} \\
&= \frac{\pi}{1-\pi}\frac{\pi^{-1}}{\pi^{-1}} \\
&= \frac{\pi\pi^{-1}}{(1-\pi)\pi^{-1}} \\
&= \frac{1}{(\pi^{-1}-\pi\pi^{-1})} \\
&= \frac{1}{(\pi^{-1}-1)} \\
&= (\pi^{-1}-1)^{-1}
\end{aligned}
$$

Corollary 1.2 (Odds of a non-event)**.** *If π is the probability of event A and ω is the corresponding odds of A, then the odds of $\neg A$ are:*

$$
\begin{aligned}
\omega(\neg A) &= \frac{1-\pi}{\pi} \\
&= \pi^{-1} - 1
\end{aligned}
$$

Proof

Proof. Using Definition 1.7 (applied to the event $\neg A$) and the complement rule[a]:

$$
\begin{aligned}
\omega(\neg A) &\overset{\text{def}}{=} \frac{\Pr(\neg A)}{\Pr(\neg\neg A)} && \text{(definition of odds, applied to } \neg A) \\
&= \frac{\Pr(\neg A)}{\Pr(A)} && \text{(double negation: } \neg\neg A = A) \\
&= \frac{1-\pi}{\Pr(A)} && \text{(complement rule: } \Pr(\neg A) = 1-\pi) \\
&= \frac{1-\pi}{\pi} && (\Pr(A) = \pi \text{ by hypothesis}) \\
&= \frac{1}{\pi} - \frac{\pi}{\pi} && \text{(split the fraction)} \\
&= \frac{1}{\pi} - 1 && \text{(cancel: } \pi/\pi = 1) \\
&= \pi^{-1} - 1 && \text{(rewrite the reciprocal as a } -1 \text{ power)}
\end{aligned}
$$

□

[a] probability.qmd#cor-p-neg

Odds of rare events

Exercise 1.13. What odds value corresponds to the probability $\pi = 0.2$, and what is the numerical difference between these two values?

Solution

Solution 1.14. The odds value corresponding to $\pi = 0.2$ is:

$$
\omega = \frac{\pi}{1-\pi} = \frac{.2}{.8} = .25
$$

The numerical difference between the odds and the probability is:

$$\omega - \pi = .25 - .2 = .05$$

Exercise 1.14. Find the difference between an odds ω and its corresponding probability π, as a function of π.

Solution

Solution 1.15.

$$\begin{aligned}
\omega - \pi &= \frac{\pi}{1-\pi} - \pi \\
&= \frac{\pi}{1-\pi} - \frac{\pi(1-\pi)}{1-\pi} \\
&= \frac{\pi}{1-\pi} - \frac{\pi-\pi^2}{1-\pi} \\
&= \frac{\pi-(\pi-\pi^2)}{1-\pi} \\
&= \frac{\pi-\pi+\pi^2}{1-\pi} \\
&= \frac{\pi^2}{1-\pi} \\
&= \frac{\pi}{1-\pi}\pi \\
&= \omega\pi
\end{aligned}$$

Theorem 1.5 (Difference between odds and probability)**.** *Let* $\omega = \frac{\pi}{1-\pi}$*. Then:*

$$\omega - \pi = \frac{\pi^2}{1-\pi}$$

ℹ Proof

Proof. By Solution **??**. □

For rare events (small π), odds and probabilities are nearly equal (see Figure 1), because $1 - \pi \approx 1$ and $\pi^2 \approx 0$.

For example, in Example 1.8, the probability and odds differ by 6.777622×10^{-6}.

1.3.2 The inverse odds function

Exercise 1.15. If π is the probability of an event A and ω is the corresponding odds of A, how can we compute π from ω?
For example, if $\omega = 3/2$, what is π?

Solution

Solution 1.16. Starting from Theorem 1.2, we can solve Equation 6 for π in terms of ω:

$$
\begin{aligned}
\omega &= \frac{\pi}{1-\pi} \\
(1-\pi)\omega &= \pi \\
\omega - \pi\omega &= \pi \\
\omega &= \pi + \pi\omega \\
\omega &= (1+\omega)\pi \\
\pi &= \frac{\omega}{1+\omega}
\end{aligned}
$$

So if $\omega = 3/2$,

$$
\begin{aligned}
\pi &= \frac{3/2}{1+3/2} \\
&= \frac{3/2}{5/2} \\
&= \frac{3}{5}
\end{aligned}
$$

Theorem 1.6 (Probability as a function of odds)**.** *If π is the probability of an event and ω is the corresponding odds of that event, then:*

$$\pi = \frac{\omega}{1+\omega} \tag{11}$$

Proof

Proof. By Theorem 1.2 and Solution 1.16. □

Definition 1.11 (inverse odds function)**.**

$$\text{invodds}\{\omega\} \stackrel{\text{def}}{=} \frac{\omega}{1+\omega} \tag{12}$$

can be called the **inverse-odds function**.

Corollary 1.3 (Probability via the inverse-odds function)**.**

$$\pi = \text{invodds}\{\omega\}$$

Proof

Proof. By Definition 1.11 and Theorem 1.6. □

Corollary 1.4 (Inverse-odds function inverts the odds function)**.**

$$\text{invodds}\{\omega\} = \text{odds}^{-1}\{\omega\}$$

Proof

Proof. Using Corollary 1.1 and Theorem 1.6:

$$
\begin{aligned}
\text{invodds}\{\text{odds}\{\pi\}\} &= \text{invodds}\{\omega\} \\
&= \frac{\omega}{1+\omega} \\
&= \pi
\end{aligned}
$$

Likewise (not shown):

$$\text{odds}\{\text{invodds}\{\omega\}\} = \omega$$

□

The inverse-odds function converts odds into their corresponding probabilities (Figure 2). Its domain of inputs is $\omega \in [0, \infty)$ and its range of outputs is $\pi \in [0, 1]$.

