1.1 Binomial model

In the example of red and white balls, we described bayesian inference when only two balls were drawn and both happened to be red. In general, if N balls are drawn (with replacement) from a bag with M balls, we can observe a sequence of red and white balls. If we define

$$X_i = \begin{cases} 1 & \text{if the } i\text{th ball is red} \\ 0 & \text{if the } i\text{th ball is white} \end{cases}$$

then, the (conditional) probability for a specific sequence, e.g. $0, 1, 1, 0, 1$ can be written as

$$(1-r) \times r \times r \times (1-r) \times r = r^3(1-r)^2$$

which is the same as for another sequence of $1, 1, 1, 0, 0$. Generally:

$$P(X_1, \ldots, X_N \mid r) = r^{\sum X_i}(1-r)^{(N-\sum X_i)}$$

where r is the proportion of red balls in the bag. It is apparent that only the sum of red (or white) balls matters for the probability of the sequence, not their order of appearance in the sequence. When making classical statistical inference about r, based on this conditional probability model of the X_is given r, the above expression is seen as a function of (the unknown) r, for a given data $X_1, \ldots, X_N$. The function is called *likelihood function*, and the sum is said to be *sufficient statistic*, (tyhjentävä tunnusluku) [1]. In classical statistics a sufficient statistic contains all the information in the sample needed to compute an estimate for a parameter. In this example: $\hat{r} = \sum_i^N X_i/N$. If we only observe the sum $Y = \sum X_i$, but not the exact sequence, then

$$P(Y \mid r) = \binom{N}{Y} r^Y (1-r)^{N-Y} \;\propto\; r^Y(1-r)^{N-Y},$$

which is the binomial distribution with parameters r and N. Individual draws are said to be Bernoulli experiments, corresponding to binomial distribution with parameters r and $N = 1$. So far, the proportion r has been considered as discrete valued. But if the number of balls in the bag is very large, we can think of the limiting value

$$\lim_{M\to\infty} \frac{R(M)}{M} = r,$$

where $R(M)$ is the number of red balls among M balls. The object of inference is now a continuous valued parameter $r \in [0, 1]$ and for a bayesian statistical inference we must specify a prior *density* for this.

About notations: usually, probability is denoted as P whereas a probability density is written with a different symbol. In some cases we need to write a multivariate distribution where some of its variables are continuous and some discrete. To avoid switching symbols, below π is used loosely to denote all distributions, so that the reader should guess from the context if it means a probability mass function or probability density. Then, symbol P can be reserved to denote probabilities.

Analogous choice to the previously used discrete uniform distribution would be uniform probability density (as in the original example of reverend Bayes):

[1] $T(X)$ is sufficient for r if $P(X)$ can be written in the form $h(X)g(r, T(X))$

$$\pi(r) = 1 \; \forall r \in [0,1] \text{ and } 0 \; \forall r \notin [0,1].$$

This uniform prior is a special case of a Beta(α, β)-density, obtained by setting $\alpha = \beta = 1$ (Bayes-Laplace uniform prior):

$$\pi(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}.$$

The posterior distribution of r is then obtained again by applying Bayes's formula, but now with probability densities:

$$\pi(r \mid Y) \propto r^{(Y+\alpha-1)}(1-r)^{(N-Y+\beta-1)}.$$

For bayesian inference too, the result is the same if we have observed the exact sequence of X_i's or if we just observe the sum Y. For a given Y, the posterior density is still of the same form, regardless of the sequence. From the functional form above - taken as a density for r, and knowing that this is indeed a probability density (the remaining terms, whatever they are, must be the normalizing constant) - the posterior density of r is *recognized* to be a Beta-density, with parameters $Y + \alpha$ and $N - Y + \beta$. (You can also calculate this exactly, without 'recognizing'). The expected value of r from the posterior density is

$$E(r \mid Y, N, \alpha, \beta) = \frac{\alpha + Y}{\alpha + \beta + N},$$

which can also be written as a weighted average:

$$w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{Y}{N},$$

where $w = (\alpha+\beta)/(\alpha+\beta+N)$. The parameters of the prior can thus be chosen so that they represent some imaginary data Y_0, N_0, corresponding to $(\alpha, \beta) = (Y_0, N_0 - Y_0)$.

In this example, the posterior density could actually be solved so that the solution is among standard probability densities. This was possible because the binomial distribution of the data, and the beta-density prior are conjugate distributions. Generally, they don't have to be so, and we could choose any other prior distribution, but the resulting posterior would not be among any of the well known standard distributions. Yet, it could still be computed by using numerical methods in the absence of analytical solution.

