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). The posterior density is the basis for this too. What we get is a **posterior predictive distribution**. Assume a parametric model $\pi(X_i \mid \theta)$ for each variable in the sequence $X_1, X_2, \ldots$, so that the variables X_i are conditionally independent of each other, given the parameter θ. The goal is to predict next $X^{\text{next}} = X_{n+1}$, given the previous observed values $X = \{X_1, \ldots, X_n\}$.

$$\pi(X^{\text{next}} \mid X) = \int \pi(X^{\text{next}}, \theta \mid X)\mathbf{d}\theta = \int \underbrace{\pi(X^{\text{next}} \mid \theta, X)}_{=\pi(X^{\text{next}}\mid\theta)} \pi(\theta \mid X)\mathbf{d}\theta$$

This is the solution to the practical problem: having some probability model $\pi(X \mid \theta)$, how to compute a prediction, based on our observations X, **without knowing** the underlying value of θ? It is easy to calculate $\pi(X \mid \theta)$ and generate random values for X, when we assume some specific value for θ. In practice, this is unknown in every real application. Therefore, we use probability distribution to describe our uncertainty about θ. But the data informs us about probable values of θ. Hence, the posterior distribution is used, and the prediction distributions $\pi(X^{\text{next}} \mid \theta)$ are weighted by this posterior distribution.

Before having data, we just had the prior. From this, we can similarly compute the **prior predictive distribution**:

$$\pi(X^{\text{next}}) = \int \pi(X^{\text{next}}, \theta)\mathbf{d}\theta = \int \pi(X^{\text{next}} \mid \theta)\pi(\theta)\mathbf{d}\theta$$

In these notations, we could have written that they are conditional to the prior information, so that the prior predictive distribution actually is $\pi(X^{\text{next}} \mid I)$. This corresponds to our prior beliefs about the *observable* variables X. The parameter θ can be seen as purely a technical device, which provides a way to write this. This parameter may or may not have a close interpretation as a physical condition. Our focus is on assigning our probabilities to the actually observable quantities X. Parameter θ may have no interest in its own right.

> *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).*

3.1 Exchangeability

Consider a sequence of binary variables 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. An important 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).

In all this, 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. Hence, parameter r is just a mathematical device that allows us to update our probabilities concerning the X_i.

Similarly, exchangeability works for other sequences of variables, not just binary variables. Whenever our beliefs about the observable variables X_i are exchangeable, it follows that there must exist a parametric model $\pi(X \mid \theta)$ and a distribution $\pi(\theta)$ so that our probability of $X_1, \ldots, X_n$ can be expressed as

$$\pi(X_1, \ldots, X_n) = \int_\Theta \prod_i^n \pi(X_i \mid \theta)\pi(\theta)\mathbf{d}\theta$$

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

Therefore, the X_i are not independent of each other, **only conditionally independent**, given θ. This means that we can learn from the observed X_i to predict other X_j that are not yet observed.

Quoting Bernardo: *It is important to realise that if the observations are conditionally independent, - as it is implicitly assumed when they are considered to be a random sample from some model - , then they are necessarily exchangeable. The representation theorem, - a pure probability theory result - proves that if observations are judged to be exchangeable, then they must indeed be a random sample*

from some model and there must exist a prior probability distribution over the parameter of the model, hence requiring a Bayesian approach. Note however that the representation theorem is an existence theorem: it generally does not specify the model, and it never specifies the required prior distribution. An additional effort is necessary to assess a prior distribution for the parameter of the model.

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.

3.2 Prediction for binomial experiment

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 (but perhaps epistemic uncertainty and aleatoric uncertainty), and the resulting beta-binomial distribution reflects both types of uncertainties. The result can be either prior predictive distribution (in which case α, β represent parameters of a prior (Beta-) distribution), or posterior predictive distribution (in which case α, β represent parameters of a posterior (Beta-) distribution). Beta-binomial distribution can be used to account for **overdispersion in binomial models**: the distribution has two parameters, α, β, in place of the single parameter r of the binomial distribution.

