Statistical Models & Statistical Inference

Mark Andrews

Contents

Introduction

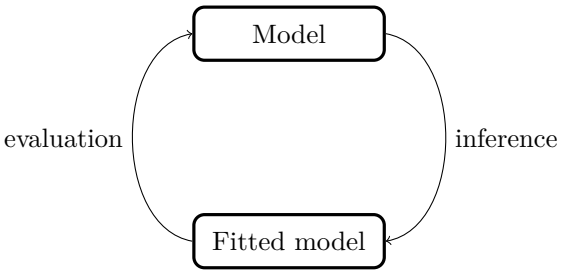
As important as exploratory analysis is, it is a fundamentally different undertaking to statistical modelling and statistical inference. Exploratory data analysis aims to describe and visualize the data and to identify possible trends and patterns in it. Statistical models, by contrast, are mathematical models of the *population* from which the data originated. The term *population* is here used in a technical sense that is specific to statistics. In general, a population is the hypothetical set from which our actual data are assumed to be a random sample. From the perspective of statistical modelling, our ultimate interest lies not in the data itself but rather in the population, or more specifically in our mathematical models of the population.

Let us assume that we have the following data set: $y_1, y_2 \dots y_n$, where each y_i is a single (scalar) value. From the perspective of statistical modelling, $y_1, y_2 \dots y_n$ is a random sample from a population. Ultimately, this population is described by a probability distribution. More specifically, we treat the values $y_1, y_2 \dots y_n$ as a *realization* of the random variables $Y_1, Y_2 \dots Y_n$. A random variable has a formal mathematical definition, but informally speaking, it is a variable that can take on different values according to some probability distribution. The probability distribution over the variables $Y_1, Y_2 \dots Y_n$ can be denoted generally by $\mathcal{P}(Y_1, Y_2 \dots Y_n)$. This probability distribution is a function over an n dimensional space that gives the probabilities of every possible

combination of values of the n variables $Y_1, Y_2 \dots Y_n$. Our observed data $y_1, y_2 \dots y_n$ is but one realization of the random variables $Y_1, Y_2 \dots Y_n$, and every possible realization is a random sample from the probability distribution $\mathcal{P}(Y_1, Y_2 \dots Y_n)$. Thus, the possible values that $Y_1, Y_2 \dots Y_n$ can take and their corresponding probabilities formally define the population.

In general, except in artifical scenarios, we do not know the nature of the population probability distribution $\mathcal{P}(Y_1, Y_2 \dots Y_n)$. Our aim, therefore, is to first propose a model of this probability distribution, and then use statistical inference to infer the properties of this model. The model that we develop is a probabilistic model, which means that it is defines a probability distribution over the variables $Y_1, Y_2 \dots Y_n$. This model is often referred to simply as the statistical model. In some contexts, it is also known as the *generative model*, the *probabilistic generative model*, the *data generating model*, and so on. Almost always, as we will see, this statistical model is defined in terms of parameters and other variables that are assumed to have fixed but unknown values. We use statistical inference, which we will discuss at length below, to infer what the values of these variables are.

The process of developing a statistical model and inferring its parameters and other unknown variables is, in fact, an iterative process, and is illustrated in the following diagram.



First, we propose or assume a statistical model. This may be done on the basis of some exploratory data analysis and visualization, as well as by using our scientific knowledge and understanding of the phenomenon being studied. Having assumed some model, we then infer its parameters and other unknowns. We are now in position to critically evaluate the resulting fitted model. Specifically, we evaluate whether or not the predictions of model make sense and whether they are consistent with the data. On the basis of this elaboration, we may now need to elaborate or possibly simplify our originally proposed model, which leads to a new proposed model. The unknowns of this model are then inferred, and the new fitted model is evaluated. This process iterates until we are satisfied that the model is sufficient for practical purposes or that no major alternative models have been overlooked. The final fitted model is then used as the model of the population. As we will see through examples, this is then effectively a mathematical or probabilistic model of the phenomenon being studied. With this model, we can explain and reason about the phenomenon, make predictions about future data, and so on.

As a brief illustrative outline of this modelling process, consider the following `housing_df` data frame.

```
housing_df <- read_csv('data/housing.csv')
housing_df
## # A tibble: 546 x 1
##    price
##    <dbl>
##  1 42000
##  2 38500
##  3 49500
##  4 60500
##  5 61000
##  6 66000
##  7 66000
##  8 69000
##  9 83800
```

```
## 10 88500
## # ... with 536 more rows
```

This gives the prices (in Canadian dollars) of a set of 546 houses in the city of Windsor, Ontario in 1987. We can denote these values as $y_1, y_2 \dots y_n$ and treat them as a realization of the random variables $Y_1, Y_2 \dots Y_n$ whose probability distribution is $\mathcal{P}(Y_1, Y_2 \dots Y_n)$. This defines the population. How best to conceive of this population in real world terms is usually a matter of debate and discussion, rather than a simple matter of fact. For example, the population might be conceived of narrowly as the set of house prices in Windsor in 1987, or more widely as the set of house prices in mid-sized cities in Canada in the late 1980s, or more widely still as the set of house prices in North America during the time period, and so on. Regardless of how we conceive of the population, we do not know the true nature of $\mathcal{P}(Y_1, Y_2 \dots Y_n)$ and so begin by proposing a model of it. Amongs other things, this initial proposal could be based on exploratory data analysis and visualization, such as that shown in Figure 1a-d. From the histograms and QQ plots shown here, we see that the logarithm of house prices appears to be distributed as a normal distribution.

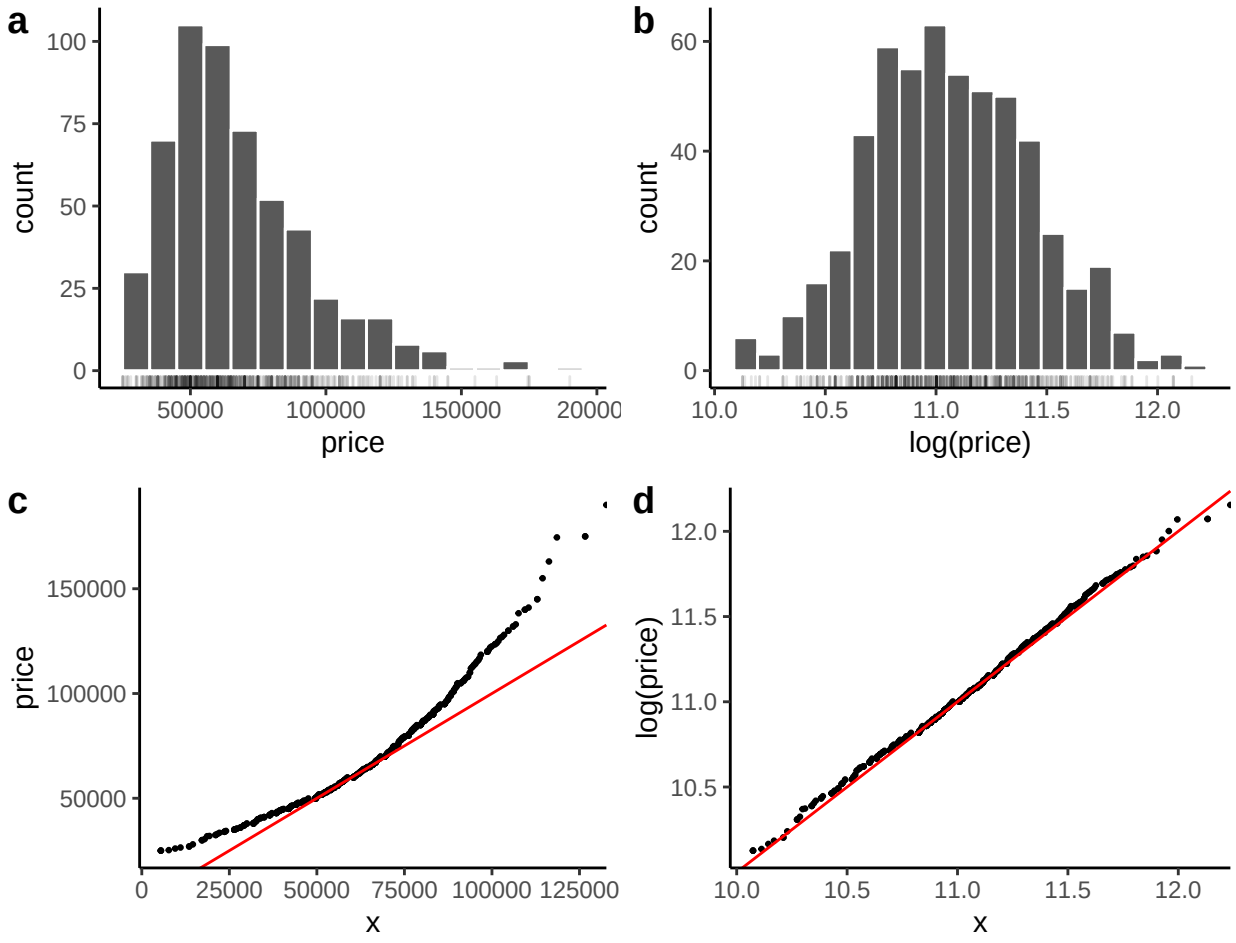


Figure 1: a) A histogram of house prices (in Canadian dollars) from the city of Windsor, Ontario, Canada, in 1987. b) A histogram of the logarithm of the house prices. c) A QQ plot of the prices compared to a normal distribution whose mean and standard deviation are set to the median and median absolute deviation (MAD) of the prices. d) The QQ plot of log of the prices, again compared to a normal distribution whose mean and standard deviation deviation are set to the median and MAD of the log prices.

This leads us to the following proposed model: for each $i \in 1 \dots n$, $Y_i \sim \text{logN}(\mu, \sigma^2)$. Note that $\text{logN}(\mu, \sigma^2)$ is shorthand to denote a log-normal distribution whose parameters are μ and σ^2, and a log-normal distribution

is a distribution of a variable whose logarithm is normally distributed. In the model statement, the $\sim$ is read as *is distributed as*. This model therefore states that each random variable Y_i, for $i \in 1 \dots N$, is distributed as a log-normal distribution whose parameters are μ and σ^2. This model is an example of an *independent and identically distributed* (IID) model: each of the n random variables is modelled as independent of one another, and each one has the same probability distribution.

Having proposed a initial model, we must now infer the values of the unknown variables μ and σ. As we will discuss at length below, there are two major approaches to doing statistical inference, but in practical terms, both methods effectively lead to estimates of μ and σ as well as measures of the uncertainty of these estimates. If we denote the estimates of μ and σ by $\hat{\mu}$ and $\hat{\sigma}$, our fitted model is then the log-normal distribution $\text{logN}(\hat{\mu}, \hat{\sigma}^2)$. At this point, we may evaluate this fitted model and determine if its assumptions or predictions are consistent with the observed data. If, on the basis of this evaluation, the assumed model is deemed atisfactory for present practical purposes, we then can use the fitted model, particularly taking into account the uncertainties in our estimates, to explain, reason about, or make predictions concerning house prices in cities like Windsor during this period.

Statistical inference

Statistical inference is the inference of the values of unknown variables in a statistical model. There are two major approaches to statistical inference. The first approach is variously referred to as the *classical*, *frequentist*, or *sampling theory* based approach. The second is the Bayesian approach. The classical approach is still the dominant one in practice, particulary for the more introductory or medium level topics. It is also the principal, or even only, approach taught in most applied statistics courses. As such, it is the only approach that most working scientists will have been formally introduced to. The Bayesian approach, on the other hand, although having its origins in the 18th century and being used in practice throughout the 19th century, had a long hiatus in statistics until around the end of the 1980s. Since then, and as we will see, largely because of the growth in computing power, it has steadily grown in terms of its popularity and widespread usage throughout statistics and data science. Throughout the remainder of this book, we will attempt to pay adequate attention to both the classical and Bayesian approaches wherever we discuss statistical models. In this section, we aim to provide a brief general introduction to both approaches. For this, we will use a simple problem, propose a statistical model for it, and then infer the values of its unknown parameters using the classical and the Bayesian approaches.

