PROBLEMS OF PROPORTIONS

Mike West*

1 Binomial Data

1.1 Surveys of public opinion: The 4% rule

The popular press and TV media routinely report public opinion surveys on all kinds of political, consumer and socio-economic issues. Such surveys are run by polling companies and media organisations themselves. Polls purporting to inform on voter support for competing political candidates are good examples and often typical in terms of the numbers of individuals polled and the methods of selecting candidates to poll. A typical political opinion poll will select several hundred or over a thousand individuals and report the numbers, or proportions, in favour or against each of a selection of propositions. Take the standard two candidate political race: Democrat versus Republican in a US election. The poll results in statements to the effect that

- *691 of the 1,100 individuals polled favour the Democratic over the Republican candidate, or*
- *the poll resulted in 63% in favour of the Democratic candidate, with a 4% margin of error.*

The outcome of 0.63 is the observed or sample proportion in favour of the Democratic candidate. It represents an estimate of the population wide proportion, say θ, of individuals in favour of the Democratic candidate. If the population is fairly large, then θ is some unknown number between 0 and 1, and 0.63 is an estimate of θ based on the poll data. The 4% *margin of error* is a standard "rule of thumb" measure of how accurate the estimate is.

1.2 Clinical trials: Which treatment is best?

A medical study reported in early 1998 concerned HIV risks for babies born of HIV positive women. A group of 400 HIV positive pregnant women were randomly separated into two groups of 200. In one group, the women received no special treatment. The women in a second group received intense AZT therapy prior to and during childbirth. The women were followed to childbirth. In the first group, 19 out of 200 babies were HIV positive; in the second group, only 9

*Sept. 1999 version, updated by Herbie Lee, September 2000 out of 200 babies were HIV positive. The key issue is whether there a significant effect of the treatment here, and what it suggests for treatment of HIV positive pregnant women in future. This is an archetype problem of comparison of proportions. Interest lies in assessing the significance, statistically and practically in terms of its relevance for future women, of the observed difference in outcome proportions: 19/200 versus 9/200.

1.3 Measuring the quality of health care

Medical institutions and organisations collect many different kinds of data relevant to assessing aspects of "quality" of health care delivery. For example, the Management Services department of the US Veteran's Administration (VA) invests considerably in the collection and assessment of data to inform on hospital and care-area specific levels of quality of care across the 170 hospitals in the VA system. This provides information relevant to evaluating patterns of variability in hospital-specific quality of care over time and across care areas, to compare and assess differences across hospitals, and to monitor the impact of internal and system-wide policy changes. One very specific example concerns the rates of return to hospital for follow-up assessment after treatment in the *Substance Abuse Psychiatric* area. Each hospital treating patients in this area reports (among many other things) how many of their patients *failed to return for an outpatient visit within 30 days of discharge* out of the total number of discharges, on an annual basis. Low return rates are indicative of low "quality" in this specific care area, and the proportions returning provide information on differences between hospitals, and changes within a hospital from year to year. Here are some data from three randomly chosen VA hospitals (named A,B and C) in 1992 and 1993:

	A	B	C
1992:	236/520	76/186	55/134
1993:	190/451	74/191	74/137

Basic questions of comparisons include:

- *Do quality levels change between years 1992 and 1993? For example, what do we make of the observed difference between years for hospital A: 236 non-returns out of 520, versus 190 non-returns out of 451?*

- *How do the quality levels differ across hospitals? For example, do the observed proportions (42% and 39%) of non-returns in hospitals A and B in 1993 suggest a real difference in quality levels?*

2 Random Sampling and Binomial Models

2.1 Distribution theory

The binomial random sampling model underlies basic methods of inference based on proportions. Consider a sequence of zero-one outcomes, $x_i = 0$ or $1 \ldots, n$, where each has the same probability θ of "success," i.e., of taking the value $x_i = 1$, and where the x_i are independent conditional on knowing θ. Then the outcomes are independent Bernoulli trials, each having density function (or probability mass function) $p(1|\theta) = Pr(x_i = 1|\theta) = \theta$ and $p(0|\theta) = Pr(x_i = 0|\theta) = 1 - \theta$, or

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} = \begin{cases} \theta & \text{if } x_i = 1, \\ 1-\theta & \text{if } x_i = 0. \end{cases}$$

The set of outcomes $X = \{x_1, \ldots, x_n\}$ forms a *random sample of size* n from the $Ber(\theta)$ distribution. The joint density function of all n outcomes is

$$p(X|\theta) = p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^y(1-\theta)^{n-y}$$

for all 2^n values of X defined by the possible values of $x_i = 0, 1$ for each i, and where $y = \sum_{i=1}^{n} x_i$ is the total number of successes, or simply the number of 1's in X.

As a random quantity, y has the binomial *sampling distribution* $Bin(n, \theta)$ with density

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

on values $y = 0, 1, \ldots, n$. The *sampling distribution* $p(y|\theta)$ describes likely values of the observed number of 1's (successes) that may arise from a binomial random sampling experiment of specified sample size n and known success probability θ. This depends on the fixed value of n as well as the population probability parameter θ, though usually this dependence is not made explicit in the notation. Some basic properties of Bernoulli trials are that $E(x_i|\theta) = \theta$ and $V(x_i|\theta) = \theta(1-\theta)$, which lead to $E(y|\theta) = n\theta$ and $V(y|\theta) = n\theta(1-\theta)$.

2.2 Assumptions

In an applied study, the assumptions underlying a binomial sampling model should be justified, at least approximately. In polling a population, if individuals are truly randomly selected then θ is interpretable as the population proportion of "successes". In fact the model applies under less stringent assumptions, so long as the individuals are judged *exchangeable* – that is, we have no additional information or data that may suggest different probabilities of success for different individuals. More than this, a Bayesian justification of binomial models underpins their use in applications when the individual outcomes x_i do indeed have differing success probabilities, and in which it is reasonable to expect that these probabilities themselves differ only randomly. A second aspect of approximation concerns the use of binomial models when the population from which the cases are drawn is not necessarily large; in such cases, the binomial is an idealised approximation to the true *finite population* sampling distribution of the total y.

For completeness, this section should include a discussion of De Finetti's Theorem and details on exchangeability. That may happen in a later update.

2.3 Sampling distribution of proportions

The sample *proportion of successes* $t = y/n$ is also the mean of the sampled x_i values (or the sample mean). As a random quantity, t has a density function with probabilities $p(y|\theta)$ on the values $t = 0, 1/n, \ldots, 1$ and with $y = nt$; thus

$$p(t|\theta) = \binom{n}{nt} \theta^{nt} (1-\theta)^{n(1-t)}$$

on values $t = 0, 1/n, \ldots, 1$. Also, $E(t|\theta) = \theta$ and $V(t|\theta) = \theta(1-\theta)/n$. Note that we can equally deal with either y or t as the summary outcome of interest; each is a *statistic* – a function of the random quantities X to be observed, and much interest lies in studying the sampling distributions of such statistics. The fact that $E(t|\theta) = \theta$ means that the sampling distribution is always "centred" at the true parameter. Also, $V(t|\theta) = \theta(1-\theta)/n$ implies that the spread about the centre is larger for small sample sizes n, indicating higher levels of *sampling variability* in limited binomial experiments. It also shows us that the spread is higher when θ is nearer 0.5 than either of the extremes of 0 or 1, so that greater variation is expected in binomial proportions with middling values of the true success probability.

The Strong Law of Large Numbers tells us that a sample mean of values in a random sample will, for very large sample sizes, be close to the population mean. Here this implies that t approaches θ for large n, partly justifying the use of t as an ad-hoc *point estimate* or guess at the value of θ. Furthermore, the Central Limit Theorem (CLT) tells us that the sampling distribution $p(t|\theta)$ is, for n rather large, well approximated by a normal distribution (in spite of the fact that it is a discrete distribution). This results in the very old and historically commonly used *asymptotic normal approximation* $p(t|\theta) \approx N(\theta, \sigma^2)$ where $\sigma^2 = \theta(1-\theta)/n$.

3 Likelihood Estimation

3.1 Likelihood functions

Statistical inference involves, in part, trying to assess the value of θ based on the data X. In opinion poll examples, we are interested in estimating the true population of Democratic supporters, θ, based on the observed sample proportion, t, and sample size n. For the HIV babies example, there are two binomial samples: one for the control group of women, one for the treated group, and these groups represent two separate populations with possibly different probabilities of HIV babies. Here the goal is to estimate and compare the two probabilities based on the two sample outcomes. This is the traditional problem of *parameter estimation* in a statistical model, and the focus here. Other inferential questions relate to *prediction:* that is, making statements about uncertain future outcomes (the actual Democrat vote, the numbers of HIV positive babies among a future group of women, etc); Prediction is discussed later on.

A key element of modern statistical inference is the concept of *likelihood* and the notion of a *likelihood function* for a parameter based on observed data. *The likelihood function is $p(X|\theta)$ viewed as a function of θ for the fixed data X.* At any specified θ, the value of the likelihood function is the probability assigned to the data value that actually occurred. Hence θ values with high likelihood are "supported" by the data, or likely values; low likelihood values are ... unlikely.

Consider the treatment group of women in the HIV example. Prior to observing the birth outcomes, uncertainty about the outcomes X are described by $p(X|\theta)$. This depends on X only through the total y, so that, for any possible value of θ, we can evaluate the probability of the actually data values observed, namely $p(X|\theta)$ at $y = 9$, or $\theta^9(1-\theta)^{191}$. For example, if $\theta = 0.05$, the data value $y = 9$ would have been assigned a probability 1.086×10^{-16}; at $p = 0.045$ the value is 1.147×10^{-16}; at $p = 0.1$ the value is 1.821×10^{-18}. It is easy to show that the maximum value of $\theta^9(1-\theta)^{191}$ as a function of θ is given at $t = 9/200 = 0.045$. One could graph the *relative likelihood function* $r(\theta) = \theta^9(1-\theta)^{191}/k$ where $k = p(X|\theta)$ evaluated at $\theta = t$, or

$$r(\theta) = \frac{p(X|\theta)}{\max_\theta p(X|\hat{\theta})} = \frac{p(X|\theta)}{p(X|\hat{\theta})}$$

as a function of θ over 0 to 1 and with $\theta = t$. Dividing by the maximum value of the likelihood simply scales the vertical axis so that the maximum value is 1, and is otherwise immaterial. This likelihood function provides a way of comparing values of θ; the value $\theta = 0.045$ is that parameter value that best predicted the actual data, in the sense that the sampling density gave higher probability to the outcome than any other value of θ. So $\theta = 0.045$ is the parameter value *most strongly supported by the observed data,* or a "most likely" value for θ. It is referred to as the *maximum likelihood estimate* of θ, often denoted *MLE.* We may compare any two values θ_1 and θ_2 the *likelihood ratio* $r(\theta_1)/r(\theta_2)$.