So, now we have seen how to obtain a posterior density for the unknown proportion r. It can be summarized in various ways, but it can also be made to work for us as a tool for many kind of scientific questions which somehow involve this parameter. When the prior was chosen as uniform density, the posterior density actually equals to the likelihood function which simply would be normalized to represent a proper probability density of r, for a given data Y. In classical statistics, a popular estimate of the parameter is the maximum likelihood estimate, which is the parameter value that gives highest probability to the data. In the special case of uniform prior, the maximum likelihood estimate coincides with the value that has maximum posterior density. Note that for a bayesian, r is viewed as random (because unknown), but in non-bayesian statistics r would be thought as fixed. In bayesian analyzes, the posterior distribution is always the primary result and not some selected point values because they would not convey the same information about the uncertainty.

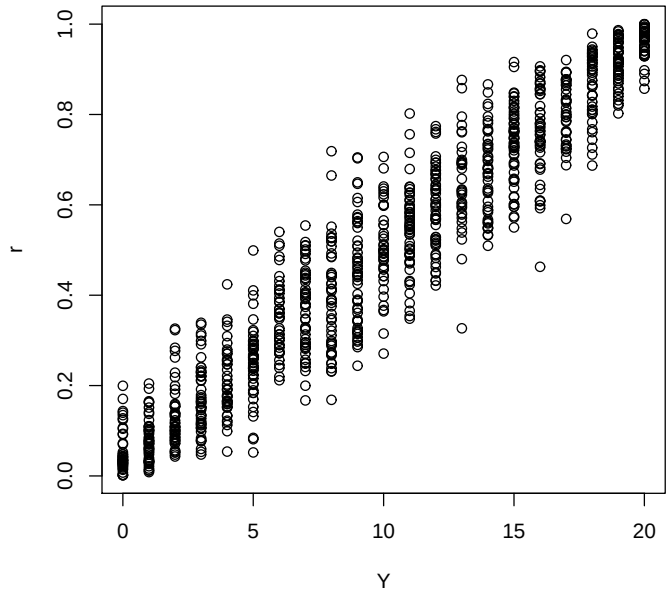

Figure 1: Simulated values from the joint distribution of $\pi(r, Y) = \pi(Y \mid r, N = 20)\pi(r)$ with uniform density $\pi(r)$. In R code: `r <- runif(1000)`, `Y <- rbinom(1000,20,r)`, `plot(Y,r)`. For any fixed r we have Binomial$(20, r)$ for Y, i.e. $\pi(Y \mid r)$. For any fixed Y we have Beta$(Y + 1, 20 - Y + 1)$ for r, i.e. $\pi(r \mid Y)$.

1.1.1 Informative priors for unknown proportion

Depending on what the prior information is, there can be different ways to formulate the prior as a density over $[0, 1]$ to reflect such prior information. the simplest case is to have a previous similar binomial experiment from which we have data Y_0, N_0 which can be directly translated to a Beta-density with parameters $Y_0, N_0 - Y_0$. Then we are assuming that the old sample and the forthcoming sample could be combined as one sample. Another source of information could be to ask from experts, or search from literature, what is the most plausible value of prevalence and call it m. Then, we should quantify also the width of the distribution by determining e.g. the standard deviation and call it s. If these can be reasonably quantified, we can then solve parameters for Beta-density:

$$\alpha = -m(mm - m + ss)/(ss) \quad , \quad \beta = (mm - m + ss)(m - 1)/(ss)$$

It may be difficult to get an opinion about s, so we could work around it by formulating the problem differently. First start by asking for the most plausible value m. This could be taken to represent the mean as above, or perhaps more accurately the mode. By looking up the formula of the mode for Beta-distribution, we write

$$m = \frac{\alpha - 1}{\alpha + \beta - 2}$$

so that the Beta-prior is then Beta$((1 + (\beta - 2)m)/(1 - m), \beta)$. Next we determine what is a value for which the expert is 95% sure that the actual value is below. Call this value u. We then have prior

probability $P(r < u) = 0.95$. By using the Beta-density which now only depends on β, we look for such value of β that we get $P(r < u \mid \beta) = 0.95$. Finally, we have solved the prior Beta(α, β). Solving the last step requires numerical techniques, e.g. using R to find percentiles. In a bayesian model, when computing posteriors can also require numerical techniques, one does not necessarily want to solve the prior numerically. Then, approximations based on normal distributions can be used to find analytical solution for the prior parameters.

In all cases, if the final prior density is Beta, we can also study what amount of prior data this would equal to. Note that Beta densities cannot represent bimodal or more complicated prior densities. However, these are rare in practice. But such prior might be obtained when combining the priors of a group of experts, as a group opinion. Then, the prior density could be expressed as a mixture of Beta-distributions.

