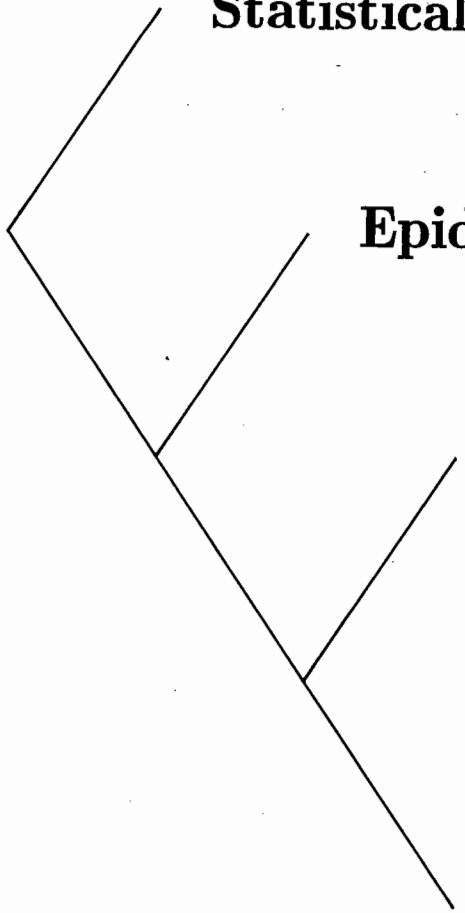
Statistical Models in Epidemiology



David Clayton
Medical Research Council, Cambridge

and

Michael Hills
London School of Hygiene and Tropical Medicine

OXFORD • NEW YORK • TOKYO
OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland Madrid
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press Inc., New York

First published 1993
Reprinted 1994 (twice)

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Clayton, David, statistician.
Statistical models in epidemiology/David Clayton and Michael Hills.
Includes bibliographical references and index.
1. Epidemiology—Statistical methods. I. Hills, Michael. II. Title.
[DNLM: 1. Epidemiology. 2. Models, Statistical. WA 105 C622s 1993]
RA652.2M8C53 1993 614.4'072—dc20 93-19448

ISBN 0 19 852221 5 (Hbk)

Printed in Great Britain by
Biddles Ltd, Guildford & King's Lynn

Preface

The aim of this book is to give a self-contained account of the statistical basis of epidemiology. The book is intended primarily for students enrolled for a masters degree in epidemiology, clinical epidemiology, or biostatistics, and should be suitable both as the basis for a taught course and for private study.

Although we anticipate that most readers will have taken a first course in statistics, no previous knowledge is assumed, and the mathematical level of the book has been chosen to suit readers whose basic training is in biology. Some of the material in the book could be omitted at first reading, either because it is rather more demanding of mathematical skills or because it deals with rather specialized points. We have been careful to gather such material either into complete chapters or complete sections and to indicate these with a marginal symbol, as here. ⋆

Epidemiologists today have ready access to computer programs of great generality, but to use these sensibly and productively it is necessary to understand the ideas which lie behind them. The most important of these is the idea of a *probability model*. All statistical analysis of data is based on probability models, even though the models may not be explicit. Only by fully understanding the model can one fully understand the analysis.

Models depend on parameters, and values must be chosen for these parameters in order to match the model to the data. In showing how this is done we have chosen to emphasize the role of likelihood because this offers an approach to statistics which is both simple and intuitively satisfying. An additional advantage of this approach is that it requires the model and its parameters to be made explicit, even in the simplest situations. More complex problems can then be tackled by natural extensions of simple methods and do not require a whole new way of looking at things.

Most of the material in this book was developed during successive residential summer courses in epidemiology and statistics, held in Florence under the auspices of the European Educational Programme in Epidemiology. We are grateful to the International Agency for Cancer Research, the Regional Office for Europe of the World Health Organization, the Commission of the European Communities, and the Tuscany Regional Government, for sponsoring the program, and to Walter Davies, Organizing Secretary, and Rodolfo Saracci, Course Director, whose respective skills ensured that the course took place each year. We also acknowledge with thanks helpful

comments on earlier drafts from Damien Jolley, Bendix Carstensen, Dave Leon, and Nick Hills.

Cambridge
London
February 1993

David Clayton
Michael Hills

Dedication

To the students of the Florence course, 1988 – 92, without whose help and encouragement this book would never have appeared.

Contents

★ Denotes a chapter which could be omitted from a first course.

Part I

Probability models and likelihood

1
Probability models

1.1 Observation, experiments and models

Science proceeds by endless repetition of a three-stage process,

1. observation;
2. building a model to describe (or 'explain') the observations; and
3. using the model to predict future observations. If future observations are not in accord with the predictions, the model must be replaced or refined.

In quantitative science, the models used are mathematical models. They fall into two main groups, *deterministic* models and probability (or *stochastic*) models. It is the latter which are appropriate in epidemiology, but the former are more familiar to most scientists and serve to introduce some important ideas.

DETERMINISTIC MODELS

The most familiar examples of deterministic models are the laws of classical physics. We choose as a familiar example *Ohm's law*, which applies to the relationship between electrical potential (or voltage), V, applied across a conductor and the current flowing, I. The law holds that there is a strict proportionality between the two — if the potential is doubled then the current will double. This relationship is represented graphically in Fig. 1.1.

Ohm's law holds for a wide range of conductors, and simply states that the line in Fig. 1.1 is straight; it says nothing about the gradient of the line. This will differ from one conductor to another and depends on the resistance of the conductor. Without knowing the resistance it will not be possible to predict the current which will flow in any *particular* conductor. Physicists normally denote the resistance by R and write the relationship as

$$I = \frac{V}{R}.$$

However, R is a different sort of quantity from V or I. It is a *parameter* — a number which we must fix in order to apply the general law to a specific case. Statisticians are careful to differentiate between observable variables

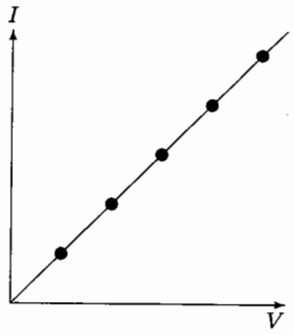


Fig. 1.1. A deterministic model: Ohm's law.

(such as V and I) and parameters (such as R) and use Greek letters for the latter. Thus, if Ohm were a modern statistician he would write his law as

$$I = \frac{V}{\rho}$$

In this form it is now clear that ρ, the resistance, is a parameter of a simple mathematical model which relates current to potential. Alternatively, he could write the law as

$$I = \gamma V$$

where γ is the conductance (the inverse of the resistance). This is a simple example of a process called *reparametrization* — writing the model differently so that the parameters take on different meanings.

STOCHASTIC MODELS

Unfortunately the phenomena studied by scientists are rarely as predictable as is implied by Fig. 1.1. In the presence of measurement errors and uncontrolled variability of experimental conditions it might be that real data look more like Fig. 1.2. In these circumstances we would not be in a position to predict a future observation with certainty, nor would we be able to give a definitive estimate of the resistance parameter. It is necessary to extend the deterministic model so that we can predict a range of more probable future observations, and indicate the uncertainty in the estimate of the resistance.

Problems such as this prompted the mathematician Gauss to develop his *theory of errors*, based on the Gaussian distribution (often also called the *Normal* distribution), which is the most important probability model for these problems. A very large part of statistical theory is concerned with this model and most elementary statistical texts reflect this. Epidemiology,

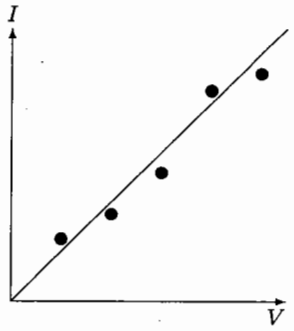


Fig. 1.2. Experimental/observational errors.

however, is more concerned with the occurrence (or not) of certain events in the natural history of disease. Since these occurrences cannot be described purely deterministically, probability models are also necessary here, but it is the models of Bernoulli and Poisson which are more relevant. The remainder of this chapter discusses a particularly important type of data generated by epidemiological studies, and the nature of the models we use in its analysis.

1.2 Binary data

Many epidemiological studies generate data in which the response measurement for each subject may take one of only two possible values. Such a response is called a *binary* response. Two rather different types of study generate such data.

COHORT STUDIES WITH FIXED FOLLOW-UP TIME

In a *cohort* study a group of people are followed through some period of time in order to study the occurrence (or not) of a certain event of interest. The simplest case is a study of *mortality* (from any cause). Clearly, there are only two possible outcomes for a subject followed, say, for five years — death or survival.

More usually, it is only death from a specified cause or causes which is of interest. Although there are now three possible outcomes for any subject — death from the cause of interest, death from another cause, or survival — such data are usually dealt with as binary data. The response is taken as death from cause of interest as against survival, death from other causes being treated as premature termination of follow-up. Premature termination of follow-up is a common feature of epidemiological and clinical follow-up studies and may occur for many reasons. It is called *censoring*, a word which reflects the fact that it is the underlying binary response which

we would have liked to observe, were it not for the removal of the subject from observation.

In *incidence studies* the event of interest is new occurrence of a specified disease. Again our interest is in the binary response (whether the disease occurred or not) although other events may intervene to censor our observation of it.

For greater generality, we shall use the word *failure* as a generic term for the event of interest, whether incidence, mortality, or some other (undesirable) outcome. We shall refer to non-failure as *survival*. In the simplest case, we study N subjects, each one being followed for a fixed time interval, such as five years. Over this time we observe D failures, so that $N - D$ survive. We shall develop methods for dealing with censoring in later chapters.

CROSS-SECTIONAL PREVALENCE DATA

Prevalence studies have considerable importance in assessing needs for health services, and may also provide indirect evidence for differences in incidence. They have the considerable merit of being relatively cheap to carry out since there is no follow-up of the study group over time. Subjects are simply categorized as affected or not affected, according to agreed clinical criteria, at some fixed point in time. In a simple study, we might observe N subjects and classify D of them as affected. An important example is serological studies in infectious-disease epidemiology, in which subjects are classified as being seropositive or seronegative for a specified infection.

1.3 The binary probability model

The obvious analysis of our simple binary data consisting of D failures out of N subjects observed is to compute the proportion failing, D/N. However, knowing the proportion of a cohort which develops a disease, or dies from a given cause, is of little use unless it can be assumed to have a wider applicability beyond the cohort. It is in making this passage from the particular to the general that statistical models come in. One way of looking at the problem is as an attempt to predict the outcome for a new subject, similar to the subjects in the cohort, but whose outcome is unknown. Since the outcome for this new subject cannot be predicted with certainty the prediction must take the form of *probabilities* attached to the two possible outcomes. This is the *binary probability model*. It is the simplest of all probability models and, for the present, we need to know nothing of the properties of probability save that probabilities are numbers lying in the range 0 to 1, with 0 representing an impossible outcome and 1 representing a certain outcome, and that the probability of occurrence of either one of two distinct outcomes is the sum of their individual probabilities (the *additive* rule of probability).

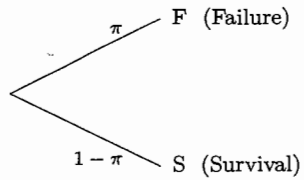


Fig. 1.3. The binary probability model.

THE RISK PARAMETER

The binary probability model is illustrated in Figure 1.3. The two outcomes are labelled F (failure) and S (survival). The model has one *parameter*, π, the probability of failure. Because the subject must either fail or survive, the sum of the probabilities of these two outcomes must be 1, so the probability of survival is $1-\pi$. In the context where π represents the probability of occurrence of an event in a specified time period, it is usually called the *risk*.

THE ODDS PARAMETER

An important alternative way of parametrizing the binary probability model is in terms of the *odds* of failure versus survival. These are

$$\pi : (1 - \pi),$$

which may also be written as

$$\frac{\pi}{1-\pi} : 1.$$

It is convenient to omit the : 1 in the above expression and to measure the odds by the fraction

$$\frac{\pi}{1-\pi}.$$

This explains why, although the word odds is plural, there is often only one number which measures the odds.

Exercise 1.1. Calculate the odds of F to S when the probability of failure is (a) 0.75, (b) 0.50, (c) 0.25.

In general the relationship between a probability π and the corresponding odds Ω is

$$\Omega = \frac{\pi}{(1-\pi)}.$$

This can be inverted to give

$$\pi = \frac{\Omega}{1+\Omega}, \quad 1-\pi = \frac{1}{1+\Omega}.$$

Exercise 1.2. Calculate the probability of failure when Ω, the odds of F to S is (a) 0.3, (b) 3.0.

RARE EVENTS

In this book we shall be particularly concerned with *rare events*, that is, events with a small probability, π, of occurrence in the time period of interest. In this case $(1-\pi)$ is very close to 1 and the odds parameter and the risk parameter are nearly equal:

$$\Omega \approx \pi.$$

This approximation is often called the *rare disease assumption*, but this is a misleading term, since even the common cold has a small probability of occurrence within, say, a one-week time interval.

1.4 Parameter estimation

Without giving a value to the parameter π, this model is of no use for prediction. Our next problem is to use our observed data to estimate its value. It might seem obvious to the reader that we should estimate π by the proportion of failures, D/N. This corresponds to estimating the odds parameter Ω by $D/(N-D)$, the ratio of failures to survivors.

It might also seem obvious that we should place more reliance on our estimate (and upon any predictions based on it) if N is 1000 than if N is 10. The formal statistical theory which provides a quantitative justification for these intuitions will be discussed in later chapters.

1.5 Is the model true?

A model which states that every one of a group of patients has the same probability of surviving five years will seem implausible to most clinicians. Indeed, the use of such models by statisticians is a major reason why some practitioners, brought up to think of each patient as unique, part company with the subject!

The question of whether scientific models are *true* is not however, a sensible one. Instead, we should ask ourselves whether our model is *useful* in describing past observations and predicting future ones. Where there remains a choice of models, we must be guided by the criterion of *simplicity*. In epidemiology probability models are used to describe past observations of disease events in study cohorts and to make predictions for future individuals. If we have no further data which allows us to differentiate subjects

in the cohort from one another or from a future individual, we have no option save to assign the same probability of failure to each subject. Further data allows elaboration of the model. For example, if we can identify subjects as exposed or unexposed to some environmental influence, the model can be extended to assign different probabilities to exposed and unexposed subjects. If additionally we know the level of exposure we can extend the model by letting the probability of failure be some increasing function of exposure.

In this book we shall demonstrate the manner in which more complicated models may be developed to deal with more detailed data. The binary model has been our starting point since it is the basic building brick from which more elaborate models are constructed.

Solutions to the exercises

1.1 (a) Odds $= 0.75/0.25 = 3$.
(b) Odds $= 0.50/0.50 = 1$.
(c) Odds $= 0.25/0.75 = 0.3333$.

1.2 (a) Probability $= 0.3/1.3 = 0.2308$.
(b) Probability $= 3/4 = 0.75$.

2
Conditional probability models

In this chapter we introduce the idea of *conditional probability*, which allows us to extend the binary model so that the probability of failure can depend on earlier events. The natural way of thinking about conditional probabilities is in terms of a tree diagram. These diagrams are used extensively throughout the book.

2.1 Conditional probability

Suppose a binary probability model assigns a probability to a subject's death during some future time period. It may be that this prediction would be better if we knew the subject's smoking habits. This would be the case if the probability of death for a smoker were 0.015 but only 0.005 for a non-smoker. These probabilities are called *conditional* probabilities; they are the probabilities of death conditional on being a smoker and a non-smoker respectively. Epidemiology is mainly concerned with conditional probability models that relate occurrence of some disease event, which we call failure, to events which precede it. These include potential causes, which we call *exposures*.

When subjects are classified as either exposed (E+) or not exposed (E−), the conditional probability model can be represented as a tree with 6 branches. The first two branches refer to E+ and E−; then there are two referring to failure and survival if the subject is exposed, and two referring to failure and survival if the subject is not exposed. An example is shown in Fig. 2.1. The tips of the tree correspond to the four possible combinations of exposure and outcome for any subject.

The probabilities on the first two branches of the tree refer to the probability that a subject is exposed and the probability that a subject is not exposed. Using the smoking example we have taken these to be 0.4 and 0.6. The probabilities in the next two pairs of branches are conditional probabilities. These are 0.015 (F) and 0.985 (S) if a subject is exposed (smokes), and 0.005 (F) and 0.995 (S) if a subject is not exposed (does not smoke).