3.2 Likelihood estimates and intervals

The MLE of θ is a common point estimate, here just the usual ad-hoc estimate $\hat{\theta} = t$. Statistical inference involves exploring uncertainty about parameter values as well as quoting estimates. Standard "rule of thumb" measures of accuracy relate closely to *interval estimates* derived from likelihood functions and related methods, including Bayesian intervals below. In this model, likelihood intervals are constructed as follows: for any value p between 0 and 1, draw the horizontal line across $r(\theta)$ at height $r(\theta) = p$; this cuts the likelihood function at two values, say $t - a_1$ and $t + a_2$ for some positive numbers a_1, a_2. The interval so defined contains all those parameter values θ whose relative likelihood is at least p, and excludes all values whose relative likelihood is less than p. Smaller values of p produce wider intervals, all straddling the MLE. Viewing a_1, a_2 as defining margins of error about the MLE, there is a trade-off between interval width and the threshold value of p.

In the HIV babies treatment group, some values of p and the resulting values of a_1, a_2 and the likelihood intervals for θ are (to two decimal places) as follows:

	a_1	a_2	$t - a_1$	$t + a_2$
$p = 0.05$:	0.03	0.05	0.02	0.10
$p = 0.10$:	0.03	0.04	0.02	0.09
$p = 0.20$:	0.02	0.03	0.03	0.08
$p = 0.50$:	0.02	0.02	0.03	0.07

Notice that the margins of error are in the 2-4% range for p in the 0.1-0.2 range. As another example, the second VA hospital has $y = 76$ out of $n = 186$ patients, so $t = 0.41$, in 1992. The error margins and likelihood intervals are (to two decimal places):

	a_1	a_2	$t - a_1$	$t + a_2$
$p = 0.05$:	0.09	0.09	0.32	0.50
$p = 0.10$:	0.08	0.08	0.33	0.49
$p = 0.20$:	0.06	0.06	0.35	0.47
$p = 0.50$:	0.04	0.04	0.37	0.45

The sample sizes n are similar in these two examples, but the proportions t are different; that in the VA example is in the mid-range near 0.5, the HIV babies value is much smaller. This difference is evident in two features. First, intervals are wider for the VA example, for each p. Second, $a_1 = a_2$ (to the accuracy quoted) in the VA example, so that likelihood functions are symmetric about t, whereas there are clear asymmetries in the HIV example.

3.3 Large sample approximations

Likelihood functions tend to concentrate about the MLE for larger values of n. Figure xxx displays four relative likelihood functions; the binomial data sets are given by (y, n) pairs $(4, 10), (20, 50), (40, 100)$ and $(400, 1000)$ respectively, so that $t = 0.4$ in each case. Note the increasing concentration about the MLE. For a given threshold p, resulting likelihood intervals are obviously shorter for larger n. *Asymptotic theory* describes how fast likelihood functions concentrate around the MLE; standard theory implies that the length of an interval defined by a fixed threshold p decreases in proportion to $1/\sqrt{n}$. This can be seen clearly via a standard asymptotic approximation using a Taylor series expansion of the *log likelihood function* $\log p(X|\theta)$ about the MLE. This leads to the important *log quadratic approximation to the relative likelihood*; as a function of θ and for large n,

$$r(\theta) \approx \exp(-(\theta - t)^2/2\tau^2)$$

where $\tau^2 = t(1-t)/n$. As a result, likelihood intervals defined by $r(\theta) > p$ are approximately of the form $t \pm a$ with $a = \sqrt{-2\log(p)}\tau$. These are obviously symmetric intervals, and the length of such an interval is, for fixed t, proportional to $1/\sqrt{n}$.

3.4 Combining proportions and sequential experiments

Independent data informing on the same parameters combine via multiplication of the resulting likelihood functions. If two separate studies on HIV positive babies are run in separate hospitals but assuming the same population of mothers generate the cases, then we obtain two independent data sets, y_1 out of n_1 and y_1 out of n_2, say, with joint probability distribution

$$p(y_1, y_2|\theta) = p(y_1|\theta)p(y_2|\theta) = \binom{n_1}{y_1}\theta^{y_1}(1-\theta)^{n_1-y_1} \times \binom{n_2}{y_2}\theta^{y_2}(1-\theta)^{n_2-y_2}$$

which is proportional to $\theta^y(1-\theta)^{n-y}$ where, simply, $n = n_1+n_2$ and $y = y_1+y_2$. As a function of θ ignoring multiplicative constants, we can work in proportional terms; then it is clear that the overall relative likelihood is

$$r(\theta) \propto r_1(\theta)r_2(\theta)$$

where $r_i(\theta)$ is that from study i alone; we multiply likelihood functions from independent sampling models. The extension to more than two data sets is immediate. Further, the data combine to simple numerical summaries by addition; $n = n_1 + n_2$ and $y = y_1 + y_2$.

Experiments, surveys and observational studies are often run sequentially in time, with data accumulating as time progresses. Binomial sampling models (and all random sampling models) do not depend on the order of data collection, and naturally provide for *sequential updating* and revision of inferences as more data are recorded. Here again, we combine old data with new by simply multiplying likelihood functions. If, after series of recordings – say, processing monthly data – we have a "current" likelihood function $r(\theta) \propto \theta^{y_c}(1-\theta)^{n_c-y_c}$, a new set of independent trials generating data (y, n) *updates* or revises the likelihood by increasing the summary quantities to $y_c + y$ and $n_c + n$, respectively.

3.5 Odds and parameter transformations

The *odds* on success in a Bernoulli trial is the ratio of the probability of success to failure, or $\omega = \theta/(1-\theta)$. Inverting this *parameter transformation* we have $\theta = \omega/(1+\omega)$. In certain applied fields, and for some individuals, working in terms of odds rather than probabilities is common. Two key features of likelihood analysis are that

- likelihood functions are invariant under monotonic parameter transformations, and
- MLEs are transformed in the same way as parameters.

In the binomial model, the MLE of the odds on success is then $\hat{\omega} = \hat{\theta}/(1-\hat{\theta}) = t/(1-t)$. The likelihood function for ω from the binomial experiment is simply $p(X|\theta)$ evaluated at $\theta = \omega/(1+\omega)$ as ω varies. Thus, over the range $\omega > 0$, we can write

$$p(X|\omega) \propto \left(\frac{\omega}{1+\omega}\right)^y \left(1 - \frac{\omega}{1+\omega}\right)^{n-y} \propto \frac{\omega^y}{(1+\omega)^n}.$$

groups in the HIV babies study.

The same ideas apply to *any* monotonic parameter transformation.

3.6 Normalised likelihood

Studying likelihood functions in terms of relative likelihoods is just one way of fixing what is otherwise an arbitrary scale for likelihood. An alternative method, and one that is apparently just as ad-hoc as relative likelihood, is suggested by noting that likelihoods, such as those displayed above, have the appearance of probability density functions. So why not simply standardise by *normalising* the functions to integrate to one over the parameter range? At one level this simply defines an alternative scale, but it provides mileage in that we can now talk about the "probabilities" of selected ranges of parameter values, at least informally. As we shall see, this notion is formalised through Bayesian interpretations, and in fact underlies the entire field of Bayesian inference. Before proceeding with this, at a purely mechanical level normalisation of the likelihood function from the binomial experiment leads to

$$g(\theta) = c\theta^y(1-\theta)^{n-y},$$

as a function of θ over 0 to 1, and where c is that positive number such that $g(\theta)$ integrates to unity over the interval. We can show that the relevant constant (constant, so far as θ is concerned, and constant now that the data y are fixed) is just $c = (n+1)\binom{n}{y}$.

We recognise $g(\theta)$ as specific *Beta* distribution, namely $Be(y+1, n-y+1)$. Figure 1 graphs the densities of the two Beta distributions $Be(10, 192)$ and $Be(20, 182)$ that arise this way for the treated and untreated groups in the HIV babies example. Notice now that, compared to relative likelihoods, these functions are densities so that they reach their maximum values at different levels. Each is normalised, so integrates to unity over the unit interval. Similarly, for any specific ranges of θ values within $(0, 1)$ we can easily compute the corresponding areas under the densities using the CDFs of Beta distributions; these areas are just the probabilities assigned to the ranges by the Beta distributions.

Consider, for example, the HIV babies treatment group, where we computed likelihood intervals of $(0.02, 0.09)$ and $(0.03, 0.08)$ at $p = 0.1$ and 0.2, respectively. Under the $Be(10, 192)$ these intervals have probabilities of 0.98 and 0.88. We are now tempted to conclude with inference statements in terms of these probabilities. Here θ is the underlying population probability of an HIV positive baby for women treated with intensive AZT therapy. The above interval probabilities suggest that we conclude *there is about a 0.98 probability that θ lies between 0.02 and 0.09*, and that *the probability that θ lies between 0.03*

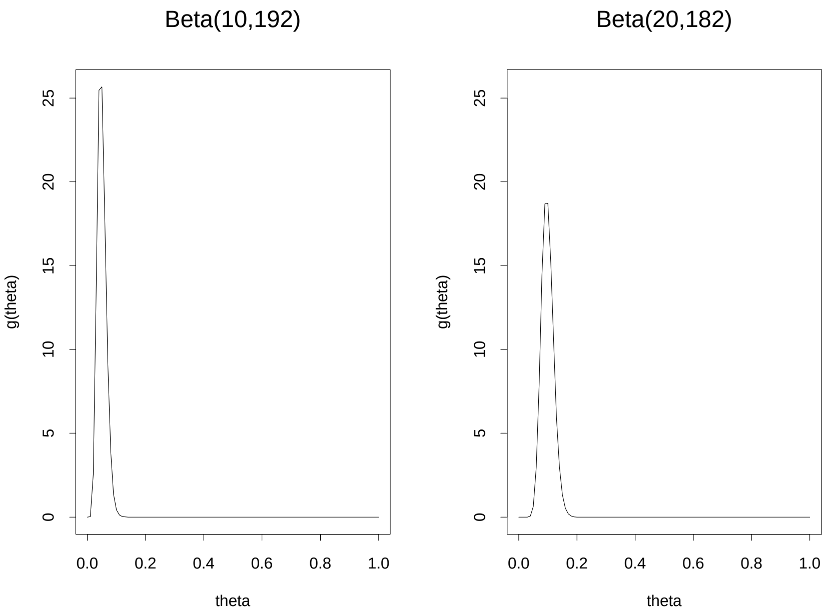

Figure 1: Beta Densities

and 0.08 is about 0.88. A common rephrasal of these statements is in terms of % or chances; for example, *there is an 88% chance that θ lies between 0.02 and 0.09.* Bayesian reasoning and the Bayesian interpretation of probabilities for fixed, though uncertain parameters such as θ provides a basis for such direct inferential statements.