I haven't seen anyone give the inverse-odds function a concise name; maybe prob() or prob() or risk()?

```r
odds_inv <- function(omega) (1 + omega^-1)^-1
library(ggplot2)
ggplot() +
  geom_function(fun = odds_inv, aes(col = "inverse-odds")) +
  xlab("Odds") +
  ylab("Probability") +
  xlim(0, 5) +
  ylim(0, 1) +
  geom_abline(aes(intercept = 0, slope = 1, col = "x=y"))
```

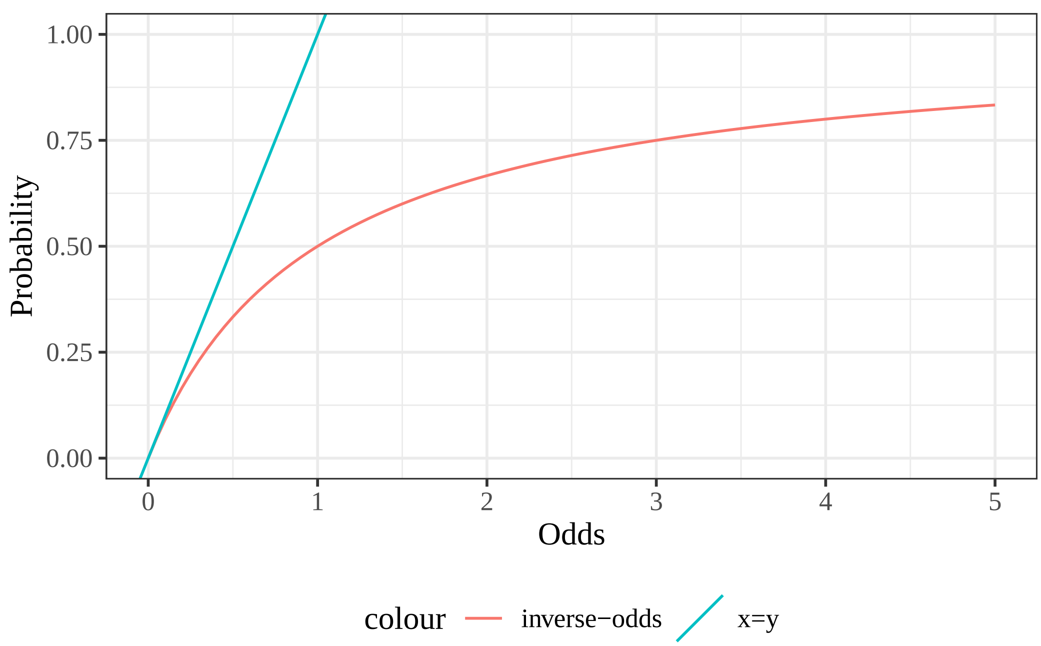

Figure 2: The inverse odds function, $\text{invodds}\{\omega\}$

Exercise 1.16. What probability corresponds to an odds of $\omega = 1$, and what is the numerical difference between these two values?

Solution

Solution.

$$
\begin{aligned}
\pi &= \text{invodds}\{1\} \\
&= \frac{1}{1+1} \\
&= \frac{1}{2} \\
&= .5
\end{aligned}
$$

$$
\begin{aligned}
\omega - \pi &= 1 - .5 \\
&= .5
\end{aligned}
$$

Lemma 1.1 (Simplified expression for inverse odds function)**.**
Equivalent expressions for the inverse odds function are:

$$
\begin{aligned}
\text{invodds}\{\omega\} &= \frac{1}{1+\omega^{-1}} \\
&= (1+\omega^{-1})^{-1}
\end{aligned} \tag{13}
$$

Exercise 1.17. Prove that Equation 13 is equivalent to Definition 1.11.

Solution

Solution 1.17. Analogous to Solution 1.13.

Lemma 1.2 (One minus inverse-odds)**.**

$$1 - \pi = \frac{1}{1+\omega}$$

Proof

Proof. By Theorem 1.6:

$$
\begin{aligned}
1 - \pi &= 1 - \frac{\omega}{1+\omega} \\
&= \frac{1+\omega}{1+\omega} - \frac{\omega}{1+\omega} \\
&= \frac{(1+\omega) - \omega}{1+\omega} \\
&= \frac{1+\omega-\omega}{1+\omega} \\
&= \frac{1}{1+\omega}
\end{aligned}
$$

□

Corollary 1.5 (One plus odds in terms of non-event probability)**.**

$$1 + \omega = \frac{1}{1-\pi}$$

1.3.3 Odds ratios

Now that we have defined odds, we can introduce another way of comparing event probabilities: odds ratios.

Definition 1.12 (Odds ratio)**.** The **odds ratio** for two conditional odds, ω_1 and ω_2, is the ratio of those odds:

$$\theta(\omega_1, \omega_2) \stackrel{\text{def}}{=} \frac{\omega_1}{\omega_2}$$

There's a 1:1 mapping between probability and odds, and according to that mapping, the odds are equal between two covariate patterns IF and ONLY IF the probabilities are also equal between those patterns.

An *odds ratio* is a *ratio of odds.* An odds is a ratio of probabilities, so odds ratios are ratios of ratios:

Theorem 1.7 (Odds ratio as a ratio of ratios)**.**

$$\begin{aligned}\theta(\omega_1, \omega_2) &= \frac{\omega_1}{\omega_2} \\ &= \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)}\end{aligned}$$

Proof

Proof.

$$\begin{aligned}\theta(\omega_1, \omega_2) &\stackrel{\text{def}}{=} \frac{\omega_1}{\omega_2} && \text{(@def-OR)} \\ &= \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)} && \text{(@thm-prob-to-odds, applied to each odds)}\end{aligned}$$

□

Exm

Example 1.10 (The OC-MI odds ratio as a ratio of ratios)**.** Using the MI risk estimates from Solution 1.1, $\hat{\pi}(OC) = 13/5000$ and $\hat{\pi}(\neg OC) = 7/10000$, Theorem 1.7 expresses the odds ratio as a ratio of the two odds, each of which is itself a ratio of probabilities:

$$\begin{aligned}\theta(\omega(OC), \omega(\neg OC)) &\stackrel{\text{def}}{=} \frac{\omega(OC)}{\omega(\neg OC)} && \text{(@def-OR)} \\ &= \frac{\left(\frac{\pi(OC)}{1-\pi(OC)}\right)}{\left(\frac{\pi(\neg OC)}{1-\pi(\neg OC)}\right)} && \text{(@thm-prob-to-odds, applied to each odds)} \\ &= \frac{\left(\frac{13/5000}{1-13/5000}\right)}{\left(\frac{7/10000}{1-7/10000}\right)} && \text{(substituting the risk estimates)} \\ &= \frac{0.002607}{7.004903 \times 10^{-4}} && \text{(computing each odds)} \\ &= 3.721361 && \text{(dividing)}\end{aligned}$$