The probability of any combination of exposure and outcome is obtained by multiplying the probabilities along the branches leading to the

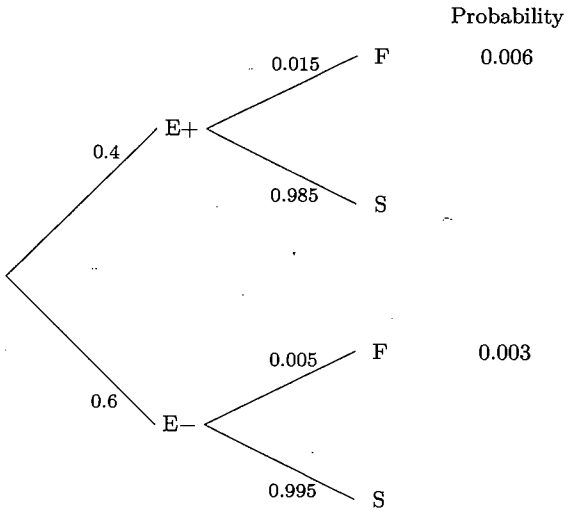


Fig. 2.1. A conditional probability tree.

tip which corresponds to that combination. For example, the probability that a subject is exposed and fails is

$$0.4 \times 0.015 = 0.006,$$

and the probability that a subject is not exposed and fails is

$$0.6 \times 0.005 = 0.003.$$

This is called the multiplicative rule.

Exercise 2.1. Calculate the probabilities for each of the remaining 2 possibilities. What is the overall probability of failure regardless of exposure?

This overall probability is usually called the *marginal* probability of failure.

STATISTICAL DEPENDENCE AND INDEPENDENCE

Fig. 2.1 illustrates a model in which the probability of failure differs according to whether an individual was exposed or not. In this case, exposure and failure are said to be *statistically dependent*. If the probability of failure is the same, whether or not the subject is exposed, then exposure and failure are said to be *statistically independent*.

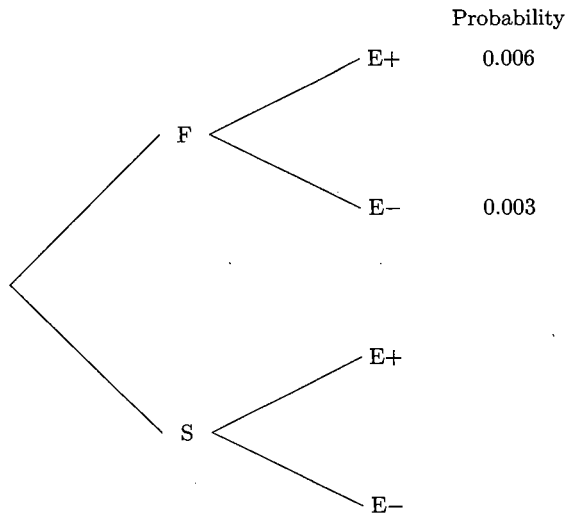


Fig. 2.2. Predicting exposure from the outcome.

2.2 Changing the conditioning: Bayes' rule

The additive and multiplicative rules are the basic building blocks of probability models. A simple application of these rules allows us to change the direction of prediction so that, for example, a model for the probability of failure given exposure can be transformed into a model for the probability of exposure given failure.

We shall demonstrate this by using the tree in Fig. 2.1, where the first level of branching refers to exposure and the second to outcome. This is turned round in Fig. 2.2, so that the first level of branching now refers to outcome and the second to exposure. The probabilities of the different combinations of exposure and outcome are the same whichever way the tree is written; our problem is to fill in the probabilities on the branches of this new tree.

Working backwards from the tips of the tree, the probability of failure regardless of exposure is $0.006 + 0.003 = 0.009$. This is the probability for the first branch of the tree to F. Since the probability corresponding to any tip of the tree is obtained by multiplying the probabilities in the branches that lead to the tip, it follows that the probability in the branch from F to E+, for example, is $0.006/0.009 = 0.667$. This is the conditional probability of being exposed given the outcome was failure. This process of reversing the order of the conditioning is called Bayes' rule, after Thomas Bayes.

Exercise 2.2. Calculate the remaining conditional probabilities.

The following exercise, inspired by problems in screening, demonstrates one of the many uses of Bayes' rule.

Exercise 2.3. A screening test has a probability of 0.90 of being positive in true cases of a disease (the *sensitivity*) and a probability of 0.995 of being negative in people without the disease (the *specificity*). The prevalence of the disease is 0.001 so before carrying out the test, the probability that a person has the disease is 0.001.
(a) Draw a probability tree in which the first level of branching refers to having the disease or not, and the second level to being positive or negative on the screening test. Fill in the probabilities for each of the branches and calculate the probabilities for the four possible combinations of disease and test.
(b) Draw the tree the other way, so that the first level of branching refers to being positive or negative on the screening test and the second level to having the disease or not. Fill in the probabilities for the branches of this tree. What is the probability of a person having the disease given that they have a positive test result? (This is called the *positive predictive value.*)

2.3 An example from genetics ★

Our next exercises illustrate a problem in genetic epidemiology. For a specified genetic system (such as the HLA system), each person's *genotype* consists of two *haplotypes*,* one inherited from the mother and one from the father. If a mother has haplotypes (a,b), then one of these is passed to the offspring with probability 0.5. Likewise for a father's haplotypes, (c,d) say. Fig. 2.3 shows the probability tree for the genotype of the offspring. The presence of haplotype (a) carries a probability of disease of 0.05 while, in its absence, the probability is only 0.01.

Exercise 2.4. Work out the probabilities for the four tips of the probability tree which end in disease (F). Hence work out the probabilities of the four possible genotypes conditional on the fact that the offspring is affected by disease (Fig. 2.4).

Exercise 2.5. In practice the probabilities of disease conditional upon genotype are not known constants but unknown parameters. Repeat the previous exercise *algebraically*, replacing the probabilities 0.01 and 0.05 by π and $\theta\pi$ respectively. How are the conditional probabilities changed if the subject's father has genotype (c,c)?

The parameter θ, described in Exercise 2.5, is a *risk ratio*,

$$\theta = \frac{\text{Risk of disease if haplotype (a) present}}{\text{Risk of disease if haplotype (a) absent}}.$$

*The word haplotype refers to a group of genetic loci which are closely linked and therefore inherited together.

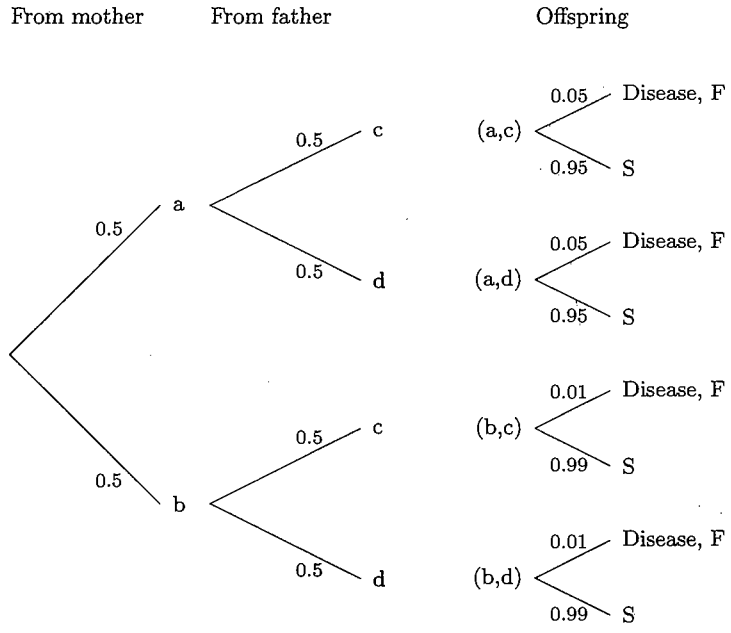


Fig. 2.3. Disease conditional upon inheritance.

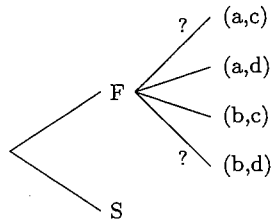


Fig. 2.4. Inheritance conditional upon disease.

It measures the strength of statistical dependence (or *association*) between the presence of haplotype (a) and occurrence of disease. The above exercise shows that the conditional probability of genotype given the presence of disease and parental genotypes depends only on this risk ratio.

Solutions to the exercises

2.1

$$\begin{aligned} \Pr(\text{E+ and S}) &= 0.4 \times 0.985 = 0.394 \\ \Pr(\text{E- and S}) &= 0.6 \times 0.995 = 0.597 \end{aligned}$$

The overall probability of failure is $0.006 + 0.003 = 0.009$.

2.2 See Fig. 2.5. The conditional probabilities of E+ and E− given survival are

$$\frac{0.394}{0.991} = 0.3976, \qquad \frac{0.597}{0.991} = 0.6024.$$

2.3 (a) See Fig. 2.6.
(b) See Fig. 2.7. The probability of disease given a positive test result is

$$\frac{0.0009}{0.005895} = 0.1527.$$

Note that this is much lower than 0.90, the sensitivity of the test. The remaining conditional probabilities are calculated in a similar manner.

2.4 The probabilities for each of the four tips are obtained by multiplying along the branches of the tree. The sum of the four probabilities is 0.0300. The *conditional* probabilities sum to 1.0.

Genotype	Disease	Probability	Conditional prob.
(a,c)	F	$0.5 \times 0.5 \times 0.05 = 0.0125$	$0.0125/0.03 = 0.417$
(a,d)	F	$0.5 \times 0.5 \times 0.05 = 0.0125$	0.417
(b,c)	F	$0.5 \times 0.5 \times 0.01 = 0.0025$	$0.0025/0.03 = 0.083$
(b,d)	F	$0.5 \times 0.5 \times 0.01 = 0.0025$	0.083
Total		0.0300	1.0

2.5 Repeating the above calculations algebraically yields:

Genotype	Disease	Probability	Conditional Prob.
(a,c)	F	$0.5 \times 0.5 \times \theta\pi = 0.25\theta\pi$	$\theta/(2\theta+2)$
(a,d)	F	$0.5 \times 0.5 \times \theta\pi = 0.25\theta\pi$	$\theta/(2\theta+2)$
(b,c)	F	$0.5 \times 0.5 \times \pi = 0.25\pi$	$1/(2\theta+2)$
(b,d)	F	$0.5 \times 0.5 \times \pi = 0.25\pi$	$1/(2\theta+2)$
Total		$0.25\pi(2\theta+2)$	1.0

If the father has genotype (c,c) then he can only pass on (c) and the possible genotypes of offspring are (a,c) and (b,c). Prior to observation of disease presence, these both have probabilities 0.5. Thus, for a subject known to have disease, we have

Genotype	Disease	Probability	Conditional Prob.
(a,c)	F	$0.5 \times \theta\pi = 0.5\theta\pi$	$\theta/(\theta+1)$
(b,c)	F	$0.5 \times \pi = 0.5\pi$	$1/(\theta+1)$
Total		$0.5\pi(\theta+1)$	1.0

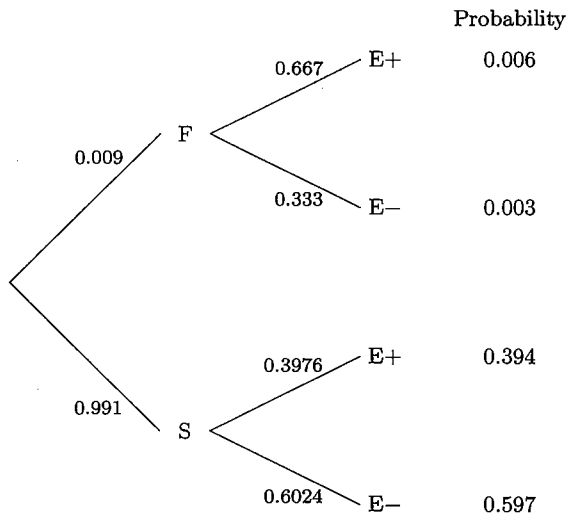


Fig. 2.5. Probability tree for exposure given outcome.

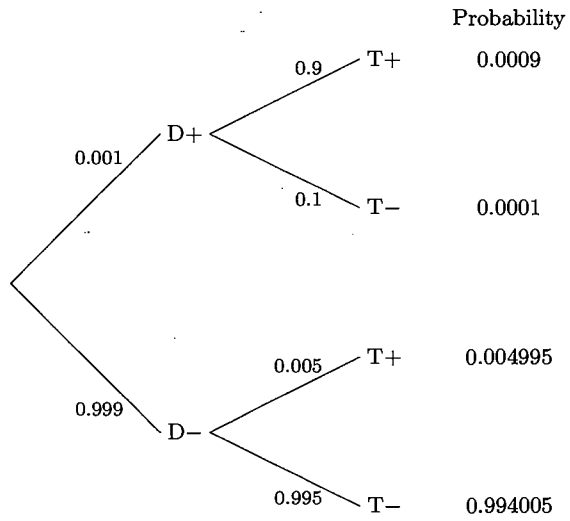


Fig. 2.6. Test results, T, given disease status, D.

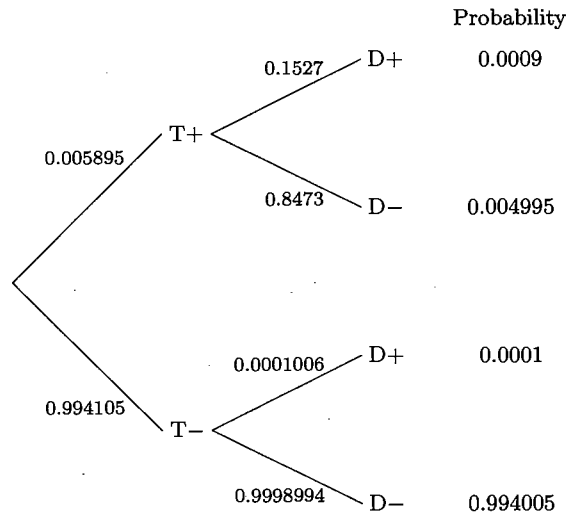


Fig. 2.7. Disease status given test results.

1 Probability models

1.1 Observation, experiments and models

Stochastic Models[1]

Normal vs *Bernoulli* and *Poisson*: We need to distinguish between *individual* observations, governed by Bernoulli and Poisson (or if quantitative rather than all-or-none or a count, Normal) and *statistics* formed by aggregation of individual observations. If a large enough number of individual independent but non-Gaussian observations are used to form a statistic, its (sampling) distribution can be described by a Gaussian (Normal) probability model. So, ultimately, the Normal probability model is very relevant.

1.1.1 Epidemiologic [subject-matter] models [JH]

We need to also make a distinction between the quantity(quantities) that is(are) of substantive interest or concern, the data from which this(these) is(are) estimated, the *statistical* models used to get to the the quantity(quantities) and the relationships of interest.

For example, of medical, public health or personal interest/concern might be the [list compiled some years ago, add your own for 2021 onwards]

- level of use of cell phones among drivers
- average and range [across people] of reductions in cholesterol with regular use of a cholesterol-lowering medication.
- amount of time taken by health care personnel to decipher the handwriting of other health care personnel.
- (average) number of times people have to phone to reach a 'live' person.
- reduction in one's risk of dying of a specific cancer if one is regularly screened for it.
- appropriate-size tracheostomy tube for an obese patient, based on easily easily obtained anthropometric measurements.
- length of central venous catheter that can be safely inserted into a child as a function of the child's height etc.
- rate of automobile accidents as a function of drivers' blood levels of alcohol and other drugs, numbers of persons in the car, cell-phone and other activities, weather, road conditions, etc.
- Psychological Stress, Negative Life Events, Perceived Stress, Negative Affect Smoking, Alcohol Consumption and Susceptibility to the Common Cold.
- The force of mortality as a function of age, sex and calendar time.
- Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction.
- Are seat belt restraints as effective in school age children as in adults?
- Levels of folic acid to add to flour, so that most people have sufficiently high blood levels, but birth defects are reduced.
- Early diet in children born preterm and their IQ at age eight.
- Prevalence of Down's syndrome in relation to parity and maternal age.

Of broader interest/concern might be

- the wind chill factor as a function of temperature and wind speed.
- how many fewer Florida votes Al Gore got in 2000 US Presidential because of a badly laid-out ballot.
- a formula for deriving one's "ideal" weight from one's height.
- yearly costs under different cell-phone plans.
- yearly maintenance costs for different makes and models of cars.
- car or life insurance premiums as a function of ...
- cost per foot^2 of commercial or business rental space as a function of ...
- Rapid Changes in Flowering Time in British Plants.
- How much money the City of New York should revover from Brink's for the losses the City incurred by the criminal activities of two Brink's employees (they collected the money from the parking meters, but kept some of it!).

