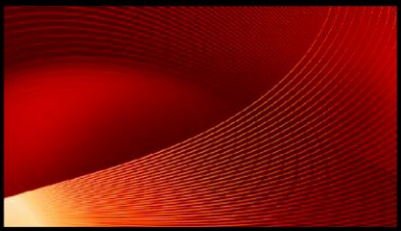
STATISTICAL TOOLS FOR EPIDEMIOLOGIC RESEARCH



STEVE SELVIN

OXFORD

Statistical Tools for Epidemiologic Research

This page intentionally left blank

Statistical Tools for Epidemiologic Research

STEVE SELVIN

OXFORD
UNIVERSITY PRESS
2011

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016
www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Selvin, S.
Statistical tools for epidemiologic research / Steve Selvin.
p. ; cm.
ISBN 978-0-19-975596-7
1. Epidemiology—Statistical methods. I. Title.
[DNLM: 1. Epidemiologic Methods. WA 950 S469sd 2010]
RA652.2.M3S453 2010
614.402′1—dc22
2010013555

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper.

for David, Liz, Ben, and, especially, Nancy

This page intentionally left blank

PREFACE

This text answers the question: After a typical first-year course in statistical methods, what next? One answer is a description of "second-year" methods that provide an introduction to intermediate biostatistical techniques without advanced mathematics or extensive statistical theory (for example, no Bayesian statistics, no causal inference, no linear algebra, and only a slight hint of calculus). Intuitive explanations richly supported with numerous examples produce an accessible presentation for readers interested in the analysis of data relevant to epidemiologic or medical research.

STATISTICAL TOOLS

A common and important statistical technique, called *smoothing*, removes the unimportant details and minimizes the influences of extraneous variation to expose the true underlying and frequently simple relationships within sampled data. This text is a kind of "smoothing." Apparently complicated statistical methods are frequently rather simple conceptually but are obscured by a forest of technical details and are often described with a specialized and sometimes confusing statistical language. The following material, however, is a blend of nonmathematical explanations, symbols, data, examples, and graphics applied to case study data. The goal is simply to explain how statistical methods work.

In most statistical texts, statistical methods are illustrated with examples. In reality, the situation is reversed. The data and questions under investigation dictate the statistical methods. A case study approach describes and illustrates statistical tools that are useful in the quest for answers to specific questions generated from sampled data. The case studies come from published research papers, doctoral dissertations, routinely reported government disease/mortality data, and clinical experimentation. The analysis of actual data leads to descriptions of statistical logic in a realistic context. For example, data collected to identify the role of lifestyle behavior in the likelihood of coronary heart disease are used to illustrate how statistical tools provide answers to important biologic/medical questions about risk and risk factors. Frequently, small subsets of the original data are analyzed for convenience of space and to allow the reader to easily repeat the calculations or explore different approaches. As part of communicating an understanding of the statistical process, minor sacrifices in rigor and notation yield important gains in accessibility to the material. For example, to avoid general notation, all confidence intervals have a level of significance of 95% because other levels are almost never used.

The text is divided into a sequence of 15 chapters starting with analysis of the fundamental 2×2 table. From this foundation, the topics gradually increase in sophistication, with particular emphasis on regression techniques (logistic, Poisson, conditional logistic, and log-linear) and then beyond to useful techniques that are not typically discussed in an applied context. Topics are chosen for two reasons. They describe widely used statistical methods particularly suitable for the analysis of epidemiologic/medical data and contribute to what might be call a "statistical tool box." These topics also illustrate basic concepts that are necessary for an understanding of data analysis in general. For example, *bootstrap estimation* is a topic that introduces an extremely useful statistical tool and, at the same time, clearly illustrates the fundamental concept of sampling variation. In addition, a concerted effort is made to trace the threads of statistical logic throughout the entire book. The same statistical issues are explored in different contexts, links among similar approaches are identified (a road map), extensions of previous methods are always noted, and the same data are occasionally analyzed from different perspectives. Thus, every attempt is made to avoid a "cookbook" kind of presentation of a series of statistical tools.

The presented material evolved from a second-year masters' degree course in epidemiologic data analysis taught over the last decade at the University of California, Berkeley, and also taught as part of the Summer Institute for Biostatistics and Epidemiology at The Johns Hopkins Bloomberg School of Public Health. Parts of the text have also been included in courses presented at the Graduate Summer Session in Epidemiology at the University of Michigan, School of Public Health. In other words, the material has been throughly "classroom tested."

The 15 chapters of this text are a confluence of two sources. Ten chapters consist of new material. The other five chapters are a substantial revision and

streamlining of the key elements from my text *Statistical Analysis of Epidemiologic Data*, third edition. They constitute a kind of a "fourth edition." The combination produces an uninterrupted description of statistical logic and methods, so that the reader does not have to refer to other sources for background. That is, the text is largely a self-contained discussion of statistical tools beyond the first-year introductory level.

COMPUTER TOOLS

In keeping with the practical rather than the theoretical nature of the text, the next question becomes: What are the computer software tools that produced the statistical analyses used to describe the case studies? In this computer-laptop-age, a text about use of statistical tools requires a parallel discussion of the computer software tools. The idea of doing statistical computations "by hand" long ago disappeared.

Although statistical software systems give the same answers, the system called *Stata* (version 10.0) is used. It is a relatively simple and widely available computer system that is certainly adequate for the statistical techniques encountered in the text and takes little investment to use effectively. Stata code (identified by dots) and the resulting output (no dots) are given for the analyses of the case studies presented and are located at www.oup.com/us/statisticaltools. The text, as might be expected, describes and explains statistical methods using only the relevant summaries and selected statistical results. The computer output contains the details. The rigors of executing computer code enhances the understanding of the analytic process and provides additional insight into the properties of statistical methods. For example, it is a simple matter to assemble the Stata commands to calculate approximate confidence interval bounds based on the normal distribution. It is also a simple matter to use a Stata command to calculate exact confidence interval bounds, easily producing an immediate idea of the accuracy of the classic approximation.

A small amount of knowledge of the Stata system is assumed. The introductory tutorial that comes with the purchase of the software (*Getting Started with Stata*) is more than sufficient. The Stata code and output are applications of the statistical tools discussed, not artificial problems. Each component of a statistical analysis system, such as the Stata system, is designed to produce a specific result. The analysis of actual data, however, is rarely an orderly process. Therefore, a Stata command frequently provides exactly the desired result. In other situations, it is necessary to construct a series of commands to get the desired result. Both situations are extensively illustrated, with the Stata statistical system applied to the case studies.

Modern computer systems are extremely user-friendly, and extensive resources exist to deal with the sometimes tedious in-and-outs of using the Stata

system. For example, three important commands for the application of the Stata system are:

1. *findit* <searches for character strings>
2. *help* <Stata command>
3. http://statcomp.ats.ucla.edu/stata/

The *findit* command searches the Stata online manual for commands associated with specific words or strings of characters. Often, the user is then able to directly access the appropriate *help* command. The *help* command then produces a detailed and complete online description of any valid Stata command, including several example applications. The third command identifies a link to a website (UCLA) that contains a large number of useful examples of the application of the Stata computer code. This code or pieces of this code frequently serve as a template for the problem at hand. The presented Stata output is slightly edited to focus directly on the analytic results. Bookkeeping messages, such as `frequency weights assumed` or `Iteration 2: log likelihood = -876.52235`, are not included.

Applications of other statistical computer systems follow similar patterns. For example, readers familiar with the major statistical systems SAS, SPSS, Splus, or R should be able to translate the Stata computer code into the system of their choice. The computer implementation is purposely made totally separate from the text material, so that no important statistical issues or methods will be missed by completely ignoring the computer input/output. The chapter topics are about understanding how the statistical tools work, and the Stata results are about how the computer tools accomplish the task.

It is important to acknowledge the debt to those who originated the statistical methods and those who taught them to me, particularly the Department of Statistics at the University of California, Berkeley. A listing of the names of these people is far to extensive to produce here. However, four mentors/colleagues stand-out from the rest, namely Elizabeth Scott, Warren Winkelstein, Nicholas Jewell, and Richard Brand.

Steve Selvin

CONTENTS

This page intentionally left blank

Statistical Tools for Epidemiologic Research

This page intentionally left blank

1

TWO MEASURES OF RISK: ODDS RATIOS AND AVERAGE RATES

Two fundamental statistical measures of association basic to exploring epidemiologic and medical data are the *odds ratio* and the *average rate*. Both statistics are easily used, but their interpretation is enhanced by an understanding of their sometimes subtle properties. This chapter describes these two measures in detail, outlines some of their important properties, and contrasts them to several other commonly used statistical measures of association.

ODDS RATIO

An odds ratio, as the name indicates, is the ratio of two values called *the odds*. The statistical origin of the odds is *a probability*. Specifically, when the probability of an event is denoted p, the odds associated with the event are $p/(1 - p)$. Odds are best understood in terms of gambling, particularly horse racing. When the probability a chosen horse will win a race is 1/3, the odds are 1 to 2 (1/3 divided by 2/3). The odds indicate the pay-off for a bet and at the same time slightly obscure the underlying probability of winning. A fair pay-off for a bet at these odds is $2 for every $1 bet. However, horse racing is not a fair game. The amount paid for a winning bet is considerably less than would be expected from a fair game. Table 1–1 is a typical racing form description of a five-horse race.

Table 1–1. Bay Meadows Handicap—One mile, four-year-olds and up, purse $9,500 (post time 12:45)

Horse	Jockey	Weight	Odds	Probability
In Love With Loot	R. Baze	119	7–2	0.22
Skip A Wink	J. Lumpkins	121	8–5	0.38
Playing'r Song	R. Gonzalez	121	9–5	0.36
Ms. Dahill	J. Ochoa	121	4–1	0.20
Rare Insight	J. Rosario	119	6–1	0.14

The last column translating the odds into probabilities is not a typical part of a racing form. The sum of these "probabilities" is 1.30. When the odds are based on a fair pay-off, these probabilities add to 1.0. However, the track has to pay the winner's purse, support the cost of the track, and make a profit. For this race and all others, the "probabilities" add to a value greater than one and identify the share of the money bet that goes directly to the track.

In statistical applications, the odds ratio (*or*) is a comparison of odds. The odds calculated under one set of conditions, $p_1/(1-p_1)$, are compared to the odds calculated under another set of conditions, $p_2/(1-p_2)$ by the ratio

$$odds\ ratio = or = \frac{odds\ under\ conditions\ 1}{odds\ under\ conditions\ 2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

The central property of an odds ratio is that when no difference exists between the conditions ($p_1 = p_2$), the odds ratio is $or = 1.0$. When $p_1 \neq p_2$, the odds ratio measures the extent of the difference between the two conditions relative to 1.0.