Example problem: The problem we will consider was described in the Guardian newspaper in January 4, 2002[1]: "Polish mathematicians Tomasz Gliszczynski and Waclaw Zawadowski... spun a Belgian one euro coin 250 times, and found it landed heads up 140 times." Here, the data is the observed number of heads, $m = 140$. The total number of spins $n = 250$ is the sample size, and is a fixed and known quantity, and so is not modelled per se. From the perspective of statistical modelling, the observed number of heads is treated as a sample from a population. In this case, the population is the set of all possible observed number of heads, and their relative frequencies, that would be obtained if Gliszczynski and Zawadowski were to infinitely repeat their study, under identical circumstances with the exact same Belgian one euro coin. The population is therefore a probability distribution over the set of possible values of the number of heads in a sample of $n = 250$ trials, which is a probability distribution over $0 \dots n$. Thus, m is a realization of a random variable Y whose possible values are $0 \dots n$ and whose probability distribution is $\mathcal{P}(Y)$. We do not know $\mathcal{P}(Y)$ and so we begin by proposing a model of it. In this case, the only viable option for this model is the binomial distribution: $Y \sim \text{Binomial}(\theta, n = 250)$. While more complex options are possible in principle, they would all go beyond what the data can tell us, and so simply can not be evaluated. In general terms, a binomial distribution gives the probability of the number of so-called "successes" in a fixed number of trials n where the probability of a success on any trial is fixed quanity θ and all trials are independent of another. Translated into the terms of the current example, the binomial distribution is the probaility distribution over the observed number of heads in $n = 250$ spins where the probability of a heads on any trial is θ. In this binomial model, the parameter θ has a fixed but unknown quantity. The value of θ is therefore the objective of our statistical inference.

[1] https://www.theguardian.com/world/2002/jan/04/euro.eu2

Classical statistical inference

Classical statistical inference begins with an *estimator* of the value of θ, denoted by $\hat{\theta}$, and then considers the *sampling distribution* of $\hat{\theta}$ for any hypothetical value of θ. Informally speaking, we can see the estimator $\hat{\theta}$ as an educated guess of what the true value of θ is. There are different methods of estimation available, and each one can be evaluated, as we will see, according to certain criteria, such as *bias*, *variance*, etc. One widely used method of estimation, perhaps the most widely used method, is *maximum likelihood estimation*.

Maximum likelihood estimation

The maximum likelihood estimator is the value of the unknown variables that maximizes the *likelihood function*. The likelihood function is an extremely important function in both classical and Bayesian approaches to statistical inference. It is a function over the space of unknown variables that gives the probability of observing the data given any particular value of these variables. In order to determine the likelihood function in the case of the binomial distribution model, we first start with the definition of the binomial distribution itself. If Y is random variable distributed as a binomial distribution with parameter θ and sample size n, which we can state succinctly as $Y \sim \text{Binomial}(\theta, n)$, then the probability that Y takes on the value of m is as follows:

$$\text{P}(Y = m|\theta, n) = \text{Binomial}(Y = m|\theta, n) = \binom{n}{m}\theta^m(1-\theta)^{n-m}.$$

Why the binomial distribution has the probability mass function shown on the very right hand side here is not something we will derive here, and we will just take it as a given, but it is relatively straightforward to derive from the above definition of a binomial problem. Note that this probability mass function is a function that maps values of $m \in 0 \ldots n$ to the interval $[0, 1]$ for fixed values for θ and n. The corresponding likelihood function takes the same formula, i.e. $\binom{n}{m}\theta^m(1-\theta)^{n-m}$, and treats it as a function of θ for fixed values of m and n. In other words, the likelihood function is as follows:

$$L(\theta|m, n) = \binom{n}{m}\theta^m(1-\theta)^{n-m},$$

where θ is assumed to vary from $(0, 1)$ and where n and m are fixed. In other words, the likelihood function gives the probability of the observed data for every possible value of the unknown variable θ. Technically speaking, any likelihood function is defined up to a proportional constant. What this means is that multiplying, or dividing, any likelihood function by a fixed constant value results in the same likelihood function. In practice, therefore, when writing a likelihood function, we usually drop any constant multipliers. In the above likelihood function, the binomial coefficient $\binom{n}{m}$ is a constant term that does not change for any value of θ, and so it can be dropped, leading to the likelihood function being written as

$$L(\theta|m, n) = \theta^m(1-\theta)^{n-m}.$$

The likelihood function for $m = 140$ and $n = 250$ is shown in Figure 2a. As we can see from this plot, the likelihood function is concentrated from around $\theta = 0.45$ to around $\theta = 0.65$. Outside of that range, the likelihood of any value of θ becomes negligble. In itself, this is very informative. It tells us that the probability of observing the data that we did, i.e., $m = 140$ heads in $n = 250$ spins, is neglible if θ is less than around 0.45 or greater than around 0.65. Therefore, only values in the range of approximately 0.45 to 0.65 have evidential support from the data.

The value of θ that maximizes the likelihood function is obtained by the usual procedure of optimizing functions, namely calculating the derivative of the function with respect to θ, setting the derivative equal to zero, and solving for the value of θ. We can do this more easily by using the logarithm of the likelihood function, rather than the likelihood function itself. The logarithm of the likelihood function is shown in Figure 2b. Because the logarithm is a monotonic transformation, the value of θ that maximizes the logarithm of the likelihood also maximizes the likelihood function. The derivative of the log of the likelihood function

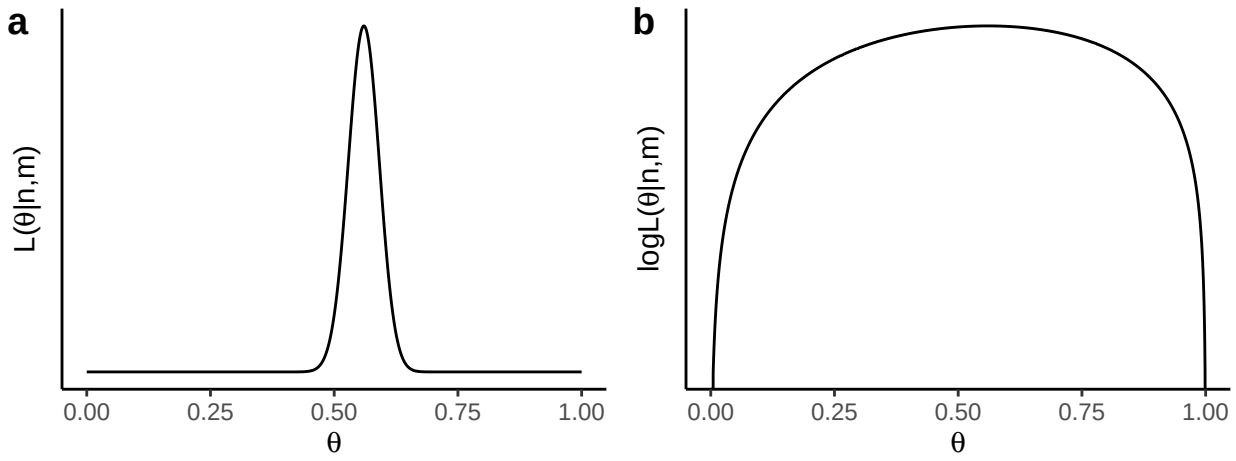


Figure 2: a) The binomial likelihood function for θ when $m = 140$ and $n = 250$. b) The logarithm of the likelihood function.

with respect to θ is as follows:

$$\begin{aligned}\frac{d}{d\theta}\log\left(\theta^m(1-\theta)^{n-m}\right) &= \frac{d}{d\theta}\left[m\log(\theta) + (n-m)\log(1-\theta)\right], \\ &= \frac{m}{\theta} - \frac{n-m}{1-\theta}\end{aligned}$$

Setting this derivative equal to zero and solving for θ gives us the following.

$$\begin{aligned}\frac{m}{\theta} - \frac{n-m}{1-\theta} &= 0, \\ \theta &= \frac{m}{n}.\end{aligned}$$

Thus, the maximum likelihood estimator for θ is $\hat{\theta} = \frac{m}{n} = \frac{140}{250} = 0.56$. This is obviously a very simple and intuitive result. It tells us that the best guess for the value of θ, which is the probability of obtaining heads on any given spin, is simply the relative number of heads in the n spins so far.

Sampling distribution of $\hat{\theta}$

The maximum likelihood estimator can be seen as a random variable. It is a deterministic function of the observed data m, but m would vary were we repeat the experiment even under identical circumstances. For example, if we knew the true value of θ, and we spun the coin n times, and did so ad infinitum, then we know that the distribution of m (the observed number of heads) would be the binomial distribution with sample size n and parameter value of θ. Given that the maximum likelihood estimator is always $\hat{\theta} = \frac{m}{n}$, the set of possible values of $\hat{\theta}$ are $\frac{0}{n}, \frac{1}{n} \ldots \frac{n}{n}$, the probability that $\hat{\theta}$ takes on the value of $\frac{m}{n}$ for any given value of m is

$$\binom{n}{m}\theta^m(1-\theta)^{n-m}.$$

In general, the sampling distribution for $\hat{\theta}$ can be written as follows.

$$\mathrm{P}(\hat{\theta}|\theta, n) = \binom{n}{\hat{\theta}n}\theta^{\hat{\theta}n}(1-\theta)^{n-\hat{\theta}n}.$$

The sampling distribution of $\hat{\theta}$ when $n = 250$ and for the example value of $\theta = 0.64$ is shown in Figure 3.

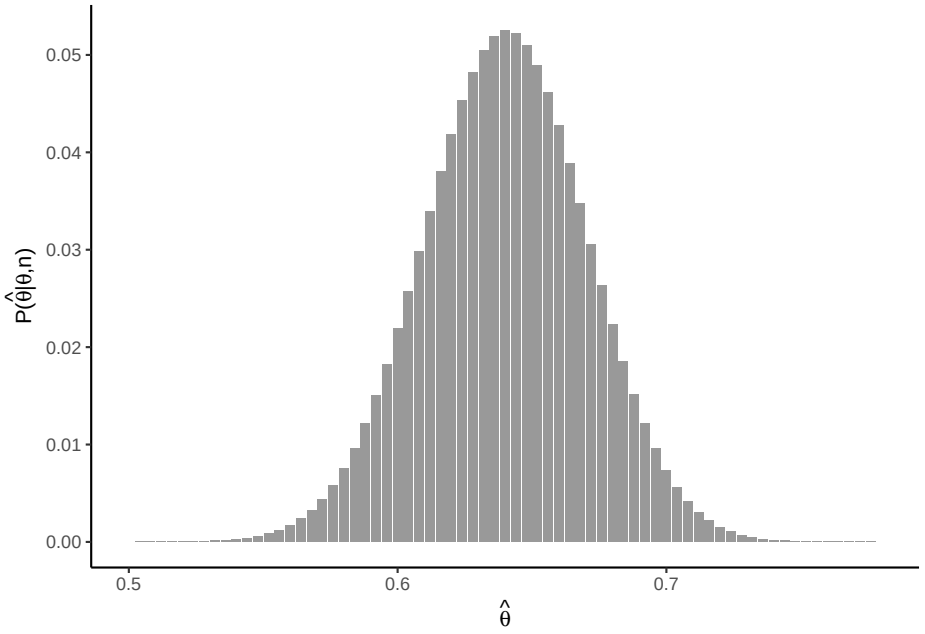


Figure 3: Sampling distribution of the binomial maximum likelihood estimator $\hat{\theta}$ when $n = 250$ and $\theta = 0.64$. Here, we limit the x axis to just beyond those values of $\hat{\theta}$ that have nontrivial probabilities.

The expected value of the sampling distribution of $\hat{\theta}$ is θ. This tells us that the binomial maximum likelihood estimator is an *unbiased* estimator of the true value of θ. In other words, on average, $\hat{\theta}$ is equal to the true value of θ. The variance of the distribution of $\hat{\theta}$ is $\frac{1}{n}\theta(1-\theta)$. The lower the variance of the estimator, the less sampling variability there will be in the value of the estimators. Here, we see that the variance decreases as n increases, and so as sample size increases, there is less variability in the estimator's values. The standard deviation of the sampling distribution is $\frac{1}{\sqrt{n}}\sqrt{\theta(1-\theta)}$. This is known as the *standard error*, and often plays an important role in classical statistical inference, as we will see in later examples.

p-values

Informally speaking, a p-value tells us whether the value of an estimator, or more generally, the value of any statistic or function of the observed data, is consistent with some hypothetical value of the unknown variable. If the p-value is low, the estimator's value is not consistent with the hypothesized value. The higher the p-value is, the more consistent the estimator's value is with the hypothesized value. If the p-value is sufficiently low, we can, practically speaking, rule out or reject the hypothesized value. In general, the p-value takes values from 0 to 1 and so is an inherently continuous measure of support for a hypothesis. Where we "draw the line" on this continuum between low and not-low is a matter of convention, but the almost universally held convention is that a p-value must be lower than at most 0.05 to be considered sufficiently low for the corresponding hypothesized value to be rejected. This threshold for rejection/non-threshold is usually signified by α.