[1] 'Stochastic' http://www.allwords.com/word-stochastic.html French: stochastique(fr) German: stochastisch(de). Etymology: From Ancient Greek (polytonic,), from (polytonic,) "aim at a target, guess", from (polytonic,) "an aim, a guess". Parzen, in his text on Stochastic Processes .. page 7 says: <<The word is of Greek origin; see Hagstroem (1940) for a study of the history of the word. In seventeenth century English, the word "stochastic" had the meaning "to conjecture, to aim at a mark." It is not clear how it acquired the meaning it has today of "pertaining to chance." Many writers use the expression "chance process" or "random process" as synonyms for "stochastic process." >>

1.1.2 From behaviour of statistical 'atoms' to statistical 'molecules'

'1 condition' or '1 circumstance' or 'setting' ["1-sample problems"]

The smallest statistical element or unit (? atom): the quantity of interest might have a Y distribution that under sampling, could be represented by a discrete random variable with '2-point' support (Bernoulli), 3-point support, $k-$point support, etc. or interval support (Normal, gamma, beta, log-normal,)

The *aggregate* or summary of the values associated with these elements is often a sum or a count: with e.g., a Binomial, Negative Binomial, gamma distribution. Or the summary might be more complex – it could be some re-arrangement of the data on the individuals (e.g., the way the tumbler longevity data were summarized). This brings in the notion of "sufficient statistics".

More complex: t, F, ...

'2 or conditions' or 'circumstances' or 'settings', indexed by possible values of 'X' variable(s). Think of the 'X' variable(s) as 'covariate patterns' or 'profiles,' not as a 'random' variable.

Unknown conditions or circumstances: Sometimes we don't have a measurable (or measured) 'X' variable(s) to explain the differences in Y from say family to family or person to person. There instead of the usual multiple regression approach, we use a hierarchical or random-effects or latent class or mixture model. [case in point]: JH's numbers of steps, 2017-2020]

1.2 Binary data

It is worth recalling from bios601 in earlier years [2] the concepts of **states** and **events** (transitions from one state to another).

COHORT STUDIES WITH FIXED FOLLOW-UP TIME

Recall: *cohort* is another name for a closed population, with membership (entry) defined by some event, such as birth, losing one's virginity, obtaining one's first driver's permit, attaining age 21, graduating from university, entering the 'ever-married' state, undergoing a certain medical intervention, enrolling in a follow-up study, etc. Then the *event of interest* is the *exit* (transition) from a/the state that prevailed at entry. So *death* is the transition from the *living* state to the *dead* state, receiving a *diagnosis* of cancer changes one's state from 'no history of cancer since entry/birth' to 'have a history of cancer', a conviction for traffic offense changes one's state from 'clean record' to 'have a history of traffic offences.' We can also envision more complex situations, with a transition from 'never had a stroke,' to 'have had 1 stroke,' to 'have had 2 strokes,' ... or 'haven't yet had a cold this winter,' to 'have had 1 cold,' to 'have had 2 colds,' etc.

Censoring: to be distinguished from *truncation*. Truncation implies some observations are missed by the data-gathering process, i.e., that the observed distribution is a systematic distortion of the true distribution. Note that we can have censoring of *any* quantitative variable, not just one that measures the *time duration* until the event of interest. For example, the limits on say a thermometer or a weight scale or a chemical assay may mean that it cannot record/detect values below or above these limits. Also, the example in C&H implicitly refers to *right* censoring: one can have *left* censoring, as with lower limits of detection in a chemical assay, or *interval* censoring, as – in repeated cross-sectional examinations – with the date of sero-conversion to HIV.

Incidence studies: the word *new* means a change of state since entry.

"*Failure*": It is a pity that C&H didn't go one step more and use the even more generic term "*event*". That way, they would not have to think of graduating with a PhD (i.e., *getting out of – exiting from – here*) as "*failure*" and "still pursuing one" as "*survival.*" This simpler and more general terminology would mean that we would not have to struggle to find a suitable label of the 'y' axis of the $1 - F(t)$, usually called $S(t)$, function. One could simply say *"proportion still in initial state,"* and substitute the term for the initial state, i.e., proportion still in PhD program, proportion event-free, etc.

N or n? D or d? JH would have preferred lower case, at least for the denominator. In *sampling* textbooks, N usually denotes the *population* size, and n the *sample* size. In the style manual used in *social sciences*, n is the sample size in each stratum, whereas N is the overall sample size: thus, for example, a study might report on a sample of $N = 76$ subjects, composed of $n = 40$ females and $n = 36$ males.

Cohort studies with variable follow-up time: If every subject entered a study at least 5 years ago, then, in principle, one should be able to determine D and $N - D$, and the 5-year survival proportion. However, *losses to follow-up* before 5 years, and before the event of interest, lead to observations that are typically regarded as censored at the time of the loss. Another phenomenon that leads to censored observations is *staggered entry*, as in the JUPITER trial.[3] Unfortunately, some losses to follow-up may be examples of '*informative*' censoring.

[2] https://jhanley.biostat.mcgill.ca/bios601/Epidemiology1/epi-notes-bios601-2009.pdf

[3] https://jhanley.biostat.mcgill.ca/c634/JUPITER/

Cross-sectional prevalence data

Recall again that prevalence refers to a *state*. Examples would include the proportion (of a certain age group, say) who wear glasses for reading, or have undetected high blood pressure, or have high-speed internet at home, or have a family history of a certain disease, or a certain 'gene' or blood-type.

From a purely *statistical* perspective, the analysis of *prevalence* proportions of the form D/N and *incidence* proportions of the form D/N takes the same form: the underlying statistical 'atoms' are N Bernoulli random variables.

Important: Concepts and terms in Epidemiology

- **State**[4] vs. **Event**[5] [the *transition* (rapid) from one state to another] [6]

- Population An aggregate of people, defined by a membership-defining...
  - event → "cohort" (closed population i.e., closed for exit)
    or
  - state – one is a member just for duration of state → Open population (open for exit) / dynamic / turnover

- Prevalence (of a state) : The existence (as opposed to the inception or termination) of a particular state among the members of the population.

- Prevalence Rate: the proportion of a population that is in a particular state.

- Population-time: The amount of population experience in terms of the integral of population size over the period of observation.

- Incidence: The appearance of events of a particular kind in a population (of candidates over time)
  - *Incidence density (ID)*: The ratio of the number of events to the corresponding population time (candidate time). If we subdivide time into very short spans, *ID* becomes a function of time, $ID(t)$; otherwise *ID* refers to the average over the entire span of time.
  - *Hazard*: limiting case of *ID* as we narrow the span of time. More commonly used w.r.t. *closed* population, with a natural "t_0."
  - *Force of morbidity/mortality (Demography).*

- **Case**: Medicine – episode of illness, ("a case of gonorrhea"). Epidemiology – a person representing a case (in medical sense) of some state or event.[7]

- Incident cases: Cases that appear (as against those that exist or prevail).

- Cumulative Incidence (CI): The *proportion* of a cohort (of candidates) experiencing the event at issue over a particular risk period if time-specific incidence density is considered to operate over that period.

[4] Google: The way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state". State of matter: (chemistry) the three traditional states of matter are solids (...) liquids (...) and gases (...).

[5] Most of the definitions below are adapted from the glossary in the textbook *Theoretical Epidemiology: Principles of Occurrence Research in Medicine* by O.S. Miettinen (Wiley 1985). See also, by same author, *Epidemiological Research: Terms and Concepts* https://link-springer-com.proxy3.library.mcgill.ca/book/10.1007%2F978-94-007-1171-6 Google: something that happens at a given place and time | a phenomenon located at a single point in space-time; the fundamental observational entity in relativity theory | In the Unified Modeling Language, an event is a notable occurrence at a particular point in time. Events can, but do not necessarily, cause *state transitions* from one state to another ... | An event in computer software is an action which can be initiated either by the user, a device such as a timer or Keyboard (computing), or even by the operating system. | ~~In probability theory, an event is a set of outcomes and a subset of the sample space where a probability is assigned. Typically, when the sample space is finite, any subset of the sample space is an event (i.e. all elements of the power set of the sample space are defined as events).~~ | An occurrence. | A runtime condition or change of state within a system. | A thing which happens, like a button is pressed. Events can by low-level (such as button or keyboard events), or they can be high level (such as when a new dataset is available for processing). | A means by which the server notifies clients of *changes of state*. An event may be a side effect of a client request, or it may have a completely asynchronous cause, such as the user's pressing a key or moving the pointer. In addition, a client may send an event, via the server, to another client.

[6] In epidemiology, some authors reserve the word "*occur*" for an event (Google: happen; take place; come to pass; "Nothing occurred that seemed important") But, both in epidemiology and in lay use, it is and can also be used for a *state* (to be found to exist; "sexism occurs in many workplaces"; "precious stones occur in a large area in Brazil"). Miettinen [European J of Epi. (2005) 20: 11-15] makes this point in his reply to one of the several authors who commented on his article *Epidemiology: Quo vadis?* ibid, 2004; 19: 713-718.

> Walker's commentary was devoted to teaching me that the concept of occurrence has to do with outcome events only; that it thus does not encompass outcome *states*; and that etiologic occurrence research therefore does not encompass the important study of causal *prevalence* functions. As I now consult The New Oxford Dictionary of English (1998 edn), I find as the meanings of 'occurrence' (as a mass noun) these: 'the fact or frequency of something happening' and 'the fact of something *existing* or being found . . .,' as in 'the occurrence of natural gas fields.' And in my Perspective article I find 'state' or 'prevalence' occurring as many as eight times, 'event' or 'incidence' no more than nine times. The verb 'occur,' I might need to add, means 'happen; take place; *exist or to be found to be present* . . . ,' as in 'radon *occurs* naturally in rocks . . . ' [italics added by JH]

[7] *Google*: an occurrence or instance of something; "a case of bad judgment"'; "another *instance* occurred yesterday"; *Merriam-Webster*: noun, Middle English cas, from Anglo-French, from Latin *casus* fall, chance, from *cadere* to fall. 1 a: a set of circumstances or conditions b (1): a situation requiring investigation or action 6 a: an *instance* of disease or injury <a *case* of pneumonia> .

  - The relation between ID and CI [8] can be expressed mathematically as

$$CI_T = CI_{0\to T} = 1 - \exp\left\{ -\int_0^T ID(t)dt \right\}.$$

  - As a function of t, the *complement*, $1-CI_{0\to t}$ is called the "*Survival*" function, $S(t)$, since it is the proportion of the cohort that, at time t, remains (continues, "survives") in the initial state.

- Risk: The *probability* that an event (untoward) will occur.

- **Case Fatality Rate**: (Rothman 1986, p31) The cumulative incidence of death among those who develop an [acute] illness [e.g., SARS, influenza, COVID-19]. The time period for measuring the case fatality rate is often unstated. [9]
  In 2020, whether the denominator is limited to recognized cases, or includes all cases no matter where recognized or not, became contentious. Thus, we saw the emergence of a new term, which estimates the proportion of deaths among all infected individuals. **infection fatality ratio (IFR)**. See also footnote to Q28 in 2021 version of Measurement Error Notes.

1.3 The binary probability model

JH presumes they use this heading as a shorthand for 'the probability model for binary responses' (or 'binary outcomes' or binary random variables)

... to "*predict* the outcome" : JH takes this word *predict* in its broader meaning. If we are giving a patient the probability that he will have a certain *future* event *say within the next 5 years*, we can talk about predicting[10] the outcome: we are speaking of *pro*gnosis; but what if we are giving a woman the probability that the suspicious finding on a mammogram does in fact represent an existing breast cancer, we are speaking of the *present*, of whether a phenomenon already *exists*, and we use a prevalence proportion as an estimate of the *dia*gnostic probability. Note that prevalence and incidence refer to aggregates.

The risk parameter

Risk typically refers to the *future*, and can be used when speaking to or about one person; we don't have a comparable specialized term for *the probability that a state exists* when speaking to or about one person, and would therefore just use the generic term probability.

The odds parameter

The sex-ratio is often expressed as an odds, i.e., as a ratio of males to females. If the proportion of males is 0.51, then the male:female ratio is 51:49 or (51/49):1, i.e., approximately 1.04:1. This example is a good reason why C&H should have used a more generic pair of terms than failure and survival (or success and failure).

In betting on horse races (at least where JH comes from), odds of 3:1 are the odds *against* the horse winning; i.e., the probability of winning is 1/4.[11] When a horse is a heavy favourite so that the probability of winning was 75%, the "bookies" would give the odds as "3:1 *on*."[12]

Rare events

One of the tricks to make events *rare* will be to slice the time period into small slices or windows.

Death, the first of the two only sure events (taxes is the other) is also rare - in the short term!

Also, it would be more correct to speak of a *rare events*, since disease is often used to describe a process, rather than a transition. And since most transitions are rapid, the probability of a transition (an event) occurring within a given short sub-interval will usually be small.

If the state of interest being addressed with cross-sectional data is uncommon (or rare), then yes, the prevalence odds and the prevalence proportion will be very close to each other.

Supplementary Exercise 1.1. If one rounds probabilities or risks or prevalences (π's), or their corresponding odds, $\Omega = \pi/(1-\pi)$, to 1 decimal place, at what value of π will the rounded values of π and Ω be differerent? Also,

[8] So fundamental JH puts it in red

[9] From Miettinen's Terms and Concepts: Case-fatality rate (synonyms: fatality rate, death rate) – Concerning cases of an illness in general, or recognized cases of it (ones with rule-in diagnosis about the illness), the proportion in which the illness is fatal; that is, such that the outcome of the course of the illness is fatality from it. (Cf. 'Survival rate.')
Note: For the concept to be truly meaningful, it commonly is to be specific to particulars of the case (broadly at least) and to the choice of treatment; and it also is to be conditional on absence of intercurrent death from some other, 'competing' cause.

[10] The term 'Risk Prediction' has led to further confusion. Risk is by definition anout the future, and is a probability. It is the probability that (a future) Y=1. The Y is unknown, but the Risk may be well or poorly 'known'.

[11] Think of the 3:1 as the 'bookie' putiing $3 in an envelope, and the better butting $1, and when the race result is known, the bookie or the bettor taking the envelope with the $4.

[12] Now the bookie puts in 1 and the bettor 3

why use lowercase π for proportion, and uppercase Ω for odds?

1.4 Parameter Estimation

Should you be surprised if the estimate were π were other than D/N? Consult Google or Wikipedia on "the rule of succession," and on Laplace's estimate of the probability that the sun will rise tomorrow, given that it has unfailingly risen ($D = 0$) for the past 6000 years, i.e., $N \approx 365 \times 6000$.

Supplementary Exercise 1.2. Suppose one has 2 independent observations from the model

$$E[y|x] = \beta \times x \quad \text{['no intercept' model]}.$$

The y's might represent the total numbers of typographical errors on x randomly sample pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document. We gave this 'estimation of β' problem to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

How can this be? [If it still works] You might run the applet '2 datapoints and a model' `https://jhanley.biostat.mcgill.ca/2DatapointsAndAModel`

1.5 Is the model true?

I wonder if they were aware of the quote, attributed to the statistician George Box that goes something like this

> "**all models are wrong; but some are more useful than others**"

Box also said

> **Statisticians, like artists, have the bad habit of falling in love with their models**.

`http://en.wikiquote.org/wiki/George_E._P._Box`

2 Conditional probability models

2.1 Conditional probability

JH is surprised at how few textbooks use trees to explain conditional probabilities. Probability trees make it easy to see the **direction** in which one is preceeding, or looking, where simply (and often arbitrarily chosen) algebraic symbols like A and B can not; they make it easier to distinguish 'forward' from 'reverse' probabilities. try to order letters so it is $A \to B$ not $B \to A$.