Although the odds ratio applies to a number of situations, a good place to start is a description of the odds ratio used to evaluate an association observed between two binary variables summarized in a 2×2 *table* (such as Tables 1–2 and 1–3).

The symbols a, b, c, and d represent the counts of the four possible outcomes from a sample made up of $a+b+c+d = m_1+m_2 = n_1+n_2 = n$ individuals. For example, in the study of a binary risk factor (denoted F) and the presence/absence of a disease (denoted D), the symbol a represents the number of individuals who have both the risk factor and the disease (Table 1–2).

Table 1–2. Notation for a 2×2 table containing the counts from two binary variables (disease: D = present and $\overline{D}$ = absent and risk factor: F = present and $\overline{F}$ = absent)

	D	$\overline{D}$	**Total**
F	a	b	n_1
$\overline{F}$	c	d	n_2
Total	m_1	m_2	n

Table 1–3. Breast cancer (D) among military women who served in Vietnam (F) and who did not served in Vietnam ($\overline{F}$) during the war years 1965–1973 [1]

	D	$\overline{D}$	**Total**
F	170	3222	3392
$\overline{F}$	126	2912	3038
Total	296	6134	6430

Data from a study of female military veterans published in 1990 [2] produce a 2×2 table (Table 1–3) and an odds ratio contrasting the breast cancer risk between women who served in Vietnam ($n_1 = 3392$) and women who did not served in Vietnam ($n_2 = 3038$) [1]. The odds ratio measure of the association between a disease ($D =$ breast cancer and $\overline{D} =$ no breast cancer) and a risk factor ($F =$ served in Vietnam and $\overline{F} =$ did not served in Vietnam) is

$$\begin{aligned} odds\ ratio = or &= \frac{odds(Vietnam)}{odds(not\ Vietnam)} \\ &= \frac{P(D|F)/\left[1-P(D|F)\right]}{P(D|\overline{F})/\left[1-P(D|\overline{F})\right]} = \frac{P(D|F)/P(\overline{D}|F)}{P(D|\overline{F})/P(\overline{D}|\overline{F})} \end{aligned}$$

and contrasts two conditions, namely those women with the risk factor (F) and those without the risk factor ($\overline{F}$).

The odds of breast cancer among women who served in Vietnam (first row, Table 1–3) are

$$odds(F) = \frac{P(D|F)}{P(\overline{D}|F)} \text{ and are estimated by } \frac{a/n_1}{b/n_1} = \frac{a}{b} = \frac{170}{3222}\ .$$

The odds of breast cancer among women who did not serve in Vietnam (second row, Table 1–3) are

$$odds(\overline{F}) = \frac{P(D|\overline{F})}{P(\overline{D}|\overline{F})} \text{ and are estimated by } \frac{c/n_2}{d/n_2} = \frac{c}{d} = \frac{126}{2912}.$$

The odds ratio is, therefore, estimated by

$$\hat{or} = \frac{a/b}{c/d} = \frac{170/3222}{126/2912} = \frac{ad}{bc} = \frac{170(2912)}{126(3222)} = 1.219.$$

For all estimated values, the assessment of the influence of random variation is basic to their interpretation. For the breast cancer data, an important question becomes: Is the odds ratio estimate of $\hat{or} = 1.219$ substantially different from 1.0 (no association $= or = 1.0$) in light of its associated sampling variation? In other

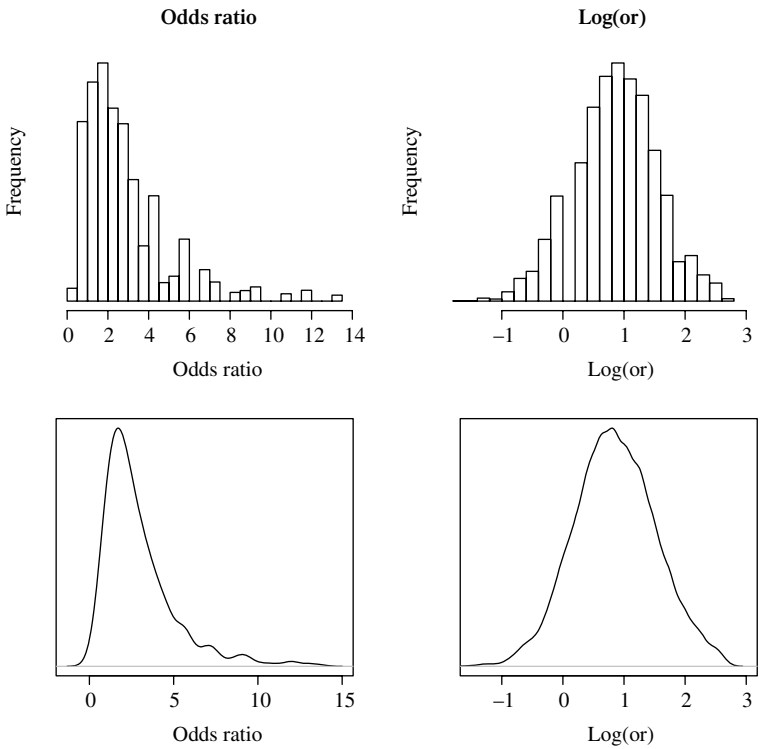


Figure 1–1. Example distributions of the odds ratio and the logarithm of the odds ratio.

words: Does the estimated odds ratio provide clear evidence that the underlying risk of breast cancer differs between military women who served and did not serve in Vietnam?

The vast majority of statistical assessments require either the knowledge or the assumption that the estimated value has at least an approximate normal distribution (more details in Chapter 3). This requirement is frequently not the case for an estimated odds ratio (Figure 1–1, left side).

A typical statistical strategy to assess an estimate with a non-normal distribution begins with a transformation designed so that it has an approximate normal distribution. Then, the usual statistical tools based on the normal distribution apply accurately (for example, statistical tests and confidence intervals). The logarithm of the odds ratio is such an empirically successful transformation (Figure 1–1, right side). *Note*: all logarithms are natural logarithms (base $e = 2.718281828\ldots$), sometimes called *Napier logarithms* in honor of John Napier (*b*. 1550), who pioneered the use of logarithms. For the example (Figure 1–1) based on 5000 samples of size $n = 30$, the underlying probabilities $p_1 = 0.6$ and $p_2 = 0.4$

($or = 2.250$) produce a mean value of the distribution of the odds ratio of 2.709. This bias (2.250 vs. 2.709) substantially decreases when the logarithm of the odds ratio is used. The mean value of the more normal-like distribution of the $log(or)$ values is 0.841, and the odds ratio becomes $or = e^{0.841} = 2.320$. Furthermore, this bias diminishes as the sample size n increases.

The estimated variance of the distribution of the log-transformed odds ratio is

$$estimated\ variance = variance[log(\hat{or})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

The origins and some of the theory underlying the determination of this expression are presented in Chapter 3.

A notable property of the estimate of the odds ratio is evident from a close look at the expression for the variance of the logarithm of the odds ratio. The precision of the estimate is determined by the sum of the reciprocal values of each of the four cell frequencies (a, b, c, and d). These frequencies are often not even close to evenly distributed among the four cells of a 2×2 table. Consequently, the precision of the estimated odds ratio is then largely determined by the smallest cell frequencies, and not the entire sample. For the Vietnam example data, because $a = 170$ and $c = 126$ are relatively small cell frequencies, the other two frequencies are not particularly important in determining the variance. The other cell frequencies only slightly increase the variance. That is, the effective "sample size" is close to $a + c = 170 + 126 = 296$ and not $n = 6430$ women.

From the Vietnam breast cancer data (Table 1–3), the estimated variance of the log-transformed odds ratio is

$$variance[log(\hat{or})] = \frac{1}{170} + \frac{1}{3222} + \frac{1}{126} + \frac{1}{2912} = 0.0145,$$

making the standard error equal to 0.120 ($\sqrt{variance[log(\hat{or})]} = \sqrt{0.0145} = 0.120$). The dominate contributions to the variance are from the two smallest frequencies, where $1/170 + 1/120 = 0.0142$. Using this estimated variance and the normal distribution produces an approximate 95% confidence interval calculated from a 2×2 table based on the estimated value, $log(\hat{or})$.

Approximate 95% confidences interval bounds derived from the normal-like distribution of the estimate $log(\hat{or})$ are

$$lower\ bound = \hat{A} = log(\hat{or}) - 1.960\sqrt{variance[log(\hat{or})]}$$

and

$$upper\ bound = \hat{B} = log(\hat{or}) + 1.960\sqrt{variance[log(\hat{or})]}.$$

The approximate 95% confidence interval bounds based on the estimate, $log(\hat{or}) = log(1.219) = 0.198$, from the breast cancer data are then

$$lower\ bound = \hat{A} = 0.198 - 1.960(0.120) = -0.037$$

and

$$upper\ bound = \hat{B} = 0.198 + 1.960(0.120) = 0.434.$$

These bounds are primarily an intermediate step in constructing a confidence interval for an odds ratio. They directly translate into approximate but generally accurate 95% confidence bounds for the odds ratio estimated by $\hat{or} = e^{0.198} = 1.219$. The lower bound becomes $e^{\hat{A}} = e^{-0.037} = 0.963$, and the upper bound becomes $e^{\hat{B}} = e^{0.434} = 1.544$, creating the 95% confidence interval (0.963, 1.544).

The process that produces a confidence interval for an odds ratio is a specific application of a general technique. A confidence interval for a function of a summary statistic is typically created by applying the same function to the bounds of the original confidence interval. Specifically, for the odds ratio, when the confidence interval bounds for the summary statistic $log(\hat{or})$ are $(\hat{A}, \hat{B})$, then the bounds for the odds ratio based on the function $\hat{or} = e^{log(\hat{or})}$ are $(e^{\hat{A}}, e^{\hat{B}})$.

The confidence interval for the odds creates another example. When the confidence interval bounds for the probability p are $\hat{A}$ and $\hat{B}$, based on the estimated probability $\hat{p}$, then the confidence interval bounds for the odds based on the function $\hat{o} = \hat{p}/(1-\hat{p})$ are $\hat{A}/(1-\hat{A})$ and $\hat{B}/(1-\hat{B})$. In general, when two confidence interval bounds are $\hat{A}$ and $\hat{B}$, based on an estimated value $\hat{g}$, in many cases, the confidence interval bounds for the function $f(\hat{g})$ are the same function of the original bounds, namely $f(\hat{A})$ and $f(\hat{B})$.

