10
Likelihood, probability, and confidence

The supported range for a parameter has so far been defined in terms of the cut-point -1.353 for the log likelihood ratio. Some have argued that the scientific community should accept the use of the log likelihood ratio to measure support as *axiomatic*, and that supported ranges should be reported as 1.353 unit supported ranges, or 2 unit supported ranges, with the choice of how many units of support left to the investigator. This notion has not met with widespread acceptance because of the lack of any intuitive feeling for the log likelihood ratio scale — it seems hard to justify the suggestion that a log likelihood ratio of -1 indicates that a value is supported while a log likelihood ratio of -2 indicates lack of support. Instead it is more generally felt that the reported plausible range of parameter values should be associated in some way with a *probability*. In this chapter we shall attempt to do this, and in the process we shall finally show why -1.353 was chosen as the cut-point in terms of the log likelihood ratio.

There are two radically different approaches to associating a probability with a range of parameter values, reflecting a deep philosophical division amongst mathematicians and scientists about the nature of probability. We shall start with the more orthodox view within biomedical science.

10.1 Coverage probability and confidence intervals

Our first argument is based on the frequentist interpretation of probability in terms of relative frequency of different outcomes in a very large number of repeated "experiments". With this viewpoint the statement that there is a probability of 0.9 that the parameter lies in a stated range does not make sense; there can only be one correct value of the parameter and it will either lie within the stated range or not, as the case my be. To associate a probability with the supported range we must imagine a very large number of repetitions of the study, and assume that the scientist would calculate the supported range in exactly the same way each time. Some of these ranges will include the true parameter value and some will not. The relative frequency with which the ranges include the true value is called the *coverage probability* for the range, although strictly speaking

it is the coverage probability for the method of choosing the range.

We shall start with Gaussian probability model and consider the estimation of the mean μ, from a single observation x, when the standard deviation, σ, is known. The log likelihood ratio for μ is

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2.$$

We saw in Chapter 8 that the range of values for μ with log likelihood ratios above the cut-point of -1.353 is

$$x \pm 1.645\sigma.$$

We shall now show that the coverage probability of this range is 0.90 by imagining an endless series of repetitions of the study with the value of μ remaining unchanged at the true value. Each study will yield a different observation, X, and hence a different range (see Fig. 10.1). The range for any particular repetition will contain the true value of μ provided the true value is judged to be supported by the data X — in other words, provided that

$$-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 > -1.353,$$

where μ now refers to the true value. Writing

$$z = \left(\frac{X-\mu}{\sigma}\right)$$

this condition is equivalent to $(z)^2$ being less than 2.706, and since $(z)^2$ has a chi-squared distribution this occurs with probability 0.90. Hence the coverage probability is 0.90.

Exercise 10.1. In a computer simulation of repetitions of a study in which a single observation is made from a Gaussian distribution with $\mu = 100$ and $\sigma = 10$, the first four repetitions produced the observations 104, 115, 82, and 92. Calculate the log likelihood ratio for $\mu = 100$ for each of these four observations. In which repetitions would the true value of μ have been supported?

The idea of coverage probability has allowed us to attach a frequentist probability, such as 0.90, to a range of parameter values, but we cannot say that the probability of the true value lying within the stated range is 0.90, because the stated range either does or does not include the true value. To avoid having to say precisely what is meant every time the probability for a range is reported, statisticians took refuge in an alternative word and professed themselves 90% *confident* that the true value lies in the reported interval. Not surprisingly the distinction between probability and confidence is rarely appreciated by scientists.

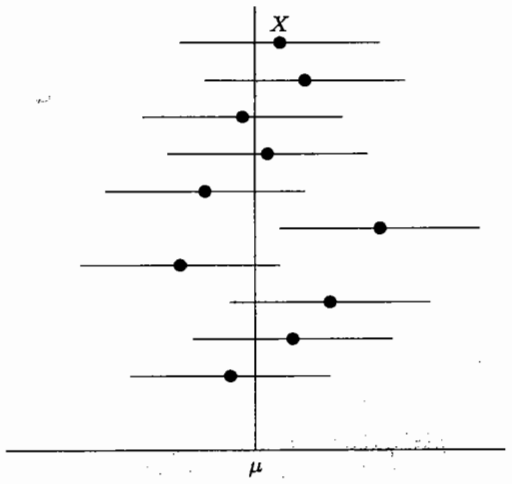


Fig. 10.1. Repeated studies and their supported ranges.

Exercise 10.2. Use tables of the chi-squared distribution to work out the cut-point for the log likelihood ratio which leads to a 95% coverage probability for the corresponding supported range, and give the formula for this range.

We have demonstrated the correspondence between the -1.353 cut-point for the log likelihood ratio and 90% coverage, but only for the case of the Gaussian log likelihood where the standard deviation is known. Fortunately the relationship also holds approximately for other log likelihoods such as the Bernoulli and Poisson. With increasing amounts of data these log likelihoods approach the quadratic shape of the Gaussian log likelihood and the coverage probability for the supported range based on the -1.353 cut-point is approximately 90%. In other words, if M is the most likely value of a parameter and S is the standard deviation of the Gaussian approximation to the likelihood, then the supported range

$$M \pm 1.645S$$

is also, at least approximately, a 90% confidence interval.

This raises the question of how much data is needed to use this approximate theory. For the Bernoulli likelihood, a reasonable guide is that the approximations are good if both D and $N - D$ are larger than 10, but can be misleading if either count is less than 5. In the Poisson case the observed number of events, D, should be larger than 10; there is no restriction on the number of person-years since this is irrelevant to the shape of the log

likelihood curve. In Chapter 12 we discuss what to do when there are too few data to use the approximate theory.

The only likelihood for which the relationship between the supported range and the 90% confidence interval holds *exactly* is Gaussian likelihood, and even here we have made the assumption that the parameter σ is known. In the early years of this century it was shown that the practice of *estimating* the standard deviation using the data and thereafter pretending that this estimate is the true value, leads to intervals with *approximately* the correct coverage probability, providing N is large enough (more than 15).

The intervals we have chosen to present correspond to 90% confidence intervals but 95% intervals are more usually reported in the scientific literature. The routine use of 90% intervals in the epidemiological literature has recently been proposed on the grounds that they give a better impression of the range of plausible values. If you prefer 95% intervals these can be obtained by replacing 1.645 by 1.960 in the calculations.

[★] 10.2 Subjective probability

The second approach to the problem of assigning a probability to a range of values for a parameter is based on the philosophical position that probability is a subjective measure of ignorance. The investigator uses probability as a measure of subjective *degree of belief* in the different values which the parameter might take. With this view it is perfectly logical to say that there is a probability of 0.9 that the parameter lies within a stated range.

Before observing the data, the investigator will have certain beliefs about the parameter value and these can be measured by *a priori* probabilities. Because they are subjective every scientist would be permitted to give different probabilities to different parameter values. However, the idea of scientific objectivity is not completely rejected. In this approach objectivity lies in the rule used to modify the *a priori* probabilities in the light of the data from the study. This is Bayes' rule and statisticians who take this philosophical position call themselves Bayesians.