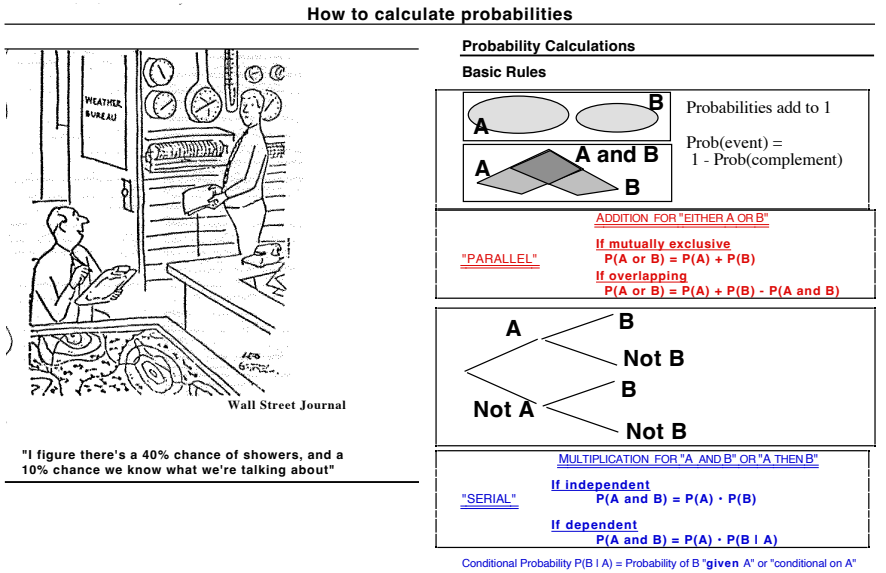


Figure 1: From JH's notes for EPIB607, introductory biostatistics for epidemiology

Trees show that the probability of a particular sequence is always *a fraction of a fraction of a fraction ..* , and that if we start with the full probability of 1 at the single entry point on the extreme left, then we need at the right hand side to account for ('conserve') all of this (i.e., the 'total') probability.

Statistical dependence and independence

JH likes to say that with independence, one doesn't have to look over one's shoulder to the previous event to know which probability to multiple by. The illustrated example on the gender composition of 2 independent births, and of a sample of 2 persons sampled (without replacement) from a pool of 5 males and 5 females, show this distinction: in the first example, when one comes to the second component in the probability product, $Pr(y_2 = male)$ is the same

whether one has got to there via the 'upper' path, or the 'lower' one.

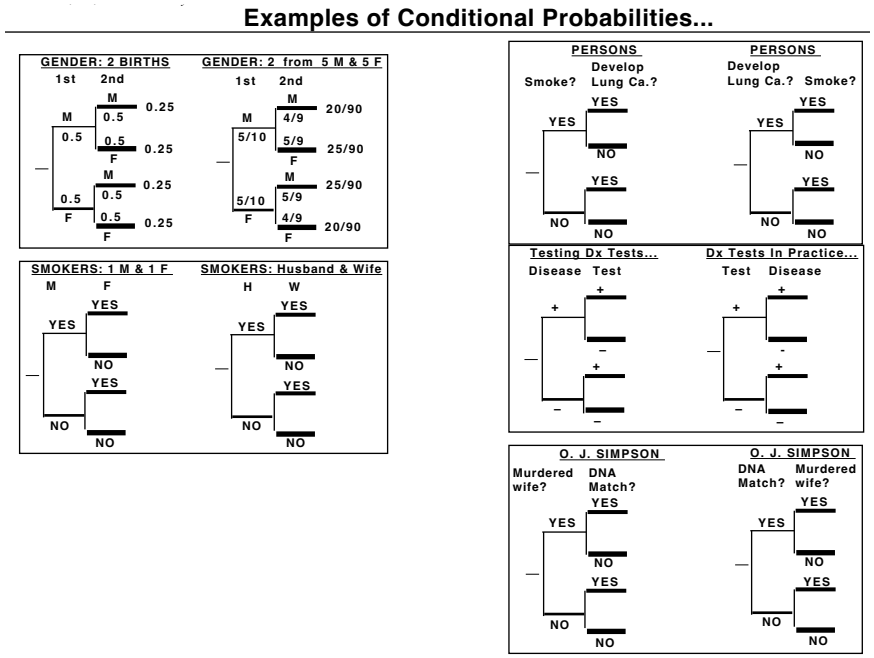


Figure 2: JH examples of ***independence/dependence*** [2 panels on left] and ***'forward'/'reverse'*** probabilities [3 panels on right]

2.2 Changing the conditioning: Bayes' rule

The panels on the right hand column of JH Figure 2 shows 3 examples of 'forward' probabilities (on the left) and 'reverse' probabilities (on the right).

The difference between 'forward' and 'reverse' probabilities distinguishes frequentist p-values (probabilities) from Bayesian posterior probabilities.

$$Probability[data \mid Hypothesis] \neq Probability[Hypothesis \mid data]$$

or, if you prefer something that rhymes,

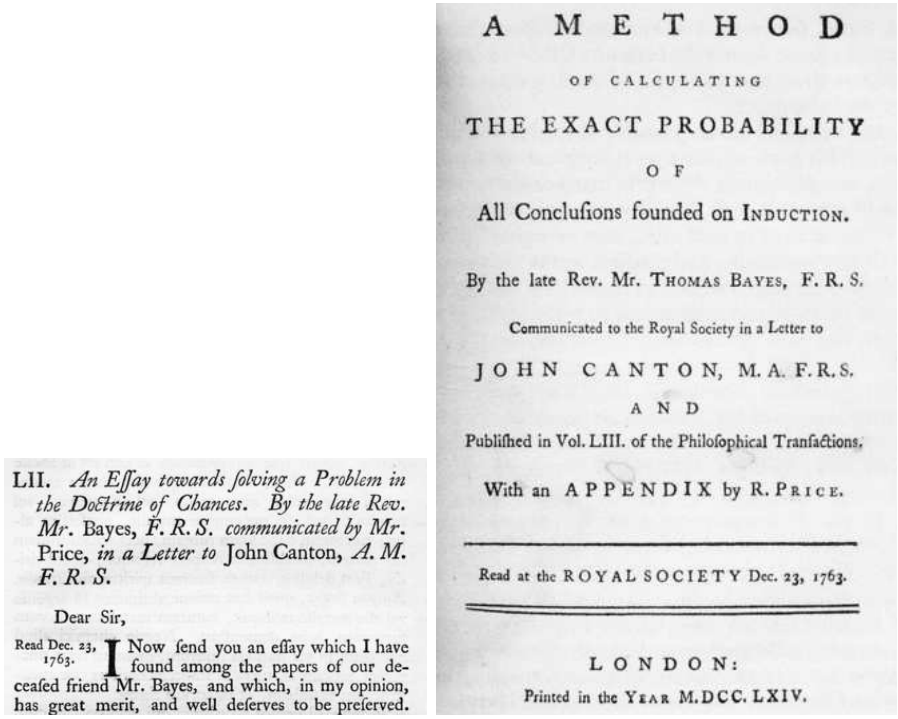
$$Probability[\ data \mid theta] \neq Probability[\ theta \mid data].$$

Two striking – and frightening – examples of misunderstandings about them are given on the next page.

The True Title of Bayes's Essay

Today's students are told that the Bayes essay was published after his death under the title "An Essay toward solving a Problem in the Doctrine of Chances". But when he spoke in Montreal at the end of 2013, Stephen Stigler gave us the inside story on the very concrete reason the person who published it, Richard Price, had for being interested in this work, and why it was advertised elsewhere under a very different title: **'A method of calculating the exact probability of all conclusions based on induction'** Read about Stigler's fascinating detective work in his captivating article Statistical Science 2013, Vol. 28, No. 3, 283-288 (Resources website) or here:

http://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/StiglerBayesTitle.pdf



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved.

A METHOD

OF CALCULATING

THE EXACT PROBABILITY

OF

All Concluſions founded on INDUCTION.

By the late Rev. Mr. THOMAS BAYES, F. R. S.

Communicated to the Royal Society in a Letter to

JOHN CANTON, M. A. F. R. S.

AND

Publiſhed in Vol. LIII. of the Philoſophical Tranſactions,

With an APPENDIX by R. PRICE.

Read at the ROYAL SOCIETY Dec. 23, 1763.

LONDON:

Printed in the YEAR M.DCC.LXIV.

Figures 1 and 3 from Stigler 2013

U.S. National Academy of Sciences under fire over plans for new study of DNA statistics: Confusion leads to retrial in UK.[13]

[...] He also argued that one of the prosecution's expert witnesses, as well as the judge, had confused two different sorts of probability.

One is the probability that DNA from an individual selected at random from the population would match that of the semen taken from the rape victim, a calculation generally based solely on the frequency of different alleles in the population. The other is the separate probability that *a match between a suspect's DNA and that taken from the scene of a crime could have arisen simply by chance* – **in other words that the suspect is innocent despite the apparent match.**[14] This probability depends on the other factors that led to the suspect being identified as such in the first place.

During the trial, a forensic scientist gave the first probability in reply to a question about the second. Mansfield convinced the appeals court that the error was repeated by the judge in his summing up, and that this slip – widely recognized as a danger in any trial requiring the explanation of statistical arguments to a lay jury – justified a retrial. In their judgement, the three appeal judges, headed by the Lord Chief Justice, Lord Farquharson, explicitly stated that their decision "should not be taken to indicate that DNA profiling is an unsafe source of evidence."

Nevertheless, with DNA techniques being increasingly used in court cases, some forensic scientists are worried that flaws in the presentation of their statistical significance could, as in the Deen case, undermine what might otherwise be a convincing demonstration of a suspect's guilt.

Some now argue, for example, that quantified statistical probabilities should be replaced, wherever possible, by a more descriptive presentation of the conclusions of their analysis. "The whole issue of statistics and DNA profiling has got rather out of hand," says one. Others, however, say that the Deen case has been important in revealing the dangers inherent in the '**prosecutor's fallacy**'. They argue that this suggests the need for more sophisticated calculation and careful presentation of statistical probabilities. "The way that the prosecution's case has been presented in trials involving DNA-based identification has often been very unsatisfactory," says David Balding, lecturer in probability and statistics at Queen Mary and Westfield College in London. "**Warnings about the prosecutor's fallacy should be made much more explicit. After this decision, people are going to have to be more careful.**"

[13] NATURE p 101-102 Jan 13, 1994.

[14] italics by JH. The wording of the italicized phrase is imprecise; the text in bold wording is much better .. if you read "despite" as "given that" or "conditional on the fact of"

"The prosecutor's fallacy": Who's the DNA fingerprinting pointing at? [15]

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician **Peter Donnelly** opened a new area of debate. He remarked that

> "forensic evidence answers the question
>
> **What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?**
>
> while the jury must try to answer the question
>
> **What is the probability that the defendant is innocent, [in the light of ALL of the OTHER EVIDENCE and] assuming that the DNA profiles of the defendant and the crime sample match? "**

Apparently, Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000 *.

The error in mixing up these two probabilities is called the "the prosecutor's fallacy," and it is suggested that newspapers regularly make this error.

Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial.

[* Comment by JH: This is a very nice example of the advantages of Bayesian over Frequentist inference .. it lets one take one's prior knowledge (the fact that he is the Archbishop) into account.

The book 'Statistical Inference" by Michael W. Oakes is an excellent introduction to this topic and the limitations of frequentist inference.] See also https://nautil.us/the-flawed-reasoning-behind-the-replication-crisis-237493/

[15] New Scientist item by David Pringle, 1994.01.29, 51-52; cited in Vol 3.02 Chance News

2.3 Examples

2.3.1 Example from genetics

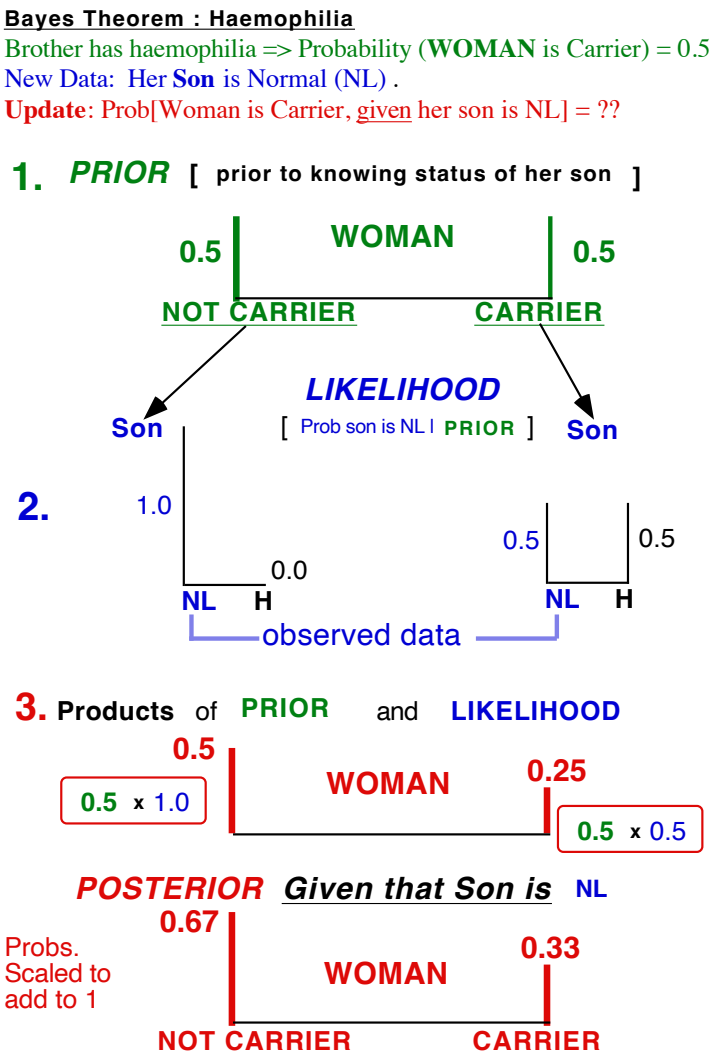


Figure 3: simpler [older] example – nowadays, direct tests mean women don't have to wait to have a son to be probabilistically sorted into definite/possible carriers.

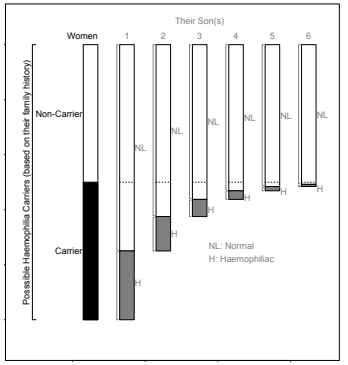


Figure 4: At the outset, each woman had a 50:50 chance of being a haemophilia carrier. **Accumulating information from their sons increasingly 'sorts' or segregates** them by moving their probability of being a carrier **to** 100% or **towards** 0%.

Probabilities: Diagnostic and Screening Tests

Try https://jameshanley.shinyapps.io/FromPreTestToPostTestProbabilities/, while noting that what JH calls

- *detection rate* is called *sensitivity* in medicine, and *power* in statistics
- *false alarm rate* is alpha (medical people call it the false positive rate, but focus on its complement, 1-alpha, and call it specificity)
- *pre-test probability* is sometimes referred to as the *prior* probability or '*prevalence*'

JH borrowed the nomogram from Fagan (p. 11, below). Fagan put the pre-test probability on the right and worked from right to left; his middle column has the Likelihood ratios (LR +ve $>$ 1 and LR- $<$ 1); his post-test probabilities are on left. In JH's nomogram, pre-test is at bottom, then LRs, and then post-test. Here is an older introduction to terminology/concepts in medical diagnosis https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/PrimerMedicalDecisionMaking.pdf

See also the very interesting '*When doctors meet numbers*' https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/Berwick1981WhenDoctorsMeetNumbers.pdf

This link https://jhanley.biostat.mcgill.ca//bios601/CandH-ch0102/ has several newer articles under PERFORMANCE AND INTERPRETATION OF DIAGNOSTIC TESTS. The one by Steurer – where he tries to improve matters by proving a user-friendly explanation of the Likelihood ratio – is of note.

SCREENING for HIV https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/MeyerPaukerHIVscreening.pdf Can we afford the False Positive Rate? MDs tell Pres. Reagan: '5/15 +ve results will be in people who are not infected.'

