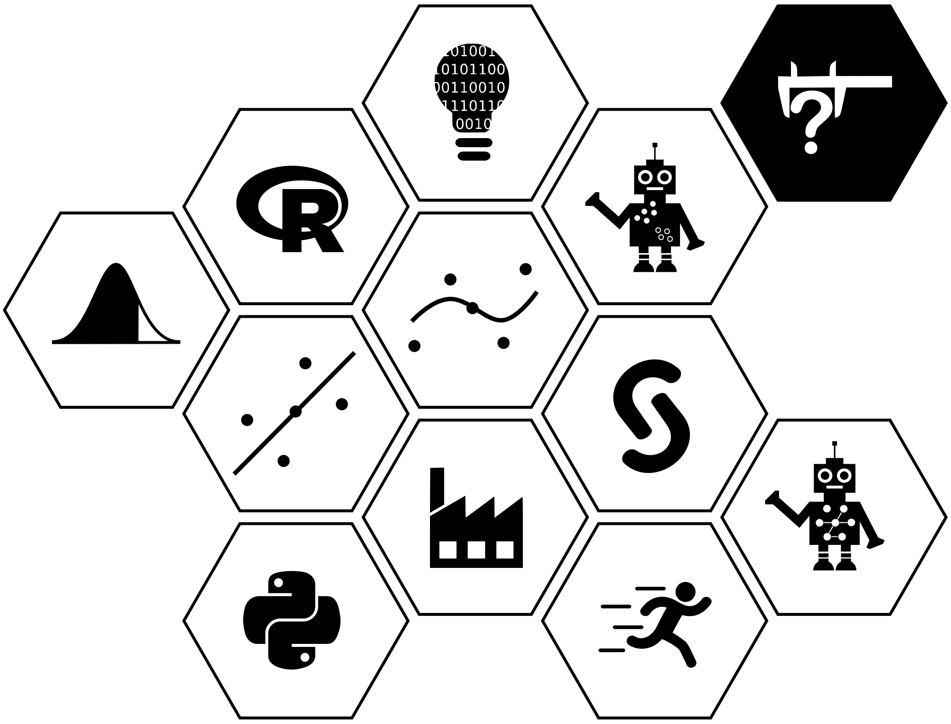
Uncertainty Assessment and Bayesian Computation

Vlad Vyshemirsky

Academic Year 2021-22

Week 2:

Inference, Prediction, Hypotheses Testing. Binomial Model. Conjugate Priors.



Inference

UABC - Inference

https://youtu.be/O4oZW2VISEY

Duration: 34m16s

In the previous part we covered that in the Bayesian framework we use probability to quantify belief. Here we will discuss how beliefs are updated when observing new evidence.

Our beliefs change as we observe new evidence. This is the essence of the process of learning. Formally, the process of learning is described by the concept of Inference — updating probability distributions corresponding to our beliefs in light of new data.

Consider the problem of establishing the age of planet Earth. Mere 300 years ago the majority of people believed that Earth is about 6000 years old. In the middle of the 18th century geologists considered strata, the layering of rocks and earth, that suggested that Earth has likely gone through several periods of development and must be much older than previously assumed. In 1779 the Comte du Buffon tried to obtain a value for the age of Earth using an experiment: He created a small globe that resembled Earth in composition and then measured its rate of cooling. This led him to estimate that Earth was about 75,000 years old. In 1862, the prominent physicist William Thomson, 1st Baron Kelvin, and a professor at the University of Glasgow, published calculations that suggested the age of Earth at between 20 million and 400 million years. His calculations assumed that Earth started off as a molten ball and similarly relied on some cooling down calculations. Few people in the general public accepted these arguments. While critics in the scientific world were doubting these results in both directions — some considering a hypothesis of a younger Earth, while others suggesting an older one. In 1895, John Perry produced an age-of-Earth estimate of 2 to 3 billion years using a model of a convective mantle and thin crust. Discovery of radioactivity invalidated all previous estimates based on the rates of cooling down, as heat produced via radioactive decay can replenish the energy lost through cooling down. Radioactivity, which had overthrown the old calculations, yielded a bonus by providing a basis for new calculations, in the form of radiometric dating. Current belief is that Earth and the rest of the Solar System formed at around 4.53 to 4.58 billion years ago. This belief is also accepted by the majority of the general public.

This story emphasises an important requirement for any framework of working with beliefs - it must allow changing the belief in the light of new evidence. Bayesian Statistics achieves this by operating upon conditional distributions.

First of all, our scientific hypotheses are expressed with parametric statistical models. These models impose **likelihoods** of a form $p(D|\theta)$ where D is data that can be observed, while θ is the parameter of the model that answers the question of interest. The likelihood defines the probability of observing the data D when model parameters are fixed at certain values.

We describe our initial belief (before observing any data) with a distribution of model parameters called **the prior distribution** $p(\theta)$.

After observing a particular data set D (our evidence), we will update our belief about θ to **the posterior distribution** $p(\theta|D)$.

This update is performed using Bayes' theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

The most difficult part is evaluating the integral in the denominator. Only in simplest cases this integral can be evaluated directly. Consider the first example, where such evaluation in performed analytically.

Example 1 (Tossing a Coin).

We begin with the simplest example of performing statistical inference. Imagine, we want to establish the probability of tossing heads with some given coin. We do not know if the coin is fair, and therefore we cannot assume any preference for this probability.

We begin with establishing a model for our planned experiments. We will be tossing the coin repeatedly, and observing what proportion of the tosses results in heads. This is a classical example for the binomial model. The binomial distribution is parametrised with the probability of success θ, in our case it is the probability of tossing heads. The number of successes in n independent tosses is measured using the following probability mass function:

$$p(D|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k},$$

where $D = (k, n)$ is the data that corresponds to observing k heads among n consecutive tosses.

Next, we define the prior distribution of θ that describes our belief in plausible values of θ before we perform any tosses. We know that θ is the probability of tossing heads, and therefore must be between zero and one. As we don't know if the coin is fair or not, we cannot prefer a particular value of θ to any other value. Therefore we will be using a uniform distribution from zero to one to express our initial belief in possible values of θ. This distribution is depicted in Figure 1.

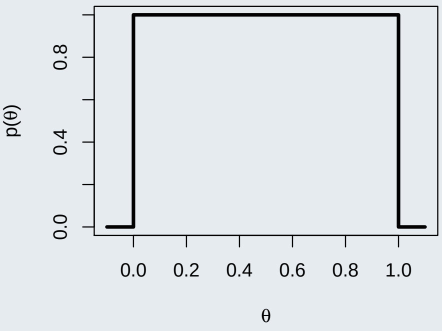


Figure 1: The prior distribution for θ.

Now we can toss the coin once, and observe the result. Assume this is a head. Our dataset in this case is $D = (1, 1)$ — one success in one trial.