Generally, a mixture prior distribution is a mixture of densities each specified by some parameter β_i:

$$\pi(\theta) = \sum_{i=1}^{k} \alpha_i \pi(\theta \mid \beta_i) = \sum_{i=1}^{k} \alpha_i \pi_i(\theta).$$

The weights α_i are the mixing weights of the component distributions ($\sum \alpha_i = 1$). Denote the model for data x as $\pi(x \mid \theta)$. The posterior distribution is then

$$\pi(\theta \mid x) = \frac{\sum_{i=1}^{k} \alpha_i \pi_i(\theta) \pi(x \mid \theta)}{\pi(x)}$$

which unfortunately is no longer recognized as a standard distribution, but this could be handled with numerical methods, e.g. in BUGS.

1.1.2 Uninformative priors for unknown proportion

If an uninformative prior is required for binomial proportion r, there are actually several choices. They are all uninformative, but in different ways.

Bayes-Laplace prior: Beta(1,1)

Jeffreys' prior: Beta(1/2,1/2)

Haldane's (improper) prior: Beta(0,0)

The Bayes-Laplace prior reflects the idea of 'insufficient reason', which says that unless there is specific reason to assign unequal probabilities, they should be equal for all possible values of r. But the problem is that the uniform prior is not uniform for all transformations. This seems to be a problem because one could say that if I'm completely uncertain about r, I should be similarly uncertain about r^2 - if that happens to be of interest too. The original Bayes-Laplace prior $r \sim \mathrm{U}(0, 1)$ would not imply a uniform prior for r^2, and vice versa. (The density of $q = r^2$ would be $\pi(q) = 0.5q^{-0.5}$, by using the transformation of variables rule, if $\pi(r) = \mathrm{U}(0, 1)$).

The priors can be interpreted as being equivalent to some amount of 'prior data'. The uniform prior Beta$(1, 1)$=U(0,1) corresponds to having 2 prior experiments, one of which was a 'red ball' and the

other 'white ball'. The Jeffreys' prior equals to having only one prior experiment in which one ball was 'drawn' and it was 'half red', 'half white'. In this sense, Haldane's prior corresponds to having no prior data at all, but the prior is actually concentrated at two points: zero and one. Moreover, with Beta$(0,0)$ prior the posterior is not defined if the observed data happens to be either 0 or N under a Binomial(N,r) model.

The Jeffreys' prior is based on the principle that an uninformative prior should be such that it does not depend on which parameter transformation is used: it should be the same for all transformations. For single parameters, the Jeffreys' prior is sometimes used but for multiparameter problems the results are more controversial, and a hierarchical modeling approach is more common. Generally, for some single parameter, r, the Jeffreys' prior is chosen so that

$$\pi(r) \propto [J(r)]^{1/2},$$

where $J(r)$ is so called *Fisher information* for r.

$$J(r) = E\Big[\Big(\frac{\mathbf{d}\log\pi(X \mid r)}{\mathbf{d}r}\Big)^2 \mid r\Big] = -E\Big[\frac{\mathbf{d}^2\log\pi(X \mid r)}{\mathbf{d}r^2} \mid r\Big].$$

(This equality is borrowed, without proof, from classical texts where the Fisher information is more used, and these two equivalent forms are 'well known' parlance).

It can be shown that for a transformation $\psi = h(r)$, with $r = h^{-1}(\psi)$, the following equation can be obtained:

$$J(\psi)^{1/2} = J(r)^{1/2} \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

and the Jeffreys' prior is defined as proportional to $J(\cdot)^{1/2}$ which makes it invariant under transformation. This means that if we calculate the prior $\pi(\psi)$ for some transformation ψ of the original parameter r, we get, using the variable transformation rule:

$$\pi(\psi) = \pi(r) \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

and if the original prior $\pi(r)$ is chosen as Jeffreys, then $\pi(\psi)$ is proportional to

$$\propto \sqrt{E\Big[\Big(\frac{\mathbf{d}\log L}{\mathbf{d}r}\Big)^2\Big]\Big(\frac{\mathbf{d}r}{\mathbf{d}\psi}\Big)^2} = \sqrt{E\Big[\Big(\frac{\mathbf{d}\log L}{\mathbf{d}r}\frac{\mathbf{d}r}{\mathbf{d}\psi}\Big)^2\Big]} = \sqrt{E\Big(\frac{\mathbf{d}\log L}{\mathbf{d}\psi}\Big)^2} = \sqrt{J(\psi)}$$

where L denotes the likelihood $\pi(\text{data} \mid \text{parameter})$.

So, the prior of the transformed parameter $\pi(\psi) \propto \sqrt{J(\psi)}$, if the prior of the original parameter $\pi(r) \propto \sqrt{J(r)}$.