2.4 Some cool interactive covid infographics from the British Medical Journal

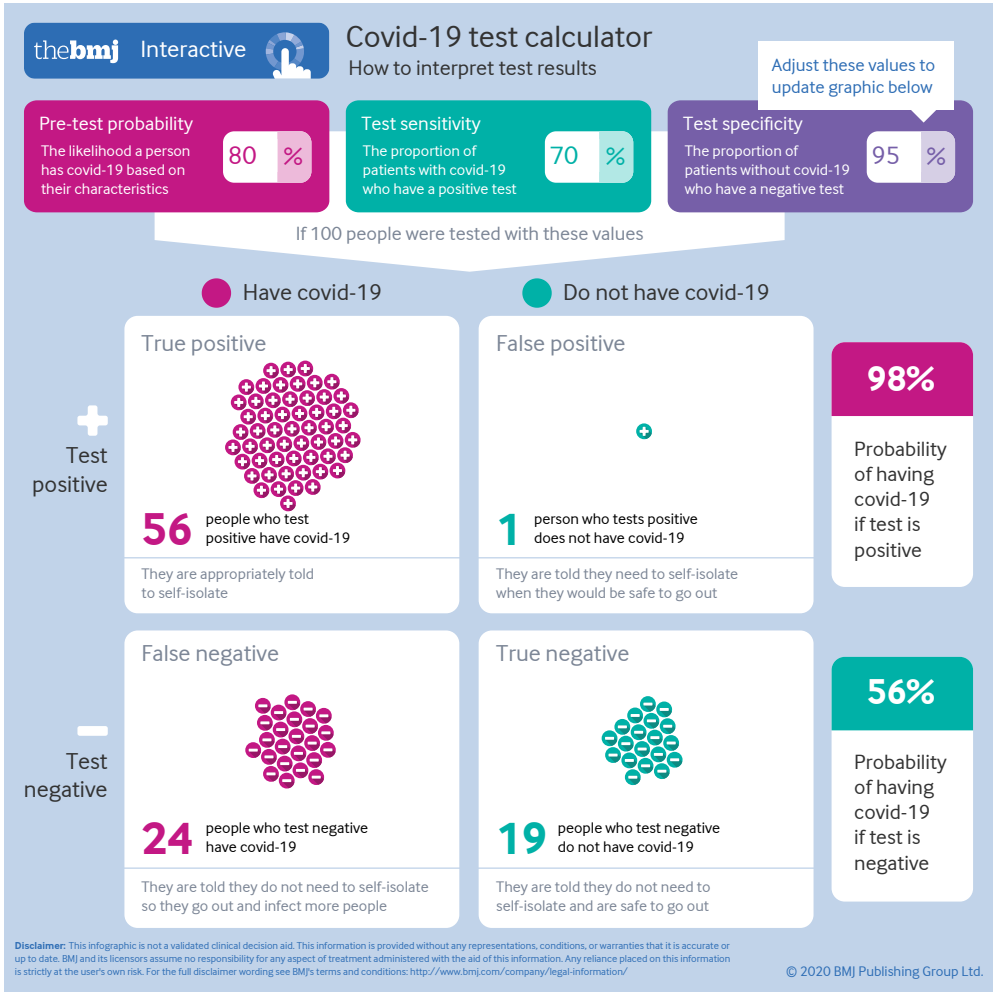
posted by Andrew Gelman on May 14, 2023

JH agrees that the BMJ Covid-19 test calculator is easier to use than the (Likelihood-Ratio-based) Fagan tools we still use. But one played around with it, would one learn the formula/basis for the predictive values? As is evident in the (1981) study of When Doctors Meet Numbers, this calculation is not easy to do in one's head, but the Likelihood Ration (LR) 'bridge' does simplify it somewhat, albeit at the cost of having to transfer at the end from the posterior *odds* to the posterior *probability*.

The comments at the end of the blog are valuable as well.

The tool is available at

`https://sandpit.bmj.com/graphics/2020/c19test/` →



→

.

2.4.1 Twins: Excerpt from an article by Bradley Efron

MODERN SCIENCE AND THE BAYESIAN-FREQUENTIST CONTROVERSY

> Here is a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked if the doctor had given her any more information. "Yes", she said, "he told me that the proportion of identical twins was one third". This is the population proportion of course, and my friend wanted to know the probability that her twins would be identical.
>
> Bayes would have lived in vain if I didn't answer my friend using Bayes' rule. According to the doctor the prior odds ratio of identical to nonidentical is one-third to two-thirds, or one half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of Identical. **Bayes' rule says to multiply the prior odds by the likelihood ratio to get the current odds**: in this case 1/2 times 2 equals 1; in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50-50 (wishing the answer had come out something else, like 63-37, to make me seem more clever.) Incidentally, the twins are a couple of years old now, and "couldn't be more non-identical" according to their mom.

Supplementary Exercise 2.1. Depict Efron's calculations using a probability tree.

Supplementary Exercise 2.2 Use a probability tree to determine the best strategy in the Monty Hall problem `http://en.wikipedia.org/wiki/Monty_Hall_problem`

Supplementary Exercise 2.3 A man has exactly two children: you meet the *older* one and see that it's a boy. A woman has exactly two children; you meet *one* of them [don't know if its the younger/older] and see is a boy. What is the probability (a) of the man's younger child being a boy, and (b) **[be careful!]** what is the probability of the woman's "other" child being a boy?

For many years JH insisted that the answer to (b) is 1/3. Early in 2023, he came across extensive writings on this (poorly-posed) problem. See Gardner (1959) and vos Savant (1997).

Wikipedia. Boy or Girl paradox

Gardner M. (1959) Mathematical Games: The Two Children Problem. Scientific American.

Gardner, M (1961). The Second Scientific American Book of Mathematical Puzzles & Diversions. New York : Simon and Schuster, 1961. University of Chicago Press 1987.

os Savant, M. (1991), Ask Marilyn. Parade Magazine. October 13, 1991 [January 5, 1992; May 26, 1996; December 1, 1996; March 30, 1997; July 27, 1997; October 19, 1997].

Carlton, M. A., and Stansfield, W.D. (2005), "Making Babies by the Flip of a Coin?," The American Statistician, 59(2), 180-182. DOI: 10.1198/000313005X42813

Garenne, M. (2009), "The Sex Composition of Two-Children Families: Heterogeneity and Selection for the Third Child: Comment on Stansfield and Carlton (2009), Human Biology, 81(1): 97-100.

Stansfield, W. D., and Carlton, M.A. (2009), The Most Widely Publicized Gender Problem in Human Genetics, Human Biology, 81(1): 3-11.

Falk, R. (2011). When truisms clash: Coping with a counterintuitive problem concerning the notorious two-child family, Thinking & Reasoning, 17 (4): 353-366. doi:10.1080/13546783.2011.613690. S2CID 145428896

Paindaveine, D. and Spindel, P, (2023), Revisiting the Name Variant of the Two-Children Problem. The American Statistician.

Senn, S. (2023), Dicing with Death: Living by Data. Second Edition. Cambridge, United Kingdom ; New York, NY : Cambridge University Press.

Supplementary Exercise 2.4

Refer to the article `https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/BBCNewsAmandaKnoxAndBadMathsInCourt.pdf`

Specifically look at the highlighted section **"why are two tests better than one?"** and in particular, the statement that

> ***"The probability that the coin is fair – given this outcome – is about 8%"***

This statement and the subsequent one involving the phrase "Now the probability for a fair coin" both seem to come out of nowhere.

Questions:

- Is this a well posed problem, or does one need to specify more context in order to do the calculations?
- Are they using a p-value somehow?
- (After you have first thought about it for a while) read the relevant portion of pages 61-62 and pages 85-86 of the book chapter

http://ebookcentral.proquest.com/lib/mcgill/reader.action?docID=991081&ppg=74 Math Error Number 4: Double Experiment: the test that wasn't done (Amanda Knox case) and find out what information was missing from the BBC article. Then verify the 92:8 posterior odds given in the chapter. Repeat the calculation, but assuming only a 5% prior probability that the coin is biased and a 95% probability that it is fair. Comment.

Supplementary Exercise 2.5

Refer to the Economist article 'Problems with scientific research: HOW SCIENCE GOES WRONG' https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/HowScienceGoesWrong.pdf

It has a very nice graphical explanation of why some many studies get it wrong, and cannot be reproduced – the topic of the Reproducibility Project in Psychology referred to on same page.

One reason is that even if all studies were published, regardless of whether the p-value was less than 0.05 (a common screening/filtering criterion) or greater than 0.05, then, of all the hypotheses tested, only a small percentage of the hypotheses are 'true'. Thus many or most of the 'positive' tests (*published* results) will be false positives. It is just like when using mammography to screen for breast cancer: in maybe 4 of every 5 women referred for biopsy, the biopsy will come back negative.

1. Represent the information in their Figure as a tree. Then present the same information in a different tree, with data on left, and hypothesis on the right (rather than the conventional 'theta → data' direction) – as JH has done in the three rightmost instances in Figure 2 on page 6 above.
2. What percentage of positive tests would be correct/not if, instead, 1 in 2 of the hypotheses interesting enough to test were true?
3. Come up with a general formula for what in medicine is called the '*positive predictive value*' of a positive medical test.
4. Try to simplify it so that the characteristics of the test (α and β) are isolated in one factor, and the testing context (the 1 in 10 or 1 in 2, etc) is in another. *Hint: use odds rather than probabilities, so that you are addressing the ratio of true positives to false positives, and the ratio of true hypothesis to false hypotheses. And use the Likelihood Ratio*
5. On the same Resources web page is another (but longer) attempt to explain these concepts graphically to left brain and right brain doctors. https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/RightSideLeftSide.pdf. JH was impressed with this, and wanted to share it with the Court for Arbitration in Sport, when explaining the interpretation of positive doping tests. But he found that the 'teaser' sentence immediately following the title

   > Can you explain why a test with 95% sensitivity might identify only 1% of affected people in the general population?

   is misleading, and so he make his own diagram (available on request).

   Exercise: Revise this misleading phrase.

see http://shinyapps.org/apps/PPV/

Supplementary Exercise 2.6[16] How many offspring do I need to test?
Background:[17]
A researcher is trying to develop a strain of "transgenic" mice, by introducing an altered gene (transgene) into the genome. In order to breed true, the animals must be made to be homozygous, i.e., to have two copies of the introduced gene (+ +) . Molecular biology techniques can detect whether the transgene is present in an individual animal (without having to sacrifice the animal), but *cannot* distinguish a hemizygote, with one copy of the gene (+ -), from a homozygote (+ +). This difference can only be detected by breeding strategies. But, time and resources are pressing.

First generations:
A copy of the transgene is injected into the pronucleus of a newly fertilized ovum, prior to fusion with the male pronucleus. Thus all animals that develop from these zygotes can have at most one copy of the gene, from the ovum. After birth, screening is performed to detect these "positive" animals, called founders. After sexual maturation, all founders are bred to normal "wild type" (WT) animals, to ensure that the transgene has been incorporated in such a way as to be heritable. Pairs of positive (hemizygous) animals in this F1 generation are then bred to each other. By Mendelian genetics, the distribution of F2 offspring should be 1:2:1, homozygous transgenic : hemizygous transgenic : homozygous normal. The homozygous normal animals are not used. The question is, how to tell the homozygous transgenic mice (the desired ones) from the hemizygous transgenic ones? Note that the mix in this reduced population is 1 homozygous transgenic to 2 hemizygous transgenic.

F2 breeding:
All 'positive' F2 animals (i.e. all homozygous and hemizygous animals) are bred to wild type. **Possible F3 genotypes** are as follows: (by Mendelian genetics)

- Hemizygous (which comprise 2/3 of the F2 animals used) x wild type = 50:50, hemizygous (and therefore 'positive') : normal (and therefore 'negative'),
- Homozygous (which comprise 1/3 of the F2 animals used) x wild type = all hemizygous (and therefore 'positive').

[16] New this year, so wording may need some polishing. Also, JH developed this question in 1991; it may well be that technology since then has made the task easier.
[17] If in a hurry, skip to the **Possible F3 genotypes** later in the piece.

That is, while only half of the offspring from a Hemi x WT pair will be 'positive' when screened, all of the offspring of a Homo x WT pair will be 'positive'.

The question:

How many F3 offspring from a particular pairing does the researcher have to screen before declaring the positive parent as homozygous? Note: as soon as an offspring is screened as 'negative', one knows the parent must have been hemizygous.

The point is to check the least number of offspring and do as few repeated breedings as possible to detect homozygous animals. How many consecutive 'positive' F3 offspring does one need to observe to be convinced (and with what probability) that the positive F2 parent is homozygous for the transgene?

1. Calculate the probability before the positive F2 parent has any offspring, and after observing 1, 4, 8, 11 consecutive 'positive' F3 offspring. Give a general rule for the probability after K consecutive 'positives'.
2. This probability/odds problem has a structure similar to the hemophilia one in §2.3. So redraw the diagram provided there, using the transgenic testing example, and making the necessary changes to the prior probabilities and to the labels. Do so first for the "1 inconclusive offspring" case: think of this 1 inconclusive offspring as *somewhat* helpful, *better* than where the probabilities stood *before* any offspring were observed. Then extend the diagram to the general "K inconclusive offspring" case; think of this as 'not quite certain but closer to it than where one stood before any offspring were observed. *[Many students argue that even after the suspected carrier has a NL son, the probability she is a carrier is unaltered, at 0.5. When asked again once she has had 3 or more consecutive NL sons that they begin to realize that each NL son pushes the probability that she is a carrier closer to 0.] For the transgenic testing, the longer the run of 'positives' the more the probability is moved close to 1.*
3. Denote the probability of being transgenic, or a haemophilia carrier, after observing K , as the post-test' probability (think of the offspring as providing a test for the status of the parent). Obviously, if at any stage the next offspring provides conclusive evidence, the probability immediately goes to 0 (transgenic) or 1 (haemophilia). But if the K consecutive offspring are inconclusive but still informative, it merely moves the probability of interest further in the other direction.
   The formula for the post-test probability of being transgenic, or a haemophilia carrier, i.e., after observing K consecutive inconclusive but still informative offspring, is awkward. It has the form A/(A + B). But, what if we switch from pre-test and post-test *probabilities* to pre-test and post-test '*odds*'?[18] Redo the calculations in terms of pre-and post-test odds, and characterize (give a more familiar name to) the ??? in the expression

$$\text{Post-test } odds = \text{Pre-test } odds \times ??? \ ,$$

   or

$$\log[\text{Post-test } odds] = \log[\text{Pre-test } \ odds] + \log[???].$$

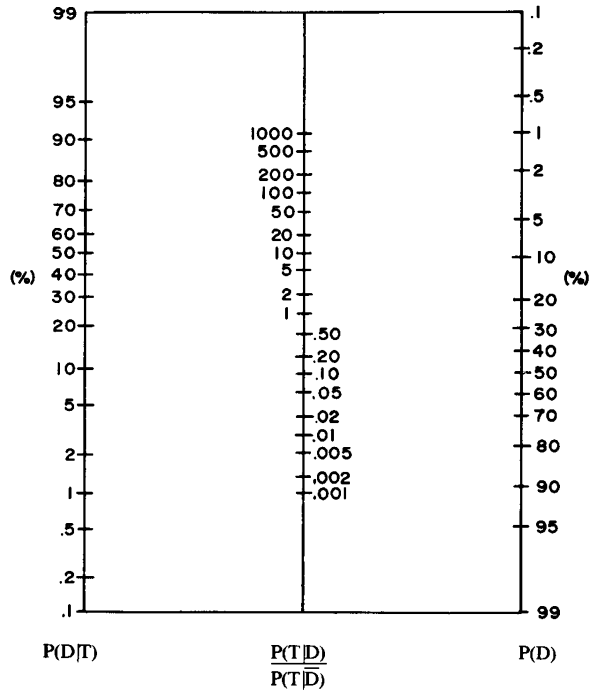
   or

$$logit[post.test.prob] = logit[pre.test.prob] + \log[???].$$

[18] As C&H tells us, $\Omega = \pi/(1-\pi)$; Odds = Prob(+)/Prob(-).

NOMOGRAM FOR BAYES'S THEOREM

To the Editor: The interest in Dr. Katz's probability graph (N Engl J Med 291:1115, 1974) causes me to offer a solution to the Bayes's rule in the form of a nomogram (Fig. 1). P(D) is the probability that the patient has the disease before the test.



P(D|T) is the probability that the patient has the disease after the test result. P(T|D) is the probability of the test result if the patient has the disease, and P(T|$\overline{D}$) is the probability of the test result if the patient does not have the disease. With this terminology the usefulness of both positive and negative test results can be assessed. A line drawn from P(D) on the right of Figure 1 through the ratio of P(T|D) to P(T|$\overline{D}$) in the center of Figure 1 gives P(D|T) on the left of Figure 1.

With Dr. Katz's renovascular hypertension example: P(D) = 10%, P(Positive IVP|D) = 100%, and P(Positive IVP|$\overline{D}$) = 10%. The ratio of P(Positive IVP|D) to P(Positive IVP|$\overline{D}$) is 10. A line drawn from P(D) = 10% through the ratio of 10 gives P(D|Positive IVP) = 53%.

The above approach can also be used to examine the usefulness of a negative test result. The ratio of P(negative test|D) to P(negative test|$\overline{D}$) is calculated and marked on the center line of Figure 1. A line drawn from P(D) through this new ratio gives P(D|negative test) on the left of Figure 1.