4 Bayesian Inference

4.1 Basic concepts and Bayes' Theorem

Bayesian statistics is predicated on the view that all uncertainties be measured and represented in terms of probability. This applies to representations of uncertainty or state of knowledge about model parameters that are fixed but uncertain quantities, as well as to outcomes of experiments and other random variables. Probability statements about θ represent measures of uncertainty, describing probable or improbable values, or ranges of values, based on knowledge and information at hand. The parameters θ in the HIV babies example are not random variables; they are fixed but uncertain quantities and direct probability statements about them are interpretable via reference to standard randomisation devices, such as coin tossing. For example, a statement that *the probability that θ is less than 0.03 is about 0.5*, indicates that uncertainty about whether or not $\theta < 0.03$ is the same as that about a head versus tail outcome on the toss of a fair coin. For continuous parameters such as the binomial probability θ, uncertainties are represented via continuous probability distributions. Thus, for example, Bayesian inferences about θ in the HIV babies treatment group are based on a summary distribution with density $p(\theta|X)$ where X is the set of observed Bernoulli outcomes and the notation explicitly reflects the *conditioning* of the density for θ on the observed and now fixed data. In standard Bayesian terminology, $p(\theta|X)$ is the density of the *posterior distribution* for θ, posterior to observing the data X. This is usually calculated via the central tool in probabilistic inference, namely Bayes' theorem. In terms of density functions over the range of values of θ,

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)},$$

or

$$p(\theta|X) = cp(\theta)p(X|\theta),$$

or

$$p(\theta|X) \propto p(\theta)p(X|\theta)$$

where $c = 1/p(X)$ is a positive constant (constant since X is now fixed) and is dropped in the final expression which is *Bayes' Theorem in proportional form*. In these expressions:

- $p(X|\theta)$ is the familiar likelihood function in θ at the fixed X values.
- $p(\theta)$ is the *initial* or *prior* density function for θ, representing the probabilistic summary of information about θ *prior* to observing the data X. This must be specified in each analysis.
- $p(\theta|X)$ is the *final* or *posterior* density function for θ conditional on X, and represents a *revised* or *updated* summary of information relative to the prior. The process of revising distributions via Bayes' theorem is a succinct summary of the process of scientific learning: a current (prior) summary is updated to a revised (posterior) summary through the likelihood function of the (new) data or information.
- $p(X)$ is the value at X of the marginal distribution for X implied by the prior and likelihood combined, namely

  $$p(X) = \int p(\theta)p(X|\theta)d\theta$$

  where the integral is over the range of θ, or 0-1 here. For evaluating and summarising the posterior density, the interpretation of $p(X)$ is not important; adopt the proportional form of Bayes' theorem and simply find c to normalise the product $p(\theta)p(X|\theta)$ to convert to a density for θ.

It is important to bear in mind that assignment of a specific probability distribution explicitly recognises the data and information on which the statements are based, as well as dependence on the individual making the probability assessments. This latter point is reflected in subjective Bayesian terminology, which tends to use the terms *belief* and *degrees of belief* for probabilities assigned to uncertain quantities. Dependence on context, data and information is often made explicit via the arguments inserted after the conditioning bar "|" of posterior densities; thus X is explicitly included in conditioning of the posterior density. Of course, prior densities are conditional on prior information, so that, formally, we might write $p(\theta|H)$ for prior density, where H stands for all prior data and information, and then the posterior is $p(\theta|X, H)$, conditional on both X and H combined.

4.2 Reference posteriors and Beta distributions

Bayesian inference gives meaning and interpretation to the otherwise ad-hoc normalisation of likelihood functions discussed earlier. Suppose that the prior is uniform, $\theta \sim U(0, 1)$ with $p(\theta) = 1$ over $0 < \theta < 1$. Then

$$p(\theta|X) = cp(X|t) = c\theta^y(1 - \theta)^{n-y}$$

over $0 < \theta < 1$, where c normalises the density. This is exactly the normalised likelihood function, with $c = (n + 1)\binom{n}{y}$. Now, however, the Bayesian interpretation adds to the framework in important ways:

- The normalised likelihood is now explicitly a Bayesian posterior density, and so has the formal interpretation as a probabilistic summary of information that is lacking in the pure likelihood framework.

- $p(\theta|X)$ is the posterior for θ relative to a uniform prior. This may be viewed as a "reference posterior" in the sense that the uniform prior represents an initially "vague" or "uninformative" position on θ. Under the uniform prior, all values of θ are treated equally; for example, the prior probability that $0 < \theta < 0.1$ is 0.1, the same as that of $0.5 < \theta < 0.6$, and of any other interval of length 0.1. This uniformity is one natural representation of an initial uninformed prior, and the resulting reference posterior is a natural first start in data analysis. Posteriors based on other priors, as we shall explore later, modify this reference solely through the form of the alternative prior.

For a fixed binomial data set (y, n), the reference posterior is a member of the family of Beta distributions, namely

$$(\theta|X) \sim Be(y + 1, n - y + 1)$$

for any sequence of n Bernoulli trials X with observed total y. Recall that Figure 1 graphs the corresponding posterior densities for the two HIV babies groups treated separately.

4.3 Posterior estimates and intervals

Posterior inference is based on various methods of summarising such posteriors. For example, *posterior modes* and *posterior means* are often used as summary point estimates. Under the $Be(y+1, n-y+1)$ posterior, the mode and mean are $t = y/n$ and $m = (y+1)/(n+2)$, respectively. For large n, $t \approx m$ as the posterior is more symmetric about its mode; for small n the posterior may be quite asymmetric. Posterior intervals provide more useful summaries, and are easily evaluated using the cumulative distributions and quantile functions of the Beta family. These are often called Bayesian *credible intervals* or simply *posterior (probability) intervals*. For example, if a is the upper 90% point of the posterior distribution, then $(0, a)$ is a *90% posterior interval for* θ : quite simply, there is a posterior probability of 0.9 that θ lies in this interval. *Equal tailed-intervals* are often used: a 95% equal-tailed posterior interval is that interval whose endpoints are the 0.025 and 0.975 percentage points of the posterior, respectively.

Highest posterior density intervals (HPD intervals) are credible intervals closely related to likelihood intervals. For example, a 95% HPD interval is such that

- all points inside the interval have higher posterior density than all points outside, and
- the posterior probability of the interval is 0.95.

This has obvious generalisations to any specified posterior probability other than 0.95.

4.4 Large sample approximations

Any Beta distribution, say $Be(a, b)$, is well approximated by a normal distribution when (a, b) are both large. This means that the reference posterior $p(\theta|X)$ may be so approximated when (y, n) are both large or, equivalently, when n is large and both $t = y/n$ and $m = (y+1)/(n+2)$ are not close to 0 or 1. The variance of the posterior is $m(1-m)/(n+2)$ so it becomes small as n increases, whereupon the posterior concentrates more and more around m, t. The likelihood asymptotics discussed earlier lead most easily to the standard asymptotic, or large sample, normal approximation to the posterior. That is,

$$p(\theta|X) \approx k \exp(-(\theta - t)^2/2\tau^2)$$

where $\tau^2 = t(1-t)/n$. Recognising the functional form of a normal density, we see that the constant k is simply $k = 1/\sqrt{2\pi\tau^2}$ and the normal approximation is

$$(\theta|X) \sim N(t, \tau^2).$$

This is a standard and easily calculable rule-of-thumb approximation for exploratory inference. By symmetry, posterior HPD intervals are of the form

$$\theta \pm z\tau$$

where z is an upper quantile of the standard normal distribution. Thus a 95% HPD interval is $\theta \pm 1.96\tau$, and so forth.

A final comment on such approximations. The exact Beta posterior is defined over $0 < \theta < 1$ whereas any normal distribution is defined over the real line. From a practical viewpoint the approximation is useful as almost all of the probability under the normal approximation is concentrated well within the unit interval. It is easy to check on this issue: compute, say, the 99% posterior HPD interval under the normal, and check whether or not the endpoints lie between 0 and 1. The approximation is likely to be invalid if this is not the case.

4.5 Updating with new data

In the VA hospital monitor study, hospital patient return data are collected sequentially over time. In hospital #1, for example, records were reported initially at the end of March 1992, when the data for that hospital for January, February and March were reported as 41 returns for follow-up care out of a total of 112 individuals. Write X_1 for the full sequence of 112 0/1 outcomes for these patients, $y_1 = 41$ for the total successes out of $n_1 = 112$. The subscripts "1" indicate that this is a first data set, based on the three months in question. At the end of March, therefore, the data relevant to estimation of the underlying probability of return θ, assumed fixed over the year, is just X_1. A standard Bayesian analysis would base inferences about θ on the implied reference posterior

$$(\theta|X_1) \sim Be(y_1+1, n_1-y_1+1) = Be(42, 72).$$

This has mean (mode) 0.37, and a 95% equal tails interval $(0.28, 0.46)$. (all to two decimal places). Inferences and comparisons with other hospitals, for example, would be based on this distribution.

At the end of June 1992, hospital #1 reported the additional data for the three months April, May and June. The records X_2 had $y_2 = 68$ successes out of a total of $n_2 = 151$ patients. Two statistical calculations are now possible.

- Use the "old" posterior $(\theta|X_1)$ as the "new prior" in Bayes' theorem, to compute an updated posterior; or
- Redo the entire calculation, combining the two data sets into one: set $X = \{X_1, X_2\}$ with binomial outcome $y = y_1 + y_2$ out of the total $n = n_1 + n_2$ trials, and compute the reference posterior $p(\theta|X)$.