Using the same logarithmic transformation, a test statistic to compare the estimated odds ratio to the value $or = 1.0$ (no association between serving in Vietnam and breast cancer) is

$$z = \frac{log(\hat{or}) - log(1.0)}{\sqrt{variance[log(\hat{or})]}} = \frac{0.198 - 0}{0.120} = 1.649.$$

The test statistic z has an approximate standard normal distribution when $log(or) = log(1) = 0$. The probability of a more extreme value occurring by chance alone is the *significance probability* = p-value = $P(|Z| \geq 1.649 | or = 1) = 0.099$. Such a p-value indicates that the elevated value of the observed odds ratio comparing the odds of breast cancer between women who served and did not serve in Vietnam is plausibly due to random variation. From another point of view, the approximate 95% confidence interval (0.963, 1.544) supports the same inference. A 95% confidence interval contains the likely values of the underlying odds ratio

estimated by $\hat{or} = 1.219$, and the value 1.0 (exactly no association) is one of these values.

An optimal and general approach to assessing evidence of an association from data classified into a 2×2 table employs the *chi-square distribution*. The Pearson chi-square test statistic is applied to compare the four observed values (denoted o_{ij}) to four values generated as if no association exists (expected values, denoted e_{ij}). For data from a 2×2 table, some algebra produces the short-cut expression

$$X^2 = \sum\sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{n(ad-bc)^2}{n_1 n_2 m_1 m_2}$$

where the expected values are $e_{ij} = n_i m_j / n$ for $i = 1, 2$ and $j = 1, 2$ (Table 1–2). The four expected values represented by e_{ij} are theoretical cell frequencies calculated as if risk factor and disease are exactly independent (details follow in Chapter 2). For the breast cancer data, the four expected values are given in Table 1–4. At this point, note that the ratios of $D/\overline{D}$ counts (expected odds) are identical regardless of Vietnam status ($156.1/3235.9 = 139.9/2898.1 = 296/6134 = 0.048$) indicating exactly no association. Thus, the odds ratio is exactly 1.0 for these theoretical values. In a statistical context, the term *expected value* has a technical meaning. An expected value is a theoretical value calculated to conform exactly to specified conditions and is treated as a fixed (nonrandom) quantity. The test statistic X^2 has a chi-square distribution with one degree of freedom when no association exists ($or = 1$). For the Vietnam/cancer data, the value of the chi-square test statistic is

$$X^2 = \frac{n(ad-bc)^2}{n_1 n_2 m_1 m_2} = \frac{6430[170(2912) - 126(3222)]^2}{3392(3038)(296)6134} = 2.726$$

yielding a p-value $= P(X^2 \geq 2.726 | no\ association) = 0.099$. The parallel statistical test based on the logarithm of the odds ratio is not the same as the

Table 1–4. Expected numbers of breast cancer cases (D) among military women who served in Vietnam (F) and those who did not serve in Vietnam (F) when the odds ratio or $= 1$ (no association*)

	D	$\overline{D}$	**Total**
F	$e_{11} = 156.1$	$e_{12} = 3235.9$	$n_1 = 3392$
$\overline{F}$	$e_{21} = 139.9$	$e_{22} = 2898.1$	$n_1 = 3038$
Total	$m_1 = 296$	$m_1 = 6134$	$n = 6430$

$$* = or = \frac{156.1/3235.9}{139.9/2898.1} = 1.0.$$

chi-square test but rarely differs by an important amount (previously, $X^2 = (1.649)^2 = 2.719$).

An odds ratio of 1.0 generates a variety of equivalent expressions for the relationships between the probability of a disease D and a risk factor F. Some examples are:

$$P(F|D) = P(F|\overline{D}),\ P(D|F) = P(D|\overline{F}),$$
$$P(D \text{ and } F) = P(D)P(F) \text{ and } P(D|F) = P(D).$$

The last expression shows in concrete terms that an odds ratio of $or = 1$ means that the risk factor (F) is unrelated to the likelihood of disease (D).

An advertisement and a brochure produced by the March of Dimes (1998) states:

> Debbie and Rich Hedding of Pittsford Vt. were devastated when they lost two babies to neural tube defects (NTDs). When Debbie read about folic acid in a March of Dimes brochure, she was astonished when she learned about the role of folic acid in preventing NTDs. "I was in tears by the time I finished reading the material. I haven't been taking folic acid nor had I been told about it. I couldn't believe that I could have reduced the risk of recurrence by 70 percent. I immediately began taking folic acid and telling every women I could about it."

The published odds ratio was $\hat{or} = 0.28$ (the 95% confidence interval is (0.12, 0.71), *Lancet*, 1996). The 70% reduction in risk refers to $(1 - 0.280) \times 100 = 72\%$. However, the odds ratio, as the name indicates, is a ratio measure of association. It measures association in terms of a multiplicative scale and behaves in an asymmetric way on an additive scale. Odds ratios between zero and 1 indicate a decrease in risk, and odds ratios greater than 1 (1.0 to infinity) indicate an increase in risk. From a more useful point of view, an odds ratio of $1/or$ measures the same degree of association as the odds ratio or, but in the opposite direction, which is a property of a ratio scale. Thus, an odds ratio of 0.28 reduces risk of an NTD by a factor of $1/0.28 = 3.57$, a reduction considerably greater than 70%. Comparisons on a ratio scale are necessarily made in terms of ratios.

From the perspective of a 2×2 table, interchanging the rows (or columns) changes the odds ratio from or to $1/or$ but does not influence the degree of association between variables because the order of the rows (or columns) is purely arbitrary. The result of a statistical test of association, for example, is unchanged (same p-value). It should be noted that the logarithms of values measured on a ratio scale produce values that are directly comparable on an additive scale. An odds ratio of 0.28 is equivalent to an odds ratio of 3.57, relative to 1.0 on a ratio scale. Thus, the value $\log(0.28) = -1.273$ is equivalent to the value $\log(3.57) = 1.273$ relative to $\log(1) = 0$ on an additive scale. The transformation of an odds

Table 1–5. Four measures of association from a 2×2 table and expressions for their estimation

	Measures	Symbols	Estimates*
Difference in probabilities (rows)	$P(D\|F) - P(D\|\overline{F})$	$\hat{p}_1 - \hat{p}_2$	$\frac{a}{a+b} - \frac{c}{c+d}$
Difference in probabilities (columns)	$P(F\|D) - P(F\|\overline{D})$	$\hat{P}_1 - \hat{P}_2$	$\frac{a}{a+c} - \frac{c}{b+d}$
Relative risk	$\frac{P(D\|F)}{P(D\|\overline{F})}$	$\hat{rr}$	$\frac{a/(a+b)}{c/(c+d)}$
Odds ratio	$\frac{P(D\|F)/P(\overline{D}\|F)}{P(D\|\overline{F})/P(\overline{D}\|\overline{F})}$	$\hat{or}$	$\frac{a/b}{c/d}$

Note: * = notation from Table 1–2.

ratio to the logarithm of an odds ratio can be viewed as a change from a ratio scale to the more familiar and intuitive additive scale.

The odds ratio is one of a number of possible measures of association between two binary variables described by a 2×2 table. Table 1–5 lists three other common measures of an association where, again, $D/\overline{D}$ could represent the presence/absence of disease and $F/\overline{F}$ could represent the presence/absence of a binary risk factor. Although these summary values reflect the magnitude of an association in different ways, their statistical evaluation follows much the same pattern.

Normal distribution–based statistical tests and 95% confidence intervals associated with the evaluation of measures of association require estimates of their *variances*. Expressions for estimates of the variances of the distributions of these measures of association and their values from the Vietnam/cancer data are given in Table 1–6.

Each measure of association can be assessed with a normal distribution–based test statistic of the form

$$z = \frac{\hat{g} - g_0}{\sqrt{variance(\hat{g})}}$$

where $\hat{g}$ represents any one of the estimated measures of association and g_0 represents a theoretical value. The previous chi-square statistic, however, is easily applied and typically used to evaluate the influence of random variation in a 2×2 table.

Approximate confidence intervals for these measures of association (Tables 1–5 and 1–6) based on the normal distribution are created from a single expression. An approximate but again generally accurate 95% confidence interval is

$$\hat{g} \pm 1.960\sqrt{variance(\hat{g})}.$$

Table 1–6. Expressions for estimates of the variances of the distributions of six measures of association* applied to the comparison of the risk of breast cancer between female veterans who served and did not serve in Vietnam (Table 1–3)

	Symbols	Variances	Std. errors
Difference in proportions (rows)	$\hat{p}_1 - \hat{p}_2$	$\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$	0.0052
Difference in proportions (columns)	$\hat{P}_1 - \hat{P}_2$	$\frac{\hat{P}_1(1-\hat{P}_1)}{m_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{m_2}$	0.0294
Relative risk	$\hat{rr}$	$\hat{rr}^2\left[\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} + \frac{1}{n_2}\right]$	0.1388**
Logarithm of the relative risk	$log(\hat{rr})$	$\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}$	0.1149**
Odds ratio	$\hat{or}$	$\hat{or}^2\left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]$	0.1467
Logarithm of the odds ratio	$log(\hat{or})$	$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$	0.1203

Note: * = Chapter 3 contains a discussion of the statistical origins of these expressions.
** = another version of the estimated variance excludes the terms $1/n_1$ and $1/n_2$ (Chapter 6).

Table 1–7. Six measures of association applied to evaluate the risk of breast cancer among female veterans who served in Vietnam (statistical tests and 95% confidence intervals)

	Symbols	Estimates	p-Values	Lower	Upper
Difference in proportions (rows)*	$\hat{p}_1 - \hat{p}_2$	0.009	0.099	−0.002	0.019
Difference in proportions (columns)*	$\hat{P}_1 - \hat{P}_2$	0.049	0.099	−0.009	0.107
Relative risk	$\hat{rr}$	1.208	0.099	0.964	1.513
Logarithm of the relative risk	$log(\hat{rr})$	0.189	0.099	−0.036	0.414
Odds ratio	$\hat{or}$	1.219	0.099	0.963	1.544
Logarithm of the odds ratio	$log(\hat{or})$	0.198	0.099	−0.037	0.434

Note: * Identical to the previous chi-square test, $X^2 = 2.726$.

As with the odds ratio, approximate tests and 95% confidence intervals for the relative risk measure of association (a ratio measure of association) are more accurate when the logarithm of the relative risk is used (an additive measure of association). Examples from the Vietnam/cancer data are displayed in Table 1–7.