We can now calculate the posterior probability of θ:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

As our $k = 1, n = 1$, then

$$\int p(D|\theta)p(\theta)d\theta = \int_0^1 \theta d\theta = \frac{1}{2},$$

and the posterior probability density function will be:

$$p(\theta|D) = \begin{cases} \dfrac{1\cdot\theta^1\cdot(1-\theta)^0\cdot 1}{\frac{1}{2}}, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 2\theta, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

This posterior probability density is depicted in Figure 2. Note the difference between Figures 1 and 2 — we went from believing that any value of θ has the same probability to believing that larger θ are more probable than the smaller ones, with the most likely value being 1. We still preserve some possibility for θ to be less than 1, because observing one toss of a coin provides only limited amount of information about the problem. Compare this result to the maximum likelihood estimation performed in classical statistics, where the conclusion would have been made that $\theta = 1$ with no other options, and any further data would have to completely invalidate this conclusion.

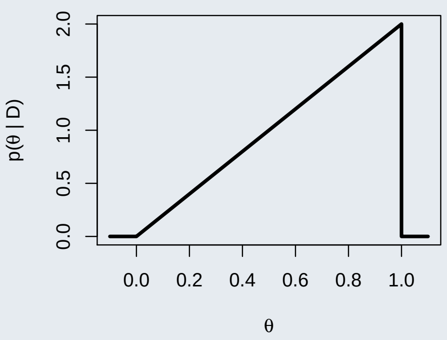


Figure 2: The posterior distribution $p(\theta|D)$ when $D = (1, 1)$ corresponding to one head in one toss.

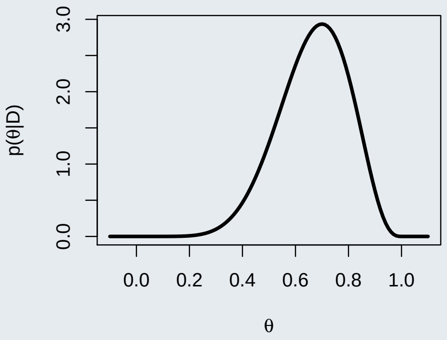


Figure 3: The posterior distribution $p(\theta|D)$ when $D = (7, 10)$ corresponding to seven heads in ten tosses.

Now, consider a case where we tossed the coin ten times, and observed seven heads. In this case $D = (7, 10)$.

First, we need the integral:

$$\int p(D|\theta)p(\theta)d\theta = \int_0^1 \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}d\theta = 120\int_0^1 \theta^7(1-\theta)^3 d\theta = \frac{1}{11}.$$

Now, the posterior density is going to be:

$$p(\theta|D) = \begin{cases} \frac{120\cdot\theta^7\cdot(1-\theta)^3\cdot 1}{\frac{1}{11}}, & \text{for } 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1320\,\theta^7(1-\theta)^3, & \text{for } 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

which is plotted in Figure 3.

We can continue the experiments. In this case, the information about θ that we learned in the first experiment, and that is expressed as the distribution in Figure 3, will be the prior for the next experiment. Assume, we toss the coin 10 more times, and observe 6 more heads, so the second data set is $D_2 = (6, 10)$. We can now calculate the posterior for both of the data sets:

$$P(\theta|D, D_2) = \frac{p(D_2|\theta)p(\theta|D)}{\int p(D_2|\theta)p(\theta|D)d\theta}$$

Note, that $p(\theta|D)$ from Figure 3 is now used as a prior in this new calculation.

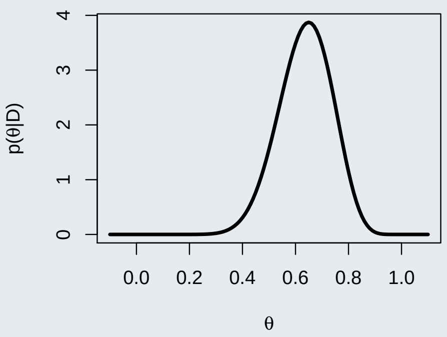


Figure 4: The posterior distribution $p(\theta|D, D_2)$ when $D = (7, 10)$ and $D_2 = (6, 10)$.

So, the integral in the denominator becomes:

$$\int p(D_2|\theta)p(\theta|D)d\theta = 277200\int_0^1 \theta^{13}(1-\theta)^7 d\theta = 277200\,B(14, 8) = \frac{277200\cdot 13!\cdot 7!}{21!} = \frac{55}{323}$$

Where $B(\alpha, \beta)$ is the beta function which is by definition the solution to that integral. The resulting posterior after observing both data sets:

$$P(\theta|D, D_2) = \begin{cases} \frac{210\cdot\theta^6\cdot(1-\theta)^4\cdot 1320\cdot\theta^7\cdot(1-\theta)^3}{\frac{55}{323}}, & \text{for } 0 \le \theta \le 1 \\ o, & \text{otherwise} \end{cases} = \begin{cases} 1627920\,\theta^{13}(1-\theta)^7, & \text{for } 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

which is now plotted in Figure 4.

Notice, that the difficult integral in the denominator always resolves to be a real number, as it acts as a scaling constant to make the posterior a proper distribution, i.e. ensuring that

$$\int p(\theta|D)d\theta = 1.$$

In low dimensional problems, when we are performing inference over one or two parameters, we can get away with evaluating the unscaled posterior density (the numerator only) over a fine grid over model parameters, and then renormalising the resulting distribution to integrate to 1. This "trick" is demonstrated in our second example.

Example 2 (Weight of a Cat).

Imagine that we are trying to establish how heavy adult domestic cats are on average.

First, we need to formulate a hypothesis about possible observations of cat weights.

A cat's weight depends on the breed of the cat, on the cat's diet, on the cat's age, some genetic traits, and so on. It is fair to assume that there is a very large number of possible factors that impact the cat's weight. By the virtue of the *central limit theorem*, we can therefore assume that the observations for the weight of adult cats will be normally distributed. Let D be the weight of a randomly selected cat, μ will be the average weight of an adult cat (that we are most interested to find), and σ^2 will be the variance of the normal distribution used in our model. In this case the *likelihood* for our model is going to be

$$p(D|\mu,\sigma^2) = \mathcal{N}_D(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(D-\mu)^2}{2\sigma^2}\right\}$$

This function defines the probability of observing D under the assumption of certain μ and σ^2.

Next, we need to define the prior for μ and σ^2 to reflect our knowledge about these before weighting any cats. We can be certain, that the average weight of a cat μ is strictly positive. Similarly, we can be quite sure that on average domestic cats are not bigger than elephants or even humans, so we can safely assume some upper limit on the weight of a cat. A quick google search shows that the heaviest cat on record weighted about 47 *lb*. So, 50 *lb* is safely greater than the weight of any cat we can observe. Finally, let's assume that we do not prefer any value within the proposed range and therefore we will use the uniform distribution from 0 *lb* to 50 *lb* as the prior for parameter μ. Parameter σ^2 is the variance for the normal population of cat weights. σ^2 is again strictly positive, and 99.7% of the observations from a normal distribution will be within 6 standard deviations σ of this distribution. So, we can put a limit on σ to be $50/6 = 8.33\,lb$, and therefore the upper limit for σ^2 will be 69.4. As we have no reasons to justify any dependency between μ and σ^2 before observing experimental data, we will keep these priors independent:

$$\begin{aligned} p(\mu,\sigma^2) = p(\mu)p(\sigma^2) &= \begin{cases} \frac{1}{50} \times \frac{1}{69.4}, & 0 < \mu < 50, 0 < \sigma^2 < 69.4 \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{3470}, & 0 < \mu < 50, 0 < \sigma^2 < 69.4 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

We plot the heatmap for this prior distribution $p(\mu,\sigma^2)$ in Figure 5.