Exm

Example 1.11 (Calculating odds ratios)**.** In Example 1.1, the odds ratio for OC users versus OC-non-users is:

$$\begin{aligned}\theta(\omega(OC), \omega(\neg OC)) &= \frac{\omega(OC)}{\omega(\neg OC)} \\ &= \frac{0.0026}{7 \times 10^{-4}} \\ &= 3.714286\end{aligned}$$

A shortcut for calculating odds ratio estimates

The general form of a two-by-two table is shown in Table 2.

Table 2: A generic 2x2 table

	Event	Non-Event	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

From this table, we have:

- $\hat{\pi}(Event|Exposed) = a/(a+b)$
- $\hat{\pi}(\neg Event|Exposed) = b/(a+b)$
- $\hat{\omega}(Event|Exposed) = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}$
- $\hat{\omega}(Event|\neg Exposed) = \frac{c}{d}$ (see Exercise 1.18)
- $\theta_{\omega}(Exposed, \neg Exposed) = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$

Exercise 1.18. Given Table 2, show that $\hat{\omega}(Event|\neg Exposed) = \frac{c}{d}$.

Properties of odds ratios

Odds ratios have a special property: we can swap a covariate with the outcome, and the odds ratio remains the same.

Theorem 1.8 (Odds ratios are reversible)**.** *For any two events A, B:*

$$\theta_{\omega}(A|B) = \theta_{\omega}(B|A)$$

ℹ Proof

Proof. Using Definition 1.12, Definition 1.8, the law of conditional probability[a], and the definition of conditional probability[b]:

$$
\begin{aligned}
\theta_\omega(A|B) &\stackrel{\text{def}}{=} \frac{\omega(A|B)}{\omega(A|\neg B)} && \text{(definition of odds ratio)}\\
&= \frac{\left(\frac{\mathrm{p}(A|B)}{\mathrm{p}(\neg A|B)}\right)}{\left(\frac{\mathrm{p}(A|\neg B)}{\mathrm{p}(\neg A|\neg B)}\right)} && \text{(definition of conditional odds)}\\
&= \left(\frac{\mathrm{p}(A|B)}{\mathrm{p}(\neg A|B)}\right)\left(\frac{\mathrm{p}(A|\neg B)}{\mathrm{p}(\neg A|\neg B)}\right)^{-1} && \text{(division is multiplication by the reciprocal)}\\
&= \left(\frac{\mathrm{p}(A|B)}{\mathrm{p}(\neg A|B)}\right)\left(\frac{\mathrm{p}(\neg A|\neg B)}{\mathrm{p}(A|\neg B)}\right) && \text{(the reciprocal of a fraction flips it)}\\
&= \left(\frac{\mathrm{p}(A|B)}{\mathrm{p}(\neg A|B)}\cdot\frac{\mathrm{p}(B)}{\mathrm{p}(B)}\right)\left(\frac{\mathrm{p}(\neg A|\neg B)}{\mathrm{p}(A|\neg B)}\cdot\frac{\mathrm{p}(\neg B)}{\mathrm{p}(\neg B)}\right) && \text{(multiply each factor by 1)}\\
&= \left(\frac{\mathrm{p}(A,B)}{\mathrm{p}(\neg A,B)}\right)\left(\frac{\mathrm{p}(\neg A,\neg B)}{\mathrm{p}(A,\neg B)}\right) && \text{(law of conditional probability, in each of the four terms)}\\
&= \left(\frac{\mathrm{p}(B,A)}{\mathrm{p}(B,\neg A)}\right)\left(\frac{\mathrm{p}(\neg B,\neg A)}{\mathrm{p}(\neg B,A)}\right) && \text{(joint probabilities are symmetric in their arguments)}\\
&= \left(\frac{\mathrm{p}(B,A)}{\mathrm{p}(\neg B,A)}\right)\left(\frac{\mathrm{p}(\neg B,\neg A)}{\mathrm{p}(B,\neg A)}\right) && \text{(swap the two highlighted denominators)}\\
&= \left(\frac{\mathrm{p}(B,A)/\,\mathrm{p}(A)}{\mathrm{p}(\neg B,A)/\,\mathrm{p}(A)}\right)\left(\frac{\mathrm{p}(\neg B,\neg A)/\,\mathrm{p}(\neg A)}{\mathrm{p}(B,\neg A)/\,\mathrm{p}(\neg A)}\right) && \text{(divide the top and bottom of each factor by the same quant}\\
&= \left(\frac{\mathrm{p}(B|A)}{\mathrm{p}(\neg B|A)}\right)\left(\frac{\mathrm{p}(\neg B|\neg A)}{\mathrm{p}(B|\neg A)}\right) && \text{(definition of conditional probability, in each of the four term}\\
&= \left(\frac{\mathrm{p}(B|A)}{\mathrm{p}(\neg B|A)}\right)\left(\frac{\mathrm{p}(B|\neg A)}{\mathrm{p}(\neg B|\neg A)}\right)^{-1} && \text{(a flipped fraction is the reciprocal of the original)}\\
&= \frac{\left(\frac{\mathrm{p}(B|A)}{\mathrm{p}(\neg B|A)}\right)}{\left(\frac{\mathrm{p}(B|\neg A)}{\mathrm{p}(\neg B|\neg A)}\right)} && \text{(multiplication by the reciprocal is division)}\\
&= \frac{\omega(B|A)}{\omega(B|\neg A)} && \text{(definition of conditional odds)}\\
&\stackrel{\text{def}}{=} \theta_\omega(B|A) && \text{(definition of odds ratio)}
\end{aligned}
$$

□

[a]probability.qmd#thm-law-conditional-prob
[b]probability.qmd#def-conditional-prob

Exm

Example 1.12 (Reversibility of the OC-MI odds ratio)**.** Theorem 1.8 says $\theta_\omega(MI|OC) = \theta_\omega(OC|MI)$: we can swap the roles of the outcome (MI) and the covariate (OC use) without changing the odds ratio. We can confirm this equality numerically using the counts from Example 1.1.