Houston, TX

TERRENCE J. FAGAN, M.D.
Baylor College of Medicine

N.E. J. Med, 1975

Supplementary Exercise 2.7[19]

A woman had a daughter and then 3 sons; the 3rd son has Duchenne muscular dystrophy[20]. Thus, there is a chance the daughter is a carrier. If the affected son's status is a result of a gene inherited from his mother (the affected gene lies on the X chromosome), then the probability that the daughter is [also] a carrier is 50%. If the son's status is a result of a spontaneous mutation in his genome, then the probability that the daughter is a carrier is negligible. The mother can have the daughter tested, but prefers to wait until she is grown up when she can decide for herself whether to be tested.

What additional information, if any, is provided by the data on the other 2 sons? [21] Depict the situation as a tree diagram, and state any assumptions/information you have to make/include.

Supplementary Exercise 2.8: How often does it land like this?

A thumbtack refers to 'a tack with a large, flat head, designed to be thrust into a board or other fairly soft object or surface by the pressure of the thumb'. The British tend to call it a drawing pin – a tack used to hold drawings on drawing boards. (Nowadays, a pin or map tack refers to thumb tacks used to mark locations on a map and to hold the map in place). [https://en.wikipedia.org/wiki/Drawing_pin.]

We will focus on this version and on what proportion (π) of times, if tossed in the air or dropped from a height, it would land in the position indicated in the above diagram, as opposed to on its back. Of course, this (π) might well depend on the height, or whether the surface it is dropped onto is soft (a carpet) or hard (a table, or wood floor), or even on the person tossing it.[22] For our class investigation we will choose a hard surface.

1. Before you gather any data, make your best educated guess as to the magnitude of π. Since you won't want to bet all your money on one specific value, you should give your 'distribution' in the form of a p.d.f. with a range of 0 to 1. Thus, describe your uncertainty (or degree of certainty) concerning π as a beta distribution with parameters α and β, with the parameter values chosen to reflect how concentrated or vague your estimate of π is. Remember (or look up, preferably in Cassela and Berger, or – only if stuck – Wikipedia) that the mean of a beta distribution is $\mu = \alpha/(\alpha+\beta)$ and the SD is $[\mu(1-\mu)/(\alpha+\beta+1)]^{1/2}$.

2. Then carry out a number of tosses and update the p.d.f.

Supplementary Exercise 2.9[23]

For this exercise we will take the term *fecundability* to mean the probability of pregnancy during a single menstrual cycle.

Couples attempting pregnancy are to be followed for up to K menstrual cycles, or until pregnancy occurs. We assume K is fairly small, so that aging during the follow-up interval will have negligible effects on the fecundability of any given couple. In practice, K is usually some number less than or equal to 12.

If all non-contracepting, sexually active couples had the same per-cycle conception probability, π, then the number of cycles required to achieve pregnancy would be distributed as geometric with parameter π. In fact, there is ample evidence that *couples vary in their fecundability*. About 30% of sexually active couples achieve pregnancy in their first non-contracepting cycle, a smaller proportion of the remaining couples achieve pregnancy in the second, and with each additional unsuccessful cycle, the conception rate continues to decline, as the risk sets become further depleted of the relatively fecund couples. The pronounced decrease in conception probability over time is not properly viewed as a time or age effect, but as a sorting effect in a heterogeneous population.

Thus, couples will be assumed to vary in their fecundability, so that a given couple has a per-cycle conception probability that stays constant throughout the follow-up interval, but these probabilities vary across couples. Assume that π varies according to a beta distribution, with parameters $\alpha = 3$, $\beta = 7$, i.e.,

$$\pi \sim Beta(3, 7).$$

[19] This exercise, also new in 2016, was suggested by a statistical geneticist colleague whom JH consulted as to whether using offspring to infer haemophilia carriage or trangenicity is outdated: answer YES!

[20] https://en.wikipedia.org/wiki/Muscular_dystrophy.

[21] The son of a female carrier has a 50% chance of receiving the affected X chromosome. If the son has the mutation, it is 100% probability of disease.

[22] The data in Beckett and Diaconis (Advances in Mathematics, 103, 107-128 (1994) 'involve repeated rolls of a common thumbtack, and recording whether the tack landed point up or point down. All tacks started point down. Each tack was flicked or hit with the fingers from where it last rested. A fixed tack was flicked 9 times. The data are recorded in Table I. There are 320 9-tuples. These arose from 16 different tacks, 2 "flickers," and 10 surfaces. The tacks vary considerably in shape and in proportion of ones. The surfaces varied from rugs through tablecloths through bathroom floors.

[23] This exercise, based on Weinberg & Gladen, Biometrics 42, pp.547-560 (1986), is new in 2016, so the wording may still need some polishing.

1. Show that if this is indeed the case, then indeed about 30% of sexually active couples achieve pregnancy in their first non-contracepting cycle. *It may help to think of the results as the realizations of Bernoulli random variables with differing expectations, in other words, i. ~~i.~~ d. Bernoullis.*

2. Now, exclude the *first*-cycle pregnancies and consider the couples who proceed to the *second* cycle. What is the distribution of π in these remaining couples? *Hint: to see what happens, you might want to make a graph: convert the continuous r.v. – and associated density – for cycle 1 into a discrete one with 100 probabilities centered on* $0.005, 0.015, \ldots, 0.995$ *with a rectangle erected over each one; then remove the appropriate portion from the top of each rectangle, and rescale the altered rectangles so that the frequencies again add to 1, and then convert the discrete r.v. back to a continuous r.v. The new p.d.f. should have a familiar (and remarkable!) functional form. What is it?*

3. Generalize to 12 cycles, and plot the 12 pdfs on a single graph. Then, for $k = 2, \ldots, 12$, find what % of those who undergo non-contracepting cycle k become pregnant in cycle k.

4. After 6 unsuccessful cycles, a couple asks you what is the estimated probability that – if they continue to try – they will be successful in one of the next 6 cycles. Rather than just giving a 'central' estimate, give a pessimistic[24] estimate and an optimistic[25] one.

5. Instead of 'assuming' $\alpha = 3$, $\beta = 7$, how might one estimate α and β from data? For concreteness, imagine one had the data from 500 couples followed for up to 12 cycles after discontinuing contraception.

6. Among the couples attempting pregnancy, a proportion ρ will have some hidden condition that makes $\pi = 0$. Thus, it may be more realistic to consider the distribution of π to be a beta 'contaminated by' (or 'mixed with') a second distribution degenerate at 0. In this context, ρ is called the 'mixing parameter'.

   Repeat the calculations for the cycle-specific distributions and percentages.

[24] Use the 5th percentile of the 'after-6-unsuccessful cycles' distribution.
[25] Use the 95th percentile of this 'after-6' distribution.

Supplementary Exercise 2.10[26] Two (orientational) probability examples from Alan **Turing** [Full Turing article available here `https://jhanley.biostat.mcgill.ca/bios601/CandH-ch0102/`, along with commentary by Zabell.] Each question is preceded by a • .

Zabell tells us

> In April 2012, two papers written by Alan Turing during the Second World War on the use of probability in cryptanalysis were released by GCHQ. The longer of these presented an overall framework for the use of Bayes's theorem and prior probabilities, including [in Ch. 2] four examples worked out in detail: the Vigenère cipher, a letter subtractor cipher, the use of repeats to find depths, and simple columnar transposition. (The other paper was an alternative version of the section on repeats.) Turing stressed the importance in practical cryptanalysis of sometimes using only part of the evidence or making simplifying assumptions and presents in each case computational shortcuts to make burdensome calculations manageable. The four examples increase roughly in their difficulty and cryptanalytic demands. After the war, Turing's approach to statistical inference was championed by his assistant in Hut 8, Jack Good, which played a role in the later resurgence of Bayesian statistics.

The following numbering of the Chapter 1 subsections was introduced by Ian Taylor, who reset the 'manuscript' in LateX.

```
Chapter 1. Introduction

1.1. Preamble
1.2. Meaning of probability and odds
1.3. Probabilities based on part of the evidence
1.4. A priori probabilities
1.5. The Factor Principle
1.6. Decibanage
```

1. Turing's Section 1.2 ('**Meaning of probability and odds**') is quite short

   > I shall not attempt to give a systematic account of the theory of probability, but it may be worth while to define shortly

[26] New in 2018, so wording may need some polishing.

probability and *odds*.

The *probability* of an event on certain evidence is the proportion of cases in which that event may be expected to happen given that evidence. For instance if it is known the 20% of men live to the age of 70, then knowing of Hitler only Hitler is a man we can say that the probability of Hitler living to the age of 70 is 0.2. Suppose however that we know that Hitler is now of age 52 the probability will be quite different, say 0.5, because 50% of men of 52 live to 70.

The *odds* of an event happening is the ratio P/(1-P) where P is the probability of it happening. This terminology is connected with the common phraseology 'odds of 5:2 on' meaning in our terminology that the odds are 5/2.

• Does Turing's definition of probability fit with what you have been taught, or is it a bit more qualified and specific? (cf. commentary by Zabell, and specifically his quotes from Laplace and Bertrand.)

• Later on, in another example, Turing admits that his 'facts' are 'no doubt hopelessly inaccurate.' How accurate are the 'facts' he used in the living to 52 and to 70 example? Compare them with those in the portion of the 'current' English lifetable from 1930-1932, used by Armitage, and discussed by JH in section 4.2 of his Notes on Clayton & Hills. Ch 4: Follow-up. `https://jhanley.biostat.mcgill.ca/bios601/ch04.pdf` In that lifetable, approximately what % of men of 50 live to 70?

2. Section 1.5. ('**The Factor Principle**') is illustrated with a medical example:

Nearly all applications of probability to cryptography depend on the 'factor principle' (or Bayes' theorem). This principle may first be illustrated by a simple example.

Suppose that one man in five dies of heart failure, and that of the men who die of heart failure two in three die in their beds, but of the men who die from other causes only one in four die in their beds. (My facts are no doubt hopelessly inaccurate). Now suppose we know that a certain man died in his bed. What is the probability that he died of heart failure?

Of all men numbering N say, we find that

N	×	(1/5)	× (2/3)	die in their beds of heart failure
N	×	(1/5)	× (1/3)	die elsewhere of heart failure
N	×	(4/5)	× (1/4)	die in their beds from other causes
N	×	(4/5)	× (3/4)	die elsewhere from other causes

Now as our man died in his bed we do not need to consider the cases of men who did not die in their beds, and these consist of

N	×	(1/5)	× (2/3)	cases of heart failure
N	×	(4/5)	× (1/4)	from other causes,

and therefore the odds are $1 \times (2/3) : 4 \times (1/4)$ in favour of heart failure. If this had been done algebraically the result would have been

A posteriori odds of the theory

= A priori odds of the theory

$$\times \frac{\text{Probability of the data being fulfilled if the theory is true}}{\text{Probability of the data being fulfilled if the theory is false}}.$$

In this the 'theory' is that the man died of heart failure, and the 'data' is that he died in his bed.

The general formula above will be described as the 'factor principle', the ratio $\frac{\text{Probability of the data if the theory is true}}{\text{Probability of the data if the theory is false}}$ is called the factor for the theory on account of the data.

• Use the above information to sketch 2 probability trees, along the lines of those shown in Figure 2 (p.6) of JH's Notes.

3. Section 1.6. ('**Decibanage**') adds 2 additional pieces of information to the same medical example.

Usually when we are estimating the probability of a theory there will be several **independent** [emphasis added by JH] pieces of evidence e.g. following our last example, where we want to know whether a certain man died of heart failure or not, we may know

a) He died in his bed
b) His father died of heart failure
c) His bedroom was on the ground floor

and also have statistics telling us

2/3 of men who die of heart failure	die in their beds
2/5	have fathers who died of heart failure
1/2	have their bedrooms on the ground floor

1/4 of men who died of other causes	die in their beds
1/6	have fathers who died of heart failure
1/20	have their bedrooms on the ground floor

Let us suppose that the three pieces of evidence are **independent** of one another if we know that he died of heart failure, and also if we know that he did not die of heart failure. That is to say that we suppose for instance that knowing that he slept on the ground floor does not make it any more likely that he died in his bed if we knew all along that he died of heart failure.

When we make these assumptions the probability of a man who died of heart failure satisfying all three conditions is obtained simply by multiplication, and is $(2/3) \times (2/5) \times (1/2)$ and likewise for those who died from other causes the probability is $(1/4) \times (1/6) \times (1/20)$, and the factor in favour of the heart theory failure is

$$\frac{(2/3) \times (2/5) \times (1/2)}{(1/4) \times (1/6) \times (1/20)}.$$

We may regard this as the product of three factors $(2/3)/(1/4)$ and $(2/5)/(1/6)$ and $(1/2)/(1/20)$ arising from the three independent pieces of evidence.

Products like this arise very frequently, and sometimes one will get products involving thousands of factors, and large groups of these factors may be equal. We naturally therefore work in terms of the logarithms of the factors. The logarithm of the factor, taken to the base $10^{1/10}$ is called the decibanage in favour of the theory! A 'deciban' is a unit of evidence; a piece of evidence is worth a deciban if it increase the odds of the theory in the ratio $10^{1/10} : 1$. The deciban is used as a more convenient unit that the 'ban'. The terminology was introduced in honour of the famous town of Banbury.[27]

Using this terminology we might say that the fact that our man died in bed scores 4.3 decibans in favour of the heart failure theory ($10\log(8/3) = 4.3$). We score a further 3.8 decibans for his father dying of heart failure, and 10 for his having his bedroom on the ground floor, totalling 18.1 decibans. We then bring in the a priori odds 1/4 or $10^{-6/10}$ and the result is the the odds are $10^{12.1/10}$, or as we may say '12.1 decibans up on evens'. This means about 16:1 on.

- Sketch the relevant parts of Turing's example using a probability tree.

This '***independence***' assumption that Turing invokes is called '***conditional independence***' today, and it is widely used in the statistical literature that deals with imperfect diagnostic tests. The 'classic' papers in the field are by Hui and Walter (1980) and Walter and Irwig (1988).[28] The work of McGill's Lawrence Joseph, beginning with Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard (AJE 1995) and his student Nandini Dendukuri, and their students has considerably extended the uses of this model. In some cases they have tried to relax the conditional independence assumption. Links to some of these are provided in the Resources.

One of the objections to the conditional independence assumption is that (especially in those with the disease who are the target of the diagnostic tests), the results of the various tests (physical examination, blood tests, imaging tests) may be correlated, and that one may be giving too

[27] As Zabell tells us, 'The factor of 10 was included to simplify the arithmetic, dropping everything after the first decimal place. For example, in the cases p=0.55 and p=0.9, one has $\log_{10}(0.55/0.45) = 0.08715$ and $\log_{10}(0.9/0.10) = 0.95424$, and these would be reported in decibans as 0.9 and 9.5, respectively.' Zabell also has an interesting note on the (time- and effort-saving) switch to *half-decibans*, a practical innovation introduced by I.J. Good.

[28] Hui SL, Walter SD. (1980) Estimating the error rates of diagnostic tests. Biometrics 36, 167-171; and Walter SD and Irwig L. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988;41(9):923-37. They latter 'shows how, under certain conditions, it is possible to estimate error parameters such as sensitivity, specificity, relative risk, or predictive value, even though no definitive classification (gold standard) is available. The parameter estimates are obtained by modelling the data, using maximum likelihood, with or without some constraints. The models recognize that the true classification of an individual is unknown, and so are sometimes referred to as "latent class" models. The latent class approach provides a unified framework for various methods found in a dispersed literature, characterising each by the number of populations or subgroups in the data, and the number of observations made on each individual; the statistical degrees of freedom are implied by the sampling design. Data sets with less than three replicate observations per individual necessarily require constraints for parameter estimation to be possible. Data sets with three or more replicates lead directly to estimates of the misclassification rates, subject to some simple assumptions.'

much weight to them if one combines the evidence on the assumption that the pieces of information are independent. However, some investigations (e.g., Torrance-Rynard and Walter. Effects of dependent errors in the assessment of diagnostic test performance. Stat Med. 1997 Oct 15;16(19):2157-75.) have found that 'violations' of the assumption may not matter greatly in practice. In a recent email to me, Lawrence Joseph agreed that 'conditional dependence is likely rarely exactly true in real problems' but went on to say that 'effects from dependence may be small so assuming independence may be good as an approximation. it can also be pretty difficult to evaluate conditional independence from available data.'