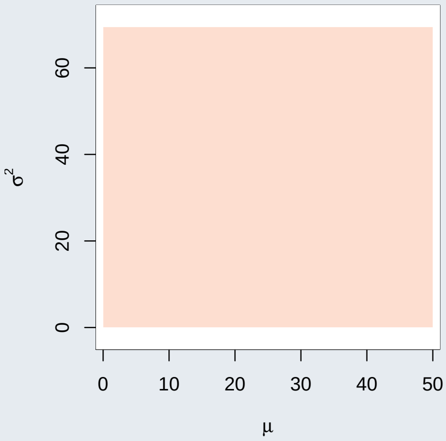


Figure 5: *The prior distribution* $p(\mu,\sigma^2)$.

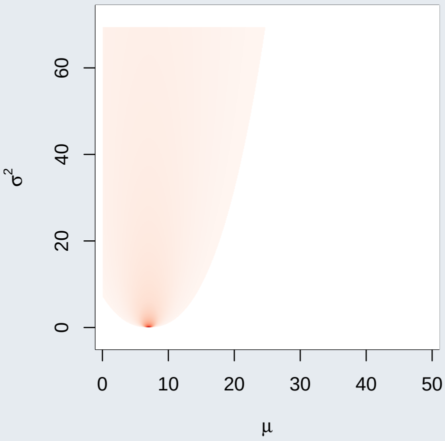


Figure 6: *The posterior distribution* $p(\mu,\sigma^2|D)$ *after observing* $D = 7$.

Now we proceed to collecting data. Imagine, we measured a cat with weight of 7 *lb*, therefore we fix $D = 7$. We can calculate *the posterior* probability of μ and σ^2 as the following:

$$p(\mu, \sigma^2|D) = \frac{p(D|\mu, \sigma^2)p(\mu, \sigma^2)}{p(D)} = \frac{p(D|\mu, \sigma^2)p(\mu, \sigma^2)}{\iint p(D|\mu, \sigma^2)p(\mu, \sigma^2)d\mu d\sigma^2}$$

This is the distribution of the belief for plausible values of μ and σ^2 after observing data D. This posterior distribution density was evaluated on a fine grid, and plotted in Figure 6. This is done using the following R code:

```
mugrid <- seq(-1,51,0.1)
sigma2grid <- seq (-5, 74.4, 0.1)
```

these two lines created a grid over μ and σ^2 with a step of 0.1. The following line creates an empty matrix to evaluate unscaled posterior density at every combination of μ and σ^2 on the grid.

```
posterior <- matrix(0,nrow=length(mugrid),ncol=length(sigma2grid))
```

Now, we will go through every element of this matrix, and evaluate the numerator of the posterior for that point on the grid using the following double loop in R:

```
for (i in 1:nrow(posterior)) {
    for (j in 1:ncol(posterior)) {
        if ((mugrid[i] > 0) & (mugrid[i] < 50) &
            (sigma2grid[j] > 0) & (sigma2grid[j] < 69.4)) {
            posterior[i,j] <- 1/3470*dnorm(7,mugrid[i],sqrt(sigma2grid[j]))
        }
    }
}
```

The points on the grid where the prior had probability zero will stay zero in the posterior. Now we need to ensure that the posterior integrates to 1. As the step on the grid was 0.1 in both dimensions, we need to make sure that the elements of the resulting matrix add up to 100 (as the integral over the whole grid must be 1, and $100 \cdot 0.1 \cdot 0.1 = 1$). This can be done the following way:

```
posterior <- posterior/sum(posterior) * 100
```

Final matrix `posterior` can now be plotted to produce the result in Figure 6 using

```
require(RColorBrewer)
image(mugrid,sigma2grid,posterior,useRaster=T,
      col=c("#FFFFFF",colorRampPalette(brewer.pal(9, "Reds"))(250)))
```

We observe in Figure 6 that the posterior distribution is significantly different from the prior in Figure 5; it is now concentrated around $\mu = 7$. As we are not interested in working out the variance of the sampling distribution, and focus on finding the average weight of an adult domestic cat, we should look at the marginal posterior distribution of μ alone: $p(\mu|D)$, where σ^2 has been integrated out. This distribution is shown with a solid line in Figure 7. This can be easily obtained with

```
mumarginal <- rowSums(posterior)*0.1
plot(mugrid,mumarginal,type="l")
```

These lines perform simple numerical integration by computing the corresponding Riemann sums on the grid for σ^2.

Comparing this posterior to the prior for μ, as depicted with the dashed line in Figure 7, we observe that our belief in the likely values for the average weight of a cat changes significantly as we observe experimental data. The divergence between the prior and the posterior corresponds to the amount of information learned from the data.

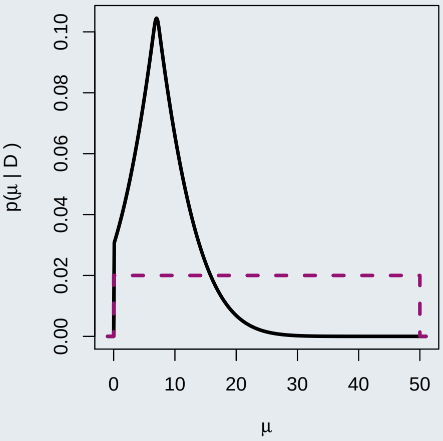


Figure 7: Solid line is the marginal posterior distribution $p(\mu|D)$ after observing $D = 7$ and integrating out σ^2. The dashed line is our prior $p(\mu)$.

Other approaches to calculating the integral in the denominator of the posterior expression include numerical integration, Monte-Carlo integration, sampling using Markov Chains that avoid calculating the integral altogether, or using a special form of the prior that simplifies the expression. We will consider most of these approaches later in the course.

Prediction

University of Glasgow

Prediction

Uncertainty Assessment a

Bayesian Computation

DATA ANALYTICS GLASGOW

UABC - Prediction

https://youtu.be/ZqQLbOZY8EA

Duration: 19m39s

Now we know how to learn from data by performing Bayesian inference and updating our prior beliefs to the posterior beliefs. A natural next step in statistical analysis is to be able to predict future observations given our current understanding of the model.

For example, if you created a novel statistical model of the stock market, the most useful application of this model will be to predict how the market will perform in the future, as these predictions will help guiding trading decisions.

Even before we observe any data, we can start predicting experiment outcomes using the information in our prior. To do that we need to evaluate the probability of the future observation $\tilde{y}$ given that the model parameters are coming from the prior:

$$p(\tilde{y}) = \int p(\tilde{y}, \theta)d\theta = \int p(\tilde{y}|\theta)p(\theta)d\theta$$

$p(\tilde{y})$ is called *the prior predictive distribution.*

Consider an example of predicting a coin toss outcome without knowing anything about the coin.