The **disease** odds ratio (the odds of MI, comparing OC users to non-users) is:

$$
\begin{aligned}
\hat{\theta}_\omega(MI|OC) &\stackrel{\text{def}}{=} \frac{\hat{\omega}(MI|OC)}{\hat{\omega}(MI|\neg OC)} && \text{(@def-OR, applied to the estimated odds)}\\
&= \frac{13/4987}{7/9993} && \text{(@thm-est-odds: estimated odds = events/non-events; counts from @exm-oc-mi)}\\
&= \frac{0.002607}{7.004903\times 10^{-4}} && \text{(dividing within each OC-use group)}\\
&= 3.721361 && \text{(dividing)}
\end{aligned}
$$

The **exposure** odds ratio (the odds of OC use, comparing MI cases to non-cases) is:

$$
\begin{aligned}
\hat{\theta}_{\omega}(OC|MI) &\stackrel{\text{def}}{=} \frac{\hat{\omega}(OC|MI)}{\hat{\omega}(OC|\neg MI)} && \text{(@def-OR, applied to the estimated odds)} \\
&= \frac{13/7}{4987/9993} && \text{(@thm-est-odds: estimated odds = events/non-events; counts from @exm-oc-mi)} \\
&= \frac{1.857143}{0.499049} && \text{(dividing within each MI-status group)} \\
&= 3.721361 && \text{(dividing)}
\end{aligned}
$$

Both equal 3.721361, confirming Theorem 1.8.

Exercise 1.19. For Table 2, show that $\hat{\theta}_{\omega}(Exposed, Unexposed) = \hat{\theta}_{\omega}(Event, \neg Event)$.

Conditional odds ratios have the same reversibility property:

Theorem 1.9 (Conditional odds ratios are reversible)**.** *For any three events A, B, C:*

$$\theta_{\omega}(A|B,C) = \theta_{\omega}(B|A,C)$$

Proof

Proof. Apply the same steps as for Theorem 1.8, inserting C into the conditions (RHS of $|$) of every expression. □

Exm

Example 1.13 (Reversibility within age strata in the WCGS data)**.** Theorem 1.9 says a conditional odds ratio is also reversible: $\theta_{\omega}(A|B,C) = \theta_{\omega}(B|A,C)$. We can illustrate this equality using data from the Western Collaborative Group Study (`wcgs`) (Rosenman et al. 1975), a prospective cohort study of coronary heart disease (CHD) in 3,154 men aged 39–59, which classified each participant's behavioral pattern as Type A (urgent, competitive) or Type B (relaxed). For this example, we take the covariate B to be behavioral pattern (Type A versus Type B), the outcome A to be CHD by 1969, and the conditioning event C to be an age stratum.
Among participants under 50, the disease odds ratio (the odds of CHD, comparing Type A to Type B) is 2.212757, and the exposure odds ratio (the odds of Type A, comparing CHD cases to non-cases) is 2.212757; the two agree, as Theorem 1.9 requires.
Among participants aged 50 or older, the disease and exposure odds ratios are 2.30791 and 2.30791, which likewise agree.

Odds Ratios vs Probability (Risk) Ratios

When the outcome is rare (i.e., its probability is small) for both groups being compared in an odds ratio, the odds of the outcome will be similar to the probability of the outcome, and thus the risk ratio will be similar to the odds ratio.

Case 1: rare events

For rare events, odds ratios and probability (a.k.a. risk, a.k.a. prevalence) ratios will be close:

$$\pi_1 = .01$$

$$\pi_2 = .02$$

```r
pi1 <- .01
pi2 <- .02
pi2 / pi1
#> [1] 2
odds(pi2) / odds(pi1)
#> [1] 2.02041
```

Exm

Example 1.14. In Example 1.1, the outcome is rare for both OC and non-OC participants, so the odds for both groups are similar to the corresponding probabilities, and the odds ratio is similar the risk ratio.

Case 2: frequent events

$$\pi_1 = .4$$

$$\pi_2 = .5$$

For more frequently-occurring outcomes, this won't be the case:

```r
pi1 <- .4
pi2 <- .5
pi2 / pi1
#> [1] 1.25
odds(pi2) / odds(pi1)
#> [1] 1.5
```

The relationship between the two spot-checks is captured by a simple algebraic identity:

Theorem 1.10 (Odds ratio as risk ratio times a correction factor)**.** *For two probabilities π_1 and π_2, the odds ratio equals the risk ratio times a correction factor:*

$$\theta(\omega_1, \omega_2) = \rho(\pi_1, \pi_2) \cdot \frac{1 - \pi_2}{1 - \pi_1} \tag{14}$$

When π_1 and π_2 are both small, the correction factor $\frac{1-\pi_2}{1-\pi_1}$ is close to 1, so the odds ratio is close to the risk ratio.

Proof

Proof.

$$\begin{aligned}
\theta(\omega_1, \omega_2) &\overset{\text{def}}{=} \frac{\omega_1}{\omega_2} && \text{(@def-OR)} \\
&= \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)} && \text{(@thm-prob-to-odds, applied to each odds)} \\
&= \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} && \text{(rearranging the compound fraction)} \\
&= \rho(\pi_1, \pi_2) \cdot \frac{1-\pi_2}{1-\pi_1} && \text{(@def-RR)}
\end{aligned}$$

□

Exm

Example 1.15 (The correction factor in the two spot-checks)**.** Theorem 1.10 explains the two spot-checks preceding it.
In **Case 1** (rare events), take $\pi_1 = 0.02$ and $\pi_2 = 0.01$. The correction factor $\frac{1-\pi_2}{1-\pi_1} = 1.010204$ is close to 1, so the odds ratio (2.020408) is close to the risk ratio (2).
In **Case 2** (frequent events), take $\pi_1 = 0.5$ and $\pi_2 = 0.4$. The correction factor $\frac{1-\pi_2}{1-\pi_1} = 1.2$ is farther from 1, so the odds ratio (1.5) and risk ratio (1.25) diverge.