Technically speaking, p-values are tail areas of the sampling distribution of the estimator corresponding to a particular hypothesized value of the unknown variable. Once we have the sampling distribution, p-values are straightforward to calculate. In the current problem, the unknown variable is θ and we can in principle hypothesize that its true value is any value between 0 and 1. If, for example, we hypothesize that the true value of θ is 0.64, the sampling distribution of $\hat{\theta}$ is that shown in Figure 3. On the basis of this sampling distribution, we can see that some values for $\hat{\theta}$ are expected and others are not. For example, we see that the $\hat{\theta}$ values are mostly from around 0.58 to around 0.7, and values of $\hat{\theta}$ much below or above those extremes rarely occur. On the basis of the estimator's value that we did obtain, namely $\hat{\theta} = \frac{140}{250} = 0.56$, we can see

that this result seems to be outside the range of values of $\hat{\theta}$ that we would expect if $\theta = 0.64$. In order to be precise in our statement of whether $\hat{\theta} = 0.56$ is beyond what we would expect if $\theta = 0.64$, we calculate the tail areas of the sampling distribution defined by values *as or more extreme* than $\hat{\theta} = 0.56$. These tail areas are shaded in Figure 4a. The total area in these tails defines the p-value. In other words, the p-value is the probability of observing a value of the estimator *as or more extreme* than $\hat{\theta} = 0.56$ if $\theta = 0.64$. If the p-value is low, then we know that $\hat{\theta} = 0.56$ is far into the tails of the sampling distribution when $\theta = 0.64$. In this particular example, the area of these tails, and so therefore the p-value, is approximately 0.01. This is clearly low according to the conventional standards mentioned above, and so therefore we say that the result $\hat{\theta} = 0.56$ is not consistent with the hypothesis that $\theta = 0.64$, and so in practical terms, we can reject the hypothesis that the true value of θ is 0.64.

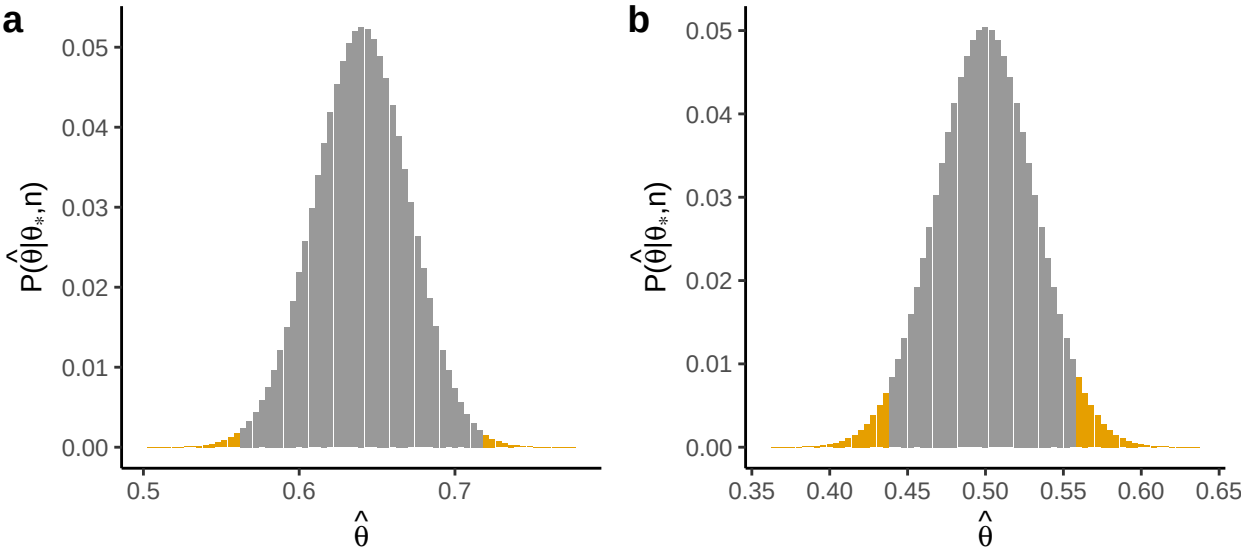


Figure 4: a) Sampling distribution of the binomial maximum likelihood estimator $\hat{\theta}$ when $n = 250$ and $\theta = 0.64$. b) Sampling distribution of the binomial maximum likelihood estimator $\hat{\theta}$ when $n = 250$ and $\theta = 0.5$. In both cases, as in Figure 3, we limit the x axis to just beyond those values of $\hat{\theta}$ that have nontrivial probabilities. The shaded tail areas correspond to values of $\hat{\theta}$ that are as or more extreme, relative to the center, than the value of the estimator we observed, which was $\hat{\theta} = \frac{140}{250} = 0.56$.

In more detail, the p-value is the sum of the two tails of the sampling distribution when $\theta = 0.64$. The lower tail is defined by all those values from $\hat{\theta} = 0$ up to $\hat{\theta} = 0.56$. These are values of $\hat{\theta}$ that are less than or equal to the observed $\hat{\theta}$ of $\frac{140}{250} = 0.56$. The upper tail is defined by all those values from $\hat{\theta} = 0.64 + |0.64 - 0.56| = 0.72$ up to 1. These are the values $\hat{\theta}$ that as extreme relative to $\theta = 0.64$ as is 0.56, but in the opposite direction. Thus, the p-value is calculated as follows.

$$p\text{-value} = \underbrace{\int_0^{0.56} \binom{n}{\hat{\theta}n} \theta^{\hat{\theta}n}(1-\theta)^{n-\hat{\theta}n}\, d\hat{\theta}}_{\text{area of lower tail}} + \underbrace{\int_{0.72}^{1} \binom{n}{\hat{\theta}n} \theta^{\hat{\theta}n}(1-\theta)^{n-\hat{\theta}n}\, d\hat{\theta}}_{\text{area of upper tail}} \approx 0.01.$$

More generally speaking, the precise definition of a p-value is as follows.

$$p\text{-value} = \mathrm{P}(|\hat{\theta} - \theta_H| \geq |\hat{\theta}_{\text{obs}} - \theta_H|).$$

Here, for clarity, we distinguish between $\hat{\theta}_{\text{obs}}$, which is the actual value of the estimator calculated from the observed data, θ_H, which is the hypothetized value of θ, and $\hat{\theta}$, which is the estimator's random variable whose distribution is the sampling distribution when $\theta = \theta_H$.

We can calculate p-values for binomial problems like this using R with the `binom.test` command. We need to pass in the observed number of "successes", which are heads in this case, as the value of the `x` parameter, and the sample size as the value of `n`. The hypothesized value of θ is passed in as the value of `p`.

```
binom_model <- binom.test(x = 140, n = 250, p = 0.64)
```

The value of the maximum likelihood estimator is given by the value of the `estimate` element of the output object.

```
binom_model$estimate
## probability of success
##                   0.56
```

The p-value is given by the value of `p.value` element.

```
binom_model$p.value
## [1] 0.01004809
```

Null hypotheses and significance

In general, we can test any hypothetical value of the unknown variable. Often some hypothetical values have a special meaning in that they correspond to values that entail that there is no effect of any interesting kind. Hypotheses of this kind are known as *null* hypotheses. In the present example, the hypothesis that $\theta = 0.5$ entails that the coin is completely unbiased; it is no more likely to come heads than tails, and so any differences in the observed numbers of heads and tails in a set of spins or flips is just a matter of chance. As such, this hypothesis would usually be seen as a null hypothesis. The sampling distribution for $\hat{\theta}$ when $\theta = 0.5$ is shown in Figure 4b. The total tail area, which is the p-value is calculated similarly to the example above.

$$p\text{-value} = \underbrace{\int_0^{0.44} \binom{n}{\hat{\theta}n} \theta^{\hat{\theta}n}(1-\theta)^{n-\hat{\theta}n}\, d\hat{\theta}}_{\text{area of lower tail}} + \underbrace{\int_{0.56}^{1} \binom{n}{\hat{\theta}n} \theta^{\hat{\theta}n}(1-\theta)^{n-\hat{\theta}n}\, d\hat{\theta}}_{\text{area of upper tail}} \approx 0.066.$$

Note that here the lower tail is defined up to $0.5 - |0.5 - 0.56| = 0.44$. From this p-value of 0.066, we see that this is not sufficiently low to reject the null hypothesis at the conventional $\alpha = 0.05$ threshold, though of course it is quite close to this threshold too. We can calculate this p-value using `binom.test`.

```
null_binom_model <- binom.test(x = 140, n = 250, p = 0.5) # p = 0.5 is default
null_binom_model$p.value
## [1] 0.06642115
```

As a general point about null hypothesis testing, given that a null hypothesis is a hypothesis of no interesting effect, if we reject that hypothesis, we say the result is *significant*. Saying that a result is significant in this sense of the term, however, is not necessarily saying much. The estimated effect may be small or even neglible in practical terms, but may still be statistical significant. Moreover, even a highly significant p-value does not necessarily mean a large effect in practical terms. As an example, there were 731,213 live births in the United Kingdom in 2018. The p-value for the null hypothesis that the probability of the birth being a male rather than female is $\approx 3.8 \times 10^{-119}$. This is a tiny p-value and so is an extremely significant result. However, the (maximum likelihood) estimator for the probability of a male birth is 0.514. Although certainly not practically meaningless, this is nonetheless quite a small effect: it corresponds to around 28 more males than females in every 1000 births. As such, an extremely statistically significant result corresponds to small effect in practical terms. In general, because the p-value for a non-null true effect will always decrease as the sample size increases, we can always construct cases of arbitrarily small p-values for arbitrarily small effects, and so even practically trivial effects may be highly statistically significant.

Confidence intervals

Confidence intervals are counterparts to p-values. As we've seen, each p-value corresponds to a particular hypothesis about the true value of θ. If the p-value is sufficiently low, we will reject the corresponding

hypothesis. If the p-value is not sufficiently low, we can not reject the hypothesis. However, not rejecting a hypothesis does not entail that we should accept that hypothesis; it just simply means that we can not rule it out. The set of hypothetical values of θ that we do *not* reject at the α p-value threshold corresponds to the $1-\alpha$ confidence interval. Practically speaking, we can treat all values in this range as the set of plausible values for θ.

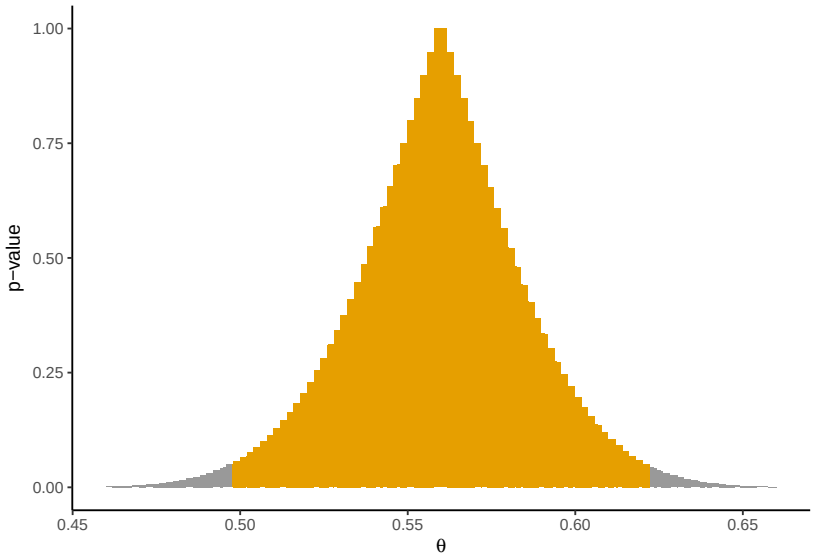


Figure 5: The p-value for each hypothetical value of θ from 0.46 to 0.66 in steps of 10^{-3}. Shaded in grey are all those values of θ that have p-values less than $\alpha = 0.05$. These values of θ would be rejected in a hypothesis test. All other values of θ have p-values greater than $\alpha = 0.05$ and so would not be rejected by a hypothesis test. These value of θ comprise the confidence interval.

In Figure 5, we shade in yellow all those values of θ that would not be rejected by a hypothesis test at $\alpha = 0.05$. Therefore, these values of θ are in the $1-\alpha = 0.95$ confidence interval. The lower and upper bounds of this interval as calculated in this figure, which is based on a discretization of the θ interval into steps of 10^{-3}, is $[0.498, 0.622]$. The interval can be calculated exactly by using a relationship between the cumulative binomial distribution and the cumulative Beta distribution, a method known as the Clopper-Pearson method (Clopper and Pearson 1934). Using R, we can calculate this Clopper-Pearson confidence interval as follows.

```
c(qbeta(0.025, 140, 250 - 140 + 1),
  qbeta(1-0.025, 140+1, 250 - 140))
## [1] 0.4960662 0.6224941
```

This will also be returned by the `binom.test`.

```
M_binom <- binom.test(x = 140, n = 250)
M_binom$conf.int
## [1] 0.4960662 0.6224941
## attr(,"conf.level")
## [1] 0.95
```

Confidence intervals have the following frequentist property. In an infinite repetition of an experiment[ˆHere, we use the term experiment in sense that it is used in probability theory, which is of a well defined procedure that generates data according to a specified probability distribution.], the 95% confidence interval will contain the true value of the unknown variable 95% of the time. What this means in the case of the present problem is that if we were to repeat ad infinitum the original coin spinning experiment under identical conditions, and so the probability of a heads outcome is a fixed though unknown value of θ, and on each repetition calculate the 95% confidence interval, then 95% of these confidence intervals would contain the true value

of θ. In Figure 6, we provide an illustration of this phenomenon. For 100 repetitions, we generate data from a binomial distribution with $n = 250$ and $\theta = 0.6$. From the data on each repetition, we calculate the 90% confidence interval. Those intervals that do not contain $\theta = 0.6$ are shaded in grey. In this set of 100 repetitions, there are 92 intervals out of 100 that contain $\theta = 0.6$.