These calculations are equivalent, and lead to the appropriate posterior

$$(\theta|X) \sim Be(y, n-y) = Be(y_1+y_2+1, n_1+n_1-y_1-y_2+1) = Be(110, 155).$$

The formal justification is as discussed in combining independent likelihood earlier. From the explicitly Bayesian viewpoint, let us work with *any* prior density (prior to January 1992 in the example) $p(\theta)$. Note that the implied posterior based on all the data is

$$p(\theta|X) \propto p(\theta)p(X|\theta) = p(\theta)p(X_1, X_2|\theta) = p(\theta)p(X_1|\theta)p(X_2|\theta)$$

since the two data sets X_1 and X_2 are conditionally independent Bernoulli sequences. Then, since $p(\theta|X_1) \propto p(\theta)p(X_1|\theta)$, we have

$$p(\theta|X) \propto p(\theta|X_1)p(X_2|\theta).$$

That is, the prior $p(\theta)$ is first updated to the March 1992 posterior $p(\theta|X_1)$, relevant at that time. With the new data X_2 at the end of June, this first posterior is now the prior, prior to X_2, and Bayes' theorem applies to update this to the required $p(\theta|X)$.

This illustrates sequential updating of posterior distributions, or sequential learning as more and more data is processed. The aphorism *"Today's posterior is tomorrow's prior"* is apt. Note the special reference case when $p(\theta)$ is uniform.

4.6 Prior specifications

In the above section the prior distribution at the beginning of April 1992 was $\theta \sim Be(42, 72)$. This is not uniform, as it was based on the data from the first quarter. The updating in light of the additional data X_2 illustrates Bayesian inference with an informed prior distribution, rather than with a uniform prior assumed as an initial "reference" position. It also illustrates the concept of *conjugate prior distributions*. The prior is $Be(42, 72)$ and the posterior is $Be(110, 155)$; both distributions are in the same parametric family, and updating from prior to posterior simply involves easy calculations to change the defining parameters. More generally, any Beta prior $\theta \sim Be(a, b)$ is updated by binomial data (y, n) to the posterior $Be(a + y, b + n - y)$, giving the general expression for conjugate prior-to-posterior updating in binomial models.

Bayesian statistics provides the facility for using priors that are based on existing (prior) information or historical data. When conjugate priors are used, they have an immediate interpretation in terms of summarising information that "looks like" it came from prior data of the same form as the experiment under analysis. A $Be(a, b)$ prior "looks like" the reference posterior from a binomial experiment generating $a - 1$ successes out of $a + b - 2$ trials. Thus the case $a = b = 1$ provides a uniform prior that reflects no such imaginary prior binomial data and is, in this sense, a suitable candidate for a prior representing "no prior information."

4.7 Odds and parameter transformations

Inference may often be couched in terms of estimates or intervals for the odds on success in the binomial model, $\tau = \theta/(1 - \theta)$. Given a posterior distribution for θ, the corresponding posterior for θ is immediately available by the standard use of transformation of densities. If $g(\theta)$ represents the posterior density for t, then the implied density for τ is

$$p(\tau) = |J(\tau)|g(\tau/(1+\tau))$$

where $J(\tau)$ is the *Jacobian* of the transformation,

$$J(\tau) = \frac{\partial\theta}{\partial\tau}.$$

Since $\theta = \tau/(1+\tau)$ here we have $J(\tau) = (1+\tau)^{-2}$. Then, if if θ has the reference $(\theta|X) \sim Be(y+1, n-y+1)$ density then $g(\theta) = c\theta^y(1-\theta)^{n-y}$ with $c = (n+1)\binom{n}{y}$. As a result,

$$p(\tau|X) = c\tau^y/(1+\tau)^{n+2}$$

over $\tau > 0$.

Rather often, we explore and summarise transformed densities quite routinely using simulation. It is trivial to simulate a very large random sample from a distribution such as the Beta posterior for θ here; a sample of several tens of thousands of values gives, for most inferential purposes, a more than adequate numerical approximation to the exact Beta distribution. Numerical summaries, such as various quantiles and approximate HPD or other intervals, are easily computed from such samples. It is then very easy to explore the implied posteriors for parameter transforms such as τ, simply by summarising the transformed samples. With a single parameter of interest, this may seem unnecessary, but the utility of working in terms of large simulations to represent posterior densities becomes very clear when we move to more complicated, and realistic models with more than a single parameter. In modern scientific studies, dealing with simulated distributions is a standard *modus operandi,* and it is often simply impossible to perform the required computations otherwise.

5 Comparison of Proportions

5.1 Comparing two proportions

Recall that Figure 1 displays the reference posteriors for the two probabilities of HIV positive babies among treated and untreated women. Write θ_0 for the probability that a baby is HIV positive among the untreated women (the control group), and θ_1 for the probability that a baby is HIV positive among the treated women (the treatment group). Our earlier likelihood comparisons had indicated that perhaps the probabilities differ, but left unanswered the question of just how different they may be. We now study inference on the ratio $\phi = \theta_0/\theta_1$ to answer this question. In studies such as this, ϕ is called the *relative risk;* here it is the risk of a newborn being HIV positive in the untreated group relative to that in the treated group.

Formally, the two reference posteriors arise from the full model for the two data sets under independent uniform priors. That is, we begin with the *joint prior density* for (θ_1, θ_0) defined by the two uniform margins, under the assumption that they are independent: namely

$$p(\theta_1, \theta_0) = p(\theta_1)p(\theta_0) = 1$$

over

$$0 < \theta_1, \theta_0 < 1.$$

Then consider the sampling model for the two binomials counts, y_1 and y_0 in the treated and untreated groups, respectively. Under the assumption that the outcomes in the two groups are conditionally independent, we have the *joint sampling density*

$$p(y_1, y_0|\theta_1, \theta_0) = p(y_1|\theta_1)p(y_0|\theta_0) = \binom{n_1}{y_1}\theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\binom{n_0}{y_0}\theta_0^{y_0}(1-\theta)^{n_0-y_0}$$

where $n_0 = n_1 = 200, y_0 = 19$ and $y_1 = 9$. Then, by Bayes's theorem for the two θ parameters,

$$p(\theta_1, \theta_0|y_1, y_0) \propto p(\theta_1, \theta_0)p(y_1, y_0|\theta_1, \theta_0)$$

which simplifies to

$$p(\theta_1, \theta_0|y_1, y_0) \propto \theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\theta_0^{y_0}(1-\theta)^{n_0-y_0}$$

as a result of the above conditional independencies in both prior and sampling model. This is the product of the two unnormalised reference Beta densities from the individual analyses, so it is immediate that

$$p(\theta_1, \theta_0|y_1, y_0) = p(\theta_1, |y_1)p(\theta_0|y_0),$$

the product of the two Beta posteriors. That is, the bivariate posterior factors as the product of the two reference posteriors; θ_1 and θ_0 are independent in the prior, the two data sets are independent random samples, and the independence structure is therefore passed through to the posterior.

One important implication of this result is that we can simulate random draws from the joint posterior by sampling the individual reference Beta distributions independently. Doing this for a large number, say $k = 20,000$ draws provides a sample-based representation of the uncertainty about the probabilities. This translates into inferences about $\phi = \theta_0/\theta_1$ simply by computing the value of ϕ for each pair of simulated probabilities; the k values so computed are a random sample from the posterior for ϕ, i.e., $p(\phi|X_1, X_2)$. Figure xxx displays a histogram of such a sample, with $k = 20,000$, for the HIV babies data. Some numerical summaries include:

- The 95% equal-tails interval for ϕ is approximately (0.99,4.46), and the mean of the posterior is approximately $E(\phi|X_1, X_2) = 2.2$.

- 97% of the simulated values exceed $\phi = 1$, implying that $Pr(\phi > 1|X_1, X_2) \approx 0.97$; equivalently, $Pr(\theta_1 > \theta_0|X_1, X_2) \approx 0.97$.

It is now clear that the data strongly support the hypothesis that intensive HIV therapy prior to and during childbirth significantly reduces the chance that the child will be HIV positive. The posterior odds on $\theta_1 > \theta_0$ are roughly 97:3, or about 33:1, indicating the strength of evidence. The range of ϕ values supported indicated that the increased risk among untreated women is likely to be around 2-3 times relative to treated women.

5.2 Prospective and retrospective studies

The HIV babies study is an example of a clinical trial run *prospectively*, or a *prospective study*. All HIV positive, pregnant women enrolled in the study were randomly split into the two comparison groups, one to receive the intense HIV treatment of interest, the other to act as an untreated *control* group. Outcomes – the numbers of HIV positive babies – were then recorded, and the binomial models apply to the outcomes condition on the numbers of women in each group. Similar randomised studies abound in medicine. Recent advances in understanding and treating breast cancer has involved ranges of studies of women with cancer and women at risk, clinical trials to compare the effects of new or experimental drugs with existing treatments, and observational and designed studies of breast cancer genetics. A standard prospective study involves observation of a selected group of women over a period of years to observe whether or not they develop breast cancer by a specified age, say 55 years. At the start of the study, all women are diagnosed as cancer free. Consider a study in which family history is the main risk factor. Two groups of women are observed: women in the first group have a family history (mother or a sister with breast cancer), women in the second have no such history. The question of interest is to compare the risks of developing breast cancer, and we will do this by comparing the two probabilities of breast cancer in the two groups. This is modelled as two binomials,

$$(y_F|\theta_F) \sim Bin(n_F, \theta_F) \quad \text{and} \quad (y_{\bar{F}}|\theta_{\bar{F}}) \sim Bin(n_{\bar{F}}, \theta_{\bar{F}})$$

independently, where subscript F denotes the first group with a recorded "F"amily history, and $\bar{F}$ denotes no family history. Scientific interest lies in a comparison of the two binomial probabilities, and in exploring whether or not θ_F is *practically* significantly greater than $\theta_{\bar{F}}$. A standard measure of interest is the relative risk $\theta_F/\theta_{\bar{F}}$. The group sample sizes n_F and $n_{\bar{F}}$ are assumed fixed in advance, as part of the *experimental design*, and are uninformative about the probabilities of contracting cancer. Valid inferences can now be made about $(\theta_F, \theta_{\bar{F}})$ when the prospective outcomes $(y_F, y_{\bar{F}})$ are observed.