Calculate the Jeffreys' prior for the binomial proportion r. To begin, we have the following from the binomial model:

$$\log\pi(X \mid r) = \text{constant} + X\log(r) + (N - X)\log(1 - r)$$

$$\frac{\mathbf{d}\log\pi(X \mid r)}{\mathbf{d}r} = \frac{X}{r} - \frac{N-X}{1-r}$$
$$\frac{\mathbf{d}^2\log\pi(X \mid r)}{\mathbf{d}r^2} = \frac{-X}{r^2} - \frac{N-X}{(1-r)^2},$$

and taking the negative of expected value, $-E(\cdot \mid r)$, gives

$$J(r) = -\Big(\frac{-rN}{r^2} - \frac{N-rN}{(1-r)^2}\Big) = \frac{N}{r(1-r)}.$$

The Jeffreys' prior for binomial proportion r is thus

$$\pi(r) \propto [J(r)]^{1/2} \propto r^{-1/2}(1-r)^{-1/2}$$

which is Beta(1/2,1/2).

What does all this mean for some transformation of r? For example $\psi(r) = \sqrt{r}$, with inverse transform $r(\psi) = \psi^2$, and $\mid \mathbf{d}r/\mathbf{d}\psi \mid = 2\psi$. If we want the posterior density of ψ, we can obtain it in two ways:

(1). Compute the posterior density $\pi(r \mid X) \propto \pi(X \mid r)\pi(r)$ using ***Jeffreys' prior for*** r, **and then use transformation of variables** to get the posterior density of ψ:

$$\pi(\psi \mid X) = \pi(\, r(\psi) \mid X \,) \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid \;\propto\; \pi(X \mid r(\psi)) \underbrace{\pi(r(\psi))}_{\text{Jeffreys'}} \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

$$\propto \underbrace{\psi^{2X}(1-\psi^2)^{(N-X)}}_{\propto \text{Bin}(N,\psi^2)} \times (\psi^2)^{-1/2}(1-\psi^2)^{-1/2} \times 2\psi.$$

(2). Compute **directly the posterior** $\pi(\psi \mid X) \propto \pi(X \mid \psi)\pi(\psi)$ using ***Jeffreys' prior for*** ψ. In this case, $\log\pi(X \mid \psi) = \text{constant} + 2X\log(\psi) + (N-X)\log(1-\psi^2)$, and after some calculations we get $J(\psi) = 4N/(1-\psi^2)$. Therefore, Jeffreys' prior for ψ is

$$\pi(\psi) \propto [J(\psi)]^{1/2} = \frac{2\sqrt{N}}{\sqrt{1-\psi^2}} \propto (1-\psi^2)^{-1/2}.$$

Using this prior, we calculate the posterior of ψ directly:

$$\pi(\psi \mid X) \propto \pi(X \mid \psi) \underbrace{\pi(\psi)}_{\text{Jeffreys'}}$$

$$= \underbrace{\psi^{2X}(1-\psi^2)^{(N-X)}}_{\propto \text{Bin}(N,\psi^2)} \times (1-\psi^2)^{-1/2}.$$

By comparing (1) and (2), either way, the posterior of ψ is the same!

However, Jeffreys' prior violates so called *likelihood principle* which states that whenever the likelihood function is (proportionally) the same, the inferences should be the same too. For example, the binomial model (for a sample result with fixed N) and the negative binomial model (for the number of

samples N needed before fixed number of successes X is obtained) produce (proportionally) the same likelihood function for the success probability r. Therefore any differences in posterior must be due to different priors. In this example, Jeffreys' prior leads to two different prior distributions depending on which of the two models is used in the calculations. The difference is because in the first case, the expected value in the Fisher information is taken of derivatives of log-likelihood in which the random variable is X (given r, N), but in the second case the random variable is N (given r, X). Jeffreys' prior can also lead to improper prior distributions which cannot be normalized to proper probability distributions (which should integrate to one).

Note also that if the prior of r is Beta(α, β), then the posterior will be Beta$(X + \alpha, N - X + \beta)$ and the posterior mode is then $(X + \alpha - 1)/(\alpha + \beta + N - 2)$, and posterior mean is $(X + \alpha)/(\alpha + \beta + N)$. The posterior mode becomes X/N when the Bayes-Laplace prior is used. The posterior mean becomes X/N when the Haldane's prior is used. Note that the fraction X/N is also the *maximum likelihood estimator* for r in *likelihood inference*. I.e., it is the value of $r \in [0, 1]$ that gives the highest probability for the data, X, that was observed: $\text{argmax}_{r \in [0,1]} P(X \mid N, r)$.

