# 4 Exchangeability and conditional independence

The assumption behind the binomial distribution was that the individual experiments (balls drawn) were independent events, given the true proportion $r$. This means that

$$P(X_1, \ldots, X_N \mid r) = P(X_1 \mid r) \times \ldots \times P(X_N \mid r)$$

so that the exact order of the results $X_i$ does not matter because the resulting probability will be the same as long as the sum $\sum X_i$ is the same. Hence, the binomial distribution becomes defined for the sum of 'successful events' in a series of $N$ trials. This is a special case, in which the true proportion $r$ encapsulates all 'essential' background knowledge. In order to calculate the probability, we need to know (or assume) a value for $r$. But in practice, of course, we cannot know what $r$ is.

In a more general approach, we can study probabilities of the form

$$P(X_1, \ldots, X_N \mid I)$$

where $I$ denotes all our background information (which we always have) for a given problem, when assigning our (subjective) probability for some sequence of observations $X_i$. If our probability is such that it remains the same regardless of the ordering of the sequence,

$$P(X_1, \ldots, X_N \mid I) = P(X_{s_1}, \ldots, X_{s_N} \mid I)$$

for all permutations $s$ of the indexes, then the sequence of $X_i$ is said to be (finitely) *exchangeable*. This is an important concept in bayesian modeling. A famous result (by Bruno de Finetti, 1906-1985, http://www.brunodefinetti.it/) follows from the assumption of *infinite* exchangeability. It can be shown that then the probability can be written in the form

$$P(X_1, \ldots, X_N \mid I) = \int_0^1 \prod_i^N r^{X_i}(1-r)^{1-X_i}\pi(r)\mathbf{d}r$$

The interpretation of parameter $r$ is that $r = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i$. It can also be interpreted as marginal probability of a single event, $r = P(X_i = 1)$.

Interpretation of de Finetti's theorem of subjective probability:

(**I**) Parameter $r$ can be thought *as if* it was the proportion of successful events in an infinite sequence, or the probability of an individual event.
(**II**) Parameter $r$ has to be considered as a random quantity with probability density $\pi(r)$.
(**III**) Conditionally, given $r$, the variables $X_i$ are independent and equally distributed, as Bernoulli($r$).

Note that parameter $r$ emerges only as a mathematical device when the subjective probability concerning the $X_i$ is such that it obeys exchangeability. We are still assigning probabilities for the observable events $X_i$. The density $\pi(r)$ is not a 'probability of probability'. We have just written our probability of the sequence $X_i$ as a mathematical expression that directly follows from the exchangeability assumption. In fact, parameter $r$ is just a mathematical device that allows us to update our probabilities concerning the $X_i$.

*With the predictive approach parameters diminish in importance,*
*especially those that have no physical meaning.*
*From the Bayesian viewpoint, such parameters can be regarded as*
*just place holders for a particular kind of uncertainty*
*on your way to making good predictions. (Draper 1997, Lindley 1972).*

The conditional probability $P(X_i \mid r)$ provides an important tool for parametric modeling in which we simplify our background knowledge $I$ into a one or few parameters. This is the problem of model choice that is always a subjective choice (in all modeling, not just bayesian). And the density $\pi(r)$ is an important tool in bayesian analysis, where the whole model is not just of the form $P(X \mid r)$, but it is the joint model $P(X, r)$ of both the observable part $X$ *and* the unobservable part $r$.

Therefore, the $X_i$ are not independent of each other, only conditionally independent, given $r$. This means that we can learn from the observed $X_i$ to predict other $X_j$ that are not yet observed. For example, the *prior predictive distribution* of $X_1$ is

$$P(X_1) = \int_0^1 P(X_1 \mid r)\pi(r)\mathbf{d}r$$

and after we have observed $X_1$, the *posterior predictive distribution* of $X_2$ is

$$P(X_2 \mid X_1) = \int_0^1 P(X_2 \mid r)\pi(r \mid X_1)\mathbf{d}r$$

where $\pi(r \mid X_1)$ is the posterior distribution of parameter $r$.