PROPERTIES OF THE ODDS RATIO

The term *relative risk* (denoted *rr*, Table 1–5) refers to a ratio of two probabilities and is a natural multiplicative measure to compare the probability of disease between a group with the risk factor to a group without the risk factor (Chapter 6). When the disease is rare [$P(D|F)$ and $P(D|\overline{F})$ are both less than 0.1], the odds ratio is approximately equal to the relative risk ($or \approx rr$). For the Vietnam example data, the estimated odds ratio is $\hat{or} = 1.219$ and the estimated relative risk is $\hat{rr} = 1.208$, because breast cancer is rare in both groups (probability ≈ 0.05).

The results from a survey of 720 primary care physicians conducted by Schulman and colleagues were published in *The New England Journal of Medicine* (1999). These authors pointed out that their data indicate that "women (odds ratio, 0.60; 95% confidence interval 0.4 to 0.9, ...) and blacks (odds ratio, 0.60; 95% confidence interval 0.4 to 0.9, ...) were less likely to be referred to cardiac catheterization than men and whites, respectively." This apparent inequality in care immediately became a topic of the national news coverage and the subject of the television show *Nightline*. Generally, the media reported that a recent study showed that 40% of black patients were less likely than white patients to be referred for appropriate coronary care.

The collected data showed a 84.7% correct response for black and a 90.6% correct response for white patients from physicians managing chest pain. Thus, an odds ratio of

$$odds\ ratio = \hat{or} = \frac{0.847/0.153}{0.906/0.094} = 0.574.$$

For these same data, the estimated relative risk is

$$relative\ risk = \hat{rr} = \frac{P(correct|black)}{P(correct|white)} = \frac{0.847}{0.906} = 0.935.$$

As noted, an odds ratio accurately measures risk (approximates relative risk) only when the "disease" is rare in both compared groups. Although an odds ratio measures the association between race and care, it does not reflect risk because the probability of correct care is not rare in both compared groups.

Following the publication of this article, the editors of *The New England Journal of Medicine* stated that, "we take responsibility for the media over-interpretation of the article by Schulman and colleagues. We should not have allowed the odds ratio in the abstract." Perhaps the most intuitive and easily interpreted contrast between white and black patients is the directly measured difference in proportions, where

$$difference = P(correct|white) - P(correct|black) = 0.906 - 0.847 = 0.059.$$

A small numeric example illustrates the reason for the similarity of the odds ratio to the relative risk as a measure of risk only when the disease is rare. Fictional data simply indicate the role of the "rare disease requirement."

Consider the case in which the probability of disease is $P(disease) = 8/28 = 0.29$ (not rare) and

	Disease	No disease	**Total**
Risk factor present	3	10	13
Risk factor absent	5	10	15

$$odds\ ratio = \frac{3/10}{5/10} = 0.60 \qquad relative\ risk = \frac{3/13}{5/15} = 0.69$$

Now, consider the case where the probability of disease is $P(disease) = 8/208 = 0.04$ (rare in both groups) and

	Disease	No disease	**Total**
Risk factor present	3	100	103
Risk factor absent	5	100	105

$$odds\ ratio = \frac{3/100}{5/100} = 0.60 \qquad relative\ risk = \frac{3/103}{5/105} = 0.61$$

In symbols, for the second case, both measures of association are similar because $b \approx a+b$ and $d \approx c+d$ making

$$\hat{rr} = \frac{a/(a+b)}{c/(c+d)} \approx \frac{a/b}{c/d} = \hat{or} \qquad \text{(Table 1–5)}$$

That is, the number of cases of disease 3 and 5 are not particularly small relative to 13 and 15 (approximately 40%) but are substantially smaller than 103 and 105 (approximately 5%).

At the beginning of the 18th century, Sir Thomas Bayes (*b.* 1702) noted that, for two events, denoted A and B,

$$P(A \text{ and } B) = P(A|B)P(B) \qquad \text{and} \qquad P(A \text{ and } B) = P(B|A)P(A),$$

leading to the historic expression called *Bayes' theorem* where

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

Less historic, but important to the study of disease and other binary variables, the odds ratio has the same value whether the association under study is measured by comparing the likelihoods of the risk factor between individuals with and

without the disease or is measured by comparing the likelihoods of the disease between individuals with and without the risk factor. Retrospectively and case/control collected data require the comparison of the likelihood of the risk factor between individuals with and without the disease. Prospectively and cross-sectionally collected data require the comparison of the likelihood of the disease among individuals with and without the risk factor. The key to assessing an association using these two kinds of data is the comparison of the probability of the risk factor among those with the disease $P(F|D)$ to those without disease $P(F|\overline{D})$, or the comparison of the probability of the disease among those with the risk factor $P(D|F)$ to those without the risk factor $P(D|\overline{F})$.

An application of Bayes' theorem shows that, regardless of the kind of data collected, the odds ratio measure of association is the same. Specifically, four applications of Bayes' theorem give

$$or = \frac{P(D|F)/P(\overline{D}|F)}{P(D|\overline{F})/P(\overline{D}|\overline{F})} = \frac{\left[\dfrac{P(F|D)P(D)}{P(F)} \Big/ \dfrac{P(F|\overline{D})P(\overline{D})}{P(F)}\right]}{\left[\dfrac{P(\overline{F}|D)P(D)}{P(\overline{F})} \Big/ \dfrac{P(\overline{F}|\overline{D})P(\overline{D})}{P(\overline{F})}\right]}$$
$$= \frac{P(F|D)/P(F|\overline{D})}{P(\overline{F}|D)/P(\overline{F}|\overline{D})} = or.$$

The estimated odds ratio follows the same pattern, where

$$\hat{or} = \frac{a/b}{c/d} = \frac{ad}{bc} = \frac{a/c}{b/d} = \hat{or}.$$

The property that the odds ratio and its estimate are the same for both kinds of data is also a property of comparing two proportions from a 2×2 table. That is, the test statistics produce exactly the same result. Specifically, the evaluation of the difference $P(D|F) - P(D|\overline{F})$ or $P(F|D) - P(F|\overline{D})$ produces

$$\textit{difference in proportions (rows)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$
$$= \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{m_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{m_2}}} = \textit{difference in proportions (columns)}.$$

Another notable property of the odds ratio is that it is always further from 1.0 (no association) than the relative risk calculated from the same 2×2 table. In symbols,

$$\text{when } rr \geq 1.0, \text{ then } or \geq rr \geq 1.0$$

and

$$\text{when } rr \leq 1.0, \text{then } or \leq rr \leq 1.0.$$

In this sense, the relative risk is conservative, because any association measured by relative risk is more extreme when measured by an odds ratio.

THREE STATISTICAL TERMS

A discussion of the odds ratio calculated from a 2×2 table presents an opportunity to introduce three fundamental concepts relevant to many statistical analyses. These concepts are properties of summary values calculated from data and are not properties of the sampled population. This extremely important distinction is frequently not made clear. The concepts are: *interaction, confounding*, and *independence*. These three general properties of summary values are defined and discussed for the simplest possible case, three binary variables classified into two 2×2 tables, where associations are measured by odds ratios. The example binary variables are discussed in terms of a disease D, a risk factor F, and a third variable labeled C. This introductory description is expanded in subsequent chapters.

When the odds ratios calculated from each of two tables differ, the question arises: Do the estimated values differ by chance alone or does a systematic difference exist? The failure of a measure of association to be the same within each subtable is an example of an *interaction* (Table 1–8A). An interaction, in simple terms, is the failure of the values measured by a summary statistic to be the same under different conditions.

The existence of an interaction is a central element in addressing the issues involved in combining estimates. In the case of two 2×2 tables, the first issue is the choice between describing an association with two separate odds ratios (denoted $or_{DF|C}$ and $or_{DF|\overline{C}}$) or a single summary odds ratio (denoted or_{DF}). When odds ratios systematically differ, a single summary odds ratio is usually not meaningful and, in fact, such a single value can be misleading (an example follows). However, when the odds ratios measuring association differ only by chance alone,

Table 1–8A. An odds ratio measure of association between the risk factor F and disease D that differs at the two levels of the variable C (an interaction)

C	D	$\overline{D}$	**Total**
F	20	20	40
$\overline{F}$	10	40	50
Total	30	60	90

$or_{DF|C} = 4$

$\overline{C}$	D	$\overline{D}$	**Total**
F	60	20	80
$\overline{F}$	10	20	30
Total	70	40	110

$or_{DF|\overline{C}} = 6$

Table 1–8B. The odds ratios between the risk factor *F* and disease *D* at the two levels of the variable C (no interaction) and a single odds ratio from the combined table.

C	D	$\overline{D}$	**Total**
F	40	20	60
$\overline{F}$	10	40	50
Total	50	60	110

$or_{DF|C} = 8$

$\overline{C}$	D	$\overline{D}$	**Total**
F	80	20	160
$\overline{F}$	10	20	30
Total	90	40	130

$or_{DF|\overline{C}} = 8$

$C+\overline{C}$	D	$\overline{D}$	**Total**
F	120	40	160
$\overline{F}$	20	60	80
Total	140	100	240

$or_{DF|(C+\overline{C})} = or_{DF} = 9$

the opposite is true. A single odds ratio is not only a meaningful summary but provides a useful, simpler, and more precise measure of the association between the risk factor and disease.

When a single odds ratio is a useful summary (no interaction), the second issue becomes the choice of how the summary value is calculated. The choice is between a summary value calculated by combining the two odds ratios or by combining the data into a single table and then calculating a summary odds ratio. When the table created by adding two subtables does not reflect the relationship within each subtable, it is said that the variable used to create the subtables has a *confounding influence*. Specifically (Table 1–8B), the odds ratio in the combined table $or_{DF|(C+\overline{C})} = or_{DF} = 9.0$ is not equal to odds ratios in each subtable where $or_{DF|C} = or_{DF|\overline{C}} = 8.0$, identifying the fact that the variable C has a confounding influence. Consequently, to accurately estimate the risk/disease association it is necessary to account for the variable C by combining the odds ratios from each subtable. Otherwise, combining the data into a single table fails to account for the influence of the variable C, producing a biased estimate.

When the variable C is unrelated to the disease or the risk factor. The odds ratio calculated from the 2×2 table formed by combining the data then accurately reflects the risk factor/disease relationship. Table 1–8C illustrates the case where all three odds ratios equal 4.0 ($or_{DF|C} = or_{DF|\overline{C}} = or_{DF|C+\overline{C}} = or_{DF} = 4.0$). The variable C is then said to be an *independent* summary of the risk/disease association. The collapsed table (C-table + $\overline{C}$-table) eliminates the variable C and accurately produces a simpler and more precise measure of the association.