Warning: improper priors may lead to improper posteriors. Therefore, it may be advisable to use proper priors also when aiming at an uninformative prior. Later, when using WinBUGS, it is possible to explore what happens when the prior parameters are tuned towards a nearly improper distribution. Numerical difficulties may sometimes happen even if the prior is just proper, e.g. if the parameters of beta-density are nearly zero. Sensitivity analysis is always recommended to check how sensitive the posterior results are to the choice of prior.

1.1.3 Unknown N

The usual application of binomial model Bin(N, r) involves inference about unknown r with known N. In general, any quantity could be unknown, so let's see how to make inference about N, assuming that r is known. We then would know the true proportion of red balls in a 'large' bag, and someone has done the sampling of N balls but he does not tell us what the sample size N was. Instead, we are only told how many red balls (X) there were. Again, we first have to specify a prior for N. But N could be any integer value $0, 1, 2, \ldots$ and there is no way to know how large it could be. It seems difficult to assign an uninformative probability distribution. But let's start with a simple choice that assumes some very large maximum value M, so that the prior is uniform from 0 to M:

$$P(N = i) = \frac{1}{M + 1} \forall i \in \{0, 1, \ldots, M\}$$

Now the posterior is:

$$P(N \mid X, r) \propto P(X \mid N, r)P(N) = \frac{N!}{X!(N - X)!} r^X (1 - r)^{N-X} \frac{1}{M + 1}$$

$$\propto \frac{N!}{(N - X)!} (1 - r)^N$$

$$= N(N - 1) \ldots (N - X + 1)(1 - r)^N$$

and the normalizing constant is

$$\sum_{i=X}^{M} i(i-1)\dots(i-X+1)(1-r)^i$$

This posterior distribution is not among the well known standard distributions. But it is a distribution. We just cannot find this distribution in a common statistical software. If our tools only allow to operate with a limited number of well known distributions, then we could not handle this. Therefore, it is good to have a software that allows some self-made programming in this kind of situations, e.g. in R: try the following, but be careful to use correct values: $X \leq N \leq M$.

```
p0 <- function(X,N,r){
s <- log(N)
for(i in 1:X-1){
s <- s+log(N-i)
}
s<-s+N*log(1-r)
exp(s)
}
postn <- function(X,N,M,r){
p0(X,N,r)/sum(p0(X,X:M,r))
}
```

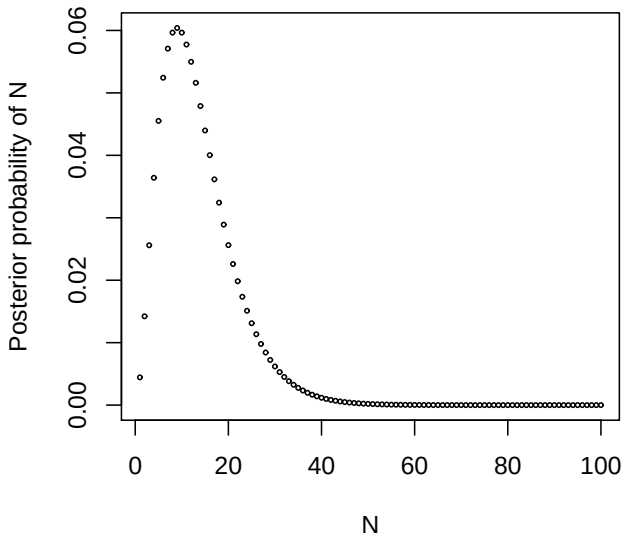

Figure 2: Posterior probability for N, given that $X = 1, r = 0.2$ with uniform prior over $0, 1, 2, \dots, M = 100$.

The estimation of unknown proportion r is a common application in many applied areas, e.g. epidemiology. Applications with unknown N are rare because usually we know the sample size. In some

situations this information may be missing. For example, if only positive results are reported in some reporting system, omitting negative results. We would not then know what the sample size was. It would also be difficult to estimate r, because all standard approaches assume N is known. In bayesian inference, unknown N just adds one more source of uncertainty to the problem (which then becomes described by a two-dimensional distribution).

In all cases, **bayesian model is the full joint distribution**. If we have the binomial model for data, $P(X \mid N, p) = \text{Bin}(N, p)$, the bayesian model is $\pi(X, p)$ as shown in a Figure before. If we also treat N as an uncertain quantity, the full model is $\pi(N, X, p)$. Depending on whatever becomes observed, the bayesian learner will compute a conditional distribution from the full model, by conditioning to the observed variables.