*D V Lindley reports that Bruno de Finetti was especially fond of the aphorism:*
**Probability does not exist**
*which conveys his idea that probability is an expression of the observer's view of the world*
*and as such it has no existence of its own.*

*Reported by D V Lindley, de Finetti insisted that*
*"random variables" should more appropriately be called "random quantities", for "What varies?"*
*Furthermore, coherently with his view of probabilistic thinking*
*as a tool to deal with uncertainty in life,*
*he thought that it should be taught to children at an early age.*

# 5 Things to do with the posterior in the City

We continue with the binomial model of red balls, which led to the posterior of the unknown proportion in the form of a beta-density. Since the expected value of a Beta$(\alpha, \beta)$-density is $\alpha/(\alpha+\beta)$, and the mode is $(\alpha-1)/(\alpha+\beta-2)$ the posterior density has mean and mode

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

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

These are often used as bayesian *point estimates*; summaries of posterior distribution. As noted earlier, the posterior mean could also be written as a weighted average:

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

showing how the prior and the data contribute to the estimate. This was the nice feature of conjugate priors which allow us to explore how much each of the two sources of information contribute to the result.

But point estimates are just summaries of the posterior distribution. These summaries could be means, modes, medians, variances and they could be compared to non-bayesian estimators in classical statistics. However, the posterior density (or marginal density) can always be displayed graphically, which can be more informative than point estimates. Assume again that the first ball drawn was red, and the second ball was also red, but the third turns out white. Next, plot the posterior density in each situation by plotting the corresponding beta-density. (We need a software, such as R).

## 5.1 Hypotheses

With continuous quantities as $r$, it is not meaningful to ask e.g. what is the probability $P(r = 0.5)$, because such probability is always zero. The mode of the density shows the value with highest probability density, and thus it provides a 'best guess'. A computable hypothesis about the value of $r$ needs to be constructed as a statement involving intervals. Using posterior density, we can then study what evidence we have to support specific hypotheses. For example, if the hypothesis is that $r < 0.5$, then the prior probability of that hypothesis is

$$P(r < 0.5) = \int_0^{0.5} \pi(r)\mathbf{d}r = 0.5 \qquad \text{(from U(0,1)-prior)}$$

but the posterior probability (when $Y = 2, N = 3$) would be

$$P(r < 0.5 \mid Y, N) = \int_0^{0.5} \text{Beta}(r \mid Y+1, N-Y+1)\mathbf{d}r$$

which is the cumulative probability of the beta-density at $r = 0.5$. The approximate value (0.3125) is obtained by typing `pbeta(0.5,Y+1,N-Y+1)`. The posterior probability summarizes the current evidence, but we may also compute posterior odds. This provides an alternative way of representing bayes formula: the posterior odds are the prior odds multiplied by the likelihood ratio. With binary variable $\theta$ describing the hypothesis (either true or false) this would be written as:

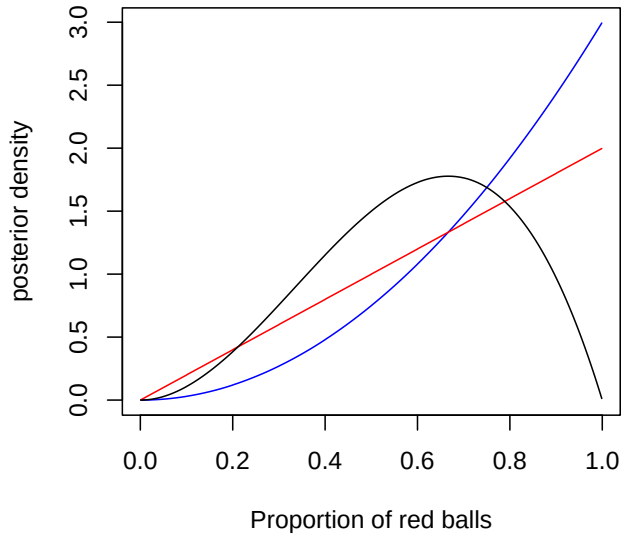


Figure 1: Posterior probability density for the proportion of red balls in an infinitely large bag of infinitely many balls, if one ball is drawn and it is red (red line), and if two balls are drawn and both are red (blue line), and if three balls are drawn and one is white (black line).

$$\frac{\pi(\theta = 1 \mid \text{data})}{\pi(\theta = 0 \mid \text{data})} = \frac{\pi(\theta = 1)\pi(\text{data} \mid \theta = 1)/\pi(\text{data})}{\pi(\theta = 0)\pi(\text{data} \mid \theta = 0)/\pi(\text{data})} = \frac{\pi(\theta = 1)}{\pi(\theta = 0)} \frac{\pi(\text{data} \mid \theta = 1)}{\pi(\text{data} \mid \theta = 0)}$$