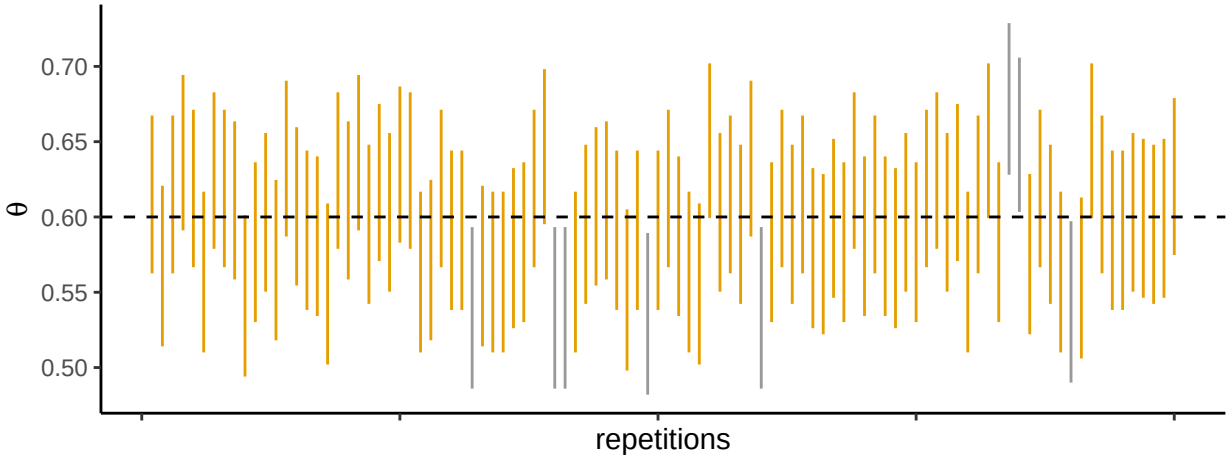


Figure 6: The 90% confidence intervals in 100 repetitions of a binomial experiment with $n = 250$ and $\theta = 0.6$. Shaded in grey are any confidence intervals that do not contain $\theta = 0.6$.

Bayesian statistical inference

Bayesian approaches to statistical inference ultimately aim to solve the same problem as that of classical approaches, namely the inference of the unknown values of variables in a statistical model. While the classical approaches is based on calculation of estimators and their sampling distributions, Bayesian approaches rely on an 18th century mathematical result known as *Bayes' rule* or *Bayes' theorem* to calculate a probability distribution over the possible values of the unknown variable. This probability distribution, known as the *posterior distribution*, give us the probability that the unknown variable takes on any given value, contingent on the assumptions of the model. To introduce Bayesian inference, we will continue with the same example problem covered above.

It is important to emphasize that the choice of the statistical model is prior to and not dependent on whether a classical or Bayesian approach to inference is used. In other words, we first assume or propose a statistical model for the data at hand and then, in principle, we can choose to use a classical or Bayesian approach to inference of the unknown variables in the model. For the problem at hand, as described above, our statistical model is that Y is a random variable with a binomial probability distribution with unknown parameter θ and sample size $n = 250$. As we've seen, we can state this as follows:

$$Y \sim \text{Binomial}(\theta, n = 250).$$

The observed number of heads, $m = 140$, is a realization of the random variable Y, and the probability that Y takes on the value of m given fixed values of θ and n is the given by the following probability mass function.

$$\text{P}(Y = m|\theta, n) = \text{Binomial}(Y = m|\theta, n) = \binom{n}{m}\theta^m(1-\theta)^{n-m}.$$

As we've also seen, the likelihood function corresponding to this probability distribution is

$$L(\theta|m, n) = \theta^m(1-\theta)^{n-m}.$$

As we'll see, the likelihood function plays a major role in Bayesian inference.

Priors

Having established our statistical model, to perform Bayesian inference on the value of θ, we must first provide a probability distribution for the possible values that θ can take on in principle. This is known as the *prior* distribution. As an example, if we assume that θ can take on any possible value in the interval $[0, 1]$ and each value has equal likelihood, our prior would be a uniform distribution over θ. On the other hand, if we assume that θ values are more likely to be equal to $\theta = 0.5$, but possibly be above or below $\theta = 0.5$ too, our prior might be a symmetrical unimodal distribution centered at $\theta = 0.5$. How wide this unimodal distribution is depends on what we think are the relative probabilities of values close to and far from $\theta = 0.5$. Likewise, if we assume that θ is likely to correspond to a bias towards heads, then the prior might be another symmetrical unimodal distribution but centered on some value of θ greater than 0.5. Examples of priors like this are shown in Figure 7.

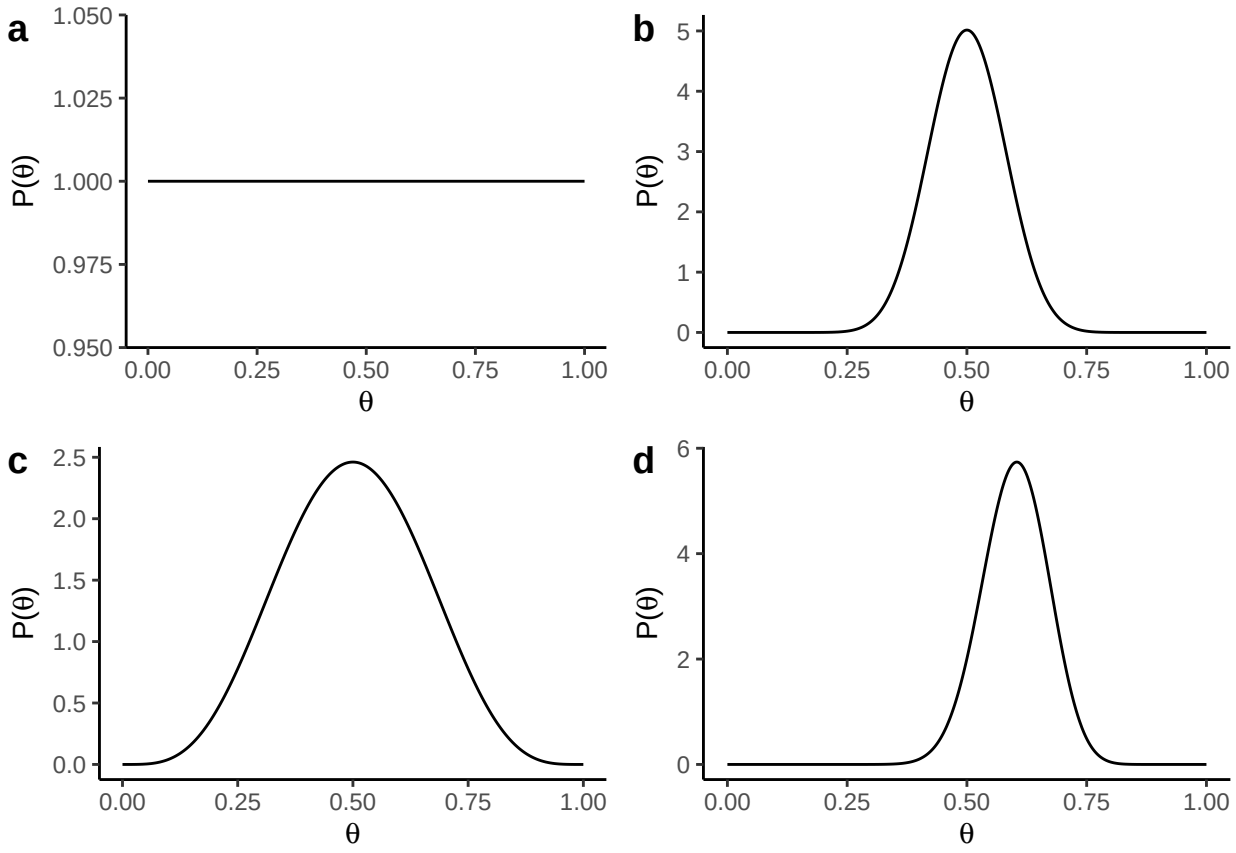


Figure 7: Examples of priors over θ. a) A uniform prior. b-c) Priors where θ is more likely to be 0.5 than otherwise but values greater and less than 0.5 are probable too, and more so in the case of c) whose variance is wider than the prior in b). d) A prior where θ is more likely to be $\theta = 0.6$ than any other way and more likely to be above rather than below 0.5.

There is a vast literature on what priors are, how to choose them, and the relative merits of, for example, subjective, objective, informative, noninformative, weakly informative, reference, maximum entropy, and other kinds of priors. We will not even attempt a summary of this vast literature here. Our general advice on the topic is that the choice of a prior should not be automatic, but rather should be treated as a modelling assumption, just like the assumptions that lead to our choice of the original statistical model. When choosing the prior, just like when we choose the statistical model, we should follow a number of guiding principles: we should use our general understanding of the problem at hand, and also of the data itself; our choices should be reasonable and justifiable rather than arbitrary; our choices should be seen as tentative and subject to revision if the assumptions or reasoning on which it was based are found to be invalid or wrong. More

specifically, following Gelman, Simpson, and Betancourt (2017), we recommend that our choices of priors should be guided by whether the prior could generate the type of data that we expect to see in the problem at hand, that it covers the range of plausible values of the unknow parameter, and most of the prior mass should be the parts of the parameter space that are likely to generate the type of data we are modelling.

It should be emphasized, however, that although in general the choice of priors is always important, it has more practical consequences in some problems than in others. In particular, in relative simple models where the ratio of observed data to unknown variables is relatively high, as is the case with the present problem, most choices of priors, unless they are extreme, will lead to practically similar conclusions. In complex models, on the other hand, or models where the amount of data relative to the number of unknown variables is low, the priors play a much more important role in the final fitted model, and so choices of priors in these contexts needs to be more careful and judicious.

For the present problem, the parameter θ is the probability that the coin lands heads up after a spin. The value of θ is a function of both the physical properties of the coin and also the manner in which it is spun. It is arguably difficult to physically bias a coin so that one side is more likely in a flip (Gelman and Nolan 2002), but experimental results with different materials can lead to outcomes that are as much as 70-30 biased (Kerrich 1946), and precise control over how it is flipped can lead to even 100% biased outcomes (Diaconis, Holmes, and Montgomery 2007). Coin spins, as opposed to flips, can be easily biased to as much as 75-25 bias (Gelman and Nolan 2002). From this, it seems to reasonable to have a unimodal and symmetrical prior centered at $\theta = 0.5$, but with a relatively wide spread to allow for relatively extreme possibilities. For this, we will use the prior displayed in Figure 7c. Clearly, the most likely value of θ according to this prior is $\theta = 0.5$, but values of $\theta = 0.25$ or $\theta = 0.75$ have substantial prior mass in their vicinity, and even extreme values of greater than $\theta = 0.9$ or less than $\theta = 0.1$ have non-trivial prior mass.

The prior that we have chosen (displayed in Figure 7c) is a beta distribution, which is a probability distribution over the interval from 0 to 1. There are two parameters in the beta distribution, which we call *hyper*parameters to distinguish them from the parameters of the statistical model, and these are conventionally denoted by α and β. In the particular beta distribution that we have chosen theire values are $\alpha = 5$ and $\beta = 5$. The density function of any beta distribution is

$$\mathrm{P}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

and its mean and variance are

$$\langle\theta\rangle = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

For our choice of prior therefore, the mean of the distribution is 0.5 and the variance is 0.023 (and so the standard deviation is 0.151).

Bayes' rule and the posterior distribution

Having specified our statistical model and our prior, we can now write out the full Bayesian model.

$$\begin{aligned} Y &\sim \mathrm{Binom}(\theta, n = 250), \\ \theta &\sim \mathrm{Beta}(\alpha = 5, \beta = 5) \end{aligned}$$

We can view this model as an expanded generative model of the observed data. In other words, to generate hypothetical data sets, first, we sample a value of θ from a beta distribution with hyperparameters $\alpha = \beta = 5$, and then we sample a value from a binomial distribution with parameter θ and sample size $n = 250$. This expanded model therefore defines a joint probability distribution over Y and θ, conditional on α, β and n, i.e. $\mathrm{P}(Y, \theta|\alpha, \beta, n)$. We will use this join probability distribution, coupled with the fact that we observe the value of Y, to caculate a probability distribution over the possible values of θ conditional on all the observed data and the assumptions of the model. This is known as the *posterior distribution*, and is the central result in Bayesian inference. To understand how the posterior distribution is calculated, we must introduce some elementary results from probability theory.