By using the two general (often useful) probability laws for total expectation and total variance:

$$E(X) = E(E(X \mid Z))$$

and

$$V(X) = E(V(X \mid Z)) + V(E(X \mid Z)),$$

the mean of beta-binomial can be found from

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

Similarly, its variance can be found from

$$V(X) = E(V(X \mid r, N)) + V(E(X \mid r, N)) = \frac{N\alpha\beta(\alpha+\beta+N)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

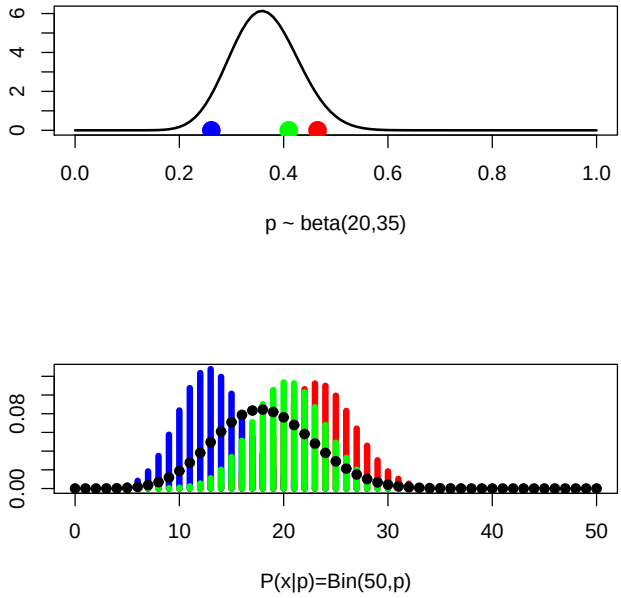

Figure 1: Upper frame: density of p (Beta(20,35)) and three randomly sampled values (red,blue,green). Lower frame: three conditional distributions for X (Bin(50,p)) with different sampled values of p (red,blue,green) corresponding to the upper frame. Integrating over all possible p according to the density of p, gives beta-binomial distribution for X (black dots). If the density of p was a posterior density based on earlier observed X_{obs}, then this gives the posterior predictive distribution of next X ($P(X \mid X_{\text{obs}})$).

```
R-code for producing the figure:
par(mfcol=c(2,1))
plot(seq(0,1,by=0.01),
dbeta(seq(0,1,by=0.01),20,35),'l',lwd=2,xlab="p ~ beta(20,35)",ylab="")
p <- rbeta(3,20,35)
points(p[1],0,cex=2,col="blue",pch=16)
points(p[2],0,cex=2,col="red",pch=16)
points(p[3],0,cex=2,col="green",pch=16)
x <- 0:50
plot(x,dbinom(x,50,p[1]),'h',lwd=5,col="blue",ylab="",xlab="P(x|p)=Bin(50,p)")
points(x,dbinom(x,50,p[2]),'h',lwd=5,col="red")
points(x,dbinom(x,50,p[3]),'h',lwd=5,col="green")
N <- 50; a <- 20; b <- 35; A <-x+a; B <- N-x+b
pr <- (gamma(N+1)*gamma(a+b)/(gamma(x+1)*gamma(N-x+1)*gamma(a)*gamma(b)))*
      gamma(A)*gamma(B)/gamma(A+B)
points(x,pr,pch=16)
```

3.2.1 Overdispersion not possible for Bernoulli variables

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?

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

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

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

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

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. **It is not possible to model overdispersion for Bernoulli variables**.

3.2.2 Example: mixture priors

Returning to the mixture priors mentioned before: the prior predictive distribution from an individual component distribution is

$$\pi_i(x) = \int_\Theta \pi_i(\theta) \pi(x \mid \theta) \, \mathbf{d}\theta,$$

and the prior predictive distribution from the whole mixture is

$$\pi(x) = \int_\Theta \pi(\theta) \pi(x \mid \theta) \, \mathbf{d}\theta = \sum_{i=1}^{k} \alpha_i \pi_i(x).$$

Now, the posterior distribution can be written as:

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

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

which is seen as a weighted average of component specific posterior distributions, with weights calculated from the predictive distributions as shown.