Bayes' rule was described in Chapter 2, where it was used to calculate the probabilities of exposure given outcome from the probabilities of outcome given exposure. Once we are prepared to assign probabilities to parameter values, Bayes' rule can be used to calculate the probability of each value of a parameter (θ) given the data, from the probability of the data given the value of the parameter.

The argument is illustrated by two tree diagrams. Fig. 10.2 illustrates the direction in which probabilities are specified in the statistical model — given the choice of the value of the parameter, θ, the model tells us the probability of the data. The probability of any particular combination of data and parameter value is then the product of the probability of the parameter value and the probability of data given the parameter value. In

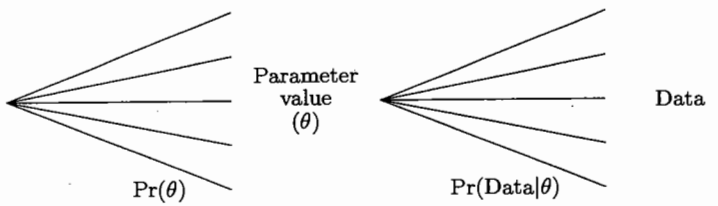


Fig. 10.2. From parameter value to data.

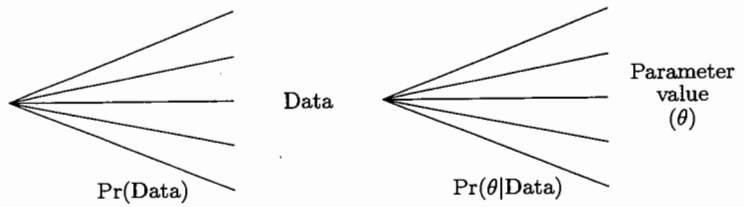


Fig. 10.3. From data to parameter value.

this product, the first term, Pr(θ), represents the *a priori* degree of belief for the value of θ and the second term, Pr(Data|θ), is the likelihood. Fig. 10.3 reverses the conditioning argument, and expresses the joint probability as the product of the overall probability of the data multiplied by the probability of the parameter given the data. This latter term, Pr(θ|Data), represents the *posterior* degree of belief in the parameter value once the data have been observed. Since the joint probability of data and parameter value is the same no matter which way we argue,

$$\Pr(\theta) \times \Pr(\text{Data}|\theta) = \Pr(\text{Data}) \times \Pr(\theta|\text{Data}),$$

so that

$$\Pr(\theta|\text{Data}) = \frac{\Pr(\theta) \times \Pr(\text{Data}|\theta)}{\Pr(\text{Data})}.$$

Thus elementary probability theory tells us how prior beliefs about the value of a parameter should be modified after the observation of data.

We shall now apply this idea to the problem of estimating the Gaussian mean, μ, given a single observation x. The likelihood for μ is

$$\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

If prior to observing x we believe that no value of μ is any more probable than any other, then the prior probability density does not vary with μ and the posterior probability density is proportional to the likelihood. Writing the likelihood as

$$\exp\left[-\frac{1}{2}\left(\frac{\mu-x}{\sigma}\right)^2\right].$$

we see that after choosing the constant of proportionality to make the total probability for μ equal to 1, the posterior distribution for μ is a Gaussian distribution which has mean x and standard deviation σ. The 5 and 95 percentiles of the standard Gaussian distribution are -1.645 and 1.645 respectively so there is a 90% probability that μ lies in the range $x \pm 1.645\sigma$. This range is called a 90% credible interval.

When the quadratic approximation

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2$$

is used for likelihoods such as the Bernoulli and Poisson, a similar argument shows that, provided the prior probability density for θ does not vary with θ, then the posterior distribution for θ is approximately Gaussian with mean M and standard deviation S. It follows that there is a 90% probability that θ lies in the range $M \pm 1.645S$.

It appears from this discussion that the frequentists and the Bayesians end up making very similar statements, differing only in their use of the words *confidence* and *probability*. But to achieve this agreement we have had to make the rather extreme assumption that *a priori* no one value of the parameter is more probable than any other. This is taking open mindedness too far and Bayesians would generally advocate the use of more realistic priors. When there is a large amount of data the posterior is more influenced by the likelihood than by the prior, and both approaches lead to similar answers regardless of the choice of prior. However, when the data are sparse, there can be serious differences between the two approaches. We shall return to this in Chapter 12.

Solutions to the exercises

10.1 When $x = 104$, the log likelihood ratio for $\mu = 100$ is

$$-\frac{1}{2}\left(\frac{104-100}{10}\right)^2 = -0.08.$$

For $x = 115, 82, 92$ the log likelihood ratio turns out to be -1.125, -1.62, and -0.32 respectively. Thus only for $x = 82$ is the support for the true value of μ less than the cut-off value of -1.353. In all other repetitions $\mu = 100$ is supported.

10.2 From tables of chi-squared, the value 3.841 is exceeded with probability 0.05, so

$$\left(\frac{x-\mu}{\sigma}\right)^2 > 3.841$$

with probability 0.05. The log likelihood ratio, which is minus one half of this quantity, is therefore less than

$$-0.5 \times 3.841 = -1.921$$

with probability 0.05. Thus the cut-point for the log likelihood ratio is -1.921.

10 Likelihood, probability, and confidence

In their preamble to this chapter, the authors tell us that

> *There are two radically different approaches to associating a probability with a range of parameter values, reflecting a deep philosophical division amongst mathematicians and scientists about the nature of probability. We shall start with the more orthodox view within biomedical science.*

A dictionary that JH consulted gave 7 definitions of 'orthodox'. The last two explained the use of the '(initial capital letter) of pertaining to, or designating the Eastern Church, esp. the Greek Orthodox Church; or 'of, pertaining to, or characteristic of Orthodox Jews or Orthodox Judaism.' The other five were:

i. of, pertaining to, or conforming to the approved form of any doctrine, philosophy, ideology, etc.

ii. of, pertaining to, or conforming to beliefs, attitudes, or modes of conduct that are generally approved.

iii. customary or conventional, as a means or method; established.

iv. sound or correct in opinion or doctrine, esp. theological or religious doctrine.

v. conforming to the Christian faith as represented in the creeds of the early church.

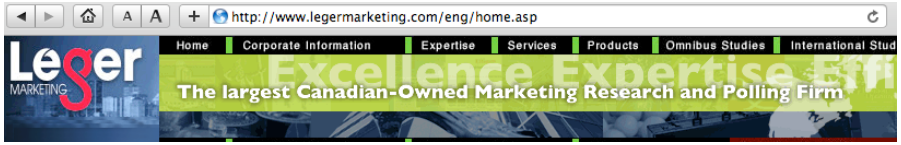
Clayton & Hills completed their book in 1993. Since then, propelled by greater computer power, and by people like Clayton's Cambridge colleague David Spiegelhalter,[1] the Bayesian 'approach' to 'associating a probability with a range of parameter values' has become more common; it has not yet reached the status of 'customary or conventional, as a means or method; established.'