The second common setup is a *retrospective study*. Such a study begins with an identified set of women in two groups: one group with diagnosed breast cancer, the other diagnosed as cancer free. The first group, the diseased women, are called *cases*, the second *controls*; retrospective studies are more often referred to as *case-control studies* in biostatistics. For valid inferences it is important that the cases (respectively, controls) be representative of the population of all cases (controls), in terms of additional risk factors (such as age cohort, family history, etc.) These issues often limit the validity of case-control studies. Assuming a valid study, this retrospective data is naturally modelled as two binomials,

$$(x_C|\pi_C) \sim Bin(m_C, \pi_C) \quad \text{and} \quad (x_{\bar{C}}|\pi_{\bar{C}}) \sim Bin(m_{\bar{C}}, \pi_{\bar{C}})$$

independently, where subscript C denotes the "C"ancer cases, and $\bar{C}$ denotes the cancer free controls. As in prospective studies, the groups sample sizes m_C and $m_{\bar{C}}$ are assumed fixed in advance as part of the design and are uninformative about the probabilities $\pi_C, \pi_{\bar{C}}$. Valid statistical inferences can now be made on $(\pi_C, \pi_{\bar{C}})$ when the data $(x_C, x_{\bar{C}})$ are reported. However, this does not address directly the key scientific and societal issue of predicting cancer risks for women who are currently cancer free. This case control data provides information about

$$\pi_C = Pr(F|C) = Pr(\text{Family History}|\text{Cancer})$$

and

$$\pi_{\bar{C}} = Pr(F|\bar{C}) = Pr(\text{Family History}|\text{No Cancer}),$$

whereas the key issues relate to the direct diagnostic probabilities

$$\theta_F = Pr(C|F) = Pr(\text{Cancer}|\text{Family History})$$

and

$$\theta_{\bar{F}} = Pr(C|\bar{F}) = Pr(\text{Cancer}|\text{No Family History}).$$

It turns out that, in fact, retrospective/case control data does provide for inferences on the direct diagnostic probabilities through inference on the diagnostic *odds ratio*. Recall that $\theta_F/(1-\theta_F)$ is the odds on cancer among women with family history, and $\theta_{\bar{F}}/(1-\theta_{\bar{F}})$ is the odds on cancer among women with no family history. Focus the scientific enquiry on inference about the *odds ratio*

$$\lambda = \frac{\theta_F}{(1-\theta_F)} / \frac{\theta_{\bar{F}}}{(1-\theta_{\bar{F}})}.$$

Clearly λ is greater than 1 if, and only if, $\theta_F > \theta_{\bar{F}}$, and λ is a natural measure in comparing chances of cancer. Now, simple application of Bayes' theorem shows that precisely the same odds ratio arises from the retrospective study; that is, λ is equivalently computed as

$$\lambda = \frac{\pi_C}{(1-\pi_C)} / \frac{\pi_{\bar{C}}}{(1-\pi_{\bar{C}})}.$$

In summary,

- Inference based on comparing two binomial probabilities is available in either prospective or retrospective studies. The two binomial probabilities estimated have quite different interpretations in the two studies.

- Nevertheless, the key scientific issue – comparison of θ_F with $\theta_{\bar{F}}$ – may be validly addressed through both studies. In the retrospective study, this is available via a focus on the odds ratio measure λ.

Such data are often displayed in 2×2 tables of counts, categorised by risk factor (columns) and outcome (rows):

	Family History	No Family History	row totals
Cancer:	a	b	$m_C = a + b$
No Cancer:	c	d	$m_{\bar{C}} = c + d$
column totals:	$n_f = a + c$	$n_{\bar{F}} = b + d$	

Then:

- Prospective studies assume *fixed column totals* and that the columns represent two independent binomials with outcomes $y_F = a$ out of n_F and $y_{\bar{F}} = b$ out of $n_{\bar{F}}$.
- Retrospective (case control) studies assume *fixed row totals* and that the rows represent two independent binomials with outcomes $x_C = a$ out of m_C and $x_{\bar{C}} = c$ out of $m_{\bar{C}}$.

5.3 Several binomial outcomes

Comparisons such as illustrated in the previous two sections lie at the heart of much of modern exploratory and confirmatory data analysis. Historically, approaches based on formal hypothesis testing have tended to dominate much of statistical teaching and therefore practice. Part of the reason for this has been purely mathematical tractability: the simulation based computations in the comparison of two binomial probabilities above involves access to computational and graphical tools that are now widely available, but were not historically. Unfortunately, testing methods that are widespread in practice often fail to address the real scientific issues: in the HIV babies example, we care about measuring the two probabilities of HIV positive outcomes, and comparing them to assess just how big a difference the treatment might make. There is really no notion of a "null hypothesis" that the probabilities are equal, and that we wish to test. Hypothesis testing methods that begin with that view are narrow and limited relative to the more global exploration of differences between parameter values in terms of estimation and posterior analysis.

These comments are more strongly relevant when we consider problems of several or many parameters characterising several related populations to be compared. Consider the VA hospital system in 1992, where we have 159 hospitals and model the "return to care" outcome data as 159 independent binomial experiments, $(y_i|\theta_i) \sim Bin(n_i, \theta_i)$ for $i = 1, \ldots, 159$, and θ_i represents "quality of care" in hospital i (remember that quality is high if θ_i is low in this study). If we begin with initial uniform distributions for the θ_i, and the assumption that the θ_i are independent in the prior, then the reasoning in Section 5.1 about independence structure in the prior, sampling model and posterior carries over to this case of $I = 159$ binomial probabilities, rather than just the $I = 2$ there. As a result, the set of θ_i quantities have independent reference posterior distributions

$$(\theta_i|y_i) \sim Be(y_i + 1, n_i - y_i + 1), \qquad i = 1, \ldots, I.$$

Formally, the product of these I univariate posteriors gives the full *joint posterior density* $p(\Theta|Y)$ where $\Theta = \{\theta_1, \ldots, \theta_I\}$ and $Y = \{y_1, \ldots, y_I\}$.

We may now explore comparisons in various ways. Figure XXX is a simple graphical display of approximate 95% posterior intervals, in this case equal-tails intervals, for the θ_i, plotted vertically against hospital index i. Here the order of the hospitals is chosen to coincide with the ordered values of the observed sample proportions $t_i = y_i/n_i$, so that t_1 is the smallest outcome and t_{159} the largest. This aids in comparisons, as we can now visually skim the graph to get a feel for the variability in θ_i values in the context of uncertainties as measured by intervals. Notice several hospitals with much wider intervals than the majority; these are hospitals with smaller numbers n_i. Hospital $i = 6$, for example, had just 8 patients in this area of care, and so has a very wide posterior interval reflecting that resulting high degree of uncertainty about θ_6. Most of the n_i are much larger, and resulting intervals more precise.

A key multi-parameter method of comparison is to explore *rankings*. The question "is $\theta_1 > \theta_2$?" may be rephrased as "what is the rank of θ_1 in the set $\{\theta_1, \theta_2\}$? The question in its first form has no direct extension to more than two parameters; in its second form, it is directly extensible. The rank of θ_i in the full set of I parameters Θ is defined as $\rho_i = k$ when θ_i is the k^{th} largest value in Θ. Ranks are directly comparative: the relative standing of any hospital in the system is deduced based on its rank. Write $R = \{\rho_1, \ldots, \rho_I\}$ for the full set of ranks. Since the parameters Θ are unknown, so are the true ranks: hence the problem of ranking hospitals is a further problem of inference about uncertain parameters, the parameter set R. The chosen ordering of the hospitals according to the observed outcomes t_i provides one data-based set of ranks, but not the true ranks; the t_i are only estimates the θ_i. For example, hospital $i = 6$ has $t_6 = 0.25$ and so appears to be ranked 6^{th} based on the t_i, but the uncertainty about the true value of θ_6 is so high that the actual rank may be very different.

With modern computational and graphical tools it is trivial to move directly to inferences about the true ranks R based on simulations of the posteriors for the θ_i. Notice that R is a direct *transformation* of Θ; if we know Θ, we know R. Hence we can imagine simulating R values from the posterior distribution $p(R|Y)$ by simulating Θ values and simply computing the ranks for each simulation. In detail:

- In the reference analysis, the joint posterior for the elements of Θ conditional on the data Y is easily simulated: to sample one full Θ vector, just sample each of the $Be(y_i + 1, n_i - y_i + 1)$ margins in turn to deliver one random draw for each θ_i. Repeat this some K times to produce a Monte Carlo sample of size K of parameter vectors. Arrange these samples in an $I \times K$ matrix: each column represents a sampled Θ vector, each row represents a sample of size K from the marginal posterior distribution for θ_i.

- For each simulated vector Θ (each column of the above matrix) compute the corresponding ranks R. This is a direct calculation, and across the K samples creates another $I \times K$ matrix of sampled ranks: each column represents a sampled R vector, and each row i represents a sample of size K from the marginal posterior distribution for the actual rank ρ_i of hospital i.

- Summarise and explore the samples to make inferences on rankings.

As a simple example, take just the first $I = 5$ hospitals from the original VA data file for 1992. After ordering according to the observed proportions t_i, the data are

y_i:	59	76	55	236	85
n_i:	162	186	134	520	175
t_i:	0.364	0.409	0.410	0.454	0.486

Notice that hospitals $i = 2$ and $i = 3$ have close outcomes t_i, suggesting that their relative rankings will be hard to infer. A sample of size $K = 10,000$ from the posterior distributions of ranks was computed as described above. The resulting rows of the resulting $5 \times 10,000$ matrix represent samples from the margins for the ranks ρ_i. For each i, the (approximate) posterior probability that $\rho_i = j$ is given by counting the proportion of times in the $k = 10,000$ samples that $\rho_i = j$; these are as follows:

Probability that $\rho_i = j$ for $i = 1, \ldots, 5$

Hospital i	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$i = 1$:	0.678	0.222	0.081	0.016	0.003
$i = 2$:	0.148	0.373	0.332	0.113	0.035
$i = 3$:	0.169	0.338	0.304	0.133	0.056
$i = 4$:	0.003	0.044	0.216	0.554	0.183
$i = 5$:	0.002	0.023	0.068	0.184	0.722

Thus, for example, there is a posterior 72% probability that hospital 5 has the largest θ_i value, and about an 68% chance that hospital 1 has the smallest θ_i value. The difficulties in ranking similar cases are apparent from the second and third rows, where the posteriors for hospitals 2 and 3 are essentially the same. This could have been anticipated and is reflected in inferences on direct comparisons of θ_2 and θ_3 as in the previous sections; the posterior probability that $\theta_2 < \theta_3$ is 0.51.

The analysis was repeated but now for the full set of $I = 159$ hospitals in 1992. This included computing approximate 95% posterior intervals for each of the ranks, and these are displayed in Figure xxx in a format similar to that for the t_i in Figure xxx. Notice the high uncertainties about ranks for a large number of hospitals in the middle of the range (as ordered by the raw data outcomes t_i). In addition, notice cases such as hospital $i = 6$, where the very low sample size n_6 is reflected in very high uncertainty about ρ_6; in particular, the empirical ranking based on t_6 alone looks suspect and potentially very misleading about the true rank, which may in fact be much higher.