4 Hypotheses

Hypotheses are usually formulated for some parameter θ so that the null hypothesis is written

$$H_0 : \theta \in \Theta_0$$

against an alternative hypothesis

$$H_1 : \theta \in \Theta_1$$

If both sets have positive probabilities, e.g. if they are intervals and θ is a continuous parameter $\in S \subset \mathbb{R}$, then the Bayesian approach is to compute a posterior probability, with data X:

$$\pi(H_0 \mid X)$$

and the posterior probability for the alternative is $1-\pi(H_0 \mid X)$. The posterior probability summarizes the current evidence. What remains is to evaluate the results and decide how large (small) probability is large (small) enough. The posterior probability is not accepting or rejecting a hypothesis, it simply provides a numerical value for its plausibility. We need to think what level of plausibility is enough, if we had to take some action. Ultimately, this would call for a loss function in order to choose the decision that minimizes the expected loss with respect to the posterior distribution.

Note: frequentist hypothesis testing with p-values gives the probability of more extreme Y than the one observed, given null hypothesis: $P(Y$ more extreme than $Yobs \mid H_0)$. The null hypothesis may thus be rejected (or not), if a more extreme observation than what we had would seem too improbable. The frequentist hypothesis testing does not give probability of the hypothesis. Harold Jeffreys (1939), commented: "an hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it. The same applies to all the current significance tests based on P-values". The same arguments tend to get repeated back and forth.

For example, hypotheses related to linear models $E(Y_i) = \alpha_0 + \alpha_1 X_i$ could e.g. focus on the slope parameter for inspecting trend. $H_0 : \alpha_1 < 0$ and $H_1 : \alpha_1 \geq 0$. Hence, the posterior distribution $P(\alpha_1 < 0 \mid Y, X)$ directly assesses the probability of this hypothesis, $P(H_0 \mid Y, X)$. Actually, in this case, the hypothesis would not be about a true physical state of the world in the sense that 'a slope' does not exist: it is only a parameter of our model and we could have written a different model with different parameters. Hence, some hypotheses may have more direct interpretation as 'state of the world' than others.

Also, posterior odds can be written: for hypothesis H_0 against H_1 we have

$$\frac{\pi(H_0 \mid X)}{\pi(H_1 \mid X)}.$$

Whenever this is > 1, it shows support for H_0. *Bayes factor* is computed as a ratio of prior and posterior odds:

$$BF = \frac{\pi(H_0 \mid X)/\pi(H_1 \mid X)}{\pi(H_0)/\pi(H_1)} = \frac{\pi(H_0 \mid X)}{\pi(H_1 \mid X)} \frac{\pi(H_1)}{\pi(H_0)}.$$

In other words:

$$\textbf{Posterior odds } = \textbf{Prior odds} \times \textbf{Bayes factor}.$$

The Bayes factor measures how much the data changes the prior odds. If the factor is bigger than one, the data gave some support for the hypothesis H_0. Bayes factor provides a scale of evidence in favor of one hypothesis against another. (But the scale is from zero to infinity, which is not as 'neat' as probability scale).