We should not take Clayton and Hills' use of the word 'more orthodox' to describe the *frequentist* approach to mean that the Bayesian approach 'does not conform to the approved form of analysis' or is in some sense 'wrong.'

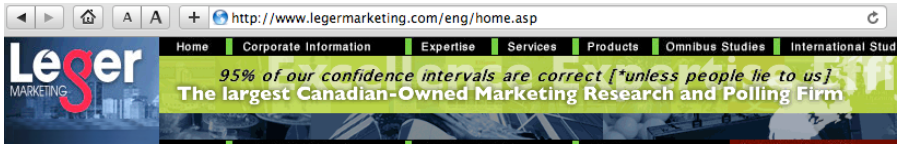
[1] A link to one of Spiegelhalter's books is given in the top right corner of the BIOS602 website.

10.1 Coverage probability and confidence intervals

The key point about the **frequentist** confidence interval is – as they state at the very end of the first paragraph (top line of p90) – that the probability is associated with the *method* of choosing (constructing) the range and not with the realized interval per se. Our 'confidence' derives from the fact that 95% (if we construct '95%' intervals) of intervals so constructed trap (contain) the true value of the parameter: it is as if we buy a product from a producer, 95% of whose products in the past 'have worked correctly' and 95% of whose products in the future 'will work correctly.' Or, we choose a surgeon based on his 'successful in 95% of cases' track record [the one difference is that if we choose this surgeon, we will usually get to know quite soon after the operation whether it was successful or not (i.e., the truth becomes evident) whereas with a specific confidence interval we may never get to know if it contained the true value of not: one situation where we do is at election times, when the election results can be compared with pre-election confidence intervals from polls. JH has often suggested that in addition to their other claims,



polling companies that use sample-surveys should add the following one...



Clayton and Hills warn about **misusing** the probability: many statisticians misuse the terminology as well.

> *The idea of coverage probability has allowed us to attach a frequentist probability, such as 0.90, to a range of parameter values,* ***but we cannot say that the probability of the true value lying within the stated range is 0.90, because the stated range either does or does not include the ttue value****. To avoid having to say precisely what is meant every time the probability for a range is reported,* ***statisticians take refuge in an alternative word*** *and professed themselves 90% confident that the true value lies in the reported interval.* ***Not surprisingly the distinction between probability and confidence is rarely appreciated by scientists****.*

One reason the two can be mixed up is that the two statements A and B

$$Prob[\ \mu - 1.96\sigma/\sqrt{n}\ <\ \bar{y}\ <\ \mu + 1.96\sigma/\sqrt{n}\ ] = 95\%\ (A)$$

i.e.,

$$Prob[\ \bar{y}\text{ 'falls' in the interval }\ \mu - 1.96\sigma/\sqrt{n}\text{ to }\mu + 1.96\sigma/\sqrt{n}\ ] = 95\%\ (A')$$

and

$$Prob[\ \bar{y} - 1.96\sigma/\sqrt{n}\ <\ \mu\ <\ \bar{y} + 1.96\sigma/\sqrt{n}\ ] = 95\%\ (B)$$

i.e.,

$$Prob[\ \mu\text{ 'falls' in the interval }\ \bar{y} - 1.96\sigma/\sqrt{n}\text{ to }\bar{y} + 1.96\sigma/\sqrt{n}\ ] = 95\%\ (B')$$

seem to be *mathematically* equivalent; after all, if we are xx% confident that Montreal is less than 4000 Km from Vancouver, then we should also be xx% confident that Vancouver is less than 4000 Km from Montreal!

But (at least in the frequentist approach) versions A and B not *logically* equivalent: $\bar{y}$ and μ do not have the same status, whereas the two cities in the distance statement do. In version A, the focus, i.e., the subject of the sentence, is $\bar{y}$, and the statement is concerned with the probabilistic behaviour of the *data* ($\bar{y}$). In version B, the focus, i.e., the subject of the sentence, is μ, and the statement gives the impression that it is concerned with the probabilistic behaviour of (or uncertainty concerning) the parameter of interest (μ). But in a frequentist approach, the parameter is regarded as a fixed but unknown quantity, and so it is difficult to think of it as 'falling' at different locations. For example, say we are concerned with the value of the physical parameter c [the speed of light]. Whereas we can think of measurements (estimates) of c as falling on both sides of c, we cannot do the reverse and think of c as moving [falling] around in the literature... it is the (data-based) estimates that move or fall around the target: the target (the speed of light) itself does not move.

Work in the early years of this century: The reference to the early years of this century at the top of page 94

> *The only likelihood for which the relationship between the supported range and the 90% confidence interval holds exactly is Gaussian likelihood, and even here we have made the assumption that the parameter σ is known.* ***In the early years of this century*** *it was shown that the practice of estimating the standard deviation using the data and thereafter pretending that this estimate is the true value, leads to intervals with approximately the correct coverage probability, providing N is large enough (more than 15).*

is to the 1908 work of Student (Gosset).

10.2 Subjective probability

At a minimum, this topic should be given an entire chapter, not just a section. A very good comprehensive book on this subject is

> *Bayesian Data Analysis, Second Edition* by Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin (Chapman & Hall/CRC Texts in Statistical Science).

It provides a full theoretical treatment of the subject, as well as a wide variety of examples. A practical application of the topic to one specific area can be found in the equally good

> *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* by David J. Spiegelhalter, Keith R. Abrams and Jonathan P. Myles (Wiley 2004; link on top right corner of BIOS602 website).

Even though its applications are limited, this book also deals with the broader and practical issues in the Bayesian approach, and covers several aspects, such as the elicitation of priors, that apply beyond Health-Care Evaluation.

One of its many attractive features is that it uses the 'normal-prior, normal-likelihood' approximations for just about all of the applications, whether they involved measured levels of blood pressure or cholesterol or blood pressure, or proportions or event rates, or ratios or differences of these; thereby, it reflects the reality that in many applications, we have enough data so that these approximations are reasonable. Where other texts might treat proportions, rates and means in separate chapters, this text treats them all in one. In so doing, it reinforces the point that the focus of the Bayesian approach is on the *parameter* of interest, and not on the *data* per se. After all, whereas Bernoulli, Binomial and Poisson data are recorded on discrete scales (0, 1, 2, ...), their *parameters* (π, λ, and, if necessary, their transforms), are all recorded on a *continuous* scale.

In addition, its first author is one of the most eloquent proponents of the Bayesian approach, and was a driving force behind the WinBugs project. Since 2007, he has been the Winton Professor of the Public Understanding of Risk, in recognition of his important and high-profile contributions to Public Policy in Britain. To get an idea of the breadth of his work, visit `http://www.statslab.cam.ac.uk/Dept/People/Spiegelhalter/davids.html` .

10.2.1 Why not start with 'Objective' probabilities? Bayes Rule works with both objective and subjective probabilities

Spiegelhalter et al. explain the subjective interpretation in the second half of page 50 of their book:

> The vital point of the subjective interpretation is that **Your** probability for an event is a property of **Your** relationship to that event, and not an objective property of the event itself. This is why, pedantically speaking, one should always refer to probabilities **for** events rather than probabilities **of** events, and the conditioning context H used in Section 2.1.1 includes the observer and all their background knowledge and assumptions.