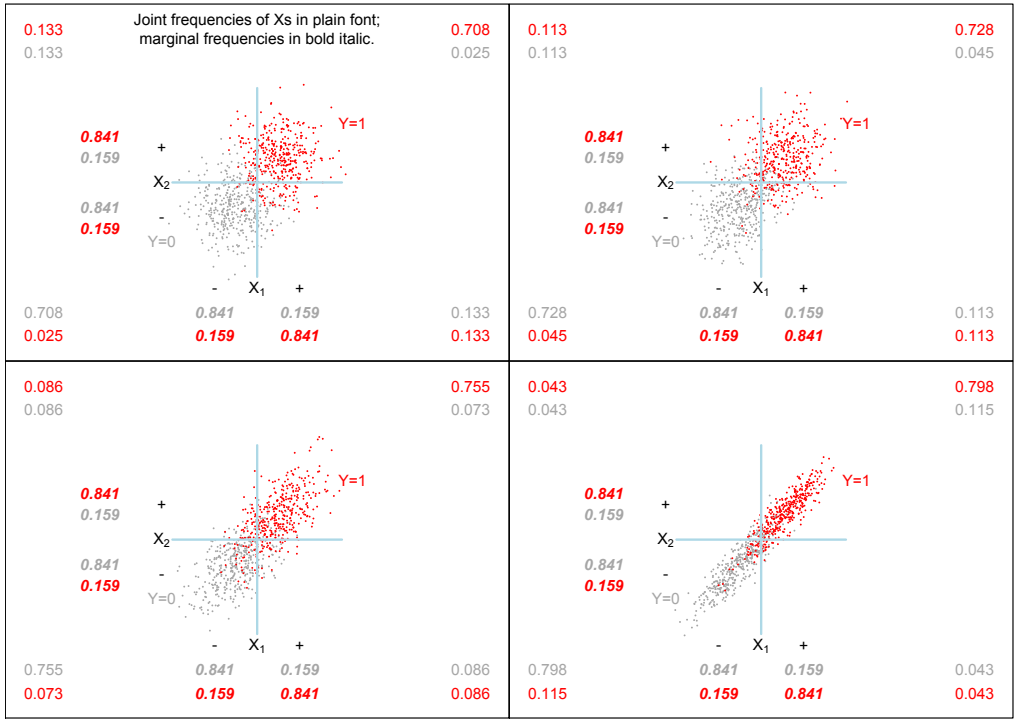
Imagine[29] 2 pieces of information (younger(Y) / older(O) age and female(F) / male(M) gender) that might help decide which cases of chest pain reported to a telephone helpline are because the patient is (a) having a heart attack, or (b) suffering from anxiety or panic.
Suppose that of 12 calls in whom it is ultimately determined that the cause is (b), the expected frequencies of the 4 profiles (YF, YM, OF, OM), are 3 : 3 : 3 : 3; and that of 12 calls in whom it is ultimately determined that the cause is (a), the frequencies of these same 4 profiles are 1 : 2 : 3 : 6.

• Do these data obey the conditional independence assumption?

• Calculate (i) the a-priori odds of heart-attack : anxiety, (ii) the (overall) 'factor' associated with each of the four profiles, and (iii) the a-posteriori odds of heart-attack : anxiety for each profile. (Don't bother converting them to decibans)

• Generically, denote the states of interest by Y=0 and Y=1, and the 2 pieces of information as X_1 and X_2, each categorized as '+' or '-'. Suppose that of 100 instances in whom it is ultimately determined that Y=0, the expected frequencies of the 4 profiles (- -, - +, + -, + +), are 75 : 9 : 9 : 7; and that of 100 instances in whom it is ultimately determined that Y=1, the frequencies of these same 4 profiles are 7 : 9 : 9 : 75.
Repeat the 2 earlier questions. How much do the 'violations' of the conditional- independence assumption affect the a-posteriori probabilities?

[29] As with Turing, these are 'made up' frequencies, so as to keep the arithmetic easy.



Note regarding the frequencies in last part of the question above:

The bottom left panel of this Figure was used to calculate the (rounded) frequencies. It uses 2 overlapping bivariate $\{X_1, X_2\}$ distributions, one for those in the $Y = 1$ state (in red) and one for those in the $Y = 0$ state (in grey). The $\{X_1, X_2\}$ correlations range from 0 (upper right panel) to 0.9 (lower left). The vertical and horizontal lines are cut-points that dichotomize the $\{X_1, X_2\}$ values into 'positive' and 'negative' results.

The assumption of **2 overlapping Multivariate Normal distributions** is the basis for the ***disciminant function*** (the linear combination βX) introduced by Ronald Fisher in 1936. If the 2 covariance matrices are equal to each other, then the linear discriminant is also (modulo an intercept) the logit of the probability that Y=1:

$$\log \frac{Prob[Y = 1|X]}{Prob[Y = 0|X]} = \beta X,$$

a form known today as ***logistic regression***.

This logistic form was introduced to epidemiology with **Cornfield's 1962 paper**[30] that used 2 variables ($X_1 = log_{10}$ cholesterol) and $X_2 = log_{10}$ (blood pressure - 75) measured in the Framingham Heart study to fit the risk (probability) of developing heart disease (Y=1) over the next 6 years. The linear discriminant function (LD) he fitted was

$$LD = -23.13 + 6.14X_1 + 3.29X_2$$

[30] `https://jhanley.biostat.mcgill.ca/c678/cornfield.pdf`

So the odds is $\exp[LD] : 1$ and the probability is odds/(odds+1)

$$\text{Probability} = \text{odds}/(1+\text{odds}) = \frac{\exp[LD]}{1+\exp[LD]}$$

The dataset had 92 instances of Y=1 and 1237 of Y=0, so the a-priori odds were 92:1237 or 0.074:1. The 'average' probability is thus approximately 7%.

Consider 4 profiles: cholesterol of 200 or 300 (log=2.30 or 2.48), and SBP of 120 or 180 (log = 2.08 or 2.26). So the 4 LDs are -23.13 + 6.14×(2.30 or 2.48) + 3.29×(2.08 or 2.26), i.e., -3.56, -2.35, -2.48 and -1.27. So the profile-specific odds are 0.028:1, 0.095:1, 0.084:1 and 0.281:1. So the probabilities are 0.03, 0.09, 0.08 and 0.22, or 3%, 9%, 8% and 22%.

50 years ago, in 1967, Truett Cornfield and Kannel[31] **relaxed the insistence on strict multivariate normality** (which could not work apply to binary variables, or to many continuous ones).

'For the multiple logistic function to provide an exact description of the relation between risk and risk factors it is sufficient that the underlying distributions be multivariate normal. It is by no means necessary, however. In fact a much weaker condition is sufficient, namely that the linear compound of risk factors be univariate normal. The circumstances under which a linear com- pound of independent variables will be normal are given by the central limit theorem, and of dependent variables by Bernstein?s theorem.'

They still used the (1-pass) method of Discriminant Analysis to fit the weights or coefficients.

That same year, in Biometrika, Walker and Duncan[32] reversed the statistical modelling focus. Remember that the focus of discriminant analysis is Prob$[X|Y]$ – the random variable is the multivariate X. But why model the joint distribution of these X variables? Walker and Duncan focused directly on what Cornfield et al. were ultimately interested in but had derived indirectly (post-fit) from the LD, namely the Prob$[Y|X]$: the random variable is now the univariate Y, and the X's are regressors.

They estimated the model coefficients 'through a least-squares argument using (iteratively) re-estimated weights, which 'as is well known' gives coefficients that 'are identical with those which would be obtained by the method of **maximum likelihood**.'

NB: Diagnostic versus Prognostic probabilities – and the 'directionalities' involved

The above examples bring out an important point that is missed by today's use of logistic regression of Y (=1/0) on X for fitting **both** diagnostic and prognostic probabilities.

In *dia-gnosis*, the disease or condition is already either present or absent, and (apart from variables such as age and sex, that act as risk factors) many of the X's (symptoms, signs, what is seen on imagining, or in blood tests) will be consequences or manifestations of Y. So the directionality is

$$Y \rightarrow X. \quad (Diagnostic)$$

In *pro-gnosis*, the disease or condition is in the future, i.e., the X's precede Y. So the directionality is

$$X \rightarrow Y. \quad (Prognostic)$$

[31] A multivariate analysis of the risk of coronary heart disease in Framingham. Truett J, Cornfield J, Kannel W. J Chronic Dis. 1967 Jul;20(7):511-24.

[32] Walker, SH, Duncan, DB. Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54:167?179.

Supplementary Exercise 2.11[33] **Lie-detection technology**

2003

The executive summary of the authoritative 2003 report *The Polygraph and Lie Detection* by the National Research Council. 2003. Washington, DC: The National Academies Press. [available for free, `https://www.nap.edu/catalog/10420/the-polygraph-and-lie-detection`] includes these conclusions

> **CONCLUSION**: Notwithstanding the limitations of the quality of the empirical research and the limited ability to generalize to real-world settings, we conclude that in populations of examinees such as those represented in the polygraph research literature, untrained in countermeasures, specific-incident polygraph tests can discriminate lying from truth telling at rates well above chance, though well below perfection. Because the studies of acceptable quality all focus on specific incidents, generalization from them to uses for screening is not justified. Because actual screening applications involve considerably more ambiguity for the examinee and in determining truth than arises in specific-incident studies, polygraph accuracy for screening purposes is almost certainly lower than what can be achieved by specific-incident polygraph tests in the field.

and

> **CONCLUSION**: Basic science and polygraph research give reason for concern that polygraph test accuracy may be degraded by countermeasures, particularly when used by major security threats who have a strong incentive and sufficient resources to use them effectively. If these measures are effective, they could seriously undermine any value of polygraph security screening.

Under the heading **Polygraph Use for Security Screening**, we read

> The proportion of spies, terrorists, and other major national security threats among the employees subject to polygraph testing in the DOE laboratories and similar federal sites presumably is extremely low. Screening in populations with very low rates of the target transgressions (e.g., less than 1 in 1,000) requires diagnostics of extremely high accuracy, well beyond what can be expected from polygraph testing.
>
> Table S-1 illustrates the unpleasant tradeoffs facing policy makers who use a screening technique in a hypothetical population of 10,000 government employees that includes 10 spies, even when an accuracy is assumed that is greater than can be expected of polygraph testing on the basis of available research. If the test were set sensitively enough to detect about 80 percent or more of deceivers, about 1,606 employees or more would be expected "fail" the test; further investigation would be needed to separate the 8 spies from the 1,598 loyal employees caught in the screen.
>
> **TABLE S-1** Expected Results of a Polygraph Test Procedure with an Accuracy Index of 0.90 in a Hypothetical Population of 10,000 Examinees That Includes 10 Spies
>
> **S-1A** If detection threshold is set to detect the great majority (80 percent) of spies

	Examinee's True Condition		
Test Result	Spy	Nonspy	Total
"Fail" test	8	1,598	1,606
"Pass" test	2	8,392	8,394
Total	10	9,990	10,000

> If the test were set to reduce the numbers of false alarms (loyal employees who "fail" the test) to about 40 of 9,990, it would correctly classify over 99.5 percent of the examinees, but among the errors would be 8 of the 10 hypothetical spies, who could be expected to "pass" the test and so would be free to cause damage.
>
> **S-1B** If detection threshold is set to greatly reduce false positive results

	Examinee's True Condition		
Test Result	Spy	Nonspy	Total
"Fail" test	2	39	41
"Pass" test	8	9,951	9,959
Total	10	9,990	10,000

> Available evidence indicates that polygraph testing as currently used has extremely serious limitations in such screening applications, if the intent is both to identify security risks and protect valued employees. Given its level of accuracy, achieving a high probability of identifying individuals who pose major security risks in a population with a very low proportion of such individuals would require

[33] New in 2019, so wording may need some polishing.

setting the test to be so sensitive that hundreds, or even thousands, of innocent individuals would be implicated for every major security violator correctly identified. The only way to be certain to limit the frequency of "false positives" is to administer the test in a manner that would almost certainly severely limit the proportion of serious transgressors identified.

CONCLUSION: Polygraph testing yields an unacceptable choice for DOE employee security screening between too many loyal employees falsely judged deceptive and too many major security threats left undetected. Its accuracy in distinguishing actual or potential security violators from innocent test takers is insufficient to justify reliance on its use in employee security screening in federal agencies. Polygraph screening may be useful for achieving such objectives as deterring security violations, increasing the frequency of admissions of such violations, deterring employment applications from potentially poor security risks, and increasing public confidence in national security organizations. On the basis of field reports and indirect scientific evidence, we believe that polygraph testing is likely to have some utility for such purposes. Such utility derives from beliefs about the procedure?s validity, which are distinct from actual validity or accuracy. Polygraph screening programs that yield only a small percentage of positive test results, such as those in use at DOE and some other federal agencies, might be useful for deterrence, eliciting admissions, and related purposes. However, in populations with very low base rates of the target transgressions they should not be counted on for detection: they will not detect more than a small proportion of major security violators who do not admit their actions.

We have thought hard about how to advise government agencies on whether or how to use information from a diagnostic screening test that has these serious limitations. We note that in medicine, such imperfect diagnostics are often used for screening, though only occasionally in populations with very low base rates of the target condition. When this is done, either the test is far more accurate than polygraph testing appears to be, or there is a more accurate (though generally more invasive or expensive) follow-up test that can be used when the screening test gives a positive result. Such a follow-up test does not exist for the polygraph. The medical analogy and this difference between medical and security screening underline the wisdom in contexts like that of employee security screening in the DOE laboratories of using positive polygraph screening results – if polygraph screening is to be used at all – only as triggers for detailed follow-up investigation, not as a basis for personnel action. It also underlines the need to pay close attention to the implications of false negative test results, especially if tests are used that yield a low proportion of positive results.

A belief that polygraph testing is highly accurate probably enhances its utility for such objectives as deterrence. However, overconfidence in the polygraph – a belief in its accuracy that goes beyond what is justified by the evidence – also presents a danger to national security objectives. Overconfidence in polygraph screening can create a false sense of security among policy makers, employees in sensitive positions, and the general public that may in turn lead to inappropriate relaxation of other methods of ensuring security, such as periodic security re-investigation and vigilance about potential security violations in facilities that use the polygraph for employee security screening. It can waste public resources by devoting to the polygraph funds and energy that would be better spent on alternative procedures. It can lead to unnecessary loss of competent or highly skilled individuals in security organizations because of suspicions cast on them by false positive polygraph exams or because of their fear of such prospects. And it can lead to credible claims that agencies that use polygraphs are infringing civil liberties for insufficient benefits to the national security. Thus, policy makers should consider each application of polygraph testing in the larger context of its various costs and benefits.

Exercises related to the 2003 report

1. Figure out what the authors mean by 'an Accuracy Index of 0.90' in their table.

2. Plot the 2 operating points (from tables S1-A and S1-B) in the (unit-square) ROC space.

3. The authors used a 'spy' prevalence of 10/10,000 or 0.1%. Develop a general equation linking the post-test odds (after a 'Fail" result) that a person is a spy to the pre-test odds that a person is a spy.

4. What (post - "Fail" result) probability would this equation yield if the equation were applied to a person for whom – on the basis of *all of the other evidence* bearing on the case – the probability of his/her having committed a very serious offence against another person is thought be be (a) 20% (b) 50% (c) 80% ?

5. What would the (minimum) pre-test probability have to be in order for the post-test probability to exceed the 50% (the balance-of-probabilities) threshold use in civil law cases?

2016

Refer to the 2016 article "Laboratory and Field Research on the Ocular-Motor Deception Test" ('ODT') by Kircher JC and Raskin DC in the journal *European Polygraph* 10(4): 159-172. You can find it here https://www.polygraph.pl/vol/2016-4/european-polygraph-2016-no4-kircher-raskin.pdf For this exercise, refer specifically to the section 'Field study of the ODT' on pages 168-169.

Exercises related to this 2016 article

1. From the reported percents, back-calculate the numerators and denominators for each of the 5 folds in Table 4, and add across folds to get an overall 'specificity' (for the 'truthful' row, consisting of 83 persons) and an overall sensitivity (for the 'deceptive' row, consisting of 71 persons) [the overall n's are given in paragraph 1]
2. Add this single operating point to the already-plotted points in ROC space.
3. Calculate the (post - 'Fail" result) probability of deception for a person for whom – on the basis of *all of the other evidence* bearing on the case – the pre-test probability is thought to be (a) 0.1% (b) 20% (c) 50% (d) 80%.
4. What would the (minimum) pre-test probability have to be in order for the post-test probability to exceed 50%?