6 An Introduction to Hierarchical Models

6.1 Basic ideas and examples

Section 4.6 introduced the notion of non-uniform priors for binomial probabilities, and the use of specific Beta prior distributions in cases where priors are based on past data. We now elaborate this discussion in connection with analysis of several or many binomial probabilities, in contexts where the set of such probabilities is naturally described by a non-uniform prior with a frequency interpretation. This introduces broader notions of *prior modelling* and *hierarchical models*, which underlie methods that are central to modern applied statistics.

Consider the first 150 of the $I = 159$ VA hospital outcomes in 1992, and recall the point estimates $m_i = (y_i + 1)/(n_i + 2)$ of θ_i based on the observed data. Figure XXX displays a histogram of the m_i, $(i = 1, \dots, 150)$; the number sequencing is arbitrary so that these are 150 hospitals randomly selected from the 159. The histogram then estimates the VA system-wide distribution of quality levels θ_i; "estimates" since the m_i are only estimates of the θ_i (although some, with large n_i values, are very precise estimates). Ignoring this estimation error, the histogram reflects the variety of θ_i values across the system, as we have studied above in other ways. The curve superimposed is actually a Beta density function, namely $Be(8.6, 11.9)$, chosen as described below. This interpolates the histogram fairly well. Viewing the m_i values as a random sample from an underlying continuous distribution, this specific Beta distribution appears to be a good candidate for this underlying distribution. Other candidates may be chosen, of course.

Now consider one of the remaining hospitals, say $i = 151$, and consider estimation of θ_{151} based on the observed outcome y_{151} of n_{151}. Maintaining the earlier assumptions of independence of the θ_i across hospitals, we proceed via Bayes' theorem. Now the reference analysis is called into question; it supposes a uniform prior for θ_{151} but, based on what we see in Figure XXXX and in this context of viewing the hospitals similarly, or exchangeably, it can be argued that inferences for θ_{151} should use this other data and that the reference analysis is therefore suboptimal. This suggests that the prior should be something like the $Be(8.6, 11.9)$ distribution. The context and data force the view that separate reference analyses, while valid and informative, do not adequately represent this specific applied context, as they are explicitly based on a common uniform prior for the θ_i whereas the true variability is apparently quite non-uniform, clustering about 40-45% and with little probability of values near 0 or 1.

The reference analysis is based on the notion of the uniform prior as representing an uninformed or vague initial view of the θ_i, one by one. Use of a non-uniform prior moves us closer to the interpretation of the prior as a frequency-based *model*; the hospital quality parameters θ_i are themselves viewed as a random sample from a prior with something like the above Beta shape. In some areas of application the θ_i, or transformations of them, are referred to as *random effects*, and the prior is the *random effects distribution*.

The specific $Be(8.6, 11.9)$ was chosen quite crudely. Making the choice of a Beta distribution in the first place, we face the question of choosing its parameters. Generally, we can work in terms of a Beta specified by its standard index parameters, as in the notation $Be(a, b)$, or equivalently in terms of the mean of the distribution $m = a/(1 + b)$ and the *precision* $M = a + b$. In the latter terms we have the distribution $Be(Mm, M(1-m))$, and the variance follows as $v = m(1 - m)/(M + 1)$ so that $M = m(1 - m)/v - 1$. Suppose now we simply match the mean m and variance v with the *sample* mean and variance of the 150 m_i values in the histogram; this leads to $m = 0.42$ and $M = 20.6$ whereupon $a = 8.6$ and $b = 11.9$. This is not a formal, optimal method of estimating these new parameters, but serves simply to develop this discussion. We could also have chosen a parametric form different to the Beta form here.

Consider now inference on θ_{151} with the actual data $y_{151} = 637$ and $n_{151} = 1013$. Whatever the parameters (a, b) chosen are, we have the updated posterior $Be(a+637, b+376)$. It is now clear that the specific prior $Be(8.6, 11.9)$ leads to only smalls difference in resulting inferences relative to the reference posterior based on $a = b = 1$, as the two posteriors, $Be(645.6, 387.9)$ and $Be(638, 377)$, are fairly similar. See this by graphing the densities and cumulative distributions of the two Beta posteriors. The reason for this is that the data is extremely informative and just dominates the prior, in either case. Suppose, however, that hospital 151 had a total of just $n = 15$ patients and $y = 9$ returns to follow-up care – a very much smaller sample size, but roughly the same outcome proportion. Now the informed posterior is $Be(17.6, 17.9)$ compared to the reference posterior $Be(10, 7)$. These are very different distributions, and the effect of the population based prior is to "shrink" the informed posterior towards favouring values of θ_{151} that are more consistent with the initial 150 hospitals. This is a key feature of prior modelling, hierarchical models and the use of random effects in general. The hospitals are viewed exchangeably, so that information measured at the first 150 hospitals is relevant to inference about other hospitals. The prior:likelihood combination is the formal mechanism for transferring the information content of the 150 past hospitals to inferences for the next one, where it is combined with the actual data in hospital 151. Relative to reference analyses, there will be small differences for hospitals with large sample sizes, as the hospital-specific data is then very informative about the parameter and "swamps" the prior. It is cases of small additional hospitals that are most influenced by the population prior distribution, for inferences are then more informed by this process of *borrowing strength* from the past hospitals. In using the conjugate Beta prior form, the relative strengths are explicit: a $Be(a, b)$ prior has the interpretation of "adding" a successes and b failures, or a successes out of $M = a + b$ in total, to the data from the current hospital. Furthermore, the mean m of the prior (around 0.42) can be interpreted as a hospital system-wide quality level, and the precision M is a measure of how concentrated the hospitals are around this mean.

6.2 More advanced treatments

More formal and extensive analysis treats (a, b) (or, equivalently, (m, M)) as parameters to be included in the study. Such uncertain parameters defining the *prior model* are often referred to as *hyperparameters* to distinguish them from the primary parameters, here $\Theta = \{\theta_1, \ldots, \theta_I\}$. In this example the model can be represented as in Figure XXX – a *graphical* or *network* representation of the model. The directed arrows between circled quantities represent conditional distributions; thus that from θ_1 to y_1 represents the binomial sampling distribution $p(y_1|\theta_1) = Bin(n_1, \theta_1)$, and that from (a, b) to any of the θ_i represents the Beta prior model $(\theta_i|a, b) \sim Be(a, b)$. The lack of arrows (links) between circled quantities corresponds to *conditional independence*; thus, conditional on (a, b) the θ_i are mutually conditionally independent, and conditional on the θ_i, the outcomes y_i are mutually conditionally independent. Also, given the values of the θ_i, there are no links from (a, b) to the y_i consistent with the lack of dependence of y_i on (a, b) when θ_i is known. A complete analysis of the full set of $I = 159$ outcomes requires us to compute and summarise a joint posterior distribution for Θ and (a, b) all together. This requires an additional specification, a prior $p(a, b)$. Assuming this, the overall model as displayed can be written mathematically as a complete joint density function

$$\begin{aligned} p(Y, \Theta, (a, b)) &= p(Y|\Theta)p(\Theta|(a, b))p(a, b) \\ &= \prod_{i=1}^{I} \{p(y_i|\theta_i)p(\theta_i|(a, b))\}p(a, b), \end{aligned}$$

where the various conditional independencies are explicitly recognised. From this the joint posterior for all parameter and hyperparameters is simply

$$p(\Theta, (a, b)|Y) = c \prod_{i=1}^{I} \{p(y_i|\theta_i)p(\theta_i|(a, b))\}p(a, b)$$

for some normalisation constant c. Analysis of this density is beyond our scope here, though there are standard methods for *simulating* sets of values and, as we now know, summarising large samples from posteriors is trivial; this provides the route to analysis of more general classes of hierarchical models.

7 Prediction

Statistical work is often motivated by decision problems that demand predictions be made about future uncertain events or quantities. The HIV babies study is of interest as it provides data relevant to decisions about treatment of HIV positive pregnant women in future. Parameter estimation is only part of the analysis; a further component of analysis is the exploration of resulting *predictive distributions* for future potential experiments, studies or outcomes.

Consider the HIV treatment group with data $y = 9$ and $n = 200$, and adopt the reference posterior $(\theta|y) \sim Be(a, b)$ with $a = 10$ and $b = 192$. Consider a single future women treated to be treated this way. If she is assumed exchangeable with the test group, then the probability that she has an HIV positive baby *conditional on* θ is, of course, just θ. The study data is relevant as it provides information on θ, now summarised in the $Be(a, b)$ posterior. The (posterior) *predictive* probability her baby is HIV positive is computed as follows. Write $x = 1$ if the baby is positive, $x = 0$ otherwise. Then

$$\begin{aligned} Pr(x = 1|y) &= \int_0^1 Pr(x = 1, \theta|y) d\theta \\ &= \int_0^1 Pr(x = 1|\theta, y) p(\theta|y) d\theta \\ &= \int_0^1 \theta p(\theta|y) d\theta \\ &= E(\theta|y) = a/(a + b) \end{aligned}$$

where the last identity is just the formula for the mean of the (posterior) $Be(a, b)$ distribution. This derivation makes use of the conditional independence of x and y when θ is known; that is, $Pr(x = 1|\theta, y) = Pr(x = 1|\theta) = \theta$. Hence the current posterior mean of θ is the *predictive probability* a future outcome is HIV positive. In our example, $a/(a + b) = (y + 1)/(n + 2) = 10/202 = 0.0495$. The general reference case is $Pr(x = 1|y) = (y + 1)/(n + 2)$ whatever (y, n) are.

Now look at a future group of n_{new} women, and ask about the likely number y_{new} who will bear HIV positive babies following the intense AZT treatment. The above argument generalises: theoretically, we have the *predictive density*

$$p(y_{new}|y) = \int_0^1 p(y_{new}|\theta) p(\theta|y) d\theta$$

where the component $p(y_{new}|\theta)$ is simply Binomial, $y_{new} \sim Bin(n_{new}|\theta)$. The predictive distribution is the *average* of the binomial sampling distributions, averaged over all possible values of θ weighted by the current posterior density at those values. This is a general formula, applied in the binomial example here. Notice that the predictive density takes into account two sources of uncertainty: the uncertainty arising from sampling variability (in $p(y_{new}|\theta)$), and the uncertainty about the value of θ (in $p(\theta|y)$). As a result, predictive distributions are more diffuse, or spread out, than any specific sampling distribution, and will be more markedly spread in cases when there is little information on θ so that the posterior is diffuse. Approximate predictions based on $p(y_{new}|\hat{\theta})$, for an estimate $\hat{\theta}$ "plugged-in" to the sampling model, have been commonly used; when the posterior is rather diffuse, such "plug-in" approximations will lead to inferences that are over-precise, ignoring uncertainty about the estimate $\hat{\theta}$. This feature leads to the remark that predictions based on formal predictive distributions are *honest* in reflecting all sources of uncertainty.