The Bayesian approach relies on a fundamental mathematical formula or rule derived from universally-agreed-upon mathematical axioms. *'Priors' are not necessarily subjective*: as we will illustrate in our first several example, they can be objective. JH argues that teachers who begin their introduction to Bayesian analysis by asking students to 'assume' certain priors but offer no documented basis for these may weaken their case for the Bayesian approach, and risk turning off their audience. That is why, below, even though the two examples (a specific person's cholesterol level, and a specific person's age) have the same mathematical structure, he begins with the one where the source for the prior is more objective.

10.2.2 Let's start with a parameter concerning an individual

Likewise, it should be easier to engage students and other clients with an application where the inference concerns a parameter related to a single individual, rather than a group. That is why, below, even though two possible applications (a specific person's mean level, and a specific a specific population's mean level) have the same mathematical structure, we begin with the one where the target parameter concerns the individual.

10.2.3 The parameter scale: categorical / interval

Just as most books do, Chapter 2 of C&H and in Chapters 2 of Spiegelhalter et al. first illustrate Bayes rule using applications involving a parameter that can take on only 2 (or just a few) possible values (e.g., genetic carrier or not) When such examples also involve binary or categorical data, they allow the entire two-dimension grid of (possibilities for prior) $\times$ (possibilities for new data) to be displayed, the probabilities conditional on a specific new-data value to be calculated, collected together and re-scaled, and thus the posterior probabilities to be derived. When we move up to a parameter that takes values on a numerical ('continuous') scale, the display becomes more complicated, and unless we are willing to represent the parameter range using bins, we have to resort to integration to perform the rescaling.

It is interesting to trace Spiegelhalter et al.'s transition from the use of objective to subjective priors, from parameters involving the individual to ones involving a population, and from categorical to continuous parameter scales.

It is interesting to trace Spiegelhalter et al.'s transition.

10.2.4 Often, the motivations for using a Bayesian approach are more practical than ideological

Whereas much ink has been spilled on arguments as to why individual studies should not incorporate outside information but rather should focus on what *new* information they *add* to the literature, the main purposes of many Bayesian analysis are merely to

i. take advantage of a very flexible and computationally-tractable tool for model-fitting,

ii. be able to communicate directly, in probabilistic terms, about a range of parameter values, something that is not possible with the frequentist approach.

Most such analyses have not used strong priors; instead, the posterior distributions are largely determined by the new data, and not by the prior.

The following are examples of parameter/data combinations where the subject of the inference is an **individual**:

Para.	Example	New Data	Example
Qual.	Haemophilia carrier	Qual.	Son affected?
..	Innocence (crime)	..	Blood type
..	Cystic Fibrosis	Quant.	Salt in sweat
..	HIV Status	..	Optical density
Quant.	Income Level	Qual.	Postal Code
..	Cholesterol Level	..	Parental History, Age Group
..	Cholesterol Level	Quant.	Chol. Measurement
..	Age	..	Anthropometry

You might wish to further refine or expand on the above tabulation, by distinguishing between the basis for and objective vs. subjective nature of the 'prior' distribution of the parameter [it might be based on a mix of both objective and subjective data; and you might likewise be able to expand the table to show examples of objectively- vs. subjectively-established new data.]

10.2.5 Inference re a parameter concerning an individual

• **Target**: state/trait, qualitative; New data: qualitative

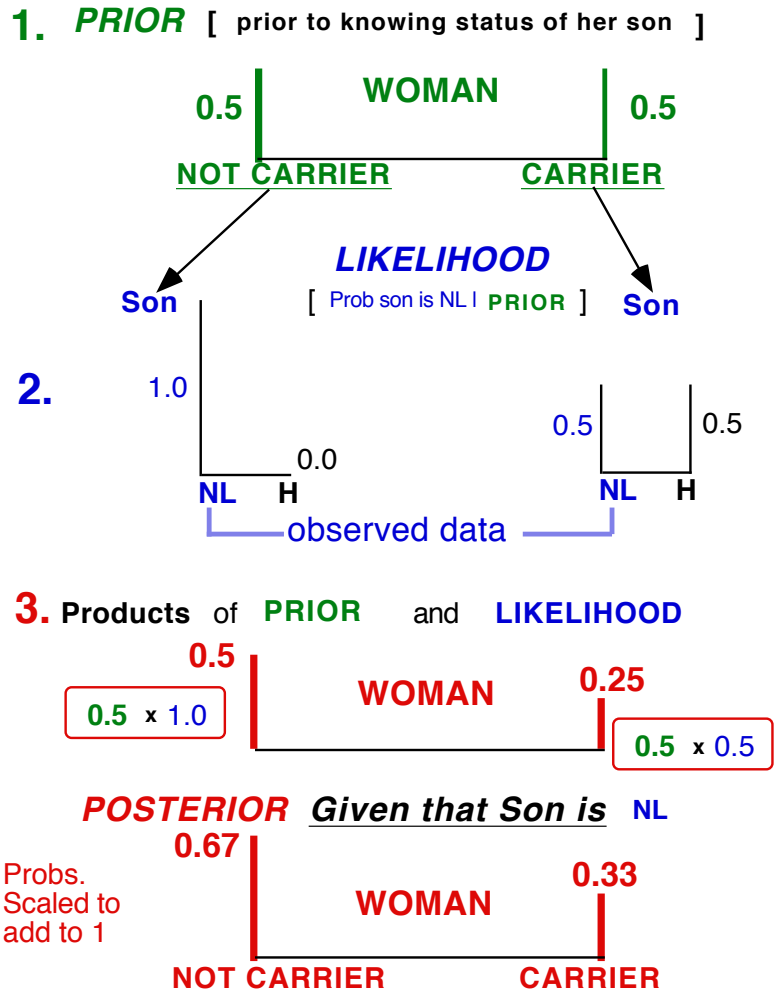
Most clinical epidemiology textbooks and courses describe the use of Bayes Rule to make inferences concerning the updating of an individual's probability *of* (Spiegelhalter would say *for*) all-or-none phenomenon [such as a genetic trait or current disease state (diagnosis context) or future disease state (prognosis context)][2] using new data of a qualitative nature and so (other than mentionning in passing the classic haemophilia carrier example, opposite)

Bayes Theorem : Haemophilia

Brother has haemophilia => Probability (**WOMAN** is Carrier) = 0.5

New Data: Her **Son** is Normal (NL) .

Update: Prob[Woman is Carrier, given her son is NL] = ??



we will not repeat these examples in detail here. Worked examples can be found in C&H Chapter 2 and in JH's accompany notes, or in example 3.1 of Spiegelhalter et al.

[2] Spiegelhalter et al., §3.2, call it "Bayes Theorem for two hypotheses."