This description of the terms interaction, confounding, and independence does not account for random variation of observed values, but consists of simple and fictional data that perfectly reflect the three concepts. However, important properties are illustrated. Namely, the presence or absence of an interaction determines

Table 1–8C. The odds ratios between the risk factor F and disease D at the two levels of the variable C, and a single odds ratio from the combined table (independence)

C	D	$\overline{D}$	**Total**
F	20	20	40
$\overline{F}$	10	40	50
Total	30	60	90

$or_{DF|C} = 4$

$\overline{C}$	D	$\overline{D}$	**Total**
F	40	40	80
$\overline{F}$	20	80	100
Total	60	120	180

$or_{DF|\overline{C}} = 4$

$C+\overline{C}$	D	$\overline{D}$	**Total**
F	60	60	120
$\overline{F}$	30	120	150
Total	90	180	270

$or_{DF|(C+\overline{C})} = or_{DF} = 4$

whether or not a single summary value is useful. When a single summary is useful, the presence or absence of a confounding influence from a third variable determines the way in which a summary value is calculated. When a confounding influence exists, the values calculated from each subtable are combined to form a summary value. When a confounding influence does not exist, the summary value is more effectively and simply calculated from a single table created by adding the subtables. That is, the third variable is ignored. The topic of creating accurate summary values in applied situations is continued in detail in Chapter 2 and beyond.

A classic statistical example [3], called *Simpson's paradox*, is not really a paradox but an illustration of the consequences of ignoring an interaction. Simpson's original "cancer treatment data" are presented in Table 1–9.

Simpson created "data" in which the summary values from each subtable have the opposite relationship from the summary value calculated from the combined data. A new treatment (denoted F) appears more successful than the usual treatment (denoted $\overline{F}$) for two kinds of cancer patients, but when the data are combined into a single table, the new treatment appears less effective.

The three relevant odds ratios are:

$$or_{SF|C} = 2.38,\ or_{SF|\overline{C}} = 39.0 \quad \text{and} \quad or_{SF|C+\overline{C}} = or_{SF} = 0.45.$$

The stage I cancer patients (C) and stage II cancer patients ($\overline{C}$) clearly differ with respect to the new treatment, reflected by their extremely different odds ratios (2.38 and 39.0). Technically, it is said that the treatment and stage measures of survival (odds ratios) have a strong (a very strong) interaction with the treatment classification. Therefore, combining the data and calculating a single summary value produces a result that has no useful interpretation. The point is, when

Table 1–9. The "data" used to demonstrate Simpson's paradox

S = Survived and $\overline{S}$ = died
F = New treatment and $\overline{F}$ = usual treatment
C = Stage I cancer (not severe) and $\overline{C}$ = stage IV cancer (severe)

C	F	$\overline{F}$	**Total**
S	95	800	895
$\overline{S}$	5	100	105
Total	100	900	1000

$P(survived|new) = P(S|F) = 0.95$
$P(survived|usual) = P(S|\overline{F}) = 0.88$

$\overline{C}$	F	$\overline{F}$	**Total**
F	400	5	405
$\overline{F}$	400	195	595
Total	800	200	1000

$P(survived|new) = P(S|F) = 0.50$
$P(survived|usual) = P(S|\overline{F}) = 0.025$

$C + \overline{C}$	F	$\overline{F}$	**Total**
S	495	805	1300
$\overline{S}$	405	295	700
Total	900	1100	2000

$P(survived|new) = P(S|F) = 0.55$
$P(survived|usual) = P(S|\overline{F}) = 0.73$

measures of association that reflect two different relationships are combined, any value of a summary statistic is possible and is rarely meaningful.

AVERAGE RATES

Disease and mortality rates are familiar and extensively used summaries of risk, particularly in public health, medicine, and epidemiology. The United States 1940 cancer mortality rate among individuals 60–64 years old was 88.7 deaths per 100,000 person-years, and 60 years later the cancer mortality rate for the same age group has decreased to 75.9 deaths per 100,000 person-years. One gets a sense of the change in risk by comparing two such rates, but the answers to a number of questions are not apparent. Some examples are: Why report the number of deaths per person-years? How do these rates differ from probabilities? What is their relationship to the mean survival time? Or, more fundamentally: Why does such a rate reflect risk? The following is a less than casual but short of rigorous description of the statistical origins and properties of an average rate. To simplify the terminology, rates from this point on are referred to as *mortality rates*, but the description applies to disease incidence rates, as well as to many other kinds of average rates.

An average rate is a ratio of two mean values. For a mortality rate, the mean number of deaths divided by the mean survival time experienced by those individuals at risk forms the average rate.

The mean number of deaths is also the *proportion* of deaths. A proportion is another name for a mean value calculated from zeros and ones. In symbols and denoted q,

$$mean\ value = proportion = \frac{0+1+0+0+\cdots+1}{n} = \frac{\sum x_i}{n} = \frac{d}{n} = q$$

where x_i represents a binary variable that takes on only the values zero or one among n individuals and d represents the sum of these n values of zeros and ones. In the language of baseball, this mean value is called a *batting average* but elsewhere it is usually referred to as a proportion or an estimated probability (Chapter 3).

The mean survival time, similar to any mean value, is the total time at risk experienced by the n individuals who accumulated the relevant survival time divided by n. In symbols, the estimated mean survival time is

$$\bar{t} = \frac{total\ time\ at\text{-}risk}{total\ number\ individuals\ at\text{-}risk} = \frac{\sum t_i}{n}$$

where t_i represents the time at risk for each of n at-risk individuals ($i = 1, 2, \cdots, n$). Then, the expression for an average rate (denoted R) becomes

$$R = average\ rate = \frac{mean\ number\ of\ deaths}{mean\ survival\ time} = \frac{q}{\bar{t}} = \frac{d/n}{\sum t_i/n} = \frac{d}{\sum t_i}.$$

Equivalently, but less intuitively, an average rate is usually defined as the total number of deaths (d) divided by the total time-at-risk ($\sum t_i$).

An average rate can be viewed as the reciprocal of a mean value. When a trip takes 4 hours to go 120 miles, the mean travel time is 2 minutes per mile ($\bar{t} = 240/120 = 2$ minutes per mile, or $\bar{t} = 4/120 = 1/30$ of an hour per mile). Also, a trip of 120 miles traveled in 4 hours yields an average rate of speed of $120/4 = 30$ miles per hour. The reciprocal of the mean time of 1/30 of an hour per mile is an average rate of 30 miles per hour.

For an average mortality rate, first consider the not very frequently occurring case in which a group of individuals is observed until all at-risk persons have died. The mean survival time is then $\bar{t} = \sum t_i/n$, and the number of deaths is $d = n$ (all died) making

$$average\ rate = R = \frac{d}{\sum t_i} = \frac{n}{\sum t_i} = \frac{1}{\bar{t}} \quad \text{or} \quad \bar{t} = \frac{1}{R}.$$

A more realistic case occurs when a group of individuals is observed for a period of time and, at the end of that time, not all at-risk persons have died. A single estimate requires that the value estimated be a single value. Thus, when a single rate is estimated from data collected over a period of time, it is implicitly assumed that the risk described by this single value is constant over the same time period or at least approximately constant. When a mortality rate is constant, then the mean survival time is also constant. Under this condition, the constant mortality rate associated with d deaths among n at-risk individuals is, as before, estimated by the average rate given by $R = d/\sum t_i$, where t_i represents the observed time alive for both individuals who died and survived. The mean time lived (denote $\hat{\mu}$) by all n individuals is again estimated by the reciprocal of this rate or

$$\hat{\mu} = \frac{1}{R} = \frac{\sum t_i}{d}.$$

Naturally, as the rate of death increases, the mean survival time decreases and vice versa.

The estimated mean survival time is the total survival time observed for all n at-risk individuals divided by d (number of deaths) and not n (number of observed individuals) for the following reason. The sum of the survival times ($\sum t_i$) is too small because not all n observed individuals died. The $n-d$ individuals who did not die would contribute additional survival time if the period of observation was extended. Specifically, the amount of "missing time," on average, is $\mu(n-d)$ because each individual who did not die would have contributed, on average, an additional time of μ to the total survival time if he were observed until his death. Including this "missing time," the expression for the mean survival time becomes

$$\hat{\mu} = \frac{total\ time}{total\ number\ individuals\ at\text{-}risk} = \frac{\sum t_i + \hat{\mu}(n-d)}{n},$$

and solving for $\hat{\mu}$ gives $\hat{\mu} = \sum t_i/d$ as the estimated mean survival time. This estimate $\hat{\mu}$ is unbiased in the sense that it compensates for the $n-d$ incomplete survival times (unobserved times to death). As long as the mean survival time μ is the same for those individuals who died during the time period under consideration and those who did not, the mortality rate and the mean survival time are again inversely related or $\hat{\mu} = \sum t_i/d = 1/R$.

Many sources of disease incidence and mortality data consist of individuals classified into categories based on their age at death or time of death. For these kinds of survival data, the exact time of death is not usually available. This is particularly true of publicly available disease and mortality data (see the National Center for Health Statistics or the National Cancer Institute or Center for Disease Control websites http://www.cdc.gov/nchs/ or http://www.nci.nih.gov

or http://www.cdc.gov). However, a generally accurate but approximate average mortality rate can be calculated when the length of the age or time interval considered is not large.

Consider a time interval with limits denoted a_i to a_{i+1} years (length of the interval $= \delta_i = a_{i+1} - a_i$ years), where l_i individuals are alive at the beginning of the interval and d_i of these individuals died during the interval. The total time alive (at-risk) is made up of two distinct contributions. First, the $l_i - d_i$ individuals who survived the entire interval (δ_i years) contribute $\delta_i(l_i - d_i)$ person-years to the total years lived. Second, over a short interval such as 5 or even 10 years, deaths usually occur approximately at random throughout the interval so that, on average, each individual who died contributes $\frac{1}{2}\delta_i$ years to the total years lived. Thus, the d_i individuals who died during the interval contribute $\frac{1}{2}\delta_i d_i$ person-years to the total years lived. Therefore, the approximate total years lived during the i^{th} interval is

$$total\ person\text{-}years\text{-}at\text{-}risk = \delta_i(l_i - d_i) + \frac{1}{2}\delta_i d_i = \delta_i\left(l_i - \frac{1}{2}d_i\right).$$

An approximate average mortality rate for the i^{th} interval (denoted R_i), based on this approximate years-at-risk, is then

$$\begin{aligned} average\ mortality\ rate = R_i &= \frac{number\ of\ deaths\ in\ the\ i^{th}\ interval}{total\ years\text{-}at\text{-}risk\ in\ the\ i^{th}\ interval} \\ &= \frac{d_i}{\delta_i(l_i - \frac{1}{2}d_i)}\ deaths\ per\ person\text{-}years. \end{aligned}$$

This approximate rate is again the mean number of deaths ($q_i = d_i/l_i$) divided by the mean survival time ($\overline{t}_i = \delta_i[l_i - \frac{1}{2}d_i]/l_i$) or, as before, $R_i = q_i/\overline{t}_i$.