If we have two random variables A and B, an elementary probability theory result show how joint probability distribution of A and B can be factored into products of *conditional* probability distributions and *marginal* probability distributions:

$$\underbrace{\mathrm{P}(A,B)}_{\text{joint}} = \underbrace{\mathrm{P}(A|B)}_{\text{conditional}}\ \underbrace{\mathrm{P}(B)}_{\text{marginal}} = \underbrace{\mathrm{P}(B|A)}_{\text{conditional}}\ \underbrace{\mathrm{P}(B)}_{\text{marginal}}$$

In other words, the joint probability of A and B is equal to the conditional probability of A given B times the probability of B, or equally, it is equal to the conditional probability of B given A times the probability of A. Another elementary result is how we calculate marginal probability distributions from summation over joint distributions.

$$\mathrm{P}(A) = \sum_{\{B\}} \mathrm{P}(A,B) = \sum_{\{B\}} \mathrm{P}(A|B)\mathrm{P}(B),$$
$$\mathrm{P}(B) = \sum_{\{A\}} \mathrm{P}(A,B) = \sum_{\{A\}} \mathrm{P}(B|A)\mathrm{P}(A),$$

where $\{A\}$ and $\{B\}$ in the summations indicate the set of all values of A and B, respectively.

We can illustrate these results easily using any joint probability distribution table. For example, we can use the following numbers, which give the number of males and females who survived or not on the *RMS Titanic*.

	Perished	Survived	
Male	1364	367	1731
Female	126	344	470
	1490	711	2201

There were 2201 people on board, and so we can divide the numbers in each category by 2201 to obtain a joint probability distribution table.

	Perished	Survived	
Male	0.620	0.167	0.786
Female	0.057	0.156	0.214
	0.677	0.323	1.000

Note that the column totals and row totals are provided here and these give the probabilties of a person on board *Titanic* being a male or feamle, or surviving or not, respectively. From the joint probability distribuiotn, we can get the two *conditional* probability distributions. The first tells us the probability of surviving or not given that we know that the person is a Male or a Female.

Sex	Perished	Survived
Male	0.788	0.212
Female	0.268	0.732

The second conditional probability distribution table tells us the probability of being a male or a female given that the person survived or not.

Sex	Perished	Survived
Male	0.915	0.516
Female	0.085	0.484

Using the joint table, we can see that the probability of being a person being both a male and a survivor is

$$\mathrm{P}(\text{Sex} = \text{Male}, \text{Survival} = \text{Survived}) = 0.167.$$

This is equal to the probability of being a male given that the person survived times the probability of being a survivor.

$$\mathrm{P}(\text{Sex} = \text{Male}|\text{Survival} = \text{Survived})\mathrm{P}(\text{Survival} = \text{Survived}) = 0.516 \times 0.323 = 0.167.$$

It is also equal to the probability of being a survivor given that the person is a male

$$\mathrm{P}(\text{Survival} = \text{Survived}|\text{Sex} = \text{Male})\mathrm{P}(\text{Sex} = \text{Male}) = 0.212 \times 0.786 = 0.167.$$

The same holds for any other element of the joint probability table. Using these tables we can also calculate marginal probabilities.

$$\begin{aligned}
\mathrm{P}(\text{Sex} = \text{Male}) &= \mathrm{P}(\text{Sex} = \text{Male}, \text{Survival} = \text{Survived}) + \mathrm{P}(\text{Sex} = \text{Male}, \text{Survival} = \text{Perished}), \\
&= \mathrm{P}(\text{Sex} = \text{Male}|\text{Survival} = \text{Survived})\mathrm{P}(\text{Survival} = \text{Survived}) \\
&\quad + \mathrm{P}(\text{Sex} = \text{Male}|\text{Survival} = \text{Perished})\mathrm{P}(\text{Survival} = \text{Perished}), \\
&= 0.516 \times 0.323 + 0.915 \times 0.677, \\
&= 0.786
\end{aligned}$$

With these elementary results from probability theory, we end up with the following uncontroversial result known as *Bayes' rule*.

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\sum_{\{B\}} \mathrm{P}(A|B)\mathrm{P}(B)}.$$

This can be easily derived as follows:

$$\begin{aligned}
\mathrm{P}(B|A)\mathrm{P}(A) &= \mathrm{P}(A|B)\mathrm{P}(B), \\
\mathrm{P}(B|A) &= \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\sum_{\{B\}} \mathrm{P}(A|B)\mathrm{P}(B)}
\end{aligned}$$

We can use this result to solve elementary probability puzzles like the following[2].

> Box A has 10 lightbulbs, of which 4 are defective. Box B has 6 lightbulbs, of which 1 is defective. Box C has 8 lightbulbs, of which 3 are defective. If we do choose a nondefective bulb, what is the probability it came from Box C?

Here, we are being asked for $\mathrm{P}(\text{Box} = C|\text{Bulb} = \text{Working})$. By Bayes' rule this is

$$\mathrm{P}(\text{Box} = C|\text{Bulb} = \text{Working}) = \frac{\mathrm{P}(\text{Bulb} = \text{Working}|\text{Box} = C)\mathrm{P}(\text{Box} = C)}{\mathrm{P}(\text{Bulb} = \text{Working})},$$

where the denominator here is

$$\begin{aligned}
\mathrm{P}(\text{Bulb} = \text{Working}) = &\mathrm{P}(\text{Bulb} = \text{Working}|\text{Box} = A)\mathrm{P}(\text{Box} = A) \\
&+ \mathrm{P}(\text{Bulb} = \text{Working}|\text{Box} = B)\mathrm{P}(\text{Box} = B) \\
&+ \mathrm{P}(\text{Bulb} = \text{Working}|\text{Box} = C)\mathrm{P}(\text{Box} = C).
\end{aligned}$$

The conditional probabilities of drawing working bulb from each of box A, B, or C, are given by knowing the number of bulbs in each box and the number of defective bulbs in each, and so are $\frac{6}{10}$, $\frac{5}{6}$, $\frac{5}{8}$, respectively.

[2]This problem is taken from *Schaum's Outlines: Probability* (2nd ed., 2000), pages 87-88.

The marginal probabilities of box A, B, and C are $\frac{1}{3}$, $\frac{1}{3}$, and $\frac{1}{3}$. From this, we have

$$\begin{aligned}
\mathrm{P}(\text{Box} = C|\text{Bulb} = \text{Working}) &= \frac{\frac{5}{8}\frac{1}{3}}{\frac{6}{10}\frac{1}{3} + \frac{5}{6}\frac{1}{3} + \frac{5}{8}\frac{1}{3}}, \\
&= \frac{\frac{5}{8}}{\frac{6}{10} + \frac{5}{6} + \frac{5}{8}}, \\
&= 0.304
\end{aligned}$$

Returning now to our joint distribution over Y and θ, i.e., $\mathrm{P}(Y, \theta|\alpha, \beta, n)$, from this, we have the following:

$$\mathrm{P}(\theta|Y, \alpha, \beta, n) = \frac{\mathrm{P}(Y|\theta, n)\mathrm{P}(\theta|\alpha, \beta)}{\int \mathrm{P}(Y|\theta, n)\mathrm{P}(\theta|\alpha, \beta)d\theta.}$$

The left hand side is the posterior distribution. There are three components on the right hand side. First, there is $\mathrm{P}(Y|\theta, n)$. Here, the value of Y and n are known and θ is a free variable, and so $\mathrm{P}(Y|\theta, n)$ is a function over θ. This is the likelihood function over θ that we have already encountered, and which we can also write as $L(\theta|Y, n)$. Second, there is $\mathrm{P}(\theta|\alpha, \beta)$, which is the prior. Like the likelihood function, this is also a function over the set of all possible values of θ. The third component is the denominator, which is the integral of the product of the likelihood function and the prior, integrated over all possible values of θ. This integral is known as the *marginal likelihood*, and is a single value that gives the area under the curve of the function that is the product of the likelihood function and the prior. We can write this as follows.

$$\underbrace{\mathrm{P}(\theta|Y, \alpha, \beta, n)}_{\text{posterior}} = \frac{\overbrace{L(\theta|Y, n)}^{\text{likelihood}}\overbrace{P(\theta|\alpha, \beta)}^{\text{prior}}}{\underbrace{\int L(\theta|Y, n)P(\theta|\alpha, \beta)d\theta}_{\text{marginal likelihood}}}.$$

What this tells us is that having observed the data, the probability distribution over the possible values of θ is the normalized product of the likelihood function and the prior over θ. The prior tells us what values θ could take on in principle. The likelihood function effectively tells us the evidence in favour of any given value of θ according to the data. We multiply these two functions together (the numerator above) and then divide by the area under the curve of this product of functions (the marginal likelihood, which is the denominator). Dividing by the marginal likelihood ensures that the area under the curve of the posterior is *normalized* so that its integral is exactly 1.

Filling out the detail of this equation, we have the following.

$$\begin{aligned}
\mathrm{P}(\theta|Y, \alpha, \beta, n) &\propto \frac{L(\theta|Y, n)P(\theta|\alpha, \beta)}{\int L(\theta|Y, n)P(\theta|\alpha, \beta)d\theta}, \\
&= \frac{\theta^m(1-\theta)^{n-m} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^m(1-\theta)^{n-m} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}, \\
&= \frac{\theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}}{\int \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}d\theta}
\end{aligned}$$

The integral evaluates as follows.

$$\int \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}d\theta = \frac{\Gamma(m+\alpha)\Gamma(n-m+\beta)}{\Gamma(n+\alpha+\beta)}$$

From this, we have

$$\mathrm{P}(\theta|Y, \alpha, \beta, n) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(m+\alpha)\Gamma(n-m+\beta)}\theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}.$$

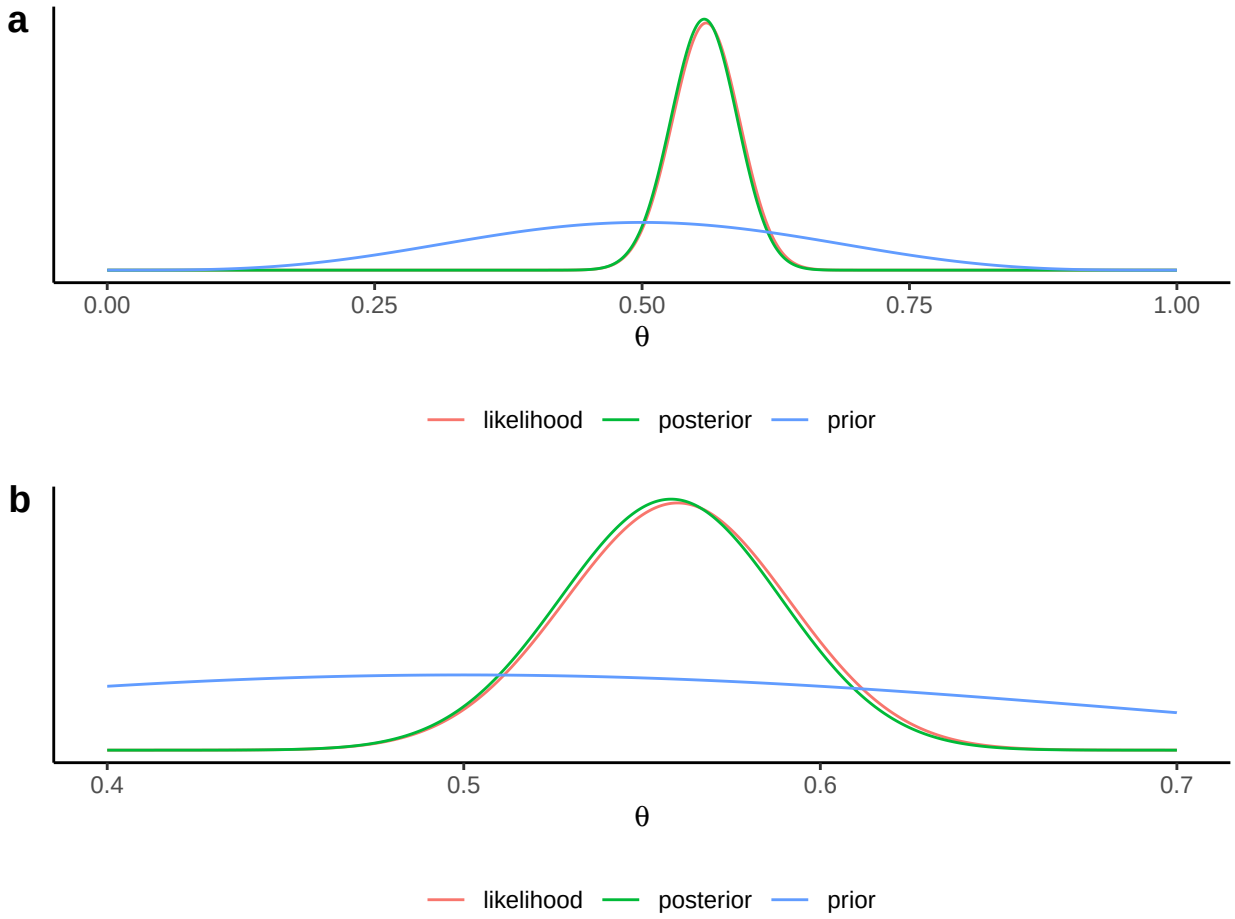


Figure 8: The posterior, likelihood, and prior in a binomial problem with $m = 140$ and $n = 250$ and where the prior is a beta distribution with $\alpha = 5$, $\beta = 5$. In b), the same functions are plotted but over a limited range of the x-axis in order to make the difference between the posterior and the likelihood more apparent. Note also that the likelihood function is scaled so that it integrates to 1. This is simply to make it easier to visualize on the same plot at the prior and posterior. Scaling the likelihood by an arbitrary positive quantity does not affect the calculations of the posterior.