Another interesting example is the case of **screening** [during pregnancy] for fetal aneuploidy. It started in the mid 1960s, using maternal age as the screening test. The major risk factor for Down Syndrome is maternal age. One of the sources for age-specific prevalences is the table above, taken from the article Chromosomal Abnormality Rates at Amniocentesis and in Live-Born Infants by Hook EB in JAMA 249:2034-2038, 1983.

Table 1.—Regression-Derived Estimated Rates per 1,000 of Cytogenetic Abnormalities by Maternal Age at Time of Amniocentesis

Maternal Age, yr	47,+21*	47,+18	47,+13†	47,XXX	47,XXY	Other Clinically Significant Abnormalities‡	All Abnormalities§
33‖	2.4	0.6	0.4	0.4	0.4	1.1	4.6-5.4
34‖	3.1	0.8	0.4	0.5	0.5	1.2	5.8-6.5
35	4.0	1.0	0.5	0.6	0.6	1.3	7.4-8.0
36	5.2	1.3	0.6	0.7	0.8	1.3	9.5-9.9
37	6.7	1.6	0.6	0.8	1.0	1.4	12.1-12.2
38	8.7	2.1	0.7	1.0	1.2	1.5	15.4-15.2
39	11.2	2.6	0.9	1.2	1.5	1.6	19.6-19.0
40	14.5	3.3	1.0	1.4	1.9	1.7	25.0-23.8
41	18.7	4.2	1.1	1.7	2.4	1.8	31.9-29.9
42	24.1	5.2	1.3	2.0	3.0	1.9	40.7-37.6
43	31.1	6.6	1.5	2.4	3.8	2.0	51.9-47.5
44	40.1	8.4	1.8	2.9	4.7	2.2	66.1-60.0
45	51.8	10.6	2.0	3.4	5.9	2.3	84.3-76.0
46	66.8	13.3	2.4	4.1	7.4	2.4	107.5-96.5
47	86.2	16.9	2.7	4.9	9.3	2.6	137.1-122.6
48	111.2	21.3	3.1	5.9	11.7	2.7	174.8-155.9
49	143.5	26.9	3.6	7.0	14.6	2.9	222.9-198.6

*A value of 0.08 per 1,000 should be added to this figure to allow for structural rearrangements associated with Down's syndrome.

†A value of 0.06 per 1,000 should be added to this figure to adjust for structural rearrangements associated with Patau's syndrome.

‡Includes structural rearrangements associated with Patau's and Down's syndromes.

§The first value of the range given is derived from an regression equation analysis on all abnormalities; the second by adding values for all abnormalities. Including abnormalities of more questionable significance would result in addition of about 2.7 per 1,000 at the lower ages (around 35 years) and about 2.1 per 1,000 at the older ages to the second values given in the range.

‖Values extrapolated from regression equation derived for ages 35 to 39 years.

New developments in maternal serum and ultrasound screening have made it possible to offer all pregnant patients a non-invasive screening test to assess their risk of having a fetus with Down syndrome or trisomy 18 to determine whether invasive prenatal diagnostic tests are necessary. The articles by Taipale P, Hiilesmaa V, Salonen R, Ylstalo P. Increased nuchal translucency as a marker for fetal chromosomal defects. New Engl J Med. 1997;337:1654-8. and by Wapner et al. First-Trimester Screening for Trisomies 21 and 18. N Engl J Med 2003;349:1405-13. (ABSTRACT only) describe how these **age-specific probabilities can be updated by data from imaging and blood tests**.

• **Target**: state/trait, qualitative; New data: quantitative

Fewer clinical epidemiology textbooks describe the use of Bayes Rule to make inferences concerning the updating of an individual's probability *of* (Spiegelhalter would say *for*) all-or-none phenomenon [such as a genetic trait or current disease state (diagnosis context) or future disease state (prognosis context)] using new data of a quantitative nature. Two texts that do so in some detail are Clinical Epidemiology: How to Do Clinical Practice Research by R. Brian Haynes, David L Sackett, Gordon H Guyatt, and Peter Tugwell. and Clinical Epidemiology: The Essentials by Robert H Fletcher and Suzanne W Fletcher.

The 'new data' in the example in Fletcher and Fletcher is xxx; The new data in the example in Sackett et al. is xxx. Rather than the more classical form based on updating the probability of the disease state [DS] of concern using the observed value (y) of the new information-item Y, both use the *odds* of that state in the updating formula,

$$PostOdds = PriorOdds \times LikelihoodRatio[y],$$

since it is simpler to remember than the more classical (algebraicly equivalent) form based on probabilities:

$$PostProb[DS|y] = \frac{PriorProb[DS] \times P[y \mid DS]}{PriorProb[\overline{DS}] \times P[y|\overline{DS}] + PriorProb[DS] \times P[y|DS]}.$$

It also separates more cleanly the two items of information, the initial $PriorOdds[DS]$ or $PriorProb[DS]$ concerning the individual in question, and the $LikelihoodRatio[y] = P[y|DS] \,/\, P[y|\overline{DS}]$ involving the quality of the new information.

In reality, all quantitative measurements are recorded with some degree of rounding. If the 'bin' associated with Y "=" y is $y - 0.5\delta y$ to $y + 0.5\delta y$, and if $f[y|DS]$ is the probability density, then $P[y|DS] \approx f(y|DS] \times \delta y$ and the δy cancels out in $LikelihoodRatio[y] = f(y|DS]/f(y|\overline{DS}]$.

The Excel spreadsheet (under Resources) has an example of using the fetal heart rate to address the probability that the baby is male or female (some might be tempted to say *predict* if the baby is male or female, but the baby is already either male of female, so it's a matter of *post* diction rather than *pre* diction.)

• **Target**: level, quantitative;[3] New data: qualitative

Since you can think of a qualitative variable as a special case of a qualitative one (it might just have the values 0 and 1), we will not work through a specific case. Below, we will work with an example where the target is a person's age, estimated visually, and the new information is number of years since the person obtained a PhD. If instead, the new information was the fact that the person is a grandparent, the updated probabilities associated with each age involve the prevalence of grandparents as a function of age.

The similarity of the calculations involved with new data items whether they are recorded on qualitative and quantitative scales emphasizes the fact that the complexity in a Bayesian analysis is the amount of calculation required: we need to calculate the scaling factor to be applied to the the $Prior(\theta) \times P(observed\ new\ data \mid \theta)$ products so that they become posterior probabilities and thus sum or integrate to 1. To do so, we sum the products over the possible θ values (or categories, or bins) of the prior. Unlike with frequentist p-values (tail areas), there is no summation over other (unobserved) values of the random variable Y that provides the new information; rather, only the observed value y of Y is considered; other 'might have been' Y values are not.

• **Target**: level, quantitative; New data: quantitative

Spiegelhalter's example: blood pressure

Spiegelhalter et al. illustrate this with Example 3.4 SBP ("Bayesian analysis for normal data"): Suppose we are interested in the long-term systolic blood pressure (SBP) in mmHg of a particular 60-year-old female. We take two independent readings 6 weeks apart, and their mean is 130. We know that SBP is measured with a standard deviation $\sigma = 5$. What should we estimate her SBP to be?

They then go on to give the frequentist ('standard') 95% confidence interval, of 123.1 to 136.9, centered on the measured value of 130. But, they continue...