The probability of death in a specific interval of time is

$$probability\ of\ death\ in\ the\ i^{th}\ interval = q_i = \frac{d_i}{l_i}.$$

That is, the probability q_i is the number events that have a specific property (for example, deaths $= d_i$) divided by the total number of events that could have occurred (for example, all persons who could have died, at-risk $= l_i$). In human data, disease or death are frequently rare events causing d_i to be much smaller than l_i ($d_i \ll l_i$). In this case, an average rate for a specific age or time interval essentially equals the probability of death divided by the length of the interval.

Or, in symbols,

$$R_i = \frac{mean\ number\ of\ deaths}{mean\ survival\ time} = \frac{d_i/l_i}{\delta_i(l_i - \frac{1}{2}d_i)/l_i}$$
$$= \frac{q_i}{\delta_i(1-\frac{1}{2}q_i)} \approx \frac{q_i}{\delta_i} \qquad \left(1 - \frac{1}{2}q_i \approx 1\right).$$

Therefore, when the interval length considered is 1 year ($\delta_i = 1$), the value of a mortality rate and a probability of death are usually interchangeable ($R_i \approx q_i$). Furthermore, a rate ratio and a ratio of probabilities are also essentially equal when applied to the same time interval. In symbols, the ratios are

$$rate\ ratio = \frac{R_1}{R_2} \approx \frac{q_1/\delta}{q_2/\delta} = \frac{q_1}{q_2} = relative\ risk\ ratio.$$

Geometry of an Average Rate

The exact geometry of an average rate is complicated and requires specialized mathematical tools [4]. However, the approximate geometry is simple and is displayed in Figure 1–2. The key to a geometric description of a rate is the proportion of surviving individuals measured at two points in time. These proportions are called *survival probabilities* and for the time interval a_i to a_{i+1} (length again δ_i) they are

$$P_i = \frac{l_i}{l_0} \qquad \text{and} \qquad P_{i+1} = \frac{l_{i+1}}{l_0}$$

where P_i and P_{i+1} then represent the proportion of individuals alive at times a_i and a_{i+1} among the original population of individuals at-risk (size denoted l_0).

The approximate mean person-years again consists of two parts, namely the mean years lived by those who survived the entire interval (a rectangle) and the mean years lived by those who died during the interval (a triangle). In more detail, for those individuals who lived the entire interval

$$rectangle = width \times height = \delta_i P_{i+1}\ person\text{-}years,$$

and for those individuals who died during the interval

$$triangle = \frac{1}{2}base \times altitude = \frac{1}{2}\delta_i(P_i - P_{i+1}) = \frac{1}{2}\delta_i D_i\ person\text{-}years,$$

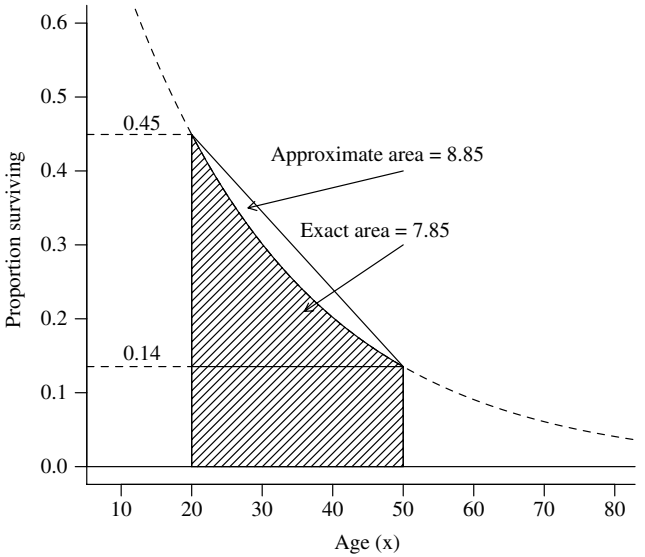


Figure 1–2. The approximate and exact areas (approximate and exact mean survival times) describing the mortality pattern of a hypothetical population.

where $P_i - P_{i+1} = D_i$ represents the proportion of deaths that occurred in the i^{th} interval. Then,

$$\begin{aligned} area = approximate\ mean\ survival\ time &= rectangle + triangle \\ &= \delta_i P_{i+1} + \frac{1}{2}\delta_i D_i = \delta_i \left(P_i - \frac{1}{2} D_i \right)\ person\text{-}years. \end{aligned}$$

Again, it is necessary to assume that the underlying mortality rate is at least close to constant, so that a single rate R accurately reflects the risk for the entire interval.

From the plot (Figure 1–2), two example survival probabilities are $P_{20} = 0.45$ and $P_{50} = 0.14$ at ages 20 and 50 years and, therefore, the area of the *rectangle* $= 30(0.14) = 4.20$ person-years. The mean number of deaths (proportion) is $D_{20} = P_{20} - P_{50} = 0.45 - 0.14 = 0.31$. Therefore, the area of the *triangle* $= \frac{1}{2}30(0.45 - 0.14) = \frac{1}{2}30(0.31) = 4.65$ person-years. The approximate mean survival time then becomes *total area* $= 4.20 + 4.65 = 8.85$ person-years for the interval 20 to 50 years ($\delta = 30$). More succinctly, the mean time survived is $30(0.45 - \frac{1}{2}0.31) = 8.85$ years. The average approximate rate is again the ratio of these two mean values or

$$R = \frac{0.31}{8.85} \times 10{,}000\ person\text{-}years = 350.3\ deaths\ per\ 10{,}000\ person\text{-}years.$$

From another but equivalent prospective, when $l_0 = 1000$ individuals are at risk, then for the interval age 20 to 50 years

$$\begin{aligned} total\ years\text{-}at\text{-}risk &= l_0(\delta_i P_{i+1}) + l_0\left(\frac{1}{2}\delta_i D_i\right) \\ &= 1000(30)(0.14) + 1000\left(\frac{1}{2}\right)(30)(0.31) \\ &= 4200 + 4650 = 8850\ person\text{-}years \qquad \text{and} \\ total\ deaths &= l_0 D_i = 1000(0.31) = 310\ deaths \end{aligned}$$

producing the identical approximate average rate

$$\begin{aligned} rate = R &= \frac{total\ deaths}{total\ person\text{-}years\text{-}at\text{-}risk} \times 10{,}000 \\ &= \frac{310}{8850} \times 10{,}000 \\ &= 350.3\ deaths\ per\ 10{,}000\ person\text{-}years. \end{aligned}$$

To calculate the exact area (mean survival time), the survival probability curve (Figure 1–2, dashed line) must be mathematically defined (left to more advanced presentations). However, for the example, the exact rate generated from the survival probabilities (Figure 1–2) is 400 death per 10,000 person-years at risk for any interval (constant mortality rate). As the interval length δ decreases, the approximate rate becomes a more accurate estimate of this constant rate because a straight line (approximate) and a curve (exact) become increasingly similar. That is, a rectangle plus a triangle more accurately estimate the mean person-years at risk (area under the curve). For the example, the approximate average rate is 395 deaths per 10,000 person-years for a 10-year interval, and the rate is 399 deaths per 10,000 person-years for a 5-year interval.

Proportionate Mortality "Rates"

Researchers Coren and Halpren found that the "average age of death for right-handers in the sample was 75 years. For left-handers, it was 66." The authors attributed this difference in age-at-death primarily to a difference in risk from accidents (*New England Journal of Medicine*,1988).

The study described by the newspaper story is based on a sample of 1,000 death certificates. All individuals sampled are dead. Comparing the frequency of deaths among individuals with different exposures or risk factors is called a *proportionate mortality study*. A fundamental property of proportionate mortality studies in general is that it is not possible to determine from the data whether

San Francisco Chronicle

sfgate.com 415-777-1111

Right-Handers Outlive Lefties By 9-Year Average, Study Says

By Malcolm Gladwell
Washington Post

Boston

In a dramatic and controversial finding, a team of psychologists has reported that left-handed people may live an average of nine years less than right-handers.

The study, based on an analysis of death certificates in the San Bernardino area, is the first to suggest that the well-documented susceptibility of left-handers to a variety of behavioral and physiological disorders can have a substantial effect on life expectancy.

Researchers suspect that left-handers are far more prone to certain kinds of diseases than right-handers and that they are disproportionately represented among those born prematurely and among the mentally retarded. Left-handers are also far more likely to suffer serious accidents.

Speculation

These differences
some researchers
to speculate
make up
por-

ence (in left-handers' life expectancy)," said Alan Searlman, a psychologist at St. Lawrence University in New York. "But I certainly can't imagine that there is a nine-year difference. Does it seem reasonable to you that there could have been a difference this large and it would have been unnoticed until 1991?"

Other researchers said Coren and Halpern did not make sure that the difference in death rates

'We were astounded. We had no idea that the difference was going to be this huge'

very stable data. There is nothing peculiar about this study."

Coren and Halpern offer several explanations for their findings. The first is straightforward: 7.9 percent of left-handers in the sample died in accidents, compared with 1.5 percent of right-handers.

This finding squares with at least one previous study that showed that living in a technological world designed for right-handers has made left-handers five times more likely to get into an accident than their counterparts.

Traffic Peril

As an exam-
handers h-
tage

the exposure increased the risk in one group or decreased the risk in another or both. For example, the increased life-time observed in right-handed individuals could be due to a decrease in their risk, or to an increase in the risk of left-handed individuals, or both.

In addition, a frequent problem associated with interpreting proportionate mortality data is that the results are confused with an assessment of the risk of death, as illustrated by the Coren and Halpren report. Rates cannot be calculated from proportionate mortality data. The individuals who did not die are not included in the collected data. From a symbolic point of view, only half of the 2×2 table necessary to estimate a rate is available or

	Exposed	Unexposed	**Total**
Died	x	$n–x$	n
Alive	?	?	?
Total	?	?	?