1.2 Exercises

2 Summarizing the posterior distribution

Often, the posterior distribution is presented graphically, possibly with the analytical mathematical expression of the density (if it could be solved) or as given in the Bayes formula (prior times likelihood). A graphical display is very informative, but sometimes we need simple summaries. In non-bayesian statistics we often deal with 'estimators', which are functions of the data and therefore 'random', conditionally to some hypothetical parameter values. The calculated values of such estimators are then taken as estimates of the (nonrandom) unknown parameters. But in Bayesian statistics, the parameters are random (i.e. uncertain), described by the posterior distribution. Therefore, the usual ways to summarize a probability distribution are directly applicable. Typically: mean, mode, or median. Also the width of the distribution is important, since it represents how uncertain we are. Therefore, variance, or standard deviation can be reported. For standard densities, these are easily calculated. For less common distributions, they may be easily available numerically in various software. Also, percentiles of the distribution can be informative. Very often, *credible intervals* (or regions for higher dimensional parameters) are reported.

The binomial model of red balls led to the posterior of the unknown proportion in the form of a beta-density. Since the expected value of a Beta(α, β)-density is $\alpha/(\alpha+\beta)$, and the mode is $(\alpha-1)/(\alpha+\beta-2)$ it is easy to summarize the posterior density by reporting the mean and mode

$$E(r \mid \alpha, \beta, N, X) = \frac{X+\alpha}{N+\alpha+\beta}$$

$$\text{Mod}(r \mid \alpha, \beta, N, X) = \frac{X+\alpha-1}{N+\alpha+\beta-2}.$$

As noted, the posterior mean can also be written as a weighted average:

$$w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{X}{N}, \qquad w = \frac{\alpha+\beta}{\alpha+\beta+N},$$

showing how the prior and the data contribute to the estimate. This is a useful way to summarize the relative importance of both sources of information. But the simple analytical expression is limited to conjugate models only.

We can draw this posterior density in each situation by simply plotting the beta-density. But then we need a software, such as R. For example, using the commands

```
X <- 2; N <- 20
p <- seq(0,1,by=0.01)
plot(p,dbeta(p,X+1,N-X+1),type="l")
```

Finally, should we summarize a posterior distribution by its mean, median, mode or something else? Eventually, this should depend on the context and the purpose of the analysis. This could be mathematically tackled by a *loss function*. This function should define how much the error 'costs' when making a decision. *The chosen loss function determines which point estimate is best.* (By 'point estimate' we mean one of the possible point values for summarizing the posterior distribution). For

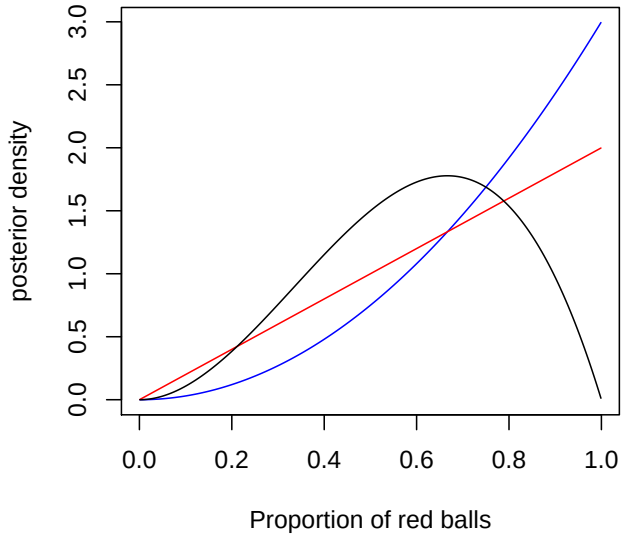

Figure 3: Posterior probability density for the proportion of red balls in an infinitely large bag of infinitely many balls, if one ball is drawn and it is red (red line), and if two balls are drawn and both are red (blue line), and if three balls are drawn and one is white (black line).

example, if we estimate some unknown parameter p by choosing a point estimate δ_x that depends somehow on our data x, and if we define a quadratic loss

$$L(p, \delta_x) = (p - \delta_x)^2$$

then the Bayes risk

$$\int \int L(p, \delta_x)\pi(p \mid x)\mathbf{d}x\mathbf{d}p$$

should be minimized. This will be minimized by minimizing the *posterior loss*

$$E(L(p, \delta_x) \mid x) = \int L(p, \delta_x)\pi(p \mid x)\mathbf{d}p$$

for each x. With the quadratic loss function, we get

$$= \int (p - \delta_x)^2 \pi(p \mid x)\mathbf{d}p = \int (p - E(p \mid x) + E(p \mid x) - \delta_x)^2 \pi(p \mid x)\mathbf{d}p$$

$$= \int (p - E(p \mid x))^2 \pi(p \mid x)\mathbf{d}p + (E(p \mid x) - \delta_x)^2$$

$$-2(\delta_x - E(p \mid x)) \underbrace{\int (p - E(p \mid x))\pi(p \mid x)\mathbf{d}p}_{=0}$$

$$= V(p \mid x) + (E(p \mid x) - \delta_x)^2$$

which is minimized when $\delta_x = E(p \mid x)$.