Figure 3 compares risk differences, risk ratios, and odds ratios as functions of the underlying probabilities being compared.

```
if (run_graphs) {
  RD <- function(p1, p2) p2 - p1
  RR <- function(p1, p2) p2 / p1
  odds <- function(p) p / (1 - p)
  OR <- function(p1, p2) odds(p2) / odds(p1)
  OR_minus_RR <- function(p1, p2) OR(p2, p1) - RR(p2, p1)

  n_ticks <- 201
  probs <- seq(.001, .99, length.out = n_ticks)
  RD_mat <- outer(probs, probs, RD)
  RR_mat <- outer(probs, probs, RR)
  OR_mat <- outer(probs, probs, OR)

  opacity <- .3
  z_min <- -1
  z_max <- 5
  plotly::plot_ly(
    x = ~probs,
    y = ~probs
  ) |>
    plotly::add_surface(
      z = ~ t(RD_mat),
      contours = list(
        z = list(
          show = TRUE,
          start = -1,
          end = 1,
          size = .1
        )
      ),
      name = "Risk Difference",
      showscale = FALSE,
      opacity = opacity,
      colorscale = list(c(0, 1), c("green", "green"))
    ) |>
    plotly::add_surface(
      opacity = opacity,
      colorscale = list(c(0, 1), c("red", "red")),
      z = ~ t(RR_mat),
      contours = list(
        z = list(
          show = TRUE,
          start = z_min,
          end = z_max,
          size = .2
        )
      ),
      showscale = FALSE,
      name = "Risk Ratio"
    ) |>
    plotly::add_surface(
      opacity = opacity,
      colorscale = list(c(0, 1), c("blue", "blue")),
      z = ~ t(OR_mat),
      contours = list(
        z = list(
          show = TRUE,
          start = z_min,
          end = z_max,
          size = .2
        )
      ),
      showscale = FALSE
```

Testing whether an odds ratio = 1 is equivalent to testing whether the corresponding risk ratio = 1, and also equivalent to testing whether the risk difference = 0. Therefore, in **hypothesis testing**, if the null hypothesis is no effect, then we can shift between RD, RR, and OR. But when we're talking about **point estimates** and **CIs**, we need to limit our conclusions to the effect measure(s) that we actually estimated, because the *sizes* of RDs, RRs, and ORs don't have a simple relationship to each other, except when $\pi_1 = \pi_2$ (as shown by Figure 3).

Odds Ratios in Case-Control Studies

Table 1 simulates a follow-up study in which two populations were followed and the number of MI's was observed. The risks are $P(MI|OC)$ and $P(MI|\neg OC)$ and we can estimate these risks from the data.

But suppose we had a case-control study in which we had 100 women with MI and selected a comparison group of 100 women without MI (matched as groups on age, etc.). Then MI is not random, and we cannot compute P(MI|OC) and we cannot compute the risk ratio. However, the odds ratio can be computed.

Definition 1.13 (Disease odds ratio)**.** For a disease D and an exposure E, the **disease odds ratio** is the odds of disease in the exposed group divided by the odds of disease in the unexposed group:

$$\theta_\omega(D|E) \stackrel{\text{def}}{=} \frac{\omega(D|E)}{\omega(D|\neg E)}$$

In a case-control study, disease status is not random, so we cannot validly compute or use the separate parts $\omega(D|E)$ and $\omega(D|\neg E)$ of the disease odds ratio.

Definition 1.14 (Exposure odds ratio)**.** For a disease D and an exposure E, the **exposure odds ratio** is the odds of exposure in the diseased group divided by the odds of exposure in the non-diseased group:

$$\theta_\omega(E|D) \stackrel{\text{def}}{=} \frac{\omega(E|D)}{\omega(E|\neg D)}$$

We can still validly compute and use the exposure odds ratio, because exposure can be treated as random.

Exm

Example 1.16 (Exposure odds ratio in the OC-MI study)**.** For the OC-MI study, the estimated exposure odds ratio is:

$$\hat{\theta}_\omega(OC|MI) = \frac{\hat{\omega}(OC|MI)}{\hat{\omega}(OC|\neg MI)}$$

And these two odds ratios, $\hat{\theta}_\omega(MI|OC)$ and $\hat{\theta}_\omega(OC|MI)$, are mathematically equivalent, as we saw in Section 1.3.3:

$$\hat{\theta}_\omega(MI|OC) = \hat{\theta}_\omega(OC|MI)$$

Exercise 1.20. Calculate the odds ratio of MI with respect to OC use, assuming that Table 1 comes from a case-control study. Confirm that the result is the same as in Example 1.11.

Solution

Solution.

```
tbl_oc_mi |> pander::pander()
```

Table 3: Simulated data from study of oral contraceptive use and heart attack risk

OC	MI	No MI	Total
OC use	13	4,987	5,000
No OC use	7	9,993	10,000
Total	20	14,980	15,000

- $\omega(OC|MI) = P(OC|MI)/(1 - P(OC|MI)) = \frac{13}{7} = 1.857143$
- $\omega(OC|\neg MI) = P(OC|\neg MI)/(1 - P(OC|\neg MI)) = \frac{4987}{9993} = 0.499049$
- $\theta_\omega(OC|MI) = \frac{\omega(OC|MI)}{\omega(OC|\neg MI)} = \frac{13/7}{4987/9993} = 3.721361$

This odds-ratio estimate is the same one we calculated in Example 1.11.

Odds Ratios in Cross-Sectional Studies

- If a cross-sectional study is a uniform probability sample of a population (which it rarely is), then we can estimate prevalence (sometimes called "prevalence risk" or just "risk") using standard methods (Lee 1994), and we can thus also estimate prevalence differences, prevalence ratios, and prevalence odds ratios comparing sub-populations.
- If the cross-sectional study is a stratified probability sample, then we can estimate prevalence, prevalence differences, prevalence ratios, and prevalence odds ratios using specialized methods for complex surveys (Lumley 2010).
- If the study has sampling biases that we cannot adjust for with survey weights, such as in a convenience sample, then we need to treat it in the same way as a case-control study, and we cannot validly estimate prevalence, prevalence differences, or prevalence ratios; we can only validly estimate prevalence *odds* ratios.