Often the computation of $p(y_{new}|y)$ is not easy mathematically, and even when it is (as in binomial the example here), simulation is an attractive and easy option. To produce a large sample of y_{new} values from the predictive distribution:

- Simulate a set of θ values from $p(\theta|y)$ (as we have been doing);
- for each θ value, simulate a corresponding y_{new} value from the sampling model $p(y_{new}|\theta)$ at that θ value.

The resulting set of y_{new} values is a Monte Carlo representation of the predictive density, and may be explored and summarised for *predictive inferences*.

The above discussion generalises. Whatever information or past data we have informing on θ, represent such information by H, and the current posterior by $p(\theta|H)$. Then, so long as the objects y_{new} of interest are conditionally independent of H when θ is given, the relevant predictive distribution is

$$p(y_{new}|H) = \int p(y_{new}|\theta)p(\theta|H)d\theta.$$

Thus, for example, we speak of the *prior predictive* density

$$p(y_{new}) = \int p(y_{new}|\theta)p(\theta)d\theta$$

when $p(\theta)$ is the assumed prior. As new data is processed, H changes, the posterior is updated, and so is the predictive for future, as yet unobserved, outcomes.

8 Sampling Theoretic Inference

Sampling-theoretic, or *frequentist* approaches to statistical inference are based on the sampling distributions of random variables constructed as statistics, such as y or t. One way this is used is to produce approximate *confidence intervals* for the unknown parameter θ, which provide rule-of-thumb estimates of *margins of error* about the estimate t. This works as follows. Consider the true value θ fixed, and look at the sampling characteristics of t under the sampling distribution $p(t|\theta)$. For easy calculations suppose that n is large enough to justify using the normal approximation from Section 2.3, namely $p(t|\theta) \approx N(\theta, \sigma^2)$ where $\sigma^2 = \theta(1-\theta)/n$. Applied work using sampling theory methods often assumes this approximation. The sampling distribution is used to construct approximate *confidence intervals* for θ which are of the form $(t-a, t+a)$, centered at the point estimate t and with *margin of error* a. For example, the normal distribution gives probability 0.95 to t lying between $\theta \pm a$ with $a = 1.96\sigma$; equivalently, the *random interval* $(t - 1.96\sigma, t + 1.96\sigma)$ covers the true value θ with a 95% chance. As θ is unknown, σ cannot be evaluated so must somehow be estimated. There are two common tricks:

- Since $\theta(1-\theta) < 1/4$ whatever θ is, then $\sigma < 1/\sqrt{4n}$ and the interval will always be no wider than $(t-a, t+a)$ for $a = 1.96/\sqrt{4n}$;

- Since t estimates θ, and for large n is very close to θ, plug-in the value t for θ in σ to get the estimated interval $(t-a, t+a)$ where $a = 1.96\sqrt{t(1-t)/n}$.

Either way, the interval $t \pm a$ delivers an approximate *95% confidence interval for* θ; the former trick delivers a wider interval that will generally, therefore, have a real confidence level greater than 95%.

There are two clear interpretations of such intervals:

- Before seeing the data t, there is (at least approximately) a 95% chance that the interval will cover the true value of θ.

- Imagine repeating the binomial experiment many times, with possibly different values of θ value; if you compute the interval for each case, then about 95% of them will cover their true θ values.

The relevance of sampling-theoretic calculations in applied statistics is questionable for several reasons. Most critical are questions about the validity of sampling theoretic probability statements about θ *after* t is observed, and the relevance of hypothetical future experiments to the current problem. In connection with the above interpretations of the confidence interval, for example, we contend the following:

- Once t is observed and fixed, any probabilities assigned to possible outcomes initially are now irrelevant. A computed interval $t \pm a$ either covers the true value or not.

- The problem at hand has one θ value to be estimated, and one observed t value; the consideration of hypothetical future experiments and data sets is irrelevant to the current inference problem.

For such reasons, we approach inference via likelihood and Bayesian methods that are explicitly *conditional* in nature: inference is based on reasoning *conditional* on the observed data at hand. That said, it is very often the case, at least in simple and standard models such as the binomial here, that the *numerical* summaries of inferences, such as likelihood or Bayesian intervals, often coincide, at least approximately, with sampling theoretic approaches. In more complicated models relevant to complex, real-world scientific problems, this is not true, and indeed sampling theoretic approaches rapidly become intractable, unlike modern Bayesian methods.

9 Some References and Additional Reading

Detailed discussion of the context, data and models in the VA quality of care study may be found in a separate report, also available from the class web site.

10 Exercises

1. Design a sampling experiment to collect data on a population proportion or probability, by asking a randomly selected group of students a relevant question. What kinds of problems of data collection might arise in trying to ensure that the assumptions of a random sampling binomial model are at least approximately satisfied?

2. In the binomial sampling model, choose a few values of θ such as 0.1, 0.3 and 0.9, and a few values of n such as 10, 50 and 500. For each combination of (n, θ), (a) graph the sampling density $p(t|\theta)$, (b) find its variance and hence standard deviation, and (c) find summary 0.005, 0.01, 0.99 and 0.995 quantiles. Briefly summarise by describing the ways in which the distribution changes it shape and spread for different values of n and θ.

3. For a range of values of n between 10 and 1000, compute the 0.005 and 0.995 quantiles of $p(t|\theta)$ when $\theta = 0.4$ and graph these against n. Interpret this graph.

4. In many statistical models *log-likelihood function* is used in derivations of the MLE and other calculations. Since the log function is increasing, the MLE maximises the log-likelihood function too. Use this to prove that, for any y, n, the MLE of θ is $\hat{\theta} = t = y/n$. To do this you should

   (a) Write down the expression for the log-likelihood $\log(p(X|\theta))$;

   (b) Differentiate and equate to zero to solve for the MLE $\hat{\theta}$;

   (c) Check that the value indeed maximises, rather than minimises, the function by checking that the second derivative is *negative* when evaluated at $\theta = \hat{\theta}$.

5. The HIV babies experiment was naturally run sequentially in time, with babies recorded as HIV positive or negative at time of birth. After the first few months, the untreated group had 2 HIV positive babies out of a total of 37 births. The treated group had no HIV positives out of 31 births. Take this as providing two initial, independent binomial experiments at this time point: $y = 2$ and $n = 37$ in the untreated group, $y = 0$ and $n = 31$ in the treated group. Compare the relative likelihoods for the two groups. Discuss the potential pitfalls of simply quoting MLEs of parameters without further features of the likelihood function.

6. Show that the relative likelihood function for θ based on the full set of Bernoulli trials data X is exactly the same as that derived from the single binomial observation y. In this sense, y is a sufficient summary of the data for inference on θ, or a *sufficient statistic.*

7. Suppose the treated group in the HIV trial progressed sequentially, and was designed to halt when exactly 9 HIV positive babies were recorded. Under this experimental design, the outcome is the number of trials, n, needed to achieve 9 "successes." Now random variation in the experiment is described by the sampling model for n; conditional on the chosen number 9 and θ, this is the *negative binomial distribution* with

$$p(n|\theta) = \binom{n-1}{8}\theta^9(1-\theta)^{n-9}$$

for possible values $n = 9, 10, \ldots,$. Show that, when n is observed, the relative likelihood function for θ is exactly as in the binomial model, even though the statistical design and sampling model are different.

8. For a few values of n, say $20, 100, 500, 1000$ and a few values of t, say 0.05, 0.1, 0.5, compute the relative likelihood functions to explore the changes in shape and concentration as n increases. On graphs of these functions overlay the corresponding log quadratic approximations; when would you trust this approximation?

9. For the 1992 VA data on hospital 1, $y = 236$ and $n = 520$, compute likelihood intervals for thresholds $p = 0.1$ and $p = 0.2$. Compare these with those based on the log quadratic approximation.

10. A study of marijuana use among Duke students was designed to assess the proportion that had used the drug in the last couple of years. Call this proportion μ, so that μ is the probability that a randomly sampled student is a user. The study randomly sampled 100 students to survey. Knowing that the straight "Have you used ..." question would generate useless data, each student was asked to:

    - Flip a coin, but don't show me the outcome;
    - If it lands head up, answer "Yes;"
    - If it lands tail up, honestly answer the question "Have you used marijuana in the last couple of years?"

    In this way, a drug-user cannot be identified from her/his response even though they answer honestly. Nevertheless, we still obtain useful information about the population-wide proportion of drug-users. Similar survey "designs" are routinely used in investigating "sensitive" personal issues.

    (a) Show that a student answers "Yes" with probability $\theta = (1+\mu)/2$.

(b) Show that if y is the total number of "Yes" responses out of 100, then $y \sim Bin(100, \theta)$.

(c) Set up a grid of values for μ over $(0, 1)$ and, over this range, graph the relative likelihood function for μ based on data $y = 54, n = 100$. Repeat the above now using data $y = 44$. Comment on, and interpret, the differences between these two likelihoods.

11. A socio-economic survey with chosen sample size $n = 1000$ results in $y = 391$. Show how the resulting statement that the *population percentage is "39% with a 3% margin of error"* may be interpreted in terms of a posterior interval for θ under the reference Beta posterior assuming a binomial sampling model.

12. Histograms provide one way of looking at the shapes of samples, and such displays generate insights into the form of the underlying *density functions* the samples represent. Sometimes we prefer to explore the shape of the cumulative distribution rather than the density, and a standard display is that of the *empirical cumulative distribution function*, or ECDF, of a sample. The ECDF of a sample $x_1, \ldots, x_n$ is just the function $F_n(x)$ that counts the proportion of values in the sample below x, as the argument x varies. Write $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ for the *ordered values* of the sample (use the S-Plus function `sort`). Then

$$F_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ 1/n & \text{if } x_{(1)} \leq x < x_{(2)}, \\ 2/n & \text{if } x_{(2)} \leq x < x_{(3)}, \\ \vdots & \quad\vdots \\ (n-1)/n & \text{if } x_{(n-1)} \leq x < x_{(n)}, \\ 1 & \text{if } x_{(n)} \leq x \end{cases}$$

Simulate a sample of, say, 100 values from the $Be(5, 10)$ distribution. Figure out how to plot the ECDF of this sample over the interval (0,1). Plot it, and overlay a line showing the exact distribution function. Repeat this exercise several times, with different samples of size 100, to explore the nature of sampling variability in ECDFs about the true CDF. Repeat with larger samples, say 5000, and briefly summarise your findings.