The prior odds for the hypothesis were

$$\frac{P(r < 0.5)}{P(r \geq 0.5)} = 1$$

but the posterior odds are only about half of that

$$\frac{P(r < 0.5 \mid Y, N)}{P(r \geq 0.5 \mid Y, N)} = \frac{0.3125}{0.6875} = 0.4545.$$

Hypotheses could also involve comparisons of two quantities. For example, we could study two different bags, each with a different proportion of red balls, $r_1$ and $r_2$, and we get some observations from both, $(Y_1, N_1)$ and $(Y_2, N_2)$. The hypothesis could then be e.g. $H_0 : r_1 < r_2$. What is the prior and the posterior probability of the hypothesis? To study this, we can create a new variable: $s = r_1 - r_2$, so that $H_0 : s < 0$. But now the distribution of $s$ is a convolution of two independent distributions and generally it may be difficult to compute.

### 5.1.1 Example: hemophilia

Example from Gelman [4]: hemophilia is a genetically inherited disease, carried in X-chromosomes. If a male (having 'XY') has this X-chromosome, he is affected. If a female (having 'XX') has this in one

X-chromosome, she is an unaffected carrier. (Having it in both X-chromosomes is rare but always fatal).

Assume a woman has an affected brother. So, her mother must be a carrier with one bad X, and one good X. If we also know that her father is not affected, then this woman has 50% chance of being carrier. Hence the prior is $P(\theta = 1) = P(\theta = 0) = 0.5$. Then we observe her two sons are not affected ($y_1 = 0, y_2 = 0$). What is the posterior probability that the woman is a carrier?

The conditional probability of such data (likelihood of $\theta$) is

$$P(y_1 = 0, y_2 = 0 \mid \theta = 1) = 0.5 \times 0.5 = 0.25$$

$$P(y_1 = 0, y_2 = 0 \mid \theta = 0) = 1 \times 1 = 1.$$

Again, using bayes formula we find that

$$P(\theta = 1 \mid y_1 = 0, y_2 = 0) = \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.2$$

The prior odds were $0.5/0.5 = 1$. The likelihood ratio is $0.25/1$. So, the posterior odds are now 0.25 (which can be converted back to probability: 0.25/(1+0.25)=0.2).

### 5.1.2 Example: analysis of birth data

Example from Gelman [4]: the proportion of female births in Germany is 0.485. In a study of a rare condition of pregnancy it was observed that in 980 of such births, 437 were female. That's 0.4459184, which is a little lower than expected. How much evidence this gives for the claim that the proportion of female births in such conditions is lower than in the large population? Assuming uniform prior probability for the female proportion $r$, the posterior density becomes

$$\pi(r \mid X = 437, N = 980) = \text{Beta}(438, 544).$$

The posterior mean of $r$ is 0.446, and the posterior standard deviation 0.016. The median is 0.446, (`qbeta(0.5,438,544)`). The probability $P(r < 0.485)$ is

$$P(r < 0.485) = \texttt{pbeta(0.485,438,544)} = 0.992826$$

which seems quite high. This result was obtained when the prior was uniform. We can check how much difference does it make if the prior would be more concentrated around population mean 0.485.

| $\frac{\alpha}{\alpha+\beta}$ | $\alpha + \beta$ | posterior median | 95%posterior interval |
|---|---|---|---|
| 0.5 | 2 | 0.446 | [0.415, 0.477] |
| 0.485 | 2 | 0.446 | [0.415, 0.477] |
| 0.485 | 5 | 0.446 | [0.415, 0.477] |
| 0.485 | 10 | 0.446 | [0.415, 0.477] |
| 0.485 | 20 | 0.447 | [0.416, 0.478] |
| 0.485 | 100 | 0.450 | [0.420, 0.479] |
| 0.485 | 200 | 0.453 | [0.424, 0.481] |

The prior mean is outside the 95% interval in all of these. In addition to $r$, an interesting quantity is the sex ratio $z = (1 - r)/r$. Distribution of $z$ could be found using the transformation of variables technique. In practice, it is easier to produce it by simulation techniques.