1.4 The logit and expit functions

1.4.1 The logit function

Definition 1.15 (log-odds)**.**
If ω is the odds of an event A, then the **log-odds** of A, which we will represent by η ("eta"), is the natural logarithm of the odds of A:

$$\eta \stackrel{\text{def}}{=} \log\{\omega\} \tag{15}$$

Theorem 1.11 (Log-odds as a function of probability)**.** *If π is the probability of an event A, ω is the corresponding odds of A, and η is the corresponding log-odds of A, then:*

$$\eta = \log\left\{\frac{\pi}{1-\pi}\right\} \tag{16}$$

Proof

Proof. Using Definition 1.15 and Theorem 1.2:

$$
\begin{aligned}
\eta &\stackrel{\text{def}}{=} \log\{\omega\} && \text{(definition of log-odds)} \\
&= \log\left\{\frac{\pi}{1-\pi}\right\} && \text{(odds as a function of probability)}
\end{aligned}
$$

□

Definition 1.16 (logit function)**.**
The **logit function** of a probability π is the natural logarithm of the odds function of π:

$$\text{logit}(\pi) \stackrel{\text{def}}{=} \log\{\text{odds}\{\pi\}\}$$

The **logit function** is a composite function[a].

[a] https://en.wikipedia.org/wiki/Function_composition

Exercise 1.21 (Compose the logit function)**.** Mathematically expand the definition of the logit function.

Solution (Compose the logit function)

Solution 1.18 (Compose the logit function).

Lem

Theorem 1.12 (Expanded expression for logit).

$$\text{logit}(\pi) = \log\left\{\frac{\pi}{1-\pi}\right\} \tag{17}$$

Proof

Proof. Using Definition 1.16 and Definition 1.10:

$$
\begin{aligned}
\text{logit}(\pi) &\stackrel{\text{def}}{=} \log\{\text{odds}\{\pi\}\} && \text{(definition of logit)} \\
&= \log\left\{\frac{\pi}{1-\pi}\right\} && \text{(definition of the odds function)}
\end{aligned}
$$

□

Corollary 1.6 (Log-odds via the logit function)**.** *If π is the probability of an event A and η is the corresponding log-odds of A, then:*

$$\eta = \text{logit}\{\pi\}$$

ℹ Proof

Proof. Using Theorem 1.11 and Theorem 1.12:

$$
\begin{aligned}
\eta &= \log\left\{\frac{\pi}{1-\pi}\right\} && \text{(log-odds as a function of probability)} \\
&= \text{logit}\{\pi\} && \text{(expanded expression for logit, read right-to-left)}
\end{aligned}
$$

□

Exm

Example 1.17 (Log-odds of an event with probability 0.75)**.** Suppose the probability of an event A is $\pi = 0.75$. Then, by Theorem 1.2, the corresponding odds of A are $\omega = \frac{0.75}{1-0.75} =$

$\frac{0.75}{0.25} = 3$, and by Corollary 1.6 and Theorem 1.12, the corresponding log-odds of A are:

$$\begin{aligned}
\eta = \text{logit}\{0.75\} &= \log\left\{\frac{0.75}{1-0.75}\right\} && \text{(expanded expression for logit)} \\
&= \log\left\{\frac{0.75}{0.25}\right\} && \text{(subtract: } 1-0.75=0.25\text{)} \\
&= \log\{3\} && \text{(divide: } 0.75/0.25=3\text{)} \\
&\approx 1.0986 && \text{(evaluate the natural logarithm)}
\end{aligned}$$

Figure 4 shows the shape of the logit() function.

```r
odds <- function(pi) pi / (1 - pi)

logit <- function(p) log(odds(p))

library(ggplot2)
logit_plot <-
  ggplot() +
  geom_function(
    fun = logit,
    arrow = arrow(ends = "both")
  ) +
  xlim(.001, .999) +
  ylab("logit(p)") +
  xlab("p") +
  theme_bw()
print(logit_plot)
```

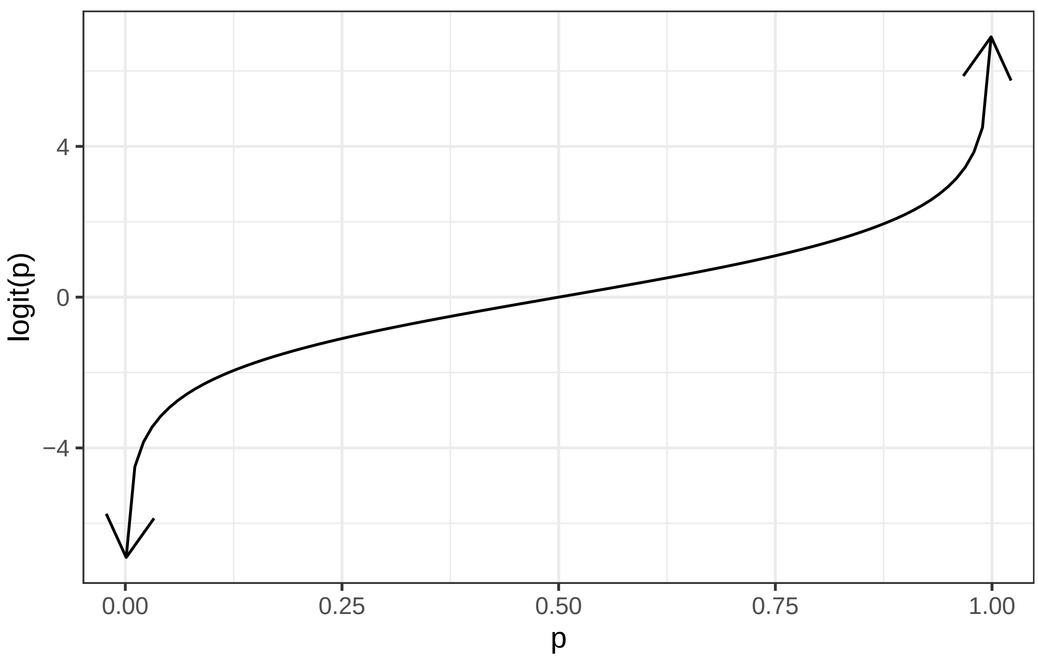

Figure 4: The logit function

1.4.2 The expit function

Lemma 1.3 (Odds from log-odds)**.**
If ω is the odds of an event A and η is the corresponding log-odds of A, then:

$$\omega = \exp\{\eta\}$$

Proof

Proof. Starting from Definition 1.15, we solve for ω:

$$\begin{aligned}
\eta &\stackrel{\text{def}}{=} \log\{\omega\} && \text{(definition of log-odds)}\\
\exp\{\eta\} &= \exp\{\log\{\omega\}\} && \text{(exponentiate both sides)}\\
\exp\{\eta\} &= \omega && \text{(the exponential function inverts the natural logarithm)}\\
\omega &= \exp\{\eta\} && \text{(swap the two sides of the equation)}
\end{aligned}$$

□

Theorem 1.13 (Probability as a function of log-odds)**.**
If π is the probability of an event A, ω is the corresponding odds of A, and η is the corresponding log-odds of A, then:

$$\pi = \frac{\exp\{\eta\}}{1 + \exp\{\eta\}}$$

Proof

Proof. Using Theorem 1.6 and Lemma 1.3:

$$\begin{aligned}
\pi &= \frac{\omega}{1+\omega} && \text{(probability as a function of odds)}\\
&= \frac{\exp\{\eta\}}{1 + \exp\{\eta\}} && \text{(substitute } \omega = \exp\{\eta\} \text{ in both places)}
\end{aligned}$$

□

Definition 1.17 (expit, logistic, inverse-logit)**.** The **expit function** of a log-odds η, also known as the **inverse-logit function** or **logistic function**, is the inverse-odds of the exponential of η:

$$\text{expit}(\eta) \stackrel{\text{def}}{=} \text{invodds}\{\exp\{\eta\}\}$$

Theorem 1.14 (Expressions for expit function)**.**

$$\begin{aligned}
\text{expit}(\eta) &= \frac{\exp\{\eta\}}{1 + \exp\{\eta\}}\\
&= \frac{1}{1 + \exp\{-\eta\}}\\
&= (1 + \exp\{-\eta\})^{-1}
\end{aligned}$$

Proof

Proof. For the first expression, apply Definition 1.17 and then Definition 1.11:

$$\begin{aligned}
\text{expit}(\eta) &\stackrel{\text{def}}{=} \text{invodds}\{\exp\{\eta\}\} && \text{(definition of expit)} \\
&= \frac{\exp\{\eta\}}{1+\exp\{\eta\}} && \text{(definition of inverse-odds, applied to } \omega = \exp\{\eta\}\text{)}
\end{aligned}$$

For the second and third expressions, apply Definition 1.17 and then Lemma 1.1:

$$\begin{aligned}
\text{expit}(\eta) &\stackrel{\text{def}}{=} \text{invodds}\{\exp\{\eta\}\} && \text{(definition of expit)} \\
&= \frac{1}{1+(\exp\{\eta\})^{-1}} && \text{(simplified inverse-odds, applied to } \omega = \exp\{\eta\}\text{)} \\
&= \frac{1}{1+\exp\{-\eta\}} && \text{(negative-exponent rule: } (\exp\{x\})^{-1} = \exp\{-x\}\text{)} \\
&= (1+\exp\{-\eta\})^{-1} && \text{(rewrite the reciprocal as a } -1 \text{ power)}
\end{aligned}$$

□

Theorem 1.15 (Probability via the expit function)**.** *If π is the probability of an event A, ω is the corresponding odds of A, and η is the corresponding log-odds of A, then:*

$$\pi = \text{expit}\{\eta\}$$

Proof

Proof. Using Theorem 1.13 and Theorem 1.14:

$$\begin{aligned}
\pi &= \frac{\exp\{\eta\}}{1+\exp\{\eta\}} && \text{(probability as a function of log-odds)} \\
&= \text{expit}\{\eta\} && \text{(first expression for expit, read right-to-left)}
\end{aligned}$$

□

Exm

Example 1.18 (Probability of an event with log-odds $\log\{3\}$)**.** In Example 1.17, a probability of $\pi = 0.75$ corresponded to a log-odds of $\eta = \log\{3\} \approx 1.0986$. By Theorem 1.15 and Theorem 1.14, the expit function recovers the probability from that log-odds:

$$\begin{aligned}
\pi = \text{expit}\{\log\{3\}\} &= \frac{\exp\{\log\{3\}\}}{1+\exp\{\log\{3\}\}} && \text{(first expression for expit)} \\
&= \frac{3}{1+3} && \text{(the exponential inverts the natural logarithm: } \exp\{\log\{3\}\} = 3\text{)} \\
&= \frac{3}{4} && \text{(add: } 1+3=4\text{)} \\
&= 0.75 && \text{(divide: } 3/4 = 0.75\text{)}
\end{aligned}$$

Figure 5 graphs the expit function.

```r
expit <- function(eta) {
  exp(eta) / (1 + exp(eta))
}
library(ggplot2)
expit_plot <-
  ggplot() +
  geom_function(
    fun = expit,
    arrow = arrow(ends = "both")
  ) +
  xlim(-8, 8) +
  ylim(0, 1) +
  ylab(expression(expit(eta))) +
  xlab(expression(eta)) +
  theme_bw()
print(expit_plot)
```