> However, we may have considerable additional information about SBPs which we can express as a prior distribution. Suppose that a survey in the same population revealed that females aged 60 had a mean long-term SBP of 120 with standard deviation 10. This population distribution can be considered as a prior distribution for the specific individual, and is shown in Figure 3.3(a):

The posterior distribution, computed from the combination of the 130 measured on the woman, and the prior, is centered on 128.9 and the 95% interval is 122.4 to 135.4.

> "This posterior distribution reveals some shrinkage towards the population mean, and a small increase in precision from not using the data alone.
>
> Intuitively, we can say that the woman has somewhat higher measure- ments than we would expect for someone her age, and hence we slightly adjust our estimate to allow for the possibility that her two measures happened by chance to be on the high side. As additional measures are made, this possibility becomes less plausible and the prior knowledge will be systematically downgraded."

It's a pity that these authors did not give an actual source for this prior, or be a bit more realistic about "a same population of females aged 60 with a mean long-term SBP of 120". This must be a somewhat selected, healthier-than-average population, since we might find such a mean of 120 in 25-year old women; in the general population of 60-year old women, it is higher than that.

Irwig's example: cholesterol

One set of authors who did go into some detail about a similar situation are Irwig et al. in a very nice medically-useful – and didactic article – in JAMA in 1991.[4] It concerns cholesterol, and begins with a single measurement on one person, before dealing with an average of several measurements on the same person. It also gives separate charts for persons of different ages, and deals not just with point and interval estimates, but also derives probability statements for the possibility that the person's true cholesterol is above some threshold that should trigger intervention. The appendix is a nice tutorial for combining information.

Their Abstract begins:

> An individual's blood cholesterol measurement may differ from the true level because of short-term biological and technical measurement variability. Using data on the within-individual and population variance of serum cholesterol, we addressed the following clinical concerns: Given a cholesterol measurement, what is the individual's likely true level? The confidence interval for the true level is wide and asymmetrical around extreme measurements because of regression to the mean. Of particular concern is the misclassification of

[3] Spiegelhalter et al., in §3.5, "the most important section in this book", refer to this as "Bayes Theorem for general quantities."

[4] "Estimating an Individual's True Cholesterol Level and Response to Intervention" by Les Irwig, Paul Glasziou, Andrew Wilson, Petra Macaskill JAMA. 1991;266:1678-1685

people with a screening measurement below 5.2 mmol/L who may be advised that their cholesterol level is "desirable" when their true level warrants further action.

The first half of the paper, which deals with two related topics, (a) Estimating the True Cholesterol Level, and (b) assesing the Probability of Misclassification shows the primary elements, and these notes will focus on the highlights. [after these, extensive excerpts will be included]

The results for (a) and (b) were presented as 2 Figures. The first gave the (posterior) credible interval for a person's true cholesterol level based on either 1, or an average of 3, measurements, using on the horizontal axis the measured value, and on the vertical one the point and 95% credible interval. Using a graph (rather than a formula) allows the clinician to use it for all possible 'what if's.

Below, we will illustrate it using one specific example, a person whose measured value was 7.15.

The second uses the (posterior) credible interval to calculate the probability that someone with a specific measured value has a true level that is above a certain threshold level used in treatment guidelines.

Thus, the key tool is the posterior distribution itself, and so we give the statistical basis for this.

Reasons to take a Bayesian approach

The reason this problem arises in the first place is because of short term biological variability in the quantity of interest in the person in question. If we were measuring a person's height, we could do so carefully at just one time-point:[5] it would not be different a week or month from now; it remains quite stable over several years. The same is not the case for a person's cholesterol level: even if we measured it very carefully at one time, it would be genuinely different a week or month later, even in the absence of any intervention of lifestyle change. (The same applies, more strikingly, for other blood levels such as C-reactive protein (CRP), which is a marker of inflammation).

Given this, and that any single measurement, or any average of a finite number of determinations, is imprecise.

So what's new? don't we meet this issue all the time in statistics?

The point of Irwig's article is that we should not rely solely on the estimate based on the person's measurements, but rather should combine it with an estimate based on outside information.

Under Resources for Bayes, the BIOS602 website has a nice example, "Bayesian integration in sensorimotor learning", which illustrates how as humans, we automatically combine estimates of different precisions into one more precise estimate, and do so using the same mathematical laws that are used in the Bayesian approach!

The same reasoning is at work when a physician repeats a measurement that seems extreme. In so doing, (s)he is not relying only on the point or interval estimate provided by the measurement itself: rather (s)he is also using knowledge of how this measurement behaves in other similar persons! And what we know about others, even if collectively, can help us with an individual.

Where to start?

So, we proceed in the same temporal sequence a physician does in a specific person of a given age and sex: we first use the estimate based just on the age and sex information; we them combine this the value obtained in a single measurement of this person. To make it concrete, we take the case where the measured cholesterol level was 7.15 mmol/L.

The best place to start is with the broader information: the distribution of true cholesterol levels for the entire population (conceptual or actual) from which the person is consider (conceptually) a randomly selected individual.

After all, as soon as we know a person's age and sex (and any other factors that determine the centre and spread of this distribution) we have some idea of where the specific individual's level is.

The key components, and a caveat about notation

We have to distinguish two distributions, and be quite careful and clear about notation:

- Each person has (even if we cannot determine it precisely) a '*true*' level (see full Irwig article). The distribution of the true levels of all of the persons in this population can be thought of as the distribution of a random variable T. Denote the mean of this population of true levels by $\mu_{T_{pop}}$ and variance by $\sigma^2_{T_{pop}}$.

  Clearly, both because we cannot study each person in the population, and because for the sample of individuals we can study, we cannot measure each sampled person often enough to 'know' their true values, we cannot directly observe this distribution or estimate the values of its 2 parameters. However, we can get at them indirectly, using 'another' distribution...

[5] it does vary slightly over the day, but, be keep it simple, we could speak of one's height at mid-day

- There is also the distribution of values one *can* observe by measuring each person (or a sample of them) once (or some twice, or some even more often).

  In this situation, for a specific person, whose true value is T, we don't 'see' T itself. Rather, because we measure the person once, on a random day, we get to observe the amalgam $T+\epsilon$. The ϵ reflects the possible fluctuations around T, caused partly genuine short-term biologic variations and partly by by any measurement 'errors' (technical). . These fluctuations are often referred to as the 'within-' or 'intra-'individual variation, and – even though it is not all 'error', their variance is often denoted by the subscript e, as in σ_e^2 (JH prefers σ_w^2, with w standing for 'within' persons; likewise, he prefers σ_b^2, with w standing for 'between' persons.)