13. A manufacturer of auto fan-belts tests samples of size 1000 belts to determine the presence or absence of a defect that may lead to premature failure if the belt is fitted. Write θ for the probability that a belt is defective. The statement made to the motor company regarding fan-belt quality is that the suppliers are *"at least 95% sure that θ is less than 0.02."*. Is this statement is justified in tests when, out of the total $n = 1000$ belts tested, the number defective is $y = 10$. What about a case in which $y = 15$? Answer this question assuming reference Beta posteriors in a binomial sampling model.

14. The binomial sampling model helps us think about uncertainties in estimating population distribution functions from sample data. Suppose we have a random sample of data from some continuous distribution, say $F(x)$ for real-valued x. This might be a Beta distribution, or a normal, but suppose we don't want to impose any such assumptions. Write $X = \{x_1, \ldots, x_n\}$ for the random sample of size n from F. Now, specify any two numbers $b > a$ and let y be the number of elements of X that lie in the interval (a, b). It follows that $(y|\theta) \sim Bin(n, \theta)$. where θ is simply the probability $\theta = F(b) - F(a)$. Hence, ignoring any information we might have about F, a reference analysis focussed only on estimating θ leads to the usual Beta posterior $\theta \sim Be(y + 1, n - y + 1)$.

    This is of use in assessing sampling variability when we use simulation to approximate distributions (that are otherwise harder to summarise). In such cases, we may choose the sample size n as large as we like to achieve high precision in estimating θ, so that accuracy can be assessed using the large sample normal approximation to the posterior distribution for θ. Use the normal approximation in the following questions.

    (a) Compute 99% probability intervals for θ given $n = 10,000$ and each of the cases (1) $y/n = 0.05$, (2) $y/n = 0.5$. Comment on the results.

    (b) Show that, whatever the value of y, a 99% interval for θ is never wider than $2.59/n^{1/2}$.

    (c) How large should n be to ensure that a 99% interval is never wider than 0.01? Discuss what this level of accuracy in estimating θ means, and how it applies to the estimation of the distribution function F more generally. Indicate how this may be used to suggest sample sizes n to use in estimating distributions via large simulation samples.

15. Is the HIV babies study prospective or retrospective? What about the data arising in the VA hospital quality monitor study?

16. This exercise is a modified version of an example kindly provided by Mark Glickman.

    A criminal suspect undergoing a polygraph test is either guilty (G) or innocent (I). His answer to the question "Are you innocent?" is "Yes." As a result of the test, the expert polygraph examiner declares the suspect to by lying, denoted by $x = 1$, or truth-telling, denoted by $x = 0$. Let $\theta_I = Pr(x = 0|I)$ and $\theta_G = Pr(x = 1|G)$ be the probabilities of correct determinations for innocent and guilty suspects, respectively.

    To provide estimates of θ_G and θ_I a pre-trial study is performed. Here the examiner administers the polygraph test to 20 "guilty" people and 18 "innocent" people, and each was asked "are you innocent?" All participants are instructed to say "yes" so that the guilty ones are lying, the innocent ones are not lying. The conditions of the test are otherwise exactly as used in testing a real suspect. Let y be the number out of the

20 guilty people that the examiner correctly identifies as lying, and let z be the number out of the 18 innocent people that the examiner correctly identifies as truth-telling.

(a) State the distribution of y given θ_G. State the distribution of z given θ_I.

(b) Assuming uniform priors $\theta_G \sim U(0,1)$ and $\theta_I \sim U(0,1)$ independently, what are the corresponding posteriors for θ_G and θ_I based on the observed pre-trial outcomes $y = 17$ and $z = 18$. What are the MLEs of θ_G and θ_I? What are the posterior means of θ_G and θ_I? Compare the posterior means with the MLEs as possible point estimates.

(c) Simulate large samples (say, $k = 10,000$) from each of the posteriors and use these to compute samples for $\phi = \theta_G/\theta_I$. Summarise the posterior for ϕ and use it to explore whether or not the examiner is in fact better at detecting truth-tellers than liars on the basis of this data. What do you conclude?

(d) Now return to the real suspect. We are really interested in the probability π that he is in fact guilty when the polygraph examiner declares him to be lying. If we suppose the prior probability he is guilty to be 0.5, then Bayes' theorem gives us $\pi = \theta_G/(\theta_G + 1 - \theta_I)$. Use the posterior simulations from (c) to compute and summarise posterior samples for π. How likely is it that the suspect is guilty?

(e) A non-Bayesian approach would be to simply estimate π by its MLE, namely $\hat{\pi} = \hat{\theta}_G/(\hat{\theta}_G + 1 - \hat{\theta}_I)$ based on the MLEs $\hat{\theta}_G$ and $\hat{\theta}_I$. Is this a good idea given our data?

17. The table below reports 14 separate 2×2 tables of counts from 14 separate clinical trials of the breast cancer drug Tamoxifen. Each was a randomised, prospective clinical trial that studied women with early breast cancer split into two groups: treatment and control. The treatment group received Tamoxifen treatment for an average of about one year, the control group received no treatment. In some of the 14 trials, both treatment and control groups also received chemotherapy. The clinical "endpoint" for the trials was breast cancer recurrence, so that each of the 14 studies delivers data in the form of "two binomials" informing on the relative risk (treatment versus control) of breast cancer recurrence. In the table, for each trial we have y_0 recurrences out of the total n_0 women in the control group, and y_1 recurrences out of n_1 in the treatment group.

TRIAL	y_1	n_1	y_0	n_0
1	55	97	67	101
2	137	282	187	306
3	505	927	590	915
4	62	123	74	140
5	99	239	118	236
6	50	130	49	107
7	185	311	200	319
8	186	303	187	307
9	148	325	178	325
10	25	79	38	86
11	223	344	224	350
12	183	937	185	936
13	2	12	0	8
14	129	434	159	449

I am grateful to Giovanni Parmigiani for kindly providing this data, which was published in a meta-analytic study by the Early Breast Cancer Trials Collaborative Group as reported in the 1998 paper "Tamoxifen for early breast cancer: an overview of the randomised trials", *The Lancet,* **351**, 1451-1467.

There is variation in recurrence rates among trials that is generally due to different follow-up times at the time of the overview, and different patient selection criteria. We are interested in exploring and summarising these variations.

Study these data, using what you've learned about inference on binomial probabilities, comparison of two binomial probabilities in terms of inference on relative risks, comparisons of several binomial probabilities, and so forth. Specific issues your study might contact include:

(a) Graphical exploration of the data, commenting on general features and specific Trials;

(b) Comments about assumptions underlying binomial models and the use of reference Beta posteriors for inference on underlying binomial probabilities;

(c) Summary inferences, using reference analyses, on the relative risks in each Trial, and graphical comparisons across all 14 Trials;

(d) Questions of global comparisons based on inferences about the *ranks of relative risk parameters* from the 14 Trials;

(e) Discussion and development of priors other than the uniform reference prior for binomial parameters. Consider issues of hierarchical modelling and questions about the relevance, potential importance and implications of non-uniform priors.

(f) Summary discussion and conclusions.

Present your study as a short scientific paper: It should begin with a section describing the data and objectives, then two or three sections describing components of the work, a final section summarising and concluding, incorporated graphs, appendices (e.g., for computer code), and any references.

18. Suppose that the HIV babies experiment was actually (as is natural) run sequentially in time, with babies recorded as HIV positive or negative at time of birth. Assume the usual binomial models for the two groups, with underlying HIV positive probabilities of θ_0 (untreated group) and θ_1 (treated group).

    (a) After the first several months of the trail, suppose that the untreated group had 5 HIV positive babies out of a total of 49 births, and the treated group had 1 HIV positive out of 43 births. Using a uniform prior for the two underlying binomial probabilities, compute and summarise the implied reference posteriors. Is there evidence here to distinguish θ_0 and θ_1?

    (b) Consider now a hypothetical group of 100 future babies born of HIV positive mothers. Conditional on the data and the reference posteriors just computed, simulate a large sample from the implied *predictive distribution* for the number, out of the 100, who would be HIV positive if the mothers *do not receive treatment*. Repeat this exercise for a group of 100 future babies whose mothers *do receive treatment*. Summarise the two predictive distributions and comment on the differences.

    (c) Now move ahead to later in the clinical trial where the totals so far give 15 HIV positive babies out of a total of 141 births, and 6 HIV positives out of 132 births in the treated group. Revise the reference posteriors to be conditional on all the data so far recorded. Repeat the predictive analysis and comparisons, with discussion.

This analysis indicates how sequentially received data revises summary inferences and predictions. Clinical trials with serious outcomes (HIV, death, etc) are often performed to a pre-specified design and the results are not analysed until a pre-determined time point. Such trials therefore ignore so-called *interim data*, and the resulting cumulating evidence that may help to distinguish positive benefits or drawbacks of a treatment being tested. Some argue (and I agree) that trials should be monitored and analysed routinely sequentially, and stopped as soon as the evidence strongly indicates benefits or drawbacks. The argument is that it is simply unethical to proceed to deliver a sub-optimal treatment/placebo to individuals remaining in the trial once the evidence is strongly indicative of the sub-optimality. For instance, it might be argued that the above trial should have been stopped at the second time point, and all remaining women in the trial switched to the intense AZT therapy prior to and during their forthcoming child birth.

Incidentally, the sequential figures quoted are hypothetical, though the final outcomes (9/200) and (19/200) are real.

19. Consider a future binomial trial $(y|\theta) \sim Bin(n, \theta)$ for some specified n, and assume the current posterior for θ is uniform, $p(\theta|H) = 1$ for $0 < \theta < 1$ where H represents all current information. Show that the implied predictive density is the discrete uniform distribution, i.e.,

$$p(y|H) = \int_0^1 p(y|\theta)p(\theta|H)d\theta = 1/(n+1)$$

for each value $y = 0, 1, \ldots, n$. Interpret this predictive distribution.