Example 3 (Prior Prediction of a Coin Toss).

Before we perform any experiments with the coin, we assume that the probability of heads θ can be anything between zero and one with equal probability, and the corresponding prior distribution is

$$p(\theta) = \begin{cases} 1, & \text{for } 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

as depicted in Figure 1.

For a fixed value of θ, the probability of tossing head is

$$p(\tilde{y} = \text{head}|\theta) = \theta$$

and the probability of tossing tail is

$$p(\tilde{y} = \text{tail}|\theta) = 1 - \theta$$

We can combine this with the prior for θ to predict the probability of tossing heads using only this prior as the source of information:

$$p(\tilde{y} = \text{head}) = \int_0^1 p(\tilde{y} = \text{head}|\theta)p(\theta)d\theta = \int_0^1 \theta d\theta = \left[\frac{\theta^2}{2}\right]_{\theta=0}^1 = \frac{1}{2} - 0 = \frac{1}{2}.$$

Similarly,

$$p(\tilde{y} = \text{tail}) = \int_0^1 p(\tilde{y} = \text{tail}|\theta)p(\theta)d\theta = \int_0^1 (1-\theta)d\theta = \left[\theta - \frac{\theta^2}{2}\right]_{\theta=0}^1 = 1 - \frac{1}{2} - 0 + 0 = \frac{1}{2}.$$

By redefining $p(\tilde{y}|\theta)$, we can make a prediction for the number of heads $\tilde{y}$ among an arbitrary number of tosses $\tilde{n}$ using the prior information. First, for a fixed θ and fixed $\tilde{n}$,

$$p(\tilde{y}|\theta) = \binom{\tilde{n}}{\tilde{y}}\theta^{\tilde{y}}(1-\theta)^{\tilde{n}-\tilde{y}}, \quad \tilde{y} = 0, 1, 2, \ldots, \tilde{n}$$

and combining this with the prior:

$$p(\tilde{y}) = \int_0^1 \binom{\tilde{n}}{\tilde{y}}\theta^{\tilde{y}}(1-\theta)^{\tilde{n}-\tilde{y}}d\theta = \binom{\tilde{n}}{\tilde{y}}B(\tilde{y}+1, \tilde{n}-\tilde{y}+1)$$

here $B(x, y)$ is the beta function, and since

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

where $\Gamma(x)$ is the gamma function, we can write

$$p(\tilde{y}) = \binom{\tilde{n}}{\tilde{y}}B(\tilde{y}+1, \tilde{n}-\tilde{y}+1) = \frac{\tilde{n}!}{\tilde{y}!(\tilde{n}-\tilde{y})!} \times \frac{\tilde{y}!(\tilde{n}-\tilde{y})!}{(\tilde{n}+1)!} = \frac{1}{\tilde{n}+1},$$

predicting all outcomes to be equiprobable, as expected from our intentions in using a flat prior.

Making predictions from the prior information may be useful if we want to verify that our hypothesis predicts correct outcomes. However, we are usually more interested in making predictions from the updated state of information - the posterior.

In this case, *the posterior predictive distribution* is

$$p(\tilde{y}|D) = \int p(\tilde{y}|\theta)p(\theta|D)d\theta$$

Consider an example of predicting a coin toss outcome based on the information learned in Example 1.

Example 4 (Posterior Prediction of a Coin Toss).

Let's consider again the probability of tossing heads, but this time we will be using the coin that we studied in Example 1. After tossing the coin 20 times, we learned that θ is distributed as depicted in Figure 4, and

$$p(\theta|D) = \begin{cases} 1627920\,\theta^{13}(1-\theta)^7, & \text{for } 0 \leq \theta \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

The probability of tossing head on the next turn is

$$p(\tilde{y} = \text{head}|\theta, D) = p(\tilde{y} = \text{head}|\theta) = \theta$$

as the coin has no memory, and all the tosses are independent.

Observing head on the 21^{st} toss now has probability:

$$p(\tilde{y}|D) = \int_0^1 \theta p(\theta|D)d\theta = \mathbb{E}\left[\theta|D\right] = \frac{14}{22}$$

Deriving a posterior prediction for the number of heads in n new tosses is a homework task for this chapter.

This example is often generalised as a famous problem of establishing whether the Sun will rise tomorrow.

Example 5 (Will the Sun Rise Tomorrow?).

Pierre-Simon Laplace introduced this example to demonstrate how inference and prediction work together. The example evaluates the probability of a statement *"The Sun will rise tomorrow"* within the Bayesian framework.

Let θ be the probability of the sunrise on any given day. Hypothetically, before observing any sunrises, we have no preference for the value of θ, expressed with the uniform prior distribution:

$$p(\theta) = \begin{cases} 1, & \text{for } 0 \leq \theta \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Given the value of θ, and no other information relevant to the question of whether the sun will rise tomorrow, the probability that the sun will rise tomorrow is $p(\tilde{y} = \text{sunrise}|\theta) = \theta$. But we don't know the value of θ. What we are given is the observed data: the sun has risen every day on record. Laplace inferred the number of days by saying that the universe was created about 6000 years ago, based on a young-earth creationist reading of the Bible.

By performing the inference for θ, and then performing the prediction for a new sunrise tomorrow, we obtain in general form:

$$p(\text{“the sun will rise tomorrow”}|\text{“it has risen every time on } k \text{ days previously”}) =$$

$$p(\tilde{y} = \text{sunrise}|\theta, D = (k,k)) = \int_0^1 \theta p(\theta|D = (k,k))d\theta = \mathbb{E}\left[\theta|D = (k,k)\right] = \frac{k+1}{k+2}.$$

as

$$p(\theta|D = (k,k)) = \frac{p(D = (k,k)|\theta)p(\theta)}{\int_0^1 p(D = (k,k)|\theta)p(\theta)d\theta} = \frac{\theta^k}{\int_0^1 \theta^k d\theta} = (k+1)\theta^k$$

and consequently

$$\int_0^1 \theta p(\theta|D = (k,k))d\theta = (k+1)\int_0^1 \theta^{k+1}d\theta = \frac{k+1}{k+2}.$$

If someone has observed the sun rising 10000 times previously, the probability it rises the next day is $10001/10002 \approx 0.99990002$. Expressed as a percentage, this is approximately a 99.990002% chance.

J.M. Keynes remarked in his book *"A Treatise on Probability"*, 1921, p. 82:

> "No other formula in the alchemy of logic has exerted more astonishing powers. For it has established the existence of God from the premiss of total ignorance; and it has measured with numerical precision the probability that the sun will rise to-morrow."

When the integrals involved in performing predictions become more complex, we can employ numerical integration. Let's see how that works with predicting the weight of another cat after considering Example 2.

Example 6 (Posterior Prediction of Cat's Weight).

In Example 2 we started with a uniform prior for the average weight of an adult domestic cat, and after measuring a cat weighting 7 *lb*, we arrived to the joint posterior distribution of the average weight of a cat and the variance of the population of cat weights depicted in Figure 6.