  With the assumption that each ϵ is independent of each T, the observable variance, across the persons, in their measured values (1 per person) will be $\sigma^2_{T_{pop}} + \sigma_e^2$ or $\sigma_b^2 + \sigma_w^2$

  Some refer to this (wider) variance as the 'total' variance, and write it as

  $$\sigma^2_{total} = \sigma^2_{T_{pop}} + \sigma_e^2$$

  or, as Irwig do, for short as

  $$\sigma^2_{total} = \sigma^2_{pop} + \sigma_e^2.$$

  From this formulation, it becomes clear that if one has an estimate of σ^2_{total}, and if one has an estimate of σ_e^2, one can subtract them and obtain an estimate of σ^2_{pop}, or what JH denotes by $\sigma^2_{T_{pop}}$. σ^2_{total} can be estimated directly as the observed variance of the values obtained by measuring a sample of persons once, and σ_e^2 can be estimated, also directly, by measuring a sample of persons more than once, and pooling the person-specific estimates of the within-person variation. [This can also be done within a single study, using classical anova to estimate the separate components of variance; such analyses can also handle the case where some persons are measured once, some twice, some more often.]

  Non-statisticians do not always appreciate these distinctions, and sometimes are fuzzy about what 'between'-variance refers to. The most common mistake is to treat the observed variance of the values obtained by measuring a sample of persons once as if it were an estimate of σ^2_{pop} or $\sigma^2_{T_{pop}}$. Yes, it looks like one is examining the variation between individuals (since one has 1 measurement for each of several persons) but in fact only some portion of that variation is caused by the fact that each person's T is different from another person's T: it also reflects the fact that *one measurement* of a specific person's T is different from *another measurement of that same person's* T. A good way to appreciate these distinctions is to examine the full version of the classic anova table, which shows not just the *observed* 'between' mean square and the *observed* 'within' mean square – in effect the calculated *statistics* – but also the *expected* 'between' mean square and the *expected* 'within' mean square, The latter are theoretical, not data-based, and are functions of the (unknown, and unknowable) *parameters*. When Irwig et al. write of "the *observed* variance of the population using a single measurement", and give it the notation σ_o^2, they are referring to what we above call σ^2_{total}.

Information needed in order to 'merge' two estimates

If the distributions of T and ϵ are reasonably Gaussian, needed are values for

- $\mu_{T_{pop}}$ and $\sigma^2_{T_{pop}}$, or μ_{pop} and σ_p^2 for short
- σ_e^2
- y, the measurement on the person

In the case of cholesterol, the distributions of T across individuals of the same age band, and of the possible measurements on a single individual about that person's T, are reasonably Gaussian, but since the population mean is higher in higher age age bands, and since individuals in these older populations tend to have higher Ts, it is not surprising that $\sigma^2_{T_{pop}}$ and σ_e^2 are also higher in these higher age bands and at higher T's. Irwig et al. avoided having to deal with this heteroskedasticity by using the distributions of *log* cholesterol levels throughout, and converting back to the levels themselves at the very end.

So, for the presentation below, the various quantities (T, θ, μ, the various σ's and y) are, unless otherwise noted, to the levels on the log scale.

Irwig et al. were able to locate several surveys that have accurately established values for σ^2_{pop} or $\sigma^2_{T_{pop}}$.

e.g., for the younger one, band "A", they established the following values:

- $\mu_p = 1.63$;[6] $\sigma_p = \sqrt{0.03347} = 0.183$
- $\sigma_e = \sqrt{0.00589} = 0.077$.[7]

[6] They never say this explicitly; instead they say that the mean level (in the 'un-logged' scale) is 5.2; since the mean of a lognormal distribution is $exp[\mu_{logs} + \sigma^2_{logs}/2]$, one can back-calculate that, in the log scale, $\mu = 1.63$ [it is also the median]. Since $\exp[1.63] = 5.1$, the distribution of 'un-logged' levels must itself be quite close to Gaussian.

[7] $0.077/1.63 \approx 0.05$,, implying a within-person coefficient of variation (CV) of 5% in the

The 'math' of merging

In the text, the authors simply state that

> The best estimate of an individual's true cholesterol level can be shown to be a combination of these two signals [estimates], giving more weight to the signal [estimate] with least noise, ie, weighting by the inverse of the variances. This weighted average provides an estimate of the regression to the mean[8] for an individual.

Before quoting from their more technical Appendix, JH has, for simplification, omitted the subscript i, used θ instead of "μ_i" to refer to the true (log)level of the person in question, and used y instead of x for the single measurement of the (log)level for that person. With these, The Appendix begins...

> Suppose the "true" values in the population have a mean μ_p and variance σ_p^2, and suppose the within-individual variance is σ_e^2. [...] If we take a single screening measurement, y, with measurement variance σ_e^2, the best estimate of that "true" value, θ, for that individual is obtained by combining the two noisy sources of information, the population and the measurement(s), with weightings equal to the inverse of their variances. The estimate of θ would be:

$$\hat{\theta} = \frac{\frac{1}{\sigma_p^2} \times \mu_p + \frac{1}{\sigma_e^2} \times y}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_e^2}} = \frac{\sigma_e^2 \times \mu_p + \sigma_p^2 \times y}{\sigma_p^2 + \sigma_e^2}$$

> The variance of this estimate is

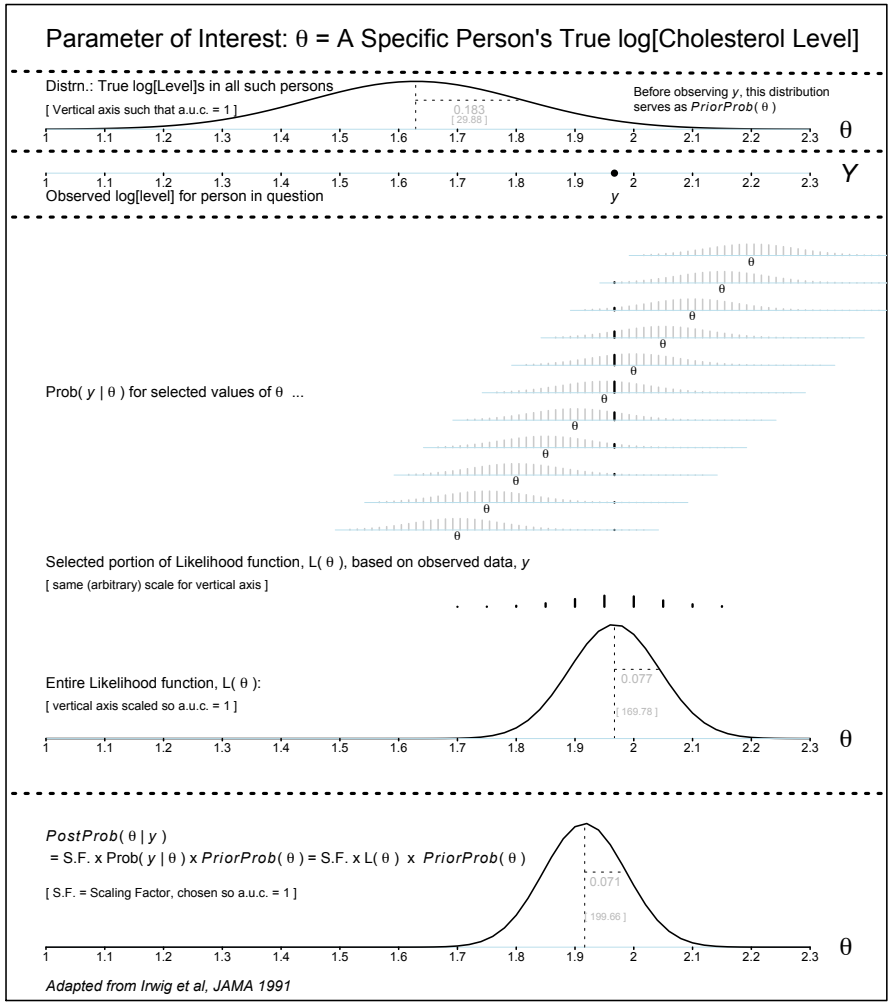
$$\sigma_{\hat{\theta}}^2 = \frac{\sigma_e^2 \times \sigma_p^2}{\sigma_e^2 + \sigma_p^2}$$