In terms of the right- and left-handed mortality data, the frequencies of surviving left- and right-handed individuals are not known. It is likely that the frequency of right-handedness is increased among the older individuals sampled, causing an apparent longer lifetime. Early in the 20th century, being left-handed was frequently treated as a kind of disability and many naturally left-handed children were trained to be right-handed. As the century progressed, this practice became less frequent. By the end of the 20th century, therefore, the frequency of right-handedness was relative higher among older individuals or conversely, relative lower among younger individuals. The death certificates collected by Coren and Halpren, therefore, likely contain older individuals who are disproportionately right-handed, increasing the observed life-time associated with right-handedness. The absence of the numbers of surviving right- and left-handed individuals makes it impossible to compare the risk of death between these two groups. More simply, to estimate a rate requires an estimate of the person-years at risk accumulated by those who die, as well as by those who survive. Similarly, to estimate the probability of death requires the number of deaths, as well as the number of those who did not die. A study measuring risk based on a sample including individuals who died and who survived allows a comparison between rates that accounts for the influence of the differing distributions of left- and right-handed individuals that is not possible in a proportionate mortality study.

2

TABULAR DATA: THE 2 × *k* TABLE AND SUMMARIZING 2 × 2 TABLES

THE 2 × *k* TABLE

The $2 \times k$ table is a summary description of a binary variable measured at k-levels of another variable. Of more importance, a complete description of the analysis of a $2 \times k$ table provides a valuable foundation for describing a number of general statistical issues that arise in the analysis of data contained in a table (continued in Chapter 13).

As noted (Chapter 1), an epidemiologic analysis frequently begins with classifying data into a 2×2 table. A natural extension is a $2 \times k$ table, in which the frequency of a binary variable, such as the presence/absence of a disease or case/control status, is recorded at k-levels of another variable. The resulting $2 \times k$ table can be viewed from the perspective of the k-level variable (risk factor) or from the perspective of the binary outcome variable (present/absent). In terms of analytic approaches, the two points of view are:

1. *Regression analysis*: What is the relationship between a k-level risk variable and a binary outcome?
2. *Two-sample analysis*: Does the mean value of the k-level risk variable differ between two sampled groups?

Table 2–1. Pancreatic cancer and coffee consumption among male cases and controls ($n = 523$)

	Coffee consumption (cups/days)				
	$X = 0$	$X = 1$	$X = 2$	$X \geq 3$	**Total**
$Y = 1$ (cases)	9	94	53	60	216
$Y = 0$ (controls)	32	119	74	82	307
Total	41	213	127	142	523
$\hat{P}_j$	0.220	0.441	0.417	0.423	0.413

These two questions generate different analytic approaches to describing the relationships within a $2 \times k$ table but, as will be seen, result in the same statistical assessment. Both approaches are distribution-free, in the sense that knowledge or assumptions about the distribution that generates the sampled data is not required. A number of texts completely develop the theory, analysis, and interpretation of discrete data classified into multiway tables (for example, [4], [5], and [6]).

Case/control data describing coffee consumption and its relationship to pancreatic cancer [7] provide an example of a $2 \times k$ table and its analysis. Men ($n = 523$) classified as cases (denoted $Y = 1$, row 1) or controls (denoted $Y = 0$, row 2) and by their reported consumption of 0, 1, 2, and 3, or more cups of coffee per day (denoted $X = j$, $j = 0, 1, 2,$ or 3, columns) produce a 2×4 table (Table 2–1). For simplicity, more than three cups consumed per day are considered as three, incurring a slight bias.

The analysis of a $2 \times k$ table consists of the answer to three basic questions. A general evaluation of the influence of the risk factor (X) on the binary status (Y) is the first task (Is coffee drinking associated in any way to case/control status?). Second, the kind of relationship between the k-level numeric variable and the binary outcome is explored (How does pancreatic cancer risk change as the amount of coffee consumed increases?). Third, the comparison of the mean value of the variable X among cases ($Y = 1$) to the mean value of the variable X among controls ($Y = 0$) indicates the magnitude of the risk/outcome association (Does the mean amount of coffee consumed differ between cases and controls?).

INDEPENDENCE/HOMOGENEITY

The general notation for a $2 \times k$ table containing n observations classified by two categorical variables (denoted X and Y) is displayed in Table 2–2 (summarized in a variety of forms at the end of this section).

The symbol n_{ij} represents the number of observations falling into both the i^{th} row and the j^{th} column; namely, the count in the $(i,\ j)^{th}$-cell. For the pancreatic

Table 2–2. Notation for a $2 \times k$ table

	$X=1$	$X=2$	$X=3$	$X=4$	. . .	$X=k$	**Total**
$Y=1$	n_{11}	n_{12}	n_{13}	n_{14}	. . .	n_{1k}	$n_{1.}$
$Y=0$	n_{21}	n_{22}	n_{23}	n_{24}	. . .	n_{2k}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	. . .	$n_{.k}$	n

Table 2–3. Notation for the probabilities underlying data classified into a $2 \times k$ table

	$X=1$	$X=2$	$X=3$	$X=4$	. . .	$X=k$	**Total**
$Y=1$	p_{11}	p_{12}	p_{13}	p_{14}	. . .	p_{1k}	q_1
$Y=0$	p_{21}	p_{22}	p_{23}	p_{24}	. . .	p_{2k}	q_2
Total	p_1	p_2	p_3	p_4	. . .	p_k	1.0

cancer data, the cell frequency n_{23} is 74 ($n_{23} = 74$, intersection of the second row and third column of Table 2–1). The *marginal frequencies* are the sums of the columns or the rows and are represented by $n_{.j}$ and $n_{i.}$, respectively. In symbols, the marginal frequencies are

$$n_{.j} = n_{1j} + n_{2j} \quad (k \text{ column sums})$$

and

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ik} = \sum n_{ij} \quad (\text{two row sums})$$

where $i = 1, 2 =$ number of rows and $j = 1, 2, \cdots, k =$ number of columns. For example, the total number of controls among the 523 study subjects (Table 2–1) is the marginal frequency of row 2 or $n_{2.} = 32 + 119 + 74 + 82 = 307$.

The properties of a $2 \times k$ table are determined by the relationships among three kinds of probabilities (Table 2–3). They are: the probability that an observation falls in a specific cell (denoted p_{ij}), the marginal probability that an observation falls in a specific column (denoted p_j, column $= j$), and the marginal probability that an observation falls in a specific row (denoted q_i, row $= i$).

The symbol n_{ij} represents an observed frequency subject to random variation, while the symbols p_{ij}, p_j, and q_i represent unobserved theoretical and fixed population probabilities.

Independence

The first issue to be addressed in the analysis of data classified into a $2 \times k$ table concerns the statistical independence of the row variable (Y) and the column variable (X). When categorical variables X and Y are unrelated, the cell probabilities are completely determined by the marginal probabilities (Table 2–4). Specifically, the probability that $Y = 1$ and $X = j$ simultaneously is then

$$p_{1j} = P(Y = 1 \text{ and } X = j) = P(Y = 1)P(X = j) = q_1 p_j \quad (\text{first row}),$$

Table 2–4. The notation for the expected values when the two categorical variables X and Y are statistically independent

	$X=1$	$X=2$	$X=3$	$X=4$	. . .	$X=k$	**Total**
$Y=1$	$nq_1\ p_1$	$nq_1\ p_2$	$nq_1\ p_3$	$nq_1\ p_4$	. . .	$nq_1 p_k$	nq_1
$Y=0$	$nq_2\ p_1$	$nq_2\ p_2$	$nq_2\ p_3$	$nq_2\ p_4$	. . .	$nq_2 p_k$	nq_2
Total	np_1	np_2	np_3	np_4	. . .	np_k	n

and similarly,

$$p_{2j} = P(Y=0 \text{ and } X=j) = P(Y=0)P(X=j) = q_2 p_j \quad \text{(second row)}$$

where $j = 1, 2, \cdots, k$.

The expected number of observations in the $(i,\ j)^{th}$-cell is $nq_i p_j$ when, to repeat, X and Y are statistically independent. Because the underlying cell probabilities (p_{ij}) are theoretical quantities (population parameters), these values are almost always estimated from the collected data. Under the conditions of independence, the probabilities p_{ij} are estimated from the marginal probabilities (denoted $\hat{p}_{ij}$) where

$$\hat{q}_i = \frac{n_{i.}}{n} \quad \text{and} \quad \hat{p}_j = \frac{n_{.j}}{n}$$

giving the estimate of p_{ij} as $\hat{p}_{ij} = \hat{q}_i \hat{p}_j$. Estimated cell counts follow because $\hat{n}_{ij} = n\hat{p}_{ij} = n\hat{q}_i\hat{p}_j$, which is usually written as $\hat{n}_{ij} = n_{i.}n_{.j}/n$. These expected values are calculated for each cell in the table and are compared to the observed values, typically using a chi-square test statistic to assess the conjecture that the categorical variables X and Y are unrelated.

For the pancreatic cancer case/control data, such expected cell frequencies ($\hat{n}_{ij}$) are given in Table 2–5. The marginal frequencies remain the same as those in the original table (Table 2.1). A consequence of statistical independence is that a table is unnecessary. The marginal probabilities exactly reflect the relationships within the row and column cell frequencies. For the case/control expected values (Table 2–5), the marginal row ratio is $216/307 = 0.706$ and is also the ratio between

Table 2–5. Estimated counts generated under the hypothesis of independence

	Coffee consumption (cups/day)				
	$X=0$	$X=1$	$X=2$	$X \geq 3$	**Total**
$Y=1$ (cases)	16.93	87.97	52.45	58.65	216
$Y=0$ (controls)	24.07	125.03	74.55	83.35	307
Total	41	213	127	142	523
P	0.413	0.413	0.413	0.413	0.413

the row frequencies in every column (Table 2–5). For example, in column 1, the ratio is $16.93/24.07 = 0.706$. Furthermore, the probabilities $P(Y = 1)$ are identical for each value of X ($\hat{P} = 0.413$, Table 2–5).