Let's predict what will be the weight of another cat that we can measure. As we assumed that the population of cat weights is normal,

$$p(\tilde{y}|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\tilde{y}-\mu)^2}{2\sigma^2}\right\}$$

Now the posterior predictive distribution of the weight of a new cat will be:

$$p(\tilde{y}|D) = \iint_0^\infty p(\tilde{y}|\mu,\sigma^2)p(\mu,\sigma^2|D)d\mu d\sigma^2 =$$

$$\iint_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\tilde{y}-\mu)^2}{2\sigma^2}\right\} p(\mu,\sigma^2|D)d\mu d\sigma^2$$

Our posterior $p(\mu,\sigma^2|D)$ was evaluated on a fine grid, and re-normalised. The result of that evaluation in R is stored in matrix `posterior`.
Now, we will introduce another grid for the possible weights of a new cat:

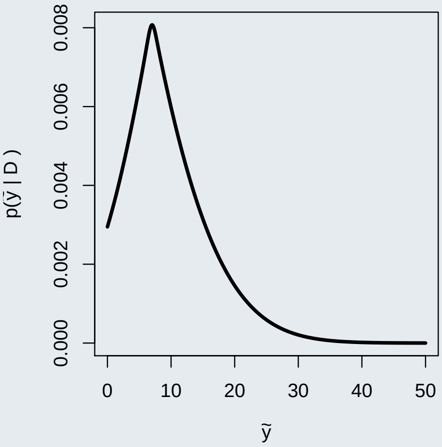


Figure 8: *The posterior predictive distribution* $p(\tilde{y}|D)$.

```
ygrid <- seq(0,50,0.1)
```

and for every value on this new `ygrid`, we will numerically integrate the posterior predictive distribution, and re-normalise the result:

```
ydensity <- rep(0,length(ygrid))
predictive <- function(mu,sigma2,weight=0)
  {ifelse(sigma2>0,dnorm(weight,mu,sqrt(sigma2)),0)}
for (i in 1:length(ygrid)) {
    pointprediction <- outer(mugrid,sigma2grid,predictive,weight=ygrid[i])
    ydensity[i] <- sum(pointprediction*posterior)
}
ydensity <- ydensity/sum(ydensity)
plot(ygrid,ydensity,type="l")
```

This evaluation is quite slow, due to the need to handle a grid approximation over three dimensions now. The plot in Figure 8 is produced as the result.

Another approximation can be achieved when a sample from the posterior distribution is available (or a sample from the prior for the prior predictive distribution). We will demonstrate this by replicating Example 4 using samples instead of analytical derivation.

Example 7 (Sample Based Posterior Prediction of a Coin Toss).

Here we replicate Example 4 using samples from the posterior, and approximating the posterior predictive distribution using these samples.

After tossing the coin 20 times, we learned that θ is distributed as

$$p(\theta|D) = \begin{cases} 1627920\,\theta^{13}(1-\theta)^7, & \text{for } 0 \le \theta \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

which is the probability density function of a beta distribution $Be(14, 8)$.

First, we can draw a large sample from this distribution:

```
postsample <- rbeta(1000,14,8)
```

Next, we will take every value in this sample, and use it as the value of θ for the probability of tossing head on the next turn:

$$p(\tilde{y} = \text{head}|\theta, D) = p(\tilde{y} = \text{head}|\theta) = \theta.$$

```
predsample <- rbinom(1000,postsample,size=1)
```

Values in `predsample` will be either 1 for head or 0 for tail, and therefore the posterior predictive probability for tossing head on the next turn will be the proportion of ones in `predsample`

```
mean(predsample)
```

```
[1] 0.637
```

Compare this to the true posterior predictive probability calculated in Example 4 as

$$p(\tilde{y}|D) = \frac{14}{22} \approx 0.6363636$$

```
# Absolute error estimation:
mean(predsample) - 14/22
```

```
[1] 0.0006363636
```

Using a larger sample helps to reduce the approximation error:

```
postsample <- rbeta(10000,14,8)
predsample <- rbinom(10000,postsample,size=1)
mean(predsample)
```

```
[1] 0.6365
```

```
# Absolute error estimation:
mean(predsample) - 14/22
```

```
[1] 0.0001363636
```

Hypotheses Testing

UABC - Hypotheses Testing

https://youtu.be/YhFzLtE6CPM

Duration: 10m21s

The next milestone in statistical analysis is performing hypotheses testing. Situations when several alternative explanations for a certain phenomenon are considered are very common. Selecting the most appropriate explanation is done by performing some experiments, collecting data, and then using the evidence of these data to find which hypothesis is more plausible than the rest.

Hypotheses testing in Bayesian framework is performed very differently to the approaches taken in classical statistics.

Every hypothesis will be represented with a separate statistical model $\mathcal{M}_1, \mathcal{M}_2, \ldots$; these models will have their own prior probabilities $p(\mathcal{M}_1), p(\mathcal{M}_2), \ldots$; and model selection will be performed by evaluating corresponding model posterior probabilities $p(\mathcal{M}_i|D)$.

Let's first consider a simple case of having just two alternative hypotheses, and therefore considering only two alternative models $\mathcal{M}_1$ and $\mathcal{M}_2$.

We will be looking for the odds of model $\mathcal{M}_1$ to be better than model $\mathcal{M}_2$ *a posteriori*:

$$\frac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_2|D)}$$

given their prior odds:

$$\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}.$$

By applying Bayes theorem, we easily obtain

$$\frac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_2|D)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \times \frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_2)}$$

which means that the posterior odds are equal to the prior odds multiplied by the model likelihood ratio. This model likelihood ratio is usually known as *the Bayes factor*. The elements in the numerator and the denominator of the Bayes factor are called the marginal likelihoods for the two alternative models, and can be obtained by marginalising out model parameters:

$$p(D|\mathcal{M}_i) = \int p(D|\mathcal{M}_i, \theta_i)p(\theta_i|\mathcal{M}_i)d\theta_i$$

Notice, that it is exactly the same integral as the one that is used for normalising the posterior of model parameters. The only difference is notational — here we made an explicit reference to a specific model.

In practice the alternative hypotheses are frequently considered to be equiprobable *a priori*, and therefore frequently

$$\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} = 1.$$

Posterior odds of the models are self explanatory — if the odds are, for example, 2:1, that means that the first model is twice as likely as the second one. If the odds are, for example, 1 000 000:1, that means that the first model is a million times as likely as the second one.

Sir Harold Jeffreys proposed an informal scale for assigning rough descriptive statements about standards of evidence in scientific investigation to the values of the Bayes factor. We consider a slightly modified version of his scale proposed by *Kass and Raftery (1995)*, here $\mathcal{B}$ is used for the value of the Bayes factor:

$2\ln(\mathcal{B})$	$\mathcal{B}$	Evidence Support
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

These categories, however, are not a calibration for the Bayes factor, as it already provides a meaningful interpretation as probability, but they are useful for informal descriptions.

Let's consider how this simple setup of comparing just two alternative hypotheses works on a simple example.

Example 8 (Gender Bias in Smoking).

The R package `MASS` comes with a data set `survey` containing the responses of 237 Statistics I students at the University of Adelaide to a number of questions.