For the case of our data and prior, i.e., $m = 140$, $n = 250$, $\alpha = \beta = 5$, the posterior, likelihood, and prior are shown in Figure 8.

This formula for the posterior distribution is, in fact, a beta distribution with hyperparameters $m + \alpha$ and $n - m + \beta$, respectively. This situation where the posterior distribution is of the same parametric family as the prior is an example of *conjugate prior*. A conjugate prior is a prior that when combined with a particular likelihood function yields a posterior distribution of the same probability distribution family. In this case, the beta distribution prior, when used with the binomial likelihood, always leads to a posterior distribution that is also a beta distribution.

Having a mathematical formula for the posterior distribution, as we do here, is an example of an *analytic solution* to the posterior distribution. This is not always possible. In other words, it is not always possible to have an mathematical expression with a finite number of terms and operations that describes the posterior distribution exactly. When an analytic solution is not possible, and in fact it is only possible in a relatively limited number of cases, we must rely on numerical methods to evaluate the posterior distribution, and in particular, *Monte Carlo* methods. We will return to this important topic in detail below.

Posterior summaries

The result of any Bayesian inference of any unknown variable is the posterior distribution. On the assumption that the statistical model and the prior are valid, the posterior distribution provides us everything that is known about the true value of the variable. For the current problem, this is $\mathrm{P}(\theta|Y = m, \alpha = 5, \beta = 5, n = 250)$, which is a beta distribution with hyperparameters $m + \alpha$ and $n - m + \beta$. From the properties of the beta distribution mentioned above, the mean and standard deviation of this posterior are as follows.

$$\langle\theta\rangle = \frac{\alpha + m}{n + \alpha + \beta} = 0.558, \quad \mathrm{Sd}(\theta) = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}} = 0.031.$$

We can also use the cumulative distribution function of the beta distribution to determine that 2.5th and 97.5% percentiles of the posterior. Between these two bounds, there is 95% of the posterior probability mass. We can calculate this using R with the `qbeta` command, which is the inverse of the cumulative distribution function of the beta distribution.

```
n <- 250
m <- 140
alpha <- 5
beta <- 5
c(qbeta(0.025, m + alpha, n - m + beta),
  qbeta(0.975, m + alpha, n - m + beta))
## [1] 0.4970679 0.6174758
```

This is a *posterior interval*, also known as a *credible interval*. More precisely, this interval is the *central* or *percentile* posterior interval. What it means in practical terms is both simple and important: there is a 95% probability that in the true value of θ is in the interval 0.497, 0.617.

Another posterior interval of interest is known as the *high posterior density* (HPD) interval. The precise definition of the HPD is as follows. The φ HPD interval for the posterior density function $f(\theta)$ is computed by finding a probability density value p^* such that

$$\mathrm{P}(\{\theta\colon f(\theta) \geq p^*\}) = \varphi.$$

In other words, we find the value p^* such that the probability mass of the set of points whose density is greater than than p^* is exactly φ. By this definition, no value of θ outside the HPD has a higher density than any value within the HPD. It will also be the shortest posterior interval containing φ. The HPD as precisely defined here is not easily calculated, and in general requires an optimization procedure to solve for p^*. In this model, using a numerical optimization technique, we calculate it to be (0.4974, 0.6177). This is obviously very close to the quantile based posterior interval defined above.

The posterior mean and the HPD interval can be compared to the maximum likelihood estimator and confidence interval in classical inference. Recall that the maximum likelihood estimator was 0.56 and the 95% confidence interval was (0.496, 0.622). By contrast, the posterior mean and 95% central posterior interval is 0.558 and (0.497, 0.617), respectively. While not identical, these are very close and for practical purposes are probably indistinguishable. This illustrates that classical and Bayesian methods can be, especially in simple models, indistinguishable. This raises the question of why we should use Bayesian methods, given that classical methods are so well established as the default methods. We will return to this issue below.

Monte Carlo sampling

In the example problem that we have been discussing, the prior we chose was a beta distribution. As we've seen, when this prior is used with a binomial likelihood function, the posterior distribution is also a beta distribution. Thus, we have a relatively simple mathematical expression for the posterior distribution. Moreover, we can now use the properties of the beta distribution to determine the posterior mean, standard deviation, posterior intervals, etc. As mentioned, this is an example of an analytic solution to the posterior distribution. This is not always possible. Consider, for example, the posterior distribution when we change

the prior to a *logit normal* distribution. This is a normal distribution over $\log\left(\frac{\theta}{1-\theta}\right)$, which is the log odds of θ. A plot of this prior for the case of a zero mean normal distribution with standard deviation $\tau = 0.5$ is shown in Figure 9a. In Figure 9b, we show a beta distribution over θ with parameters $\alpha = \beta = 8.42$, which is virtually identical to the $N(0, 0.5^2)$ distribution over $\log\left(\frac{\theta}{1-\theta}\right)$.

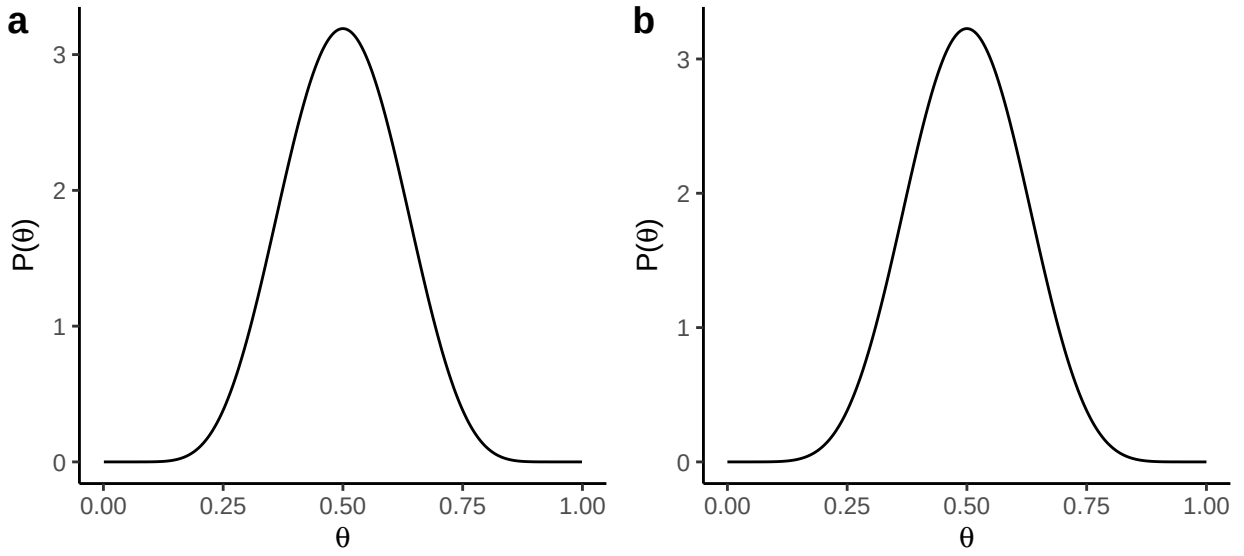


Figure 9: a) A zero mean normal distribution with standard deviation $\tau = 0.5$ over $\log(\frac{\theta}{1-\theta})$. b) A beta distribution with parameters $\alpha = \beta = 8.42$.

Despite the similarities of these two priors, if we use the normal distribution over the log odds, the posterior distribution is as follows

$$\mathrm{P}(\theta|m, n, \tau) = \frac{1}{Z}\theta^m(1-\theta)^{n-m}\frac{1}{\theta(1-\theta)}e^{-\frac{|\mathrm{logit}(\theta)|^2}{2\tau^2}},$$

where

$$Z = \int \theta^m(1-\theta)^{n-m}\frac{1}{\theta(1-\theta)}e^{-\frac{|\mathrm{logit}(\theta)|^2}{2\tau^2}}d\theta,$$

This does not simplify to an analytic expression, and it is not a known probability density function with well documented properties. The problem here arises primarily because of the integral, which does not have an analytic solution. In cases like this, we need to resort to numerical alternatives to the analytic solution. For simple models such as this one, there are many options for how to do this. However, a general approach that applies to all Bayesian models, including and especially complex and high dimensional (i.e. with large numbers of unknown variables) models, is to use *Monte Carlo* sampling methods.

Monte Carlo methods, first introduced by Metropolis and Ulam (1949), can be generally defined as numerical mehods for approximating mathematical *expectations* of random variables. If X is a random variable of dimensionality d whose probability distribution is $\mathrm{P}(X)$, and $\phi(X)$ is a function of X, then the expectation or expected value of $\phi(X)$ is

$$\langle\phi(X)\rangle = \int \phi(x)\mathrm{P}(X = x)dx.$$

This can be approximated by

$$\langle\phi(X)\rangle = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i),$$

where $x_1, x_2 \ldots x_i \ldots x_n$ are n samples from $\mathrm{P}(X)$. Particularly important is the fact that the error of approximation decreases as a function of the $\sqrt{n}$ and is independent of d, the dimensionality of X.

Quantites of interest related to the probability distribution $\mathrm{P}(X)$ such as the mean, variance, or probabilities of being above or below a certain value can all be expressed as expectations:

$$\langle X \rangle = \int x \mathrm{P}(X = x) dx, \quad \mathrm{V}(X) = \int (x - \langle X \rangle)^2 \mathrm{P}(X = x) dx, \quad \mathrm{P}(X > x_0) = \int I(x > x_0) \mathrm{P}(X = x) dx.$$

They can, therefore, be approximated as follows.

$$\langle X \rangle \approx \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \mathrm{V}(x) \approx \mathrm{var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \mathrm{P}(X > x_0) \approx \frac{1}{n} \sum_{i=1}^{n} I(x_i > x_0).$$

What this entails for Bayesian inference is that if we can draw samples from the posterior distribution, Monte Carlo methods can be used to calculate all quantities of interest related to the posterior. Moreover, this can be done in high dimensional problems without encountering the exponential rise in approximation error known as the *curse of dimensionality*.

There are many Monte Carlo methods for drawing samples from posterior probability distributions. Here, we will describe just two: the Metropolis sampler, and the Hamitonian Monte Carlo (HMC) sampler. Both of these are *Markov Chain Monte Carlo* (MCMC) samplers, as we will see. The Metropolis sampler was one of the earliest MCMC samplers, introduced by Metropolis et al. (1953), and has traditionally been one of the most widely used samplers in Bayesian analysis. The HMC sampler is an extension of the Metropolis sampler. It is the main sampler that is used the *Stan* probabilistic programming language that we will use extensively throughout the remainder of this book.

In order to discuss this topic generally, rather than just for a specific problem or model, we will assume that our observed data is $\mathcal{D}$ and that the unknown variable(s) that we are trying to infer is a d dimensional variable θ. The posterior distribution is

$$\mathrm{P}(\theta|\mathcal{D}) = \frac{L(\theta|\mathcal{D})\mathrm{P}(\theta)}{\int L(\theta|\mathcal{D})\mathrm{P}(\theta) d\theta},$$

where $L(\theta|\mathcal{D})$ is the likelihood function and $\mathrm{P}(\theta)$ is the prior. We can rewrite the posteior distribution as

$$\mathrm{P}(\theta|\mathcal{D}) = f(\theta)\frac{1}{Z}, \quad f(\theta) = L(\theta|\mathcal{D})g(\theta), \quad Z = \int L(\theta|\mathcal{D})g(\theta) d\theta.$$

Here, $f(\theta)$ is the unnormalized posterior distribution, where for notational simplicity we drop explicit reference to $\mathcal{D}$, Z is the normalization constant, and $g(\theta)$ is the unnormalized prior distribution.