A Pearson chi-square test statistic effectively summarizes the correspondence between the hypothesis-generated expected values ($\hat{n}_{ij}$) and the observed data values (n_{ij}). The test statistic for the pancreatic cancer 2×4 table becomes

$$X^2 = \sum\sum \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum\sum \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$
$$= \frac{(9 - 16.93)^2}{16.93} + \cdots + \frac{(82 - 83.35)^2}{83.35} = 7.099$$

where, again, $i = 1, 2$ and $j = 1, 2, 3$, and 4. Thus, the chi-square statistic consists of the sum of eight comparisons, one for each cell in the table. The summary X^2 value has an approximate chi-square distribution with $k - 1$ degrees of freedom when the two categorical variables used to classify the observations into a $2 \times k$ table are unrelated. For the coffee/cancer data, the degrees of freedom are three, and the significance probability (p-value) is $P(X^2 \geq 7.099 | no\ X/Y\text{-}relationship) = 0.069$. A small p-value, say in the neighborhood of 0.05, indicates that the observed cell frequencies estimated as if coffee consumption (X) and case/control status (Y) were statistically independent do not correspond extremely well to the observed values (Table 2–5 compared to Table 2–1), presenting the possibility of an underlying systematic pattern of association.

Homogeneity

Alternatively, data classified into a $2 \times k$ table can be assessed for homogeneity. The primary issue is the consistency of k conditional probabilities (denoted P_j) where P_j represents the probability that $Y = 1$ for a specific level of X or, in symbols, $P_j = P(Y = 1|X = j) = p_{1j}/p_j$. A natural estimate of the probability P_j is the frequency in the first row of the j^{th} column (n_{1j}) divided by the total in the j^{th} column ($n_{j.}$) or $\hat{P}_j = n_{1j}/n_{.j}$. From the pancreatic cancer data, these estimates are: $\hat{P}_1 = 9/41 = 0.220$, $\hat{P}_2 = 94/213 = 0.441$, $\hat{P}_3 = 53/127 = 0.417$ and $\hat{P}_4 = 60/142 = 0.423$ (Table 2–1).

To begin, a hypothesis is imposed stating that the sampled data are classified into categories (columns) where the underlying probabilities P_j are identical regardless of the level of X. In symbols, homogeneity means

$$homogeneity\ hypothesis: \quad P_1 = P_2 = P_3 = \cdots = P_k = P,$$

or equivalently,

$$homogeneity\ hypothesis: \quad P(Y = 1|X = j) = P(Y = 1) = P.$$

The k estimated probabilities $\hat{P}_j$ then differ only because of sampling variation. To evaluate this conjecture, a single probability P (ignoring the column variable X) is estimated from the tabled data ($\hat{P} = n_{1.}/n$) and compared to each probability $\hat{P}_j$, estimated from each column of the tabled data ($\hat{P}_j = n_{1j}/n_{.j}$).

For the pancreatic cancer data, the value P is estimated by $\hat{P} = 216/523 = 0.413$. The variability among the estimates $\hat{P}_j$ relative to the estimate $\hat{P}$ reflects the extent of inequality among the column probabilities (homogeneity). A chi-square statistic summarizes this variation. Thus, the test statistic

$$X^2 = \sum \left[\frac{\hat{P}_j - \hat{P}}{\sqrt{variance(\hat{P}_j)}} \right]^2 = \frac{\sum n_{.j}(\hat{P}_j - \hat{P})^2}{\hat{P}(1-\hat{P})} \qquad j = 1,\ 2,\ \cdots,\ k$$

has an approximate chi-square distribution with $k-1$ degrees of freedom, when the variation among the estimates $\hat{P}_j$ is due only to chance. For the pancreatic cancer data, the value of the test statistic is again $X^2 = 7.099$.

Notice that the chi-square statistic is a measure of the variability among a series of estimated values. The test statistic X^2, therefore, provides an assessment (p-value) of the likelihood that this variability is due entirely to chance. The purpose of calculating a chi-square statistic is frequently to identify nonrandom variation among a series of estimated values.

It is not a coincidence that the chi-square value to evaluate independence and the chi-square value to evaluate homogeneity are identical. A little algebra shows that these two apparently different assessments produce identical summary values. If P_j is constant for all levels of X, then the variable Y is not influenced by the variable X, which is another way of saying that X and Y are independent ($P = P_j$, Table 2–5). In symbols, the relationship between homogeneity and independence is

$$\begin{aligned}
&homogeneity = P(Y = i|X = j) = P(Y = i) \quad \text{implies that} \\
&P(Y = i|X = j)P(X = j) = P(Y = i)P(X = j) \quad \text{and, therefore,} \\
&P(Y = i \text{ and } X = j) = P(Y = i)P(X = j) = independence,
\end{aligned}$$

which is the relationship that generates the expected values for the chi-square test of independence.

REGRESSION

A chi-square test statistic used to assess independence does not require the categorical variables to be numeric or even ordered. Considerable gains (for

example, increased statistical power) are achieved by forming and testing specific hypotheses about relationships within a table. One such opportunity arises when the k-level categorical variables are numeric. In this setting, the key to a specific statistical hypothesis becomes k pairs of data values. The k categories of X produce the k pairs of observations $(x_j, \hat{P}_j)$ where, as before, $\hat{P}_j$ represents the estimated conditional probability that $Y = 1$ associated with each numeric value represented by x_j (categories = columns). The pancreatic cancer case/control data yield four pairs of observations. These x/y-pairs are (0, 0.220), (1, 0.441), (2, 0.417) and (3, 0.423) (Table 1–2 and Figure 2–1, circles). The estimates $\hat{P}_j$ are again simply the proportion of cases within each level of coffee consumption (x_j).

One approach to summarizing these k proportions is to estimate a straight line based on the k pairs $(x_j, \hat{P}_j)$ and to use the slope of the estimated line to reflect the strength of the relationship between the row and column categorical variables. Identical to linear regression analysis applied to a continuous variable, three quantities are necessary to estimate this line: the sum of squares of X (S_{XX}), the sum of squares of Y (S_{YY}), and the sum of cross-products of X and Y (S_{XY}).

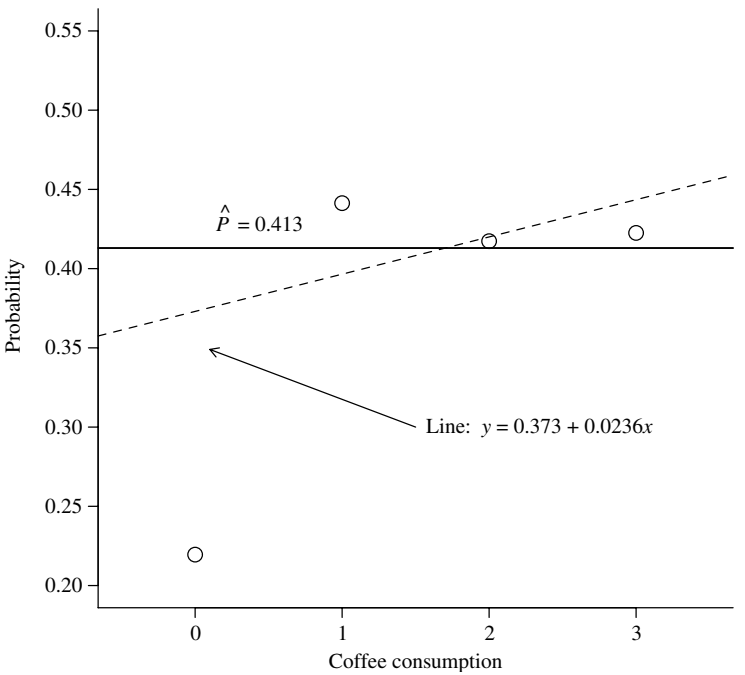


Figure 2–1. Proportion of cases of pancreatic cancer by consumption of 0, 1, 2 and 3 or more cups of coffee per day (circles).

These quantities calculated, from a $2 \times k$ table, are:

$$S_{XX} = \sum n_{.j}(x_j - \overline{x})^2 \quad \text{where} \quad \overline{x} = \sum n_{.j}x_j/n,$$
$$S_{YY} = n_{1.}n_{2.}/n, \quad \text{and}$$
$$S_{XY} = (\overline{x}_1 - \overline{x}_2)\, S_{YY} \text{ where } \quad \overline{x}_i = \sum n_{ij}x_j/n_{i.}$$

for $i = 1,\ 2$ and $j = 1,\ 2,\ \cdots,\ k$. The quantity $\overline{x}_1$ represents the mean of the X-values when $Y = 1$ based on $n_{1.}$ observations (row 1), $\overline{x}_2$ represents the mean of the X-values when $Y = 0$ based on $n_{2.}$ observations (row 2) and $n = n_{1.} + n_{2.}$ represents the total number of observations. It is not necessary to consider the data in terms of a $2 \times k$ table. The mean values and the sums of squares are the same when the data are viewed as n pairs of values $(x_i,\ y_i)$ where the table frequencies are the numbers of identical x/y-pairs (Table 2–1). From either calculation, these values for the pancreatic cancer case/control data are: $\overline{x}_1 = 1.759$, $\overline{x}_2 = 1.671$, $S_{XX} = 474.241$, $S_{YY} = 126.792$, and $S_{XY} = 11.189$.

The ordinary least-squares linear regression estimates of the slope and intercept of a summary line are given by

$$\hat{b} = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad \hat{a} = \hat{P} - \hat{b}\overline{x},$$

and the estimated line becomes $\tilde{P}_j = \hat{a} + \hat{b}x_j$.

For the pancreatic data, the estimated slope is $\hat{b} = 11.189/474.241 = 0.0236$ and intercept is $\hat{a} = \hat{P} - \hat{b}\overline{x} = 0.413 - 0.0236(1.707) = 0.373$, making the estimated regression line $\tilde{P}_j = 0.373 + 0.0236x_j$. The value $\hat{P} = 216/523 = 0.413$ is again the proportion of cases when the coffee consumption is ignored. Therefore, as before, the estimates $\hat{P}_j$ from the data are

$$\hat{P}_1 = 0.220,\ \hat{P}_2 = 0.441,\ \hat{P}_3 = 0.417,\ \hat{P}_4 = 0.423,\ \text{and}\ \ \hat{P} = 0.413.$$

The corresponding estimates $\tilde{P}_j$ from the summary straight line are

$$\tilde{P}_1 = 0.373,\ \tilde{P}_2 = 0.396,\ \tilde{P}_3 = 0.420,\ \tilde{P}_4 = 0.443,\ \text{and}\ \ \tilde{P} = 0.413.$$

An estimate of the variance of the distribution of the estimated slope $\hat{b}$ is

$$\textit{variance}(\hat{b}) = \frac{S_{YY}}{nS_{XX}} = \frac{126.792}{523(474.241)} = 0.000512.$$

To evaluate the influence of random variation on the estimated slope, a chi-square test statistic is typically used. The ratio of the squared estimated slope to its