We will consider answers to two of the questions:

- Student's sex, in this survey "male" or "female"
- Student's smoking habits, in this survey "never", "occasionally", "regularly", "heavy"

We will investigate whether the probability of smoking is the same or not the same among the two sexes.

First we will perform some preliminary processing of the data, to summarise smoking habits in two categories `smoker` and `nonsmoker`, and then generate a contingency table with the sex of the students:

```
require(MASS)
```

```
Loading required package: MASS
```

```r
sex <- survey$Sex
smoking <- factor(as.numeric(survey$Smoke == "Never"),
                  labels=c("smoker","nonsmoker"))
table(sex,smoking)
```

```
        smoking
sex      smoker nonsmoker
  Female     19        99
  Male       28        89
```

We obtain the data stating that $S_f = 19$ out of $N_f = 118$ females smoke, and $S_m = 28$ out of $N_f = 117$ males smoke. Two missing values have been omitted.

We will consider two hypotheses:

$\mathcal{H}_1$: Probability of being a smoker is the same in male and female populations

$\mathcal{H}_2$: Probability of being a smoker is different in male and female populations

We will formalise these hypotheses with two statistical models:

$\mathcal{M}_1$: $S_f \sim Bi(\theta; N_f),\ S_m \sim Bi(\theta; N_m)$

$\mathcal{M}_2$: $S_f \sim Bi(\theta_f; N_f),\ S_m \sim Bi(\theta_m; N_m)$

Note, that probability of smoking is the same in the first model, and may be different in the second one. We assign uniform priors between 0 and 1 to all of the smoking probabilities: $\theta, \theta_f, \theta_m$.

Likelihood for the first model is

$$p(S_f, S_m, N_f, N_m|\mathcal{M}_1, \theta) = \binom{N_f}{S_f}\theta^{S_f}(1-\theta)^{N_f-S_f} \cdot \binom{N_m}{S_m}\theta^{S_m}(1-\theta)^{N_m-S_m}$$

and for the second model is

$$p(S_f, S_m, N_f, N_m|\mathcal{M}_2, \theta_f, \theta_m) = \binom{N_f}{S_f}\theta_f^{S_f}(1-\theta_f)^{N_f-S_f} \cdot \binom{N_m}{S_m}\theta_m^{S_m}(1-\theta_m)^{N_m-S_m}$$

Assuming that both hypotheses are equiprobable:

$$\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} = 1$$

we compute marginal likelihoods for the two models:

$$\begin{aligned}
p(S_f, S_m, N_f, N_m|\mathcal{M}_1) &= \int_0^1 \binom{N_f}{S_f}\theta^{S_f}(1-\theta)^{N_f-S_f} \cdot \binom{N_m}{S_m}\theta^{S_m}(1-\theta)^{N_m-S_m} \cdot 1\, d\theta \\
&= \frac{N_f!}{S_f!(N_f-S_f)!} \times \frac{N_m!}{S_m!(N_m-S_m)!} \int_0^1 \theta^{S_f+S_m}(1-\theta)^{N_f+N_m-S_f-S_m} d\theta \\
&= \frac{N_f!}{S_f!(N_f-S_f)!} \times \frac{N_m!}{S_m!(N_m-S_m)!} \times B(S_f+S_m+1, N_f+N_m-S_f-S_m+1) \\
&= \frac{N_f!}{S_f!(N_f-S_f)!} \times \frac{N_m!}{S_m!(N_m-S_m)!} \times \frac{(S_f+S_m)!(N_f+N_m-S_f-S_m)!}{(N_f+N_m+1)!} \\
&= \frac{118!\,117!\,47!\,188!}{19!\,99!\,28!\,89!\,236!} = \frac{1445284062650182344369845}{8008083642837644483297741204} \approx 0.0001804781
\end{aligned}$$

$$\begin{aligned}
p(S_f, S_m, N_f, N_m|\mathcal{M}_2) &= \iint_0^1 \binom{N_f}{S_f}\theta_f^{S_f}(1-\theta_f)^{N_f-S_f} \cdot \binom{N_m}{S_m}\theta_m^{S_m}(1-\theta_m)^{N_m-S_m} d\theta_f d\theta_m \\
&= \frac{N_f!N_m!}{S_f!(N_f-S_f)!S_m!(N_m-S_m)!} \iint_0^1 \theta_f^{S_f}(1-\theta_f)^{N_f-S_f}\theta_m^{S_m}(1-\theta_m)^{N_m-S_m} d\theta_f d\theta_m \\
&= \frac{N_f!N_m!}{S_f!(N_f-S_f)!S_m!(N_m-S_m)!} B(S_f+1, N_f-S_f+1)B(S_m+1, N_m-S_m+1) \\
&= \frac{N_f!N_m!}{S_f!(N_f-S_f)!S_m!(N_m-S_m)!} \times \frac{S_f!(N_f-S_f)!}{(N_f+1)!} \times \frac{S_m!(N_m-S_m)!}{(N_m+1)!} \\
&= \frac{1}{(N_f+1)(N_m+1)} = \frac{1}{119 \cdot 118} = \frac{1}{14042} \approx 0.0000712149
\end{aligned}$$

The Bayes factor for preferring $\mathcal{M}_1$ over $\mathcal{M}_2$ is

$$\frac{p(S_f, S_m, N_f, N_m|\mathcal{M}_1)}{p(S_f, S_m, N_f, N_m|\mathcal{M}_2)} = \frac{1445284062650182344369845 \cdot 14042}{8008083642837644483297741204} \approx 2.53$$

which, informally, corresponds to quite weak evidence in preference of model $\mathcal{M}_1$. And the posterior odds of the models are:

$$\frac{p(\mathcal{M}_1|S_f, S_m, N_f, N_m)}{p(\mathcal{M}_2|S_f, S_m, N_f, N_m)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \times \frac{p(S_f, S_m, N_f, N_m|\mathcal{M}_1)}{p(S_f, S_m, N_f, N_m|\mathcal{M}_2)} \approx 2.53$$

A posteriori, Model $\mathcal{M}_1$ is about 2.5 times more plausible than model $\mathcal{M}_2$. Therefore hypothesis $\mathcal{H}_1$ is about 2.5 times more plausible than hypothesis $\mathcal{H}_2$. These odds are of relatively small scale, demonstrating relatively weak preference for hypothesis $\mathcal{H}_1$, and therefore we recommend performing further survey to collect more data.

In a more general case, when we have more than two hypotheses, and corresponding statistical models, we need to find the probability mass distribution among the alternative models:

$$p(\mathcal{M}_i|D) = \frac{p(D|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{j=1}^{k} p(D|\mathcal{M}_j)p(\mathcal{M}_j)}$$

where k models are considered in total with prior model probabilities $p(\mathcal{M}_i)$ and marginal likelihoods $p(D|\mathcal{M}_i)$.

https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf

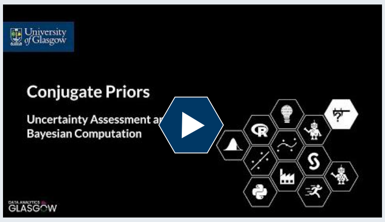
Note, that hypotheses testing in Bayesian framework is *very different* to the way it is performed in the classical Neyman-Pearson framework. Among many other benefits, the Bayesian approach allows

- Evaluating evidence *in favour* of the null hypothesis;
- Incorporating external information into evaluation of evidence about a hypothesis;
- Working with non-nested models;
- Performing variable selection and guiding evolutionary model-building process.