JH prefers to work with the precisions, $\tau_p = 1/\sigma_p^2$ and $\tau_e = 1/\sigma_e^2$ rather than variances, and so that the weights can be written directly in terms of these

$$W_p = \frac{\tau_p}{\tau_p + \tau_e}; W_e = \frac{\tau_p}{\tau_p + \tau_e}$$

and the precision of $\hat{\theta}$ written simply as

$$\tau_{\hat{\theta}} = \tau_e + \tau_p.$$



log scale. At the centre of the distribution of unlogged levels, it

Irwig et al's "within-individual variance of 0.00589 was derived from reanalysis of the Lipid Research Clinics Prevalence data, in which repeated measurements were available for almost 5000 individuals. "This corresponds to a [within-individual] about 8% for cholesterol. and 5% for log cholesterol."

[8] JH likes to distinguish between 'regression to the mean', which he thinks of as describing the behaviour of a new *measurement* in someone with an extreme (higher/lower than average, the person's own average, or the average of the population (s)he belongs to) value. He thinks of *shrinkage* as a statistician-delevoped technique for making a combined estimate that brings the estimate based on the person's data closer to the estimate based on the population.

Supplementary Exercise 10.1

Refer to the article "Road Trauma in Teenage Male Youth with Childhood Disruptive Behavior Disorders: A Population Based Analysis" by Redelmeier et al. in PLoS Medicine, November 2010, Volume 7, Issue 11, e1000369.

The following is an excerpt from the Abstract:

> A history of disruptive behavior disorders was significantly more frequent among trauma patients than controls (767 of 3,421 versus 664 of 3,812), equal to a one-third increase in the relative risk of road trauma (odds ratio = 1.37, 95% confidence interval 1.22 – 1.54, p<0.001). The risk was evident over a range of settings and after adjustment for measured confounders (odds ratio 1.38, 95% confidence interval 1.21 – 1.56, p<0.001).

In the Methods, the authors state:

> We excluded teenage girls from both groups to avoid Simpson's paradox (a spurious association created by loading on a null-null position) since this group has much lower rates of crash involvement.

In the Discussion, the authors state:

> A third limitation that causes our study to underestimate the association of disruptive behavior disorders with road trauma is that the data excluded girls [74]. To address this issue we retrieved the original databases, replicated our methods in girls rather than boys, and conducted a post hoc analysis. As anticipated, the results yielded a smaller sample (n = 4,156) and about the same estimated risk (odds ratio 1.31, 95% confidence interval 1.07 – 1.61, chi-square = 6.8, p = 0.010). Hence, the association of disruptive behavioural disorders with road trauma extended to both teenage boys and girls.

Questions:

i. <skip> Reproduce the (crude) odds ratio and CI reported in the abstract.

ii. The lower limit of the 95% CI was calculated as $\exp\{\widehat{logOR} - 1.96 \times SE[\widehat{logOR}]\}$. Use the (crude) odds ratio and CI reported in the abstract to calculate $\widehat{logOR}$ and $V_{\widehat{logOR}} = \{SE[\widehat{logOR}]\}^2$ for boys.

iii. Repeat the back-calculation for $\widehat{logOR}$ and $V_{\widehat{logOR}} = \{SE[\widehat{logOR}]\}^2$ for girls.

iv. We briefly discussed in class how one could merge(combine) the odds ratios for boys and girls to get a single point estimate and associated CI.[9] Formally merge the results in 2 ways:

   (a) Using the antilog of the weighted average of the logs of the gender-specific odds ratios (also known as Woolf's method) As part of this exercise, prove that the linear combination Woolf uses is the linear combination with the minimum variance.

   (b) Using a likelihood-based approach to estimation of $\theta = \log OR$, in which you represent each of the two items of data as normal-based log likelihoods centered on $\hat{\theta}_M$ and $\hat{\theta}_M$, then add the two log-likelihoods. Hint: since each log-likelihood is a quadratic form in θ, and since their sum is again a a quadratic form in θ, this amounts to working out where the new log-likelihood is centered, and what its curvature is. Show that its centre has the same form as the one used by Woolf.

Supplementary Exercise 10.2

Suppose that a quantity of interest, θ, has a (prior) distribution $\theta \sim N(\mu_1,\ \sigma_1^2)$ and that y is such that $y|\theta \sim N(\theta - A,\ \sigma_2^2)$, where A is a known constant.

i. Derive the (posterior) distribution of $\theta|y$.

ii. Comment on the structure of the posterior mean $E(\theta|y)$; focus on the 2 cases $\sigma_1^2 >> \sigma_2^2$ and $\sigma_1^2 << \sigma_2^2$.

iii. Comment on the influence of σ_1^2 and σ_2^2 on the posterior variance $Var(\theta|y)$; focus on the 2 cases $\sigma_1^2 >> \sigma_2^2$ and $\sigma_1^2 << \sigma_2^2$.

iv. Suppose you form a probabilistic first impression of a person's age based only on seeing the person from a distance. Suppose that you later find out that the person earned a PhD a certain number of years ago. This fact, together with the known distribution of ages at which people obtain a PhD, allows you to revise your initial impression.

   Match up the words in this verbal description with the statistical elements/concepts above. Take $\mu_1 = 62$, $\sigma_1 = 3$, $A = 30$, and $\sigma_2 = 4$.

   Would anything change if you obtained the 2 pieces of information in the opposite order?

[9] 1: *Combining the odds ratios* to get 1 new odds ratio can yield a very different result from *combining the raw frequencies* into one 2×2 table, and making one odds ratio from this one table. 2: Assume the true OR is the same in boys and girls.

11 Null Hypotheses and p-values

The following are examples where the subject is a **collective**:

Para.	Example	New Data	Example
Qual.	Haemophilia carrier	Qual.	Son affected?
Qual.	Innocence (crime)	Qual.	Blood type
Qual.	Cystic Fibrosis	Quant.	Salt in sweat
Qual.	HIV Status	Quant.	Optical density
Quant.	Proportion	Qual.	Pooled sample ($\pm$)
Quant.	Cholesterol Level	Qual.	Colour
Quant.	Proportion	Quant.	Prop. in (sub-)sample
Quant.	Age	Quant.	Biographical