### 5.1.3 Winning Monty Hall

Monty Hall problem is a famous game in which you are first offered a choice over 3 boxes, one of which contains a prize and others are empty. Once you have made your initial choice, you are not yet allowed to open your box. Instead, one of the other boxes is shown to be empty by the game master who knows exactly what was placed in each box. You are then asked to make your final choice: do you keep your initially chosen box, or do you change for the other unopened box? The hypothesis under judgement is that A='the prize is in your box already' or B='the prize is in the other box'.

Initially, the probability to make a correct choice is $P(A) = 1/3$, hence $P(B) = 2/3$. We then need to define the conditional probabilities for the data that you'll be shown. Given that the prize is already in your box, the probability that an empty box is revealed to you is surely one: $P(\text{Monty shows empty'} \mid A) = 1$. But since Monty knows exactly the contents of all boxes, there will always be at least one empty box for him that he can reveal. So: $P(\text{'Monty shows empty'} \mid B) = 1$. Now we get $P(B \mid \text{'Monty shows empty'})$

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{2}{3}}{\frac{2}{3} + \frac{1}{3}} = \frac{2}{3}.$$

But let's change the rules! Assume then that Monty is allowed to choose randomly (blindfolded) which one of his boxes he opens. Now we still have $P(\text{Monty shows empty'} \mid A) = 1$, but if the prize is in the other boxes, then $P(\text{'Monty shows empty'} \mid B) = 1/2$. This will change the result:

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{1}{2}\frac{2}{3}}{\frac{1}{2}\frac{2}{3} + \frac{1}{3}} = \frac{1}{2}.$$

We really need to know how the game is played!

## 5.2 Credible Intervals

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 estimates. 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 estimate 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 $r$ with black line in Figure (1) 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.

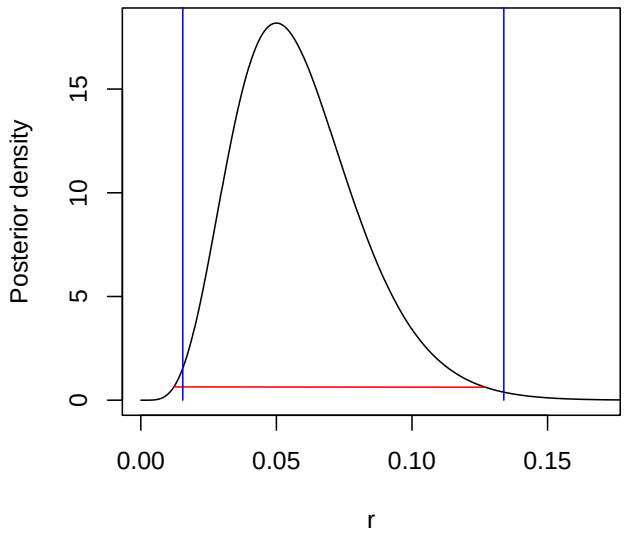


Figure 2: Comparison of HPD credible interval and simple credible interval from Beta(5+1,100-5+1) density. Red line shows 99% HPD interval. The length of 99% HPD CI is 0.1148 compared to 0.1184 of the simple 99% CI.