We recommend reading a paper by Kass and Raftery, 1995, “Bayes Factors”, *Journal of the American Statistical Association* **90**(430), pp. 773–795, to gain better understanding of the Bayesian hypotheses testing approach and how it compares to classical hypothesis testing.

Our course includes a separate chapter (in Week 10) on advanced topics in Bayesian hypotheses testing, where we will consider several approaches to approximating marginal likelihoods in the cases when analytical derivation becomes intractable.

Conjugate Priors and the Binomial Model



UABC - Conjugate Priors

https://youtu.be/HskcqW-Oi_o

Duration: 12m03s

Before modern electronic computers became widely available, numerical evaluation of posteriors, predictive distributions, and marginal likelihoods wasn’t feasible. Researchers had to make every effort to make analytical evaluation work. One of the common approaches to performing inference with guaranteed analytical solutions is to impose limitations on the types of priors used in inference. In this section we discuss the property of *conjugacy*.

Note that the utility of conjugacy is not limited to the cases when numerical evaluation is impossible, and this approach is not purely the thing of the past. Many contemporary methods rely on conjugacy to achieve astonishing performance improvement. Some sampling methods, we will talk about these later, may take days to produce a result for a Bayesian inference problem, while employing conjugate priors in these methods may decrease computational time to mere seconds.

While working with some statistical models, it is possible to find a family of distributions that is *conjugate* to the likelihood. This means that if we select a prior from this family of distributions, the posterior will also be in this family.

Consider a specific case of working with the binomial model:

$$k \sim Bi(\theta; n), \quad D = (k, n), k \in \{0, \dots, n\}, \theta \in [0, 1], n \in \mathbb{N}$$

The likelihood for this model, as introduced earlier, is

$$p(D|\theta) = p(k|\theta; n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

We will now select a prior for θ to be an arbitrary beta distribution:

$$p(\theta) = p(\theta|\alpha, \beta) = Be(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \quad \alpha > 0, \beta > 0.$$

We can now work out the posterior distribution of θ given D,

$$\begin{aligned} p(\theta|D) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^k(1-\theta)^{n-k} \\ &\propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1} \\ &\propto Be(\alpha+k, \beta+n-k) \end{aligned}$$

Notice, how we avoid the trouble of working with the constants by performing calculations down to proportionality and normalising the result in the end.

Here we demonstrated, that when the prior for the probability of success in the binomial model is a beta distribution, the posterior is also a beta distribution with different (updated) parameters. We therefore conclude that the beta distribution is a conjugate prior to the binomial likelihood. When performing inference using conjugate priors, all we need to do is to formulate a rule for updating prior parameters. In the binomial model case it is:

$$\begin{aligned} \alpha^* &= \alpha + k \\ \beta^* &= \beta + n - k \end{aligned}$$

Intuitively, $\alpha - 1$ corresponds to the number of successes and $\beta - 1$ corresponds to a number of failures in $\alpha + \beta - 2$ previous trials. The uniform prior is a particular case of this beta prior when we have no previous observations, and therefore $\alpha = \beta = 1$.

Example 9 (Conjugate Priors for Example 1).

Let's consider the problem in Example 1 once again, but this time we will rely on using conjugate priors to perform inference. As before, our likelihood is

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

The prior for θ is still going to be uniform from zero to one, but in this case we recognise that this uniform prior is a particular case of the beta prior:

$$p(\theta) = Be(1, 1)$$

First consider the case in Example 1, when we were using $D = (7, 10)$. As our prior is conjugate to the likelihood, we can perform inference using conjugacy property, and conclude that the posterior is

$$p(\theta|D) = Be(1+7, 1+10-7) = Be(8, 4)$$

which has the probability density function that looks exactly as the result in Figure 3.

Next, consider that we observe $D_2 = (6, 10)$ as the second observation. Using the same property of conjugacy we conclude that

$$p(\theta|D, D_2) = Be(8+6, 4+10-6) = Be(14, 8)$$

which is, again, exactly what is depicted in Figure 4.

In the case when we observe a series of datasets $D_1 = (k_1, n_1), D_2 = (k_2, n_2), \ldots, D_n = (k_n, n_n)$, the update of the parameters of this conjugate prior becomes:

$$\begin{aligned}\alpha^* &= \alpha + \sum_{i=1}^{n} k_i \\ \beta^* &= \beta + \sum_{1=1}^{n} n_i - \sum_{i=1}^{n} k_i\end{aligned}$$

https://en.wikipedia.org/wiki/Conjugate_prior

A comprehensive list of conjugate priors is available on Wikipedia.

We conclude this chapter with another example of inference and prediction using the binomial model that demonstrates that this model is applicable not only to coin tosses.

Example 10 (Amazon Reviews).

I was trying to buy a used copy of Gelman et al. *Bayesian Data Analysis* book from Amazon. This book is available used from several sellers for approximately the same price. Consider three sellers with the following ratings:

- Seller X: 97% positive out of 656,544 reviews
- Seller Y: 97% positive out of 1,080 reviews
- Seller Z: 99% positive out of 3,473 reviews

Which reseller is likely to provide the best service? Think for a second which one you'd intuitively select.

Do not jump into the conclusion that the seller with the highest proportion of positive reviews is necessarily the best. Consider a simpler case, say we have just two sellers: one seller (A) has 90 positive reviews out of 100, the other (B) has two reviews, both positive. You could say that one has 90% approval while the other has 100% approval, but is the one with 100% approval better?

Let θ_a be the probability of a customer being satisfied with the seller having 90 positive reviews out of 100. Assuming uniform prior on θ_a, the posterior will be (by conjugacy)

$$p(\theta_a|D_a) = Be(91, 11).$$

Let θ_b be the probability of a customer being satisfied with the seller with 2 positive reviews out of 2. Assuming the same uniform prior on θ_b we arrive to

$$p(\theta_b|D_b) = Be(3, 1).$$

These two posteriors are plotted in Figure 9: the purple line is for θ_a while the green line is for θ_b.

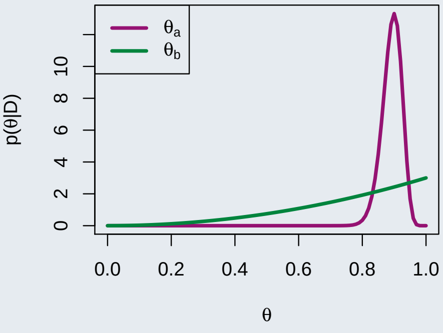


Figure 9: The posteriors for θ_a and θ_b compared.

Let's find the probability that a random sample from $p(\theta_a|D_a)$ is greater than a random sample from $p(\theta_b|D_b)$, that will be the probability that seller A is better.

$$p(\theta_a > \theta_b|D_a, D_b) = \frac{1}{B(3,1)} \int_0^1 (1 - I_{\theta_b}(91, 11))\theta_b^2 d\theta_b \approx 0.713$$

where $I_x(\alpha, \beta)$ is the regularised incomplete beta function and it is the c.d.f. for the beta distribution. So, it is more probable that seller A will provide a better experience.