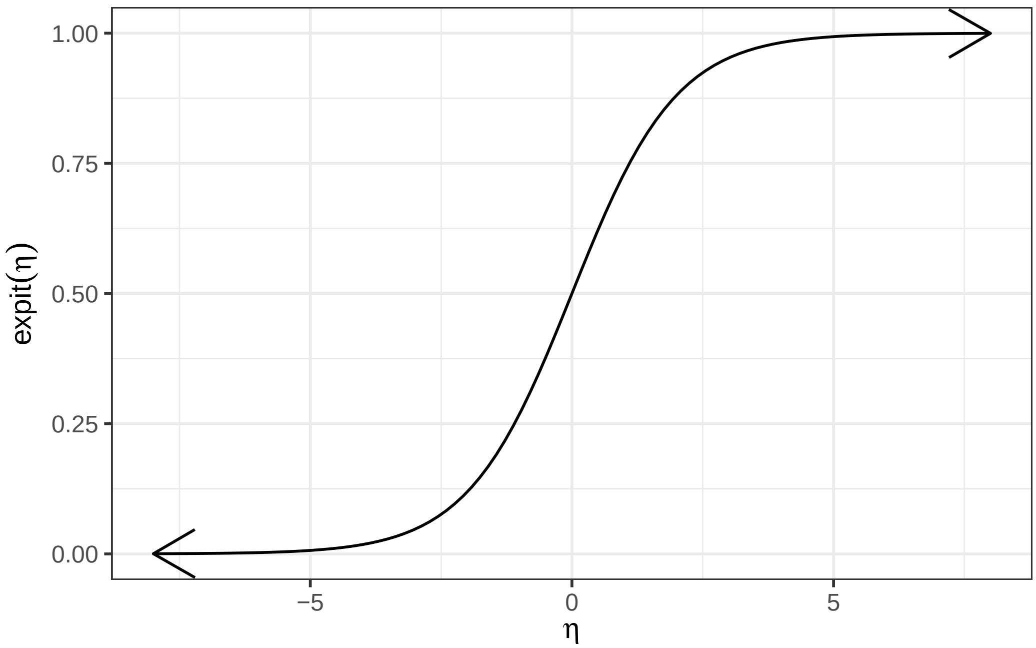

Figure 5: The expit function

Theorem 1.16 (logit and expit are each others' inverses)**.**

$$\text{logit}\{\text{expit}\{\eta\}\} = \eta$$

$$\text{expit}\{\text{logit}\{\pi\}\} = \pi$$

Proof

Proof. First, we show that $\text{logit}\{\text{expit}\{\eta\}\} = \eta$, using Definition 1.16, Definition 1.17, and Corollary 1.4:

$$
\begin{aligned}
\text{logit}\{\text{expit}\{\eta\}\} &\stackrel{\text{def}}{=} \log\{\text{odds}\{\text{expit}\{\eta\}\}\} && \text{(definition of logit, applied to } \pi = \text{expit}\{\eta\}\text{)}\\
&= \log\{\text{odds}\{\text{invodds}\{\exp\{\eta\}\}\}\} && \text{(definition of expit)}\\
&= \log\{\exp\{\eta\}\} && \text{(the odds function inverts the inverse-odds function)}\\
&= \eta && \text{(the natural logarithm inverts the exponential function)}
\end{aligned}
$$

Next, we show that $\text{expit}\{\text{logit}\{\pi\}\} = \pi$, using Definition 1.17, Definition 1.16, and Corollary 1.4:

$$
\begin{aligned}
\text{expit}\{\text{logit}\{\pi\}\} &\stackrel{\text{def}}{=} \text{invodds}\{\exp\{\text{logit}\{\pi\}\}\} && \text{(definition of expit, applied to } \eta = \text{logit}\{\pi\}\text{)}\\
&= \text{invodds}\{\exp\{\log\{\text{odds}\{\pi\}\}\}\} && \text{(definition of logit)}\\
&= \text{invodds}\{\text{odds}\{\pi\}\} && \text{(the exponential function inverts the natural logarithm)}\\
&= \pi && \text{(the inverse-odds function inverts the odds function)}
\end{aligned}
$$

□

Exm

Example 1.19 (Numeric confirmation of the inverse relationship)**.** Combining Example 1.17 and Example 1.18 confirms Theorem 1.16 numerically: $\text{expit}\{\text{logit}\{0.75\}\} = \text{expit}\{\log\{3\}\} = 0.75$.

1.4.3 Diagram of expit and logit

$$
\left[\pi \stackrel{\text{def}}{=} \Pr\left(Y = 1|\tilde{X} = \tilde{x}\right)\right]
\underbrace{\overbrace{\underset{\frac{\omega}{1+\omega}}{\overset{\frac{\pi}{1-\pi}}{\rightleftarrows}} \left[\omega \stackrel{\text{def}}{=} \text{odds}\left\{Y = 1|\tilde{X} = \tilde{x}\right\}\right] \underset{\exp\{\eta\}}{\overset{\log\{\omega\}}{\rightleftarrows}}}^{\text{logit}(\pi)}}_{\text{expit}(\eta)}
\left[\eta(\tilde{x}) \stackrel{\text{def}}{=} \eta\left(Y = 1|\tilde{X} = \tilde{x}\right)\right]
$$

Figure 6: Diagram of logistic regression link and inverse link functions

References

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Lee, James. 1994. "Odds Ratio or Relative Risk for Cross-Sectional Data?" *International Journal of Epidemiology* (England) 23 (1): 201–3. https://doi.org/10.1093/ije/23.1.201.

Lumley, Thomas. 2010. *Complex Surveys : A Guide to Analysis Using R*. Wiley Series in Survey Methodology. John Wiley. https://doi.org/10.1002/9780470580066.

Nahhas, Ramzi W. 2024. *Introduction to Regression Methods for Public Health Using R*. CRC Press. https://www.bookdown.org/rwnahhas/RMPH/.

Roach, Rachel E. J., Frans M. Helmerhorst, Willem M. Lijfering, Theo Stijnen, Ale Algra, and Olaf M. Dekkers. 2015. "Combined Oral Contraceptives: The Risk of Myocardial Infarction and Ischemic Stroke." *Cochrane Database of Systematic Reviews*, no. 8: CD011054. https://doi.org/10.1002/14651858.CD011054.pub2.

Rosenman, Ray H, Richard J Brand, C David Jenkins, Meyer Friedman, Reuben Straus, and

Moses Wurm. 1975. "Coronary Heart Disease in the Western Collaborative Group Study: Final Follow-up Experience of 8 1/2 Years." *JAMA* 233 (8): 872–77. https://doi.org/10.1001/jama.1975.03260080034016.

Rothman, Kenneth J., Timothy L. Lash, Tyler J. VanderWeele, and Sebastien Haneuse. 2021. *Modern Epidemiology*. Fourth edition. Wolters Kluwer.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.