Similarly, we can think of some function of the parameter $h(p)$, so that the posterior mean $E(h(p) \mid x)$ is again the choice which will minimize the posterior loss with quadratic loss function $(h(p) - \delta_x)^2$.

Posterior median will minimize the loss with absolute error $L(p, \delta_x) =\mid p - \delta_x \mid$, and posterior mode will minimize the loss with 'all-or-nothing' error $1_{\{p=\delta_x\}}(\delta_x)$.

In general, a full Bayesian analysis would indeed consist of a decision problem for a real life application, where the decisions and consequences have specific losses. Then we need to choose a decision which minimizes the posterior loss. Thinking of practical computation, it is easy to see that this can be hard. Firstly, the posterior density $\pi(p \mid x)$ is generally not available in closed form. Secondly, the calculation of $E(h(p) \mid x)$ is generally difficult and not possible analytically. Also, other loss functions can be even more difficult.

2.1 Credible Intervals (CI)

Mode shows where the distribution is mostly concentrated, but it does not convey information about how uncertain we are. This is always the problem with point summaries (as with point estimates in non-Bayesian statistics). Hence, variance of a distribution could be reported in addition. However, we are often required to report a region, or interval, to describe the uncertainty. From a posterior distribution we can immediately obtain intervals that contain a specific probability. The interval is usually defined so that the point summary is somewhere in the middle, but not necessarily exactly in the middle. Any interval $[a, b]$ for which

$$\int_a^b \pi(r \mid \text{data}) \mathbf{d}r = Q$$

is said to be a $Q \times 100\%$ *Credible Interval*. This is usually constructed simply by taking $Q/2$ off from both ends of the distribution. But this is not necessarily the shortest possible interval. The shortest Credible Interval is called Highest Posterior Density Interval (HPD-interval). The simple Credible Interval is computationally easier to obtain. For standard distributions, it can be calculated by using tabulated (or computerized) quantiles. For example, to compute the 95% CI for the posterior of r, shown as black line in Figure (3), in R-software:

```
> qbeta(c(0.025,0.975),2+1,3-2+1)
[1] 0.1941204 0.9324140
```

And to calculate all 95% Credible Intervals of r for all possible outcomes $x \in [0, N]$:

```
N<-100; y<-0:N
lower<-qbeta(0.025,y+1,N-y+1);
upper<-qbeta(0.975,y+1,N-y+1);
plot(c(y[1],y[1]),c(lower[1],upper[1]),'l',
xlab='Red balls in a sample of N=100',
ylab='Bayesian 95% CI',
xlim=c(0,100),ylim=c(0,1));
for(i in 2:length(y)){
```

```
points(c(y[i],y[i]),c(lower[i],upper[i]),'l');
}
```

In comparison, the corresponding HPD interval of r would contain the same probability (e.g. 0.95), but we would need to find such interval that $\pi(r^* \mid X, N) > \pi(r \mid X, N)$ when r^* and r are any values within and outside the interval, respectively.

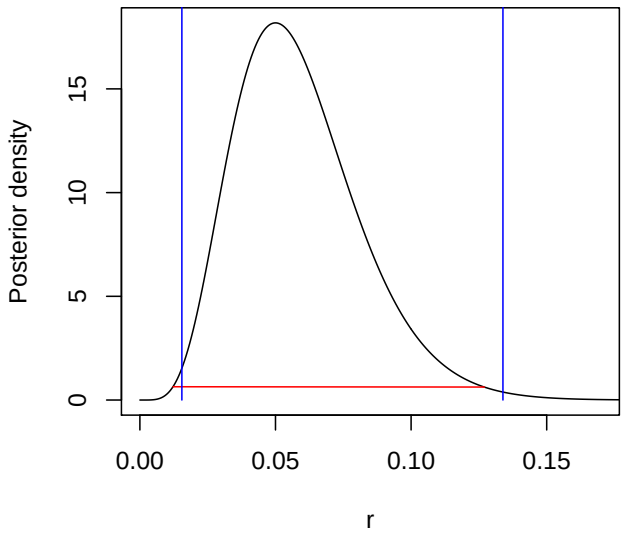

Figure 4: Comparison of HPD credible interval and simple credible interval from Beta(5+1,100-5+1) density. Red line shows 99% HPD interval. The length of 99% HPD CI is 0.1148 compared to 0.1184 of the simple 99% CI.