Now let's return to our three sellers for Gelman et al. book. We will consider three customer satisfaction probabilities $\theta_X, \theta_Y, \theta_Z$. They will have uniform priors from 0 to 1. Their posteriors are therefore going to be:

$$p(\theta_X|D_X) = Be(636849, 19697)$$
$$p(\theta_Y|D_Y) = Be(1049, 33)$$
$$p(\theta_Z|D_Z) = Be(3439, 36)$$

These distributions are plotted in Figure 10: the purple line is for θ_X, the green line is for θ_Y, and the blue line is for θ_Z. The distributions are quite clearly separated, so it is virtually impossible that sellers X or Y are better than seller Z.
In general, going by averages alone works when you have a lot of customer reviews. But when you have a small number of reviews, going by averages alone could be misleading.

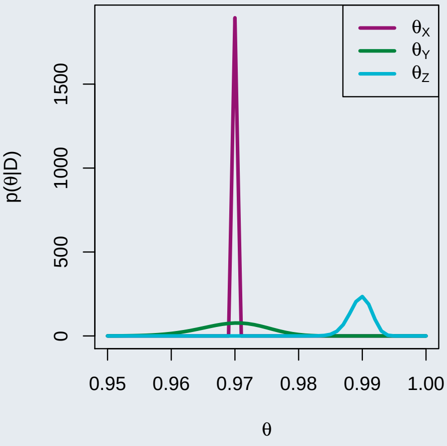


Figure 10: The posteriors for θ_X and θ_Y and θ_Z compared.

Review Exercises

Task 1.

Consider the binomial model with data $D = (k, n)$ where $k \geq 0$ is the number of successes in $n > 0$ trials:

$$k \sim Bi(\theta; n), \quad 0 \leq \theta \leq 1.$$

Assume a beta prior on θ:

$$p(\theta) = Be(\alpha, \beta),$$

and derive the posterior predictive distribution for observing new $\tilde{D} = (\tilde{k}, \tilde{n})$, where $\tilde{n}$ is a fixed constant.

Task 2.

Consider the Poisson model:

$$k \sim Po(\lambda), \quad \lambda > 0, k \in \mathbb{Z}^+$$

and demonstrate that a gamma prior for the rate parameter λ is conjugate:

$$\lambda \sim Ga(\alpha, \beta).$$

Derive corresponding parameter update for posterior inference.

Task 3 (BDA3, Exercise 2.11.1).

Suppose there is $Be(4, 4)$ prior distribution on the probability θ that a coin will yield a head when spun in a specified manner. The coin is independently spun ten times, and heads appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3. Calculate your posterior density (up to a proportionality constant) for θ and sketch it.

Answers to tasks

Answer to Task 1. We have

$$k \sim Bi(\theta; n), \quad 0 \le \theta \le 1,$$
$$p(\theta) = Be(\alpha, \beta).$$

The prior is conjugate to the likelihood, and therefore the posterior will be:

$$p(\theta|D) = Be(\alpha + k, \beta + n - k).$$

The posterior predictive distribution for observing new $\tilde{D} = (\tilde{k}, \tilde{n})$ is

$$\begin{aligned} p(\tilde{D}|D) &= \int_0^1 \binom{\tilde{n}}{\tilde{k}} \theta^{\tilde{k}} (1-\theta)^{\tilde{n}-\tilde{k}} \frac{\theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}}{B(\alpha+k, \beta+n-k)} d\theta \\ &= \binom{\tilde{n}}{\tilde{k}} \frac{1}{B(\alpha+k, \beta+n-k)} \int_0^1 \theta^{\alpha+k+\tilde{k}-1}(1-\theta)^{\beta+n-k+\tilde{n}-\tilde{k}-1} d\theta \\ &= \binom{\tilde{n}}{\tilde{k}} \frac{1}{B(\alpha+k, \beta+n-k)} B(\alpha+k+\tilde{k}, \beta+n-k+\tilde{n}-\tilde{k}) \\ &= \binom{\tilde{n}}{\tilde{k}} \frac{B(\alpha+k+\tilde{k}, \beta+n-k+\tilde{n}-\tilde{k})}{B(\alpha+k, \beta+n-k)}, \end{aligned}$$

where $k, n, \tilde{n}, \alpha$, and β are fixed constants, and $\tilde{k}$ is the random variable for which $p(\tilde{D}|D)$ defines the probability mass function. This distribution is called the beta-binomial distribution.

Answer to Task 2. We have

$$p(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$
$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

We will derive the posterior, working down to proportionality constants:

$$p(\lambda|k) \propto \lambda^k e^{-\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \propto \lambda^{\alpha-1+k} e^{-(\beta+1)\lambda} \propto Ga(\alpha+k, \beta+1)$$

As the posterior works out to be also a gamma distribution, this demonstrates that the prior is conjugate, and the parameter update is

$$\alpha^* = \alpha + k$$
$$\beta^* = \beta + 1$$

Note, that the Wikipedia page on conjugate priors lists updates for n independent observations, and this problem was concerned with only one observation.

Answer to Task 3 (BDA3, Exercise 2.11.1). The model is

$$p(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$
$$p(\theta) \propto \theta^3 (1-\theta)^3$$

With $n = 10$, we are only told that $k < 3$.
The posterior is conditioned only on the available information, i.e. $k < 3$:

$$\begin{aligned} p(\theta|k<3) &\propto p(\theta) p(k<3|\theta) \\ &\propto \theta^3(1-\theta)^3 \sum_{k=0}^{2} \binom{n}{k} \theta^k (1-\theta)^{10-k} \\ &\propto \theta^3(1-\theta)^{13} + 10\,\theta^4(1-\theta)^{12} + 45\,\theta^5(1-\theta)^{11} \end{aligned}$$

This could be evaluated on a fine grid and renormalised using the following R code:

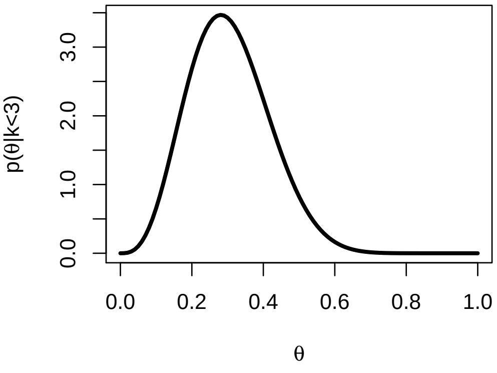


Figure 11: The posterior $p(\theta|k<3)$.

```
thetagrid <- seq(0,1,0.01)
post <- function(theta) {
    theta^3*(1-theta)^13 + 10*theta^4*(1-theta)^12+45*theta^5*(1-theta)^11
}
postgrid <- vapply(thetagrid,post,numeric(1))
postgrid <- postgrid/sum(postgrid)*100
```

and plotted

```
plot(thetagrid,postgrid,type="l")
```