If we have a point hypothesis (=simple hypothesis) where $H_0 : \theta = \theta_0$ against another point hypothesis $H_1 : \theta = \theta_1$, with some specific values of θ_0 and θ_1 (e.g. 3.5 against 5.7) then we must have positive probability for both, and the posterior odds is

$$\underbrace{\frac{\pi(\theta = \theta_0 \mid X)}{\pi(\theta = \theta_1 \mid X)}}_{\text{Posterior odds.}} = \overbrace{\underbrace{\frac{\pi(\theta = \theta_0)\times}{\pi(\theta = \theta_1)\times}}_{\text{Prior odds.}} \underbrace{\frac{\pi(X \mid \theta = \theta_0)}{\pi(X \mid \theta = \theta_1)}}_{\text{Likelihood ratio.}}}^{\text{Constant } \pi(X) \text{ cancels out}}$$

so that the likelihood ratio is the Bayes factor and this does not depend on the prior. In the expression above, it was sufficient to write the two posterior probabilities (on the left) 'proportional to' (i.e. just prior times likelihood, on the right) because the normalizing constant $\pi(X)$ cancels out from the ratio. From the equation, likelihood ratio could also be written as

$$\frac{\pi(X \mid \theta = \theta_0)}{\pi(X \mid \theta = \theta_1)} = \frac{\pi(\theta = \theta_0 \mid X)}{\pi(\theta = \theta_1 \mid X)} \frac{\pi(\theta = \theta_1)}{\pi(\theta = \theta_0)} = BF.$$

With a composite hypothesis where $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ we have a probability density of θ so that the probability for any single value of θ is zero. There is positive probability only for the set Θ_0, (and Θ_1). In a *one sided hypothesis test*: $H_0 : \theta < \theta_0$ against $H_1 : \theta \geq \theta_0$. The Bayes factor is then more complicated, and depends on prior:

$$BF = \underbrace{\frac{\int_{\Theta_0} \pi(\theta \mid X)\mathbf{d}\theta \int_{\Theta_1} \pi(\theta)\mathbf{d}\theta}{\int_{\Theta_1} \pi(\theta \mid X)\mathbf{d}\theta \int_{\Theta_0} \pi(\theta)\mathbf{d}\theta}}_{\text{post.odds}\times\text{prior odds}^{-1}} = \frac{\int_{\Theta_0} \pi(X \mid \theta)\pi(\theta)\mathbf{d}\theta}{\int_{\Theta_1} \pi(X \mid \theta)\pi(\theta)\mathbf{d}\theta} \frac{\int_{\Theta_1} \pi(\theta)\mathbf{d}\theta}{\int_{\Theta_0} \pi(\theta)\mathbf{d}\theta}$$

For a continuous parameter θ with a density function, *a two sided hypothesis* $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ does not make sense unless we place positive prior probability for the point θ_0 so that $q_0 = P(\theta = \theta_0) > 0$. With probability $q_1 = 1 - q_0 = P(\theta \neq \theta_0) > 0$. The prior is thus a mixture

$$\pi(\theta) = 1_{\{\theta=\theta_0\}}(\theta)q_0 + \pi_1(\theta)q_1$$

where a density function $\pi_1(\theta)$ is applied under H_1. This could be seen as a model choice problem where we have two competing models, M_0 and M_1 with prior probabilities for the models: q_0 and $q_1 = 1 - q_0$. Then:

$$\underbrace{\frac{P(\text{model} = M_0 \mid X)}{P(\text{model} = M_1 \mid X)}}_{\text{Post.odds}} = \frac{q_0}{q_1} \underbrace{\frac{\int \pi_0(X \mid \theta_0)\pi_0(\theta_0)\mathbf{d}\theta_0}{\int \pi_1(X \mid \theta_1)\pi_1(\theta_1)\mathbf{d}\theta_1}}_{BF}$$

where the probability of data X is either based on model M_0 as $\pi_0(X \mid \theta_0)$ or on model M_1 as $\pi_1(X \mid \theta_1)$. Parameters θ_0 and θ_1 in each model could have different dimensions. One model could be a three-parameter model, and the other a one-parameter model...