If interested, see the 2018 article https://www.wired.com/story/eye-scanning-lie-detector-polygraph-forging-a-dystopian-future/ and the 2019 one https://www.theguardian.com/technology/2019/sep/05/the-race-to-create-a-perfect-lie-detector-and-the-dangers-of-succeeding

Supplementary Exercise 2.12 Estimating Prevalence using Imperfect Tests

• Antibody surveys suggesting vast undercount of coronavirus infections may be unreliable – Science, April 21, 2020.

• COVID- 19 antibody seroprevalence in Santa Clara County, California. version 1, April 11, 2020. • version 2, April 27, 2020

• Statistical Modeling, Causal Inference, and Social Science: Gelman Blog

• Gelman's July 20 paper (with co-author Carpenter): Bayesian analysis of tests with unknown specificity and sensitivity

• Lawrence Joseph's 1995 paper (with co-authors Gyorkos and Coupal): Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Lawrence Joseph's website.

1. Summarize version 1 of the Santa Clara study in your own words, as if you were one of the authors being interviewed on national television – and had 1 minute (125 words) to do so.[34]
2. Extract the most important concerns in Gelman's (April 19) 'Concerns with that Stanford study of coronavirus prevalence'
3. Describe the main changes in version 2, and give the main points in Gelman's (April 30) 'Updated Santa Clara coronavirus report'
4. Gelman's blog has several followup items, including his May 1 'Simple Bayesian analysis' and his article with Carpenter.

   There is no reference to the 1995 article by Joseph at al. Do the methods in the Joseph at al. article apply to the Santa Clara study? If they do, apply them to the original Santa Clara data and compare the results with those of Gelman and Carpenter.

   Also, briefly describe the models Gelman and Carpenter applied to the additional data in the updated Santa Clara report.
5. How does the 'overall size of iceberg' to the 'amount visible above the water' ratio in the Santa Clara study compare with that in the US CDC data and in this report from Ireland? Suggest reasons for the differences.

[34] Journalists usually ask: why(did you do the study)? how(did you do it)? what did you find? [or, use the 5 'Ws' of journalism, Who, What, Where, When and Why] or the 3: is it new? is it true? does it matter?

Supplementary Exercise 2.13: Reverse-engineering Vaccine Efficacy from hospital statistics, and conversely. And coming up with more helpful terminology for 'Study Designs' in epidemiology

In the summer of 2021, statements such as the following became increasing common in the USA and Canada:

> "90% of COVID-19 patients admitted to our hospitals are unvaccinated."

1. Assuming, for the sake of this exercise, that 65% of the same-age population have been vaccinated, back-calculate a point estimate of Vaccine Efficacy (VE) against hospitalization.
2. In addition to doing the calculations by algebra, depict the situation using a rectangle representing the total population time, say 1 million person weeks, generated by 1 million persons (vertical axis) at risk for 1 week (horizontal axis). Now divide this rectangle horizontally into vaccinated and not vaccinated person days, and within each of the two, show the cases of COVID-19 hospitalization as randomly placed dots. See Figure 1 here as an example. Show 1 - VE as a rate ratio, using an adaptation of the calculation on the top right of the Figure.
3. If the 90% figure is based on a total of 200 hospitalized patients, compute a 95% confidence interval (VI) for the VE. (Assume that the 65% is based on a very large population, so that, by comparison, it has a negligible margin of error.)
4. Suppose, as is the case in some jurisdictions, that the percent vaccinated is only 40%. Use your VE point-estimate to (forward-)calculate what percentage of hospitalized COVID-19 patients would be unvaccinated.
5. In part 3. how would you calculate the CI if, instead, the 65% was based on a *sample* of say just 400 persons.
6. Which (colour) dots in the top half of Figure 1 correspond to the sample of 400?
7. Ask an upper-year epidemiology student what are the common (but not very helpful) names for the types of 'study' you are using in parts 1. and 5. If you were 'patenting' these types of studies, what would you have called them?[35]
8. What is the connection between the type of study in part 5. and the type of study described in the last slide of this presentation ?
9. What is the connection between the type of study in part 5. and the type of study described in this article ? Although the variance formula it uses had been derived by the statistician Yule much earlier, it tends to be called *Woolf's formula*. [We will derive it again on our chapter on proportions and logits.]
10. What is the connection between the type of study in part 5. and the type of study described in example 2 in this chapter of Gary Friedman's very readable 1974 epidemiology textbook?
11. Summarize in words the learning points in this exercise.

[35] See here, in particular, the first 3 paragraphs, Figure 1 and the 2 closing paragraphs. If interested, see also the 'Woolf/Mantel/Miettinen' and 'case control studies' sections in this website

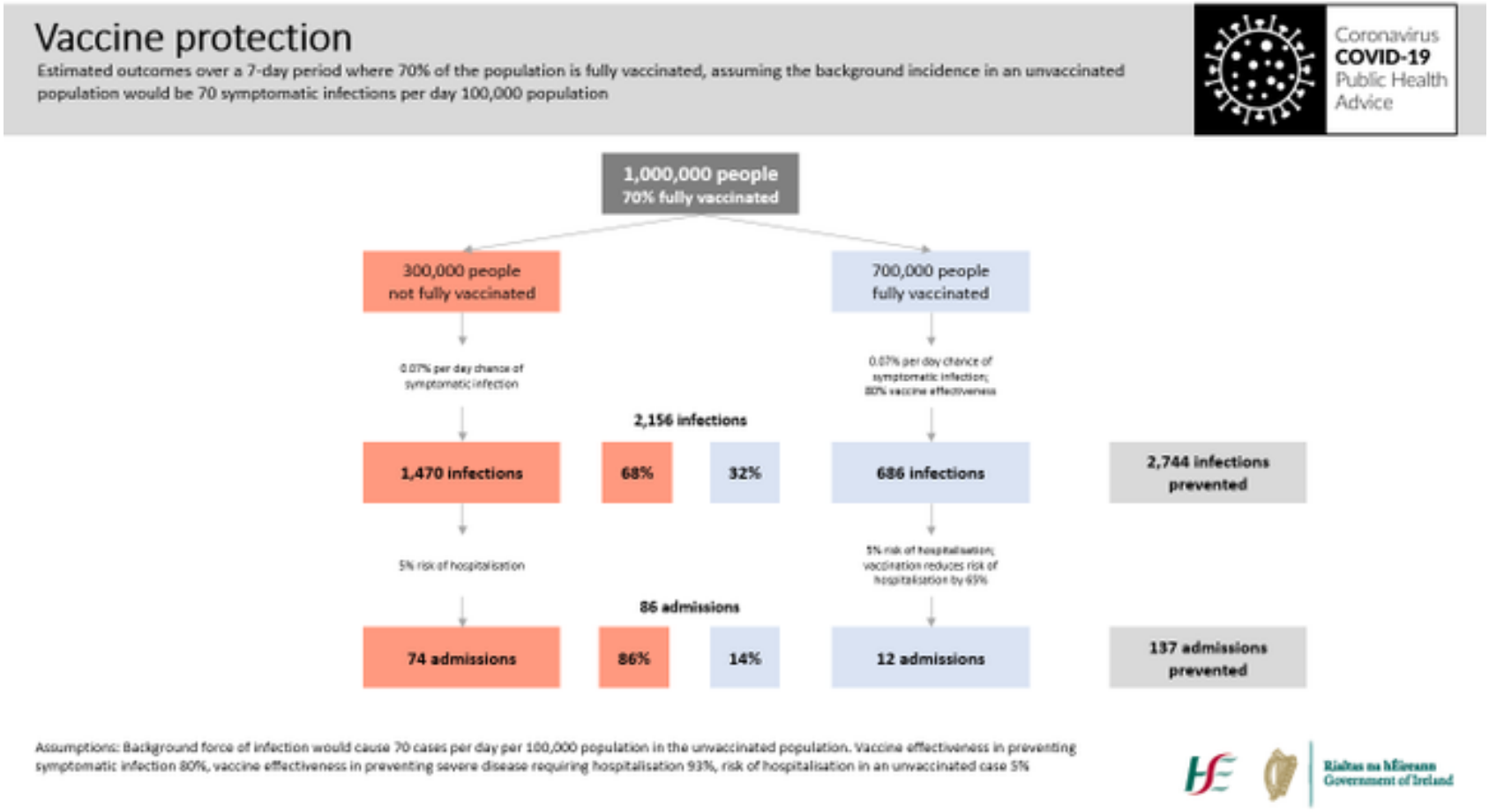
Supplementary Exercise 2.14: Translating numbers into pictures

Here is an expository article, by two by authors whom JH knows, and gave comments to. The diagram, with the relative sizes of the various rectangles within a larger rectangle/square, helps those who prefer pictures over numbers/algebra understand where the 'bottom line' numbers come from. JH had suggested to them a diagram with 'people' (or dots) drawn in, like here.

Here is another , from a newspaper in JH's alma-mater city. (There is, however, JH thinks, a logical flaw in their reasoning, one that overstates what the counterfactual would be)

We saw a nice visualization in part 5 of supplementary exercise 2.5 but it has one text error up front that, sadly, detracts from the otherwise very good article.

If anyone is interested, JH would like to work with them on translating the numbers in the boxes in their Figure .



into more helpful boxes whose dimensions show the *magnitudes* directly.

As JH had mentioned before the course, this type of exercise is designed to get you to have a much bigger (and different) view of what matters in statistics... see `https://jhanley.biostat.mcgill.ca/CommunicationCommunicationCommunication/`.

Supplementary Exercise 2.15: What's the value of a confirmatory PCR test?

[article by David Spiegelhalter and Anthony Masters in The Guardian newspaper, 17 Oct 2021]

After a wave of cases in which a positive lateral flow device (LFD) test was followed by a negative PCR test, a private laboratory handling swab tests has been suspended.

But conflicting results are not a new problem. Back in June, when secondary school students with a positive LFD were retested with a PCR check, over one in eight came back negative. And even without laboratory problems, it is unclear why a negative PCR should trump a positive LFD.

Imagine a (rather strange) legal case with the prosecution alleging that you harbour the virus. In court, it is becoming common to quote a "likelihood ratio" provided by forensic evidence — the relative support for the prosecution versus the defence.

First, the positive LFD is presented by the prosecution. If the virus were present, a recent study estimates around an 80% chance of a positive LFD – higher if you were infectious. Alternatively, if the defence is correct, there is a less than one in 1,000 chance of a false positive LFD. The likelihood ratio is therefore at least 800 (0.8/0.001). As a comparison, the curvature of the spine found on the skeleton in a Leicester car park contributed an estimated likelihood ratio of 200 in favour of the remains being those of Richard III.

The defence retorts with the negative PCR test. If you were infected, the PCR test might miss it around one in 20 times. If there were no virus, then that test is almost certain to be negative. Here, the likelihood ratio is around one in 20.

Combining these two conflicting pieces of evidence gives an overall likelihood ratio of about 40 (800 divided by 20). In a court, that might be reported as "moderate evidence" in favour of you having an infection.

As viral prevalence changes, then the probability of infection following conflicting test results also changes. At the current infection rate in England of one in 60 people, and with labs working well, out of 100 people with a positive LFD followed by a negative PCR, around 40 would actually have the virus and be falsely reassured.

The negative PCR does not outweigh the positive LFD.

[David Spiegelhalter is chair of the Winton Centre for Risk and Evidence Communication at Cambridge. Anthony Masters is statistical ambassador for the Royal Statistical Society]

Exercise:

1. Summarize the article in 50 words.
2. Enter the Fagan nomogram (on page 11) from the right, i.e. at the pre-test $[P(D)]$, at the "current infection rate in England of one in 60 people", and draw a straight line through the likelihood ratio of 800 $[P(T|D)/P(T|\bar{D})]$, ending at a $P(D|T)$ on the left side. What is this post-LFD probability?
3. Use this new probability to again enter the Fagan nomogram on page 11 from the right. and draw a straight line through the likelihood ratio of 1/20 $[P(T|D)/P(T|\bar{D})]$. Now, what is the post-LFD-post-PCR probability?
4. Does this result agree with the calculation reported in the second last paragraph of the article?
5. Are there any assumptions you might challenge? [Hint: see the 2nd column of pages 15 and 16 in relation to exercise 2.10]

Supplementary Exercise 2.16: "Probability problem involving multiple coronavirus tests in the same household"

Posted on April 28, 2021 by Andrew Gelman

Supplementary Exercise 2.17: "Here's a little problem to test your probability intuitions"

Posted on August 1, 2022 by Andrew Gelman

Supplementary Exercise 2.18: The structure/logic of the Likelihood-Ratio-based formulae used to teach post-test probabilities: derived/explained in words and pictures, rather than by algebra.

Refer to the 2019 draft of an article that attempts to do this.

1. Summarize the article in (your own) 50 words.
2. What suggestions do you have for improving it?
3. Would you be interested in being a co-author and working on getting this into a submittable manuscript?

Supplementary Exercise 2.19: How to make Sensitivity and Specificity look like functions of prevalence

Refer to Figures 1 and 2 in this report, and to the Abstract, and in particular to the 1st sentence in the "Interpretation" section of the Abstract. This 'finding' seems to run counter to the teaching concerning the operating characteristics (performance properties) of diagnostic tests, namely that "In practice, sensitivity and specificity are often treated as being *independent from* disease prevalence, defined as pre-test probability of disease or probability of the target condition in the study sample."

1. Imagine you were one of the authors being interviewed on the radio, and that the journalist wanted 1-sentence answers of each of four questions[36]
   (a) Why did you the study?
   (b) What did you do?
   (c) What did you find?
   (d) Does this matter, and if so how?

   Write out a 1-sentence answer for each of these.
2. Convert the fitted curves in Figures 1 and 2 into curves that show sensitivity itself, and specificity itself, as a function of prevalence.[37]
3. In the article, find all the instances of the phrase 'associated with' and suggest an alternative wording that avoids this phrasing.
4. In the first paragraph pf page e927, we read (bold italics added by JH)

   > For sensitivity, compared with the lowest quartile of prevalence, the second, third and fourth quartiles were associated with significantly higher odds of ***identifying a true positive case*** (odds ratio [OR] 1.17, 95% confidence interval [CI] 1.09–1.26; OR 1.32, 95% CI 1.23–1.41; OR 1.47, 95% CI 1.37– 1.58; respectively). For specificity, compared with the lowest quartile of prevalence, the second, third and fourth quartiles were associated with significantly lower odds of ***identifying a true negative case*** (OR 0.74, 95% CI 0.69–0.80; OR 0.65, 95% CI 0.60–0.70; OR 0.47, 95% CI 0.44–0.51; respectively).

   Given JH's advice to speak of *positive* and *negative* test results and persons with and without the *target condition or behaviour* (e.g. a person did of did not cheat in sport, or is or is not infected with the COVID-19 virus), try to improve on the wording of the highlighted phrases.
5. Report on what you find when you search for the term '*case*' in Miettinen's mini-dictionary of terms given in his book Epidemiological Research: Terms and Concepts available online from McGill's library. List a few other other examples of what he considers poor wording.
6. Do the 'findings' of this study *really* conflict with the so-called 'independence assumption' mentioned in the Introduction? Explain your answer.
7. How helpful do you think this article will be for practising physicians?

Supplementary Exercise 2.20:
Multi-cancer blood test shows real promise in NHS study

1. Several numbers, percentages and fractions are mentioned in the [June 2, 2023] BBC article. For each one, if appropriate, give the technical term for the measure.
2. Would you expect these percentages to apply if the test was applied in a 'screening' setting? Why/why not?
3. How does you answer relate to the sensitivity vs. prevalence relationship raised in the report examined in exercise 2.19?

Supplementary Exercise 2.21: Some cool interactive covid infographics from the British Medical Journal

{ posted by Andrew Gelman on May 14, 2003 }

See section 2.4 of the Notes above.

1. Staying for the moment with the direct (but black-box) transit from pre-test to post-test *probabilities*, can you suggest any improvements to the tool?
2. How might you incorporate/show the use of the Likelihood Ratio (LR) - based 'pre-test-odds → post-est-odds' approach?

[36] The first three are modelled after the questions that JH's former colleague, Abby Lippman, told him to expect when he was being interviewed. The fourth comes from what JH heard that a newspaper Editor (used to) ask his/her journalists was proposing to pursue a story: (a) is it *new*? (b) is it *true*? (c) does it *matter*?

[37] Instead of actually drawing a full curves, just make a table of the fitted sensitivities (specificities) at prevalences 0+ϵ, 0.2, 0.4, 0.6, 0.8, and 1-ϵ.