As a non-bayesian alternative, the exact frequentist 95% Confidence Interval (Clopper-Pearson interval) would be the set

$$\{r : P(Y \leq Y^{obs} \mid N, r) \geq 0.025\} \cap \{r : P(Y \geq Y^{obs} \mid N, r) \geq 0.025\}$$

which could be calculated for every outcome $y \in [0, N]$ as:

```
N<-100; y<-0:N
p<-seq(0,1,by=0.001);
I<-(1-pbinom(y[1]-1,N,p)>0.025)&(pbinom(y[1],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
plot(c(y[1],y[1]),c(lower,upper),'l',
xlab='Red balls in a sample of N=100',
ylab='Freq. 95% CI',xlim=c(0,N),ylim=c(0,1));
for(i in 2:length(y)){
```

```
I<-(1-pbinom(y[i]-1,N,p)>0.025)&(pbinom(y[i],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
points(c(y[i],y[i]),c(lower,upper),'l')
}
```

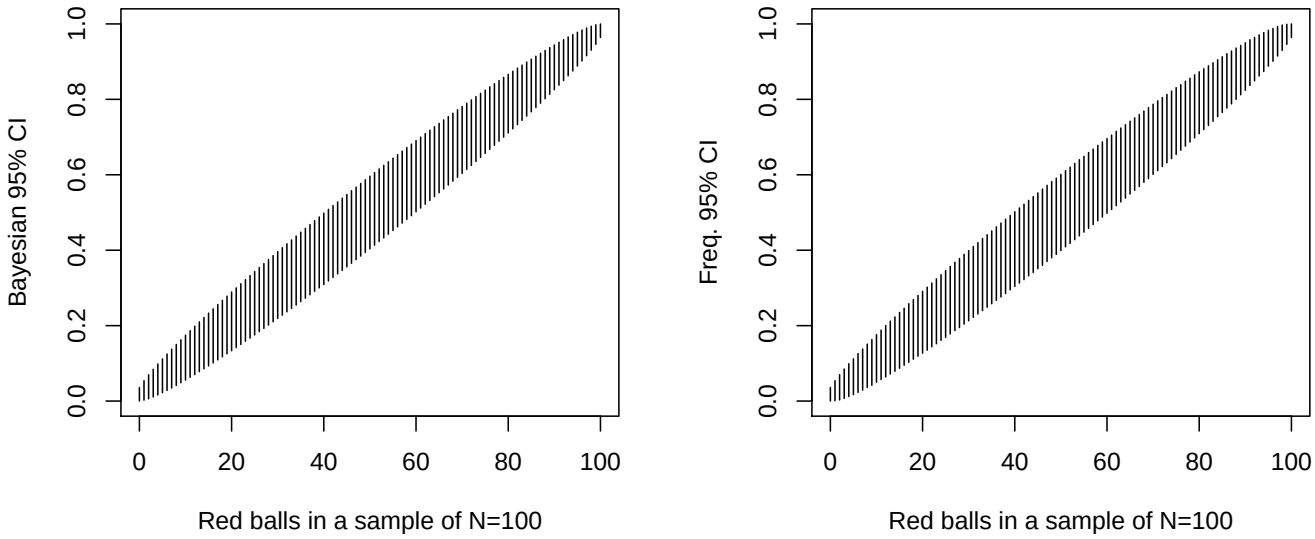

Figure 5: Bayesian Credible Intervals and frequentist Confidence Intervals.

The figure (5) looks very similar in both frequentist and bayesian calculations. Note, however, the difference of interpretation. In the bayesian approach, the unknown proportion r has distribution. In the frequentist approach, r is fixed unknown constant, and the *interval* is random, and it *would* cover the true unknown value of r in 95% of the cases if the experiment was repeated, but it says nothing about the probability that r belongs to this interval for any given sample Y that actually occurred. (See [11], page 453).

The bayesian CI was solved by finding the integration limits for the posterior, such that the required probability is achieved between $[a, b]$. In general, the HPD-CI can be a set of distinct intervals if the posterior density happens to be multimodal. Numerical techniques for solving the CI's would require that we can calculate the posterior density function accurately (which was possible above).

2.1.1 The more data, the narrower CI can be expected

Obviously, the resulting width of a CI depends on the amount of information we had. When the amount of data increases, we can expect the posterior to become more peaked, and hence the CI more narrow. On average, this is guaranteed because the prior variance of r can be written as

$$V(r) = E(V(r \mid X)) + V(E(r \mid X))$$

which shows that the posterior variance $V(r \mid X)$ is *expected* to be smaller than the prior variance. We can study the expected width of the CI with different sample sizes N and choose the value of N that gives the required expected width.

References

[1] McGrayne S B: The theorem that would not die. Yale University Press. 2011.

[2] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[3] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[4] Christensen R, Johnson W, Branscum A, Hanson E: Bayesian Ideas and Data Analysis. CRC Press. 2011.

[5] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[6] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.

[7] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[8] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[9] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[10] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[11] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.

[12] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515-533. 2006.