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')
}
```

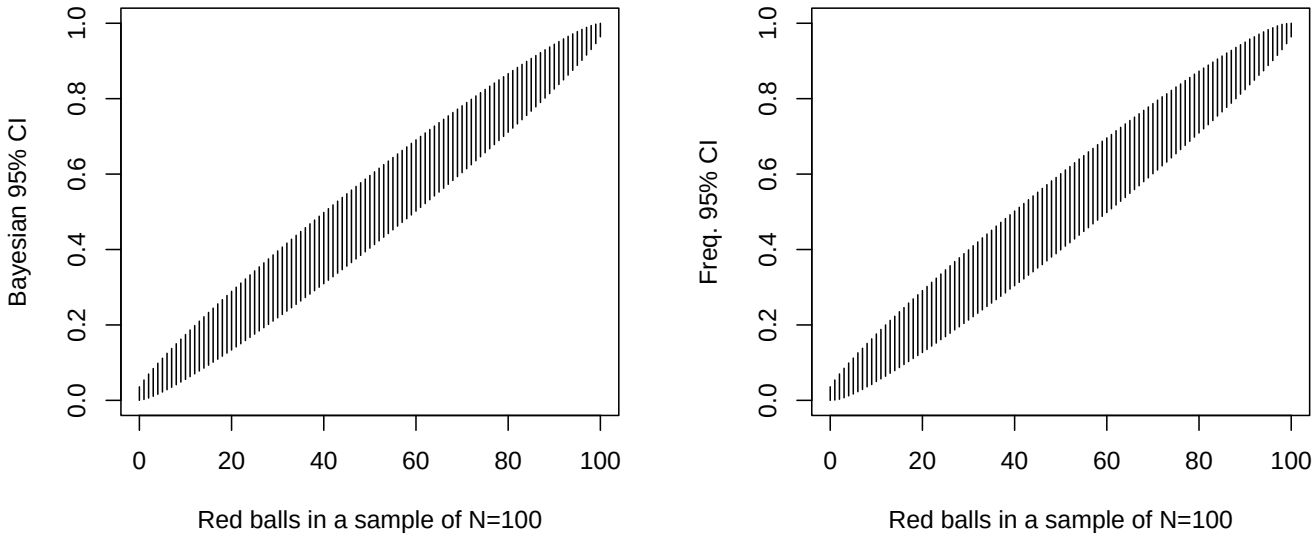


Figure 3: Bayesian Credible Intervals and frequentist Confidence Intervals.

The figure (3) 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 [7], 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).

### 5.2.1 The more data, the narrower CI to expect

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 could study the expected width of the CI with different sample sizes $N$ and choose the value of $N$ that gives the required expected width. This could be computed using Monte Carlo methods.

## 5.3 Predictions

While posterior density summarizes our current uncertainty about an unknown quantity, predictions of future experiments and events could sometimes be even more interesting. (Some have even argued that it is the ultimate purpose of modeling). For example, assume that the experiment of drawing balls is to be continued after the first three balls were picked. We should then predict the color of the next ball. Our model tells us that, conditionally on $r$, the probability of red ball in the next draw is simply $r$ (according to a parametric model and de Finetti). But the true value of $r$ was unknown (and will remain unknown, representing an infinite population). In such parametric model, we could use our current estimate for the parameter, but a fixed point estimate does not account for the fact that we are still uncertain about the parameter. The posterior predictive probability for the next ball to be red is:

$$P(\text{red} \mid Y, N) = \int_0^1 \underbrace{P(\text{red} \mid r)}_{=r} \times \underbrace{P(r \mid Y, N)}_{\text{Beta(Y+1,N-Y+1)}} \mathbf{d}r \quad = E(r \mid Y, N) = \frac{Y+\alpha}{N+\alpha+\beta}$$

which is the same as the posterior mean of parameter $r$.

Next: consider an experiment where $N$ new balls are to be picked, $X$ of them will be red, so $X \sim \text{Bin}(N, r)$, and our current uncertainty about $r$ is represented by beta-distribution $\text{Beta}(\alpha, \beta)$ (which could be the posterior of $r$, based on some earlier data). What is the predictive distribution of $X$ in this new experiment?

$$P(X \mid N, \alpha, \beta) = \int_0^1 \underbrace{P(X \mid N, r)}_{\text{Bin}(N,r)} \underbrace{\pi(r \mid \alpha, \beta)}_{\text{Beta}(\alpha,\beta)} \mathbf{d}r$$

$$= \int_0^1 \frac{\Gamma(N+1)}{\Gamma(X+1)\Gamma(N-X+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{X+\alpha-1}(1-r)^{N-X+\beta-1} \mathbf{d}r$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 r^{X+\alpha-1}(1-p)^{N-X+\beta-1} \mathbf{d}r$$

Then, write: $A = X + \alpha$, $B = N - X + \beta$, so that

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_0^1 \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} r^{A-1}(1-r)^{B-1}\mathbf{d}r}_{=1} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \binom{N}{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

which can also be written using so called *beta-functions:*

$$\binom{N}{X} \frac{\text{beta}(A,B)}{\text{beta}(\alpha,\beta)}$$

This distribution of $X$ is said to be *beta-binomial* distribution. It is sometimes used e.g. in food safety microbial risk assessments to describe e.g. the number of contaminated servings $X$ among $N$ servings, under uncertainty about the true fraction, $r$, of contaminated servings in a large (infinite) population. In risk assessment literature, the conditional distribution of $X$ (binomial distribution) is often called as the variability distribution of $X$, and the distribution of $r$ (beta distribution) as the uncertainty distribution of $r$. Hence, it is often said in RA-literature that 'variability and uncertainty are separated'. In bayesian context, both distributions are expressions of uncertainty (perhaps epistemic uncertainty and aleatoric uncertainty), and the resulting beta-binomial distribution reflects both uncertainties. This can be either prior predictive distribution, or posterior predictive distribution.