Metropolis sampler

For the Metropolis sampler, we need only be able to evaluate $f(\theta)$ at any given value of θ, and we do not need to know the value of Z. Evaluating $f(\theta)$ is often very straightforward even in complex models. If we can evaluate the likelihood function at θ and the (normalized or unnormalized) prior at θ, we simply multiply their values together to obtain $f(\theta)$. To draw samples using the metropolis sampler, we begin with an arbitrary initial value $\tilde{\theta}_0$, and then use a *proposal distribution* based around $\tilde{\theta}_0$ to propose a new point in the θ space, which we will call $\tilde{\theta}_*$. We can write this proposal distribution $Q(\tilde{\theta}_*|\tilde{\theta}_0)$. In the standard implementation of the Metropolis sampler, the proposal distribution must be symmetric such that for any two values $\tilde{\theta}_i$ and $\tilde{\theta}_j$, $Q(\tilde{\theta}_i|\tilde{\theta}_j) = Q(\tilde{\theta}_j|\tilde{\theta}_i)$. In the Metropolis-Hasting variant of the Metropolis sampler, as we will see, the proposal distribution need not be symmetric.

Having sampled θ_*, if $f(\tilde{\theta}_*) \geq f(\tilde{\theta}_0)$, we accept that proposed new point and set $\tilde{\theta}_1 = \tilde{\theta}_*$. We then propose a new point in θ space using our proposal distribution centered at $\tilde{\theta}_1$. If, on the other hand $f(\tilde{\theta}_*) < f(\tilde{\theta}_0)$, we accept $\tilde{\theta}_*$ with probability

$$\frac{f(\tilde{\theta}_*)}{f(\tilde{\theta}_0)}.$$

If we accept it, then we set $\tilde{\theta}_1 = \tilde{\theta}_*$ and propose a new point using our proposal distribution centered at $\tilde{\theta}_1$. If we reject $\tilde{\theta}_*$, we new then propose a new point using the proposal distribution centered at $\tilde{\theta}_0$, and repeat the above steps.

Continuing in this way, we produce a sequence of samples $\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2 \ldots$ that are realizations of random variables $\theta_0, \theta_1, \theta_2 \ldots$. Because each variable is dependent on the immediately preceding variable, and not dependent on any variable before that, these random variables form a first-order *Markov chain*. The marginal probability distributions of these variables are $\pi_0(\theta_0), \pi_1(\theta_1), \pi_2(\theta_2) \ldots$. The first distribution, $\pi_0(\theta_0)$, is an arbitrary starting distribution. The nature of the second distribution, $\pi_1(\theta_1)$, can be understood as follows. If $\theta_0 = \tilde{\theta}_a$, then the probability that $\theta_1 = \tilde{\theta}_b$ will be the $Q(\tilde{\theta}_* = \tilde{\theta}_b | \tilde{\theta}_b)$ if $f(\tilde{\theta}_b) > f(\tilde{\theta}_a)$ and will be $Q(\tilde{\theta}_* = \tilde{\theta}_b | \tilde{\theta}_b) \times \frac{f(\tilde{\theta}_b)}{f(\tilde{\theta}_a)}$ otherwise. Thus, $\pi_1(\theta_1)$ is

$$\pi_1(\theta_1) = \int \underbrace{\min\left(\frac{f(\theta_1)}{f(\theta_0)}, 1\right) Q(\theta_1|\theta_0)}_{T(\theta_1|\theta_0)} \pi_0(\theta_0) d\theta_0.$$

By extension, for any i, we have

$$\pi_i(\theta_i) = \int T(\theta_i|\theta_{i-1}) \pi_{i-1}(\theta_{i-1}) d\theta_i.$$

It can be shown that under some general and minimal conditions[3] of Markov chains, the sequence of probability distributions $\pi_0(\theta_0), \pi_1(\theta_1), \pi_2(\theta_2) \ldots$ will converge upon a *unique* distribution that we will label $\pi(\theta)$. This is known as the *invariant* or *stationary* distribution of the Markov chain. Furthermore, if we have a function $\phi(\theta)$ that satisfies

$$T(\theta_i = \tilde{\theta}_b | \theta_{i-1} = \tilde{\theta}_a) \phi(\theta_{i-1} = \tilde{\theta}_a) = T(\theta_i = \tilde{\theta}_a | \theta_{i-1} = \tilde{\theta}_b) \phi(\theta_{i-1} = \tilde{\theta}_b)$$

for any two states $\tilde{\theta}_a$ and $\tilde{\theta}_b$, then $\phi(\theta)$ is this invariant distribution. We can see that the $\phi(\theta)$ is the posterior distribution as follows. First, we will assume that $f(\tilde{\theta}_a) > f(\tilde{\theta}_b)$. If this is not the case, we simply reverse the labels of $\tilde{\theta}_a$ and $\tilde{\theta}_b$. Then we have the following.

$$\begin{aligned} T(\tilde{\theta}_b|\tilde{\theta}_a)\phi(\tilde{\theta}_a) &= T(\tilde{\theta}_a|\tilde{\theta}_b)\phi(\tilde{\theta}_b), \\ \frac{f(\theta_b)}{f(\theta_a)} Q(\tilde{\theta}_b|\tilde{\theta}_a)\phi(\tilde{\theta}_a) &= Q(\tilde{\theta}_a|\tilde{\theta}_b)\phi(\tilde{\theta}_b). \end{aligned}$$

Because of symmetry of the proposal distribution, we have

$$\begin{aligned} \frac{f(\theta_b)}{f(\theta_a)} Q(\tilde{\theta}_b|\tilde{\theta}_a)\phi(\tilde{\theta}_a) &= Q(\tilde{\theta}_b|\tilde{\theta}_a)\phi(\tilde{\theta}_b), \\ \frac{f(\theta_b)}{f(\theta_a)} \phi(\tilde{\theta}_a) &= \phi(\tilde{\theta}_b), \\ \frac{f(\theta_b)}{f(\theta_a)} &= \frac{\phi(\tilde{\theta}_a)}{\phi(\tilde{\theta}_b)}. \end{aligned}$$

From this we have

$$\phi(\theta) = \frac{1}{Z} f(\theta)$$

being the invariant distribution of the Markov chain.

In Figure 10, we show some of the sequence of distributions $\pi_0, \pi_1, \pi_2 \ldots$ of a Metrpolis sampler as it converges to the true posterior distribution, which is also shown. For this illustration, we use a binomial problem like above but where the data is $m = 14$ and $n = 25$, use a logit-Normal prior, and assume a uniform proposal distribution. From this illustration, we see a relatively quick convergence to the true posterior distribution.

[3]These are *irreducibility*, which is the non-zero probability of eventually transitioning from any one state to any other, *aperiodicity*, which is having no perfectly cyclic state trajectories, and *non transience*, which is the non-zero probability of returning to any given state.

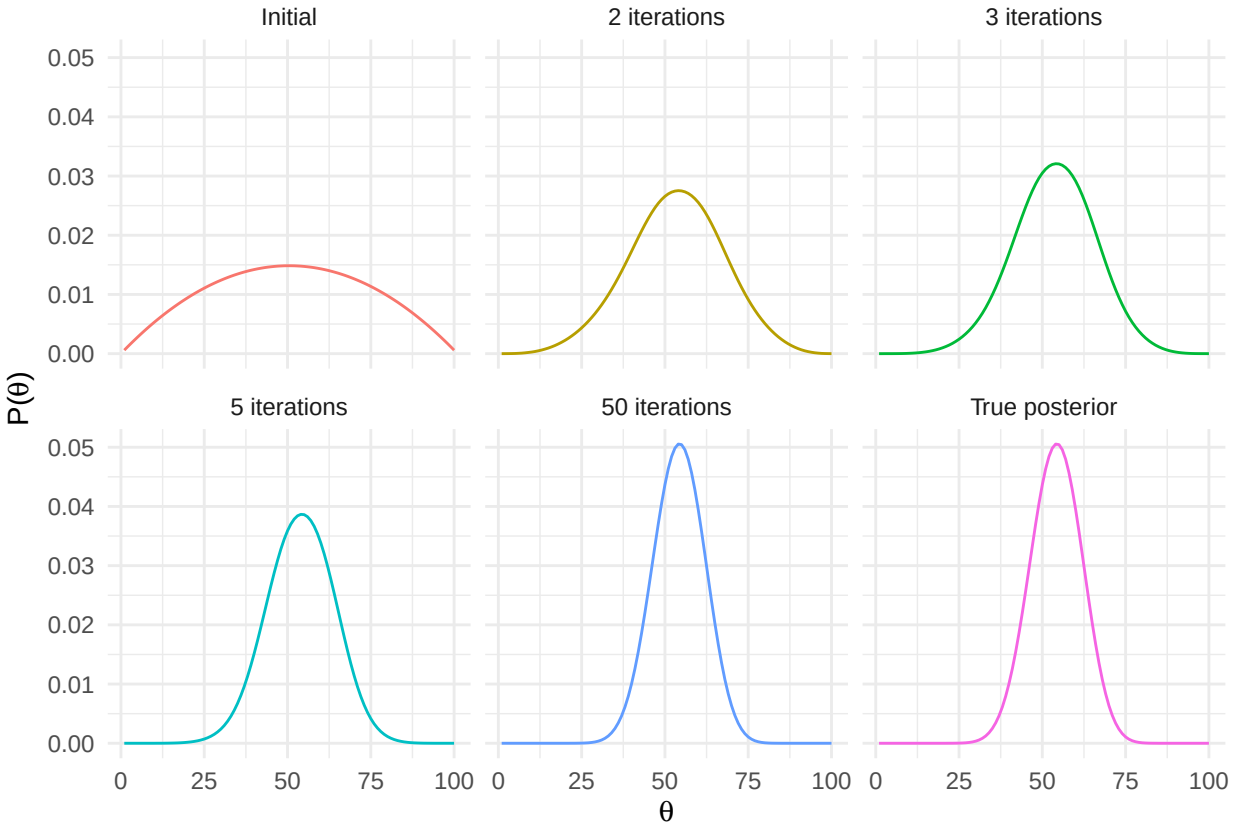


Figure 10: Convergence of a Metropolis sampler to the posterior distribution having started at an arbitrary distribution.

The Metropolis-Hastings variant of the original Metropolis sampler allows for asymmetric proposal distributions. Given an initial sample $\tilde{\theta}_i$ and proposed new sample $\tilde{\theta}_*$, sampled from the proposal distribution $Q(\tilde{\theta}_*|\tilde{\theta}_i)$, then we accept $\tilde{\theta}*$ if

$$f(\tilde{\theta}_*)Q(\tilde{\theta}_i|\tilde{\theta}_*) \geq f(\tilde{\theta}_i)Q(\tilde{\theta}_*|\tilde{\theta}_i),$$

and otherwise we accept it with probability

$$\frac{f(\tilde{\theta}_*)Q(\tilde{\theta}_i|\tilde{\theta}_*)}{f(\tilde{\theta}_i)Q(\tilde{\theta}_*|\tilde{\theta}_i)}.$$

Sampling in this way, as with the original Metropolis sampler, we converge upon an invariant probability distribution over θ that is the posterior distribution $\frac{1}{Z}f(\theta)$.

To illustrate this algorithm, we will use our binomial problem with the logit normal distribution. The likelihood function at any value of θ is

$$L(\theta|m,n) = \theta^m(1-\theta)^{n-m}.$$

The unnormalized logit normal prior is

$$g(\theta) = \frac{1}{\theta(1-\theta)}e^{-\frac{|\text{logit}(\theta)|^2}{2\tau^2}}.$$

From this, we have

$$f(\theta) = \theta^{m-1}(1-\theta)^{n-m-1}e^{-\frac{|\text{logit}(\theta)|^2}{2\tau^2}}.$$

This can be easily implemented using R as follows:

```
logit <- function(x) log(x/(1-x))
f <- function(theta, m=140, n = 250, tau = 0.5){
  theta^(m-1) * (1-theta)^(n-m-1) * exp(-logit(theta)^2/(2*tau^2))
}
```

For the proposal distribution, for simplicity, we will use uniform distribution on $(0, 1)$.

```
proposal <- function(theta_0){
  runif(1)
}
```

The following code implements a Metropolis sampler that draws 100 samples.

```
nsamples <- 100
i <- 1
theta <- numeric(length = nsamples)
theta[i] <- runif(1) # initial sample from U(0,1)

while (i < nsamples) {
  theta_star <- proposal(theta[i]) # proposed sample
  p <- min(f(theta_star)/f(theta[i]), 1) # prob. of acceptance
  if (runif(1) <= p){
    i <- i + 1
    theta[i] <- theta_star
  }
}
```

The trajectory of samples from two separate chains, each started at opposite ends of the θ space, are shown in Figure 11a. These plots are known as *trace-plots*. As can be seen, these trajectories both quickly converge upon the same area of θ space. The histogram of samples from one chain is shown in Figure 11b.