4.1 Example: evidence for population prevalence

If the hypothesis with a binomial model $\mathrm{Bin}(N, r)$ is that the large population prevalence $r < 0.5$, then the prior probability of that hypothesis is

$$P(H_0) = P(r < 0.5) = \int_0^{0.5} \pi(r)\mathbf{d}r = 0.5 \qquad \text{(from U(0,1)-prior)}$$

but the posterior probability, with data $Y = 2, N = 3$, would be

$$P(H_0 \mid Y, N) = P(r < 0.5 \mid Y, N) = \int_0^{0.5} \mathrm{Beta}(r \mid Y + 1, N - Y + 1)\mathbf{d}r$$

which is the cumulative probability of the beta-density at $r = 0.5$. The approximate value (0.3125) is obtained in R by typing `pbeta(0.5,Y+1,N-Y+1)`. The posterior probability became smaller than the prior probability.

We may also compute posterior odds to compare the difference. 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

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

Therefore, posterior odds became smaller than prior odds, i.e. there was some evidence against the hypothesis. Jeffreys (1961) suggested that a Bayes factor bigger than 10 means strong evidence for the hypothesis, whereas a Bayes factor smaller than 1/10 means strong evidence against it.

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, at least analytically. With iterative sampling methods, such posterior probabilities are routinely computed for applied problems.

4.2 Example: analysis of birth data

Example from Gelman [5]: 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

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

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

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

which seems quite high. The posterior odds is 0.992826/(1-0.992826) = 138.4, and the prior odds $0.485/(1 - 0.485) = 0.94$, giving a Bayes factor of about 147.

This result was obtained when the prior was uniform. This was somewhat against our actual prior knowledge. 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 the last case, the prior sample size equals already 200, and the prior is Beta(97, 103). From this we get prior odds of about 1.00. The posterior odds are about 78, so the Bayes factor is about 78. This is still large, but lower than 147. The choice of prior has an effect.

In addition to r, an interesting quantity is the sex ratio $z = (1-r)/r$ and some research questions could be framed about it. Distribution of z could be found using the transformation of variables technique. In practice, it is easier to produce it by simulation techniques.

4.3 Example: 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 $P(B \mid \text{'Monty shows empty'})$

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

The posterior odds for Monty having the price is 2, compared to prior odds of 2, so that the revealing of an empty box is not giving information, other than simply reducing the number of Monty's boxes by one empty box - which he always can do. But let's change the rules! Assume then that Monty is allowed to choose randomly (blindfolded) which one of his boxes he opens. (This could result into Monty's error of showing the price). 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:

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

Posterior odds for Monty winning is then 1, compared to prior odds of 2, so the revealing of empty box makes the odds for Monty having the price smaller by a factor of 2. (Bayes factor 0.5). We really need to know how the game is played!

4.3.1 Fair coin or not

In 200 tosses of a coin, 115 were heads, 85 tails. The null hypothesis is to assume a fair coin: $H_0 : p = 0.5$. Alternatively, it is something else, $H_1 : p \neq 0.5$ in which case we apply a uniform prior distribution $U(0, 1)$. This can be seen as a model choice problem, where each hypothesis corresponds to a different model, M_0 and M_1. The Bayes factor is the posterior odds divided by prior odds

$$BF = \frac{P(M_0 \mid X)/P(M_1 \mid X)}{P(M_0)/P(M_1)} = \frac{P(M_0)P(X \mid M_0)/(P(M_1)P(X \mid M_0))}{P(M_0)/P(M_1)} = \frac{P(X \mid M_0)}{P(X \mid M_1)}$$

similarly to the case of simple (point) hypothesis (the point hypothesis is now the model). Now the probability of the data X under each model needs to be computed, to get:

$$BF = \frac{P(X \mid p = 0.5)}{\int_0^1 P(X \mid p)\pi(p \mid M_1)\mathbf{d}p}$$

The probability of data X under H_0 (model M_0) and under H_1 (model M_1) is

$$P(115 \mid M_0) = \binom{200}{115} 0.5^{200} = 0.005956$$

$$P(115 \mid M_1) = \int_0^1 \binom{200}{115} p^{115}(1-p)^{85}\mathbf{d}p = \frac{1}{201} = 0.004975$$

This gives $BF = 1.197$.

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.