As a side step, consider a situation in which we pick $N$ new balls, but assuming that each of the balls is picked from a different population (e.g. different bags) so that for each draw we have Bernoulli-distribution with different parameter $r_i$. ($\text{Bin}(1, r_i)$). Our uncertainty about all $r_i$ is assumed to be described as some distribution $\pi(r_i)$, (which could be $\text{Beta}(\alpha, \beta)$). What is the distribution of $X$?

$$P(X \mid N) = \int_0^1 P(X \mid r_1, \ldots, r_N) P(r_1, \ldots, r_N) \mathbf{d}r_1 \ldots \mathbf{d}r_N$$

$$= \int_0^1 \ldots \int_0^1 \binom{N}{X} \prod_{i=1}^{X} r_{k_i} \prod_{i=X+1}^{N} (1 - r_{k_i}) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1}(1-r_{k_i})^{\beta-1} \mathbf{d}r_{k_1} \ldots \mathbf{d}r_{k_N}$$

Here, $k_1, \ldots, k_N$ is some permutation of the indices $i$. After re-arranging the terms in this expression, we get:

$$\binom{N}{X} \int_0^1 \ldots \int_0^1 \prod_{i=1}^{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha+1-1}(1-r_{k_i})^{\beta-1} \prod_{i=X+1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1}(1-r_{k_i})^{\beta+1-1} \mathbf{d}r_{k_1} \ldots \mathbf{d}r_{k_N}$$

and by integrating over each $r_i$ one by one, we get:

$$= \binom{N}{X} E(r_i)^X E(1-r_i)^{N-X} = \text{Bin}\Big(N, \frac{\alpha}{\alpha+\beta}\Big)$$

This is a distribution that depends on $N$ and the expected value of $r_i$, so the prior distribution of $r_i$ affects the result via its expected value only.

## 5.4 Approximating posterior density

Posterior density can be approximated by a normal distribution

$$\pi(\theta \mid X) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

where $\hat{\theta}$ is the posterior mode and $I(\theta)$ is so called *observed information*

$$I(\theta) = -\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X).$$

The approximation is based on Taylor series expansion of $\log \pi(\theta \mid X)$ centered at the posterior mode, $\hat{\theta}$. For a scalar valued $\theta$ this is

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \underbrace{[\frac{\mathbf{d}}{\mathbf{d}\theta} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}}}_{=0} \frac{(\theta - \hat{\theta})}{1!} + [\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)]_{\theta=\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2!} + \cdots,$$

where the first derivative at posterior mode $\hat{\theta}$ is zero. When $\theta$ is near the mode, the higher order terms are small compared to the first terms. As a function of $\theta$, the first term in the expression is constant whereas the 2nd order term is proportional to the logarithm of a normal density, which provides the approximation. For a vector valued $\theta$, the Taylor series would be

$$\log \pi(\theta \mid X) = \log \pi(\hat{\theta} \mid X) + \frac{1}{2}(\theta - \hat{\theta})^T \Big[\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(\theta \mid X)\Big]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots.$$

The normal approximation can be a useful benchmark and it gives a quick approximation of the posterior density. For final results, more accurate computations are usually needed. Even so, the first rough estimates can be obtained from the approximation, if only as realistic starting values for more complicated calculations.

## 5.5 Comment

The above examples required exact or approximate solutions to integrals and posterior densities. The purpose of the examples was to demonstrate how bayesian inference is merely a matter of applying probability calculus to practical problems of quantifying uncertainty. If exact analytical solution becomes difficult to find 'by paper and pencil', the integrals can be calculated using numerical methods available in different softwares. But there are also other ways to approximate. At this point, it is better to move to Monte Carlo simulation.

# References


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

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

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

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

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

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

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