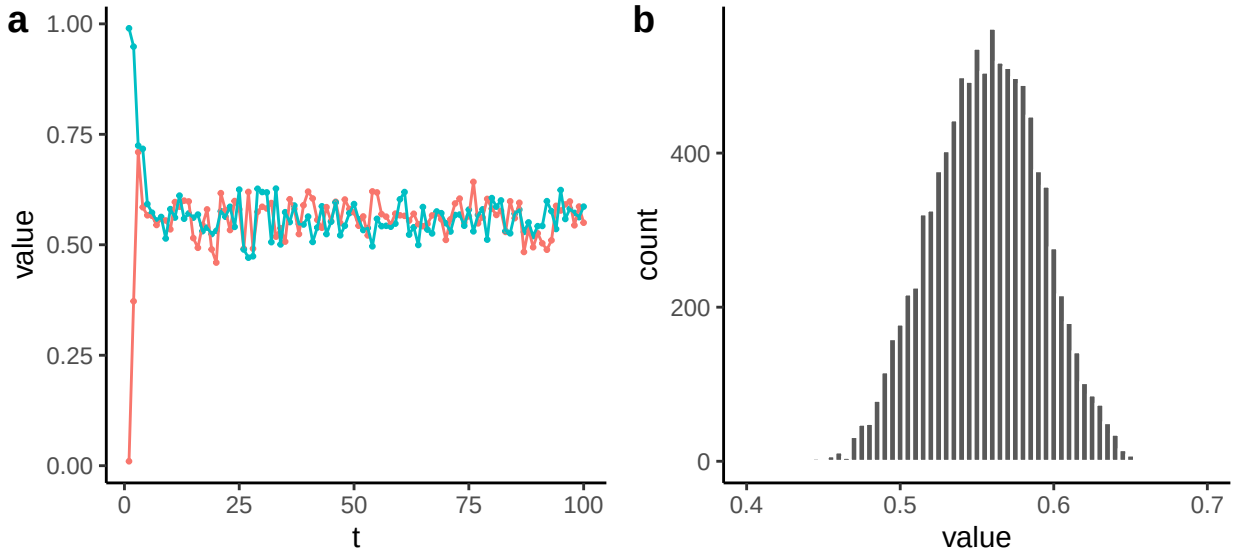


Figure 11: a) The trajectory, or *trace-plot*, of 100 samples from two separate chains of a Metropolis sampler for the binomial problem using a logit-normal prior. b) The histogram of ten thousand samples from one of the samplers.

Hamiltonian Monte Carlo

The Metropolis algorithm suffers from the fact that its proposal distribution takes random steps in parameter space. The *Hamiltonian Monte Carlo* (HMC), which was originally known as *Hybrid Monte Carlo*, aims to overcome this limitation by taking account of the gradient of the posterior distribution when making its proposals. Both theoretically and in terms of its technical details, HMC is considerably more complex than the Metropolis method (see Betancourt 2017; Neal 2011). However, HMC is extremely efficient and it is now widely used in Bayesian data analysis, and so any contemporary introduction of MCMC would be incomplete without covering it.

To appreciate HMC, we return to the fact that in any Bayesian analysis, our results are ultimately expectations of functions according to the posterior distribution. Thus, if the posterior distribution over our unknown variables is $\mathrm{P}(\theta|\mathcal{D})$, the expected value of some function $\phi(\theta)$ according to the posterior is

$$\langle \phi(\theta) \rangle = \int \phi(\theta) \mathrm{P}(\theta|\mathcal{D}) d\theta.$$

The value of this integral is largely determined by where $\mathrm{P}(\theta|\mathcal{D})$ is massed. In high dimensional spaces, most of this mass is concentrated in a thin shell far from the mode of the distribution that is known as the *typical set* (see Betancourt 2017). All MCMC methods ideally sample from this typical set. However, because the typical set is a thin shell, the random proposals made by the Metropolis sampler will be mostly to regions of low probability density outside of it, which will be rejected, leading to inefficiency. Likewise, those proposals that will be accepted will often be just those close to the current state of the sampler, and in this sense, the sampler becomes stuck in one region and does not explore the typical set.

HMC tries to overcome these problems with the random walk Metropolis by using the gradient of the posterior distribution to guide its proposals of new points to sample. For any given value of θ, the gradient of the posterior distribution is evaluated. This indicates the directions with higher or lower posterior density. Were we to simply follow the gradient, we would end up moving away from the typical set and toward the mode of the distribution. While the mode will have high density, it will have infinintesmally low volume in high dimensional problems, and so is not part of the typical set as defined above. In order to remain within the typical set, therefore, each point in θ space is given a *momentum* value denoted by r, which gives a direction and velocity to each point. This momentum speeds up as the object approaches areas of high density, and slows and reverses in regions of low density.

In order to implement this system, HMC represents the value of θ as a point in a physical space that has a *potential energy* and a *kenetic energy*. The potential energy is determined by its position, and can seen as the effect of gravity. The kenetic energy is determined by the momentum of the point. Intuitively, we can imagine this as smooth (frictionless) hilly landscape with a ball with a momentum moving over this landscape. If we kick the ball, we give it some momentum and it will move in a certain direction with a certain speed. If that direction is downwards, under the effect of gravity, it will speed up and then when it gets to the bottom of the hill, its momentum will allow it to continue moving and it will start to roll up the side of another hill. Eventually, the effect of gravity will start slow it down, and it will then reverse and roll back down the hill, and so on. The movement of the ball is physically determined by its position and its momentum and their respective potential and kenetic energies.

In more detail, the *potential energy* at each point is given by

$$U(\theta) = -\log(f(\theta)),$$

where $f(\theta) \propto \mathrm{P}(\theta|\mathcal{D})$ as defined above. This is equivalent to

$$\mathrm{P}(\theta|\mathcal{D}) = \frac{1}{Z} e^{-U(\theta)}.$$

From this perspective, the HMC sampler can seen as a physical system with an infinite number of states, each of which have a potential energy associated with it. The probability of the system being in any state is determined by the potential energy associated with that state[4]. The lower the energy, the more likely the

[4]In statistical mechanics, a physical system where the probability of being in any given state is proportional to e to the power of the negative of the energy at that state is said to have a *Boltzmann distribution*.

system is to be in that state. The higher the energy, the less likely it is to be in that state. In addition to its potential energy, each state also has a *kenetic energy*, denoted $K(r)$, which is determined by the momentum with value r. In one dimension, if the point has unit mass, the kenetic energy is $K(r) = r^2/2$ and in d dimensions, it is

$$K(r) = \frac{1}{2}\sum_{k=1}^{d} r_k^2.$$

Given this energy function, the probability distribution corresponding to it is $P(r) \propto e^{-K(r)}$, which is a d dimensional standard normal distribution. The total energy in this system is given by the *Hamitonian* function:

$$H(\theta, r) = U(\theta) + K(r).$$

The change in the position and the momentum is now determined by classical mechanics and is given the Hamiltonian equations:

$$\begin{aligned} \frac{d\theta_k}{dt} &= \frac{\partial H}{\partial r_k} &&= \frac{\partial K(r)}{\partial r_k}, \\ \frac{dr_k}{dt} &= -\frac{\partial H}{\partial \theta_k} &&= -\frac{\partial U(\theta)}{\partial \theta_k}, \end{aligned}$$

where $\frac{d\theta_k}{dt}$ and $\frac{dr_k}{dt}$ are the rate of change, at dimension k, of the position θ and momentum r, respectively. The $\frac{\partial H}{\partial r_k}$ and $\frac{\partial H}{\partial \theta_k}$ are the partial derivatives of the Hamiltonian function with respect to r_k and θ_k, respectively. Now, if we start at any point in θ space and give this point a momentum, the Hamiltonian equations will determine how the point will move through this space under the actions of both the potential and kenetic energy.

In order to simulate this continuous dynamical system, which is governed by differential equations, on a computer, we must use a discrete approximation to its dynamics. Widely used methods such as *Runge-Kutta methods* or *Euler's method* are possible, but a more suitable method is the *leapfrog* algorithm, which involves taking steps, or half-steps, of size δ as follows:

$$\begin{aligned} r_k &\leftarrow r_k - \frac{\delta}{2}\frac{\partial U(\theta)}{\partial \theta_k}, && \text{first half-step in } r \text{ space}, \\ \theta_k &\leftarrow \theta_k + \delta \frac{\partial K(r)}{\partial r_k}, && \text{step in } \theta \text{ space}, \\ r_k &\leftarrow r_k - \frac{\delta}{2}\frac{\partial U(\theta)}{\partial \theta_k}, && \text{second half-step in } r \text{ space}. \end{aligned}$$

As a simple example of Hamiltonian dynamics, let us assume that the posterior distribution is a two dimensional normal distribution with a mean μ and covariance matrix Σ. In this case, we have

$$U(\theta) = \tfrac{1}{2}(\theta - \mu)^\intercal \Sigma^{-1}(\theta - \mu),$$

and we will assume that $K(r) = \frac{1}{2}\sum_{k=1}^{d} r_k^2$. The partial derivatives are

$$\frac{\partial U(\theta)}{\partial \theta_k} = (\theta - \mu)^\intercal \Sigma^{-1}, \quad \frac{\partial K(r)}{\partial r_k} = r_k,$$

and so our leapfrog steps are

$$r_k \leftarrow r_k - \frac{\delta}{2}(\theta - \mu)^\intercal \Sigma^{-1}, \quad \theta_k \leftarrow \theta_k + \delta r_k.$$

Starting with $\theta = (0, 1.5)$, and choosing r at random from a $2d$ standard normal, we simulate the Hamiltonian dynamics for a 45 leapfrog steps with a $\delta = 0.1$. This shown in Figure 12. As we can see, the trajectories move smoothly back and forth through the posterior distribution.

The key feature of HMC that differentiates it from random walk Metropolis is that it replaces the random proposal distribution with the deterministic dynamics of the Hamiltonian system. Specifically, a point in θ

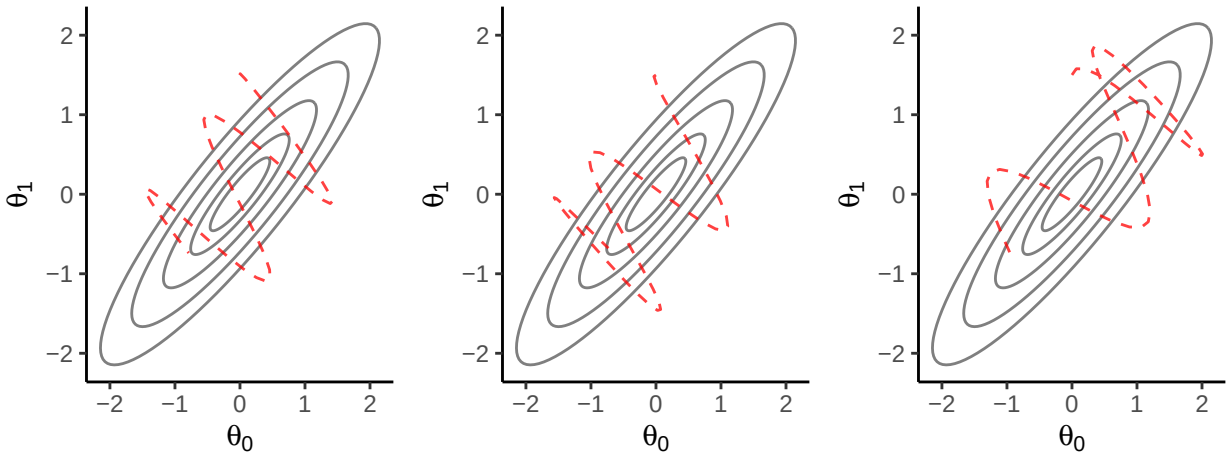


Figure 12: Three trajectories of a Hamitonian dynamical system where the posterior distribution is a $2d$ normal distribution, where each trajectory starts at $\theta = (0, 1.5)$, and where we choose the momentum r at random from a $2d$ standard normal. The system is simulated for a 45 leapfrog steps with a $\delta = 0.1$.

space is chosen at random, and a value of the momentum variable r is chosen from the probability distribution corresponding to $K(r)$, i.e. $P(r) \propto e^{-K(r)}$, the Hamiltonian system deterministically evolves for period of time to arrive at a new point $\theta_\star$ and new momentum $r_\star$. The new point $\theta_\star$ is then accepted if $H(\theta_\star, r_\star) > H(\theta, r)$. Otherwise, $\theta_\star$ is accepted with probability

$$e^{H(\theta,r) - H(\theta_\star, r_\star)} = \frac{\mathrm{P}(\theta_\star|\mathcal{D})}{\mathrm{P}(\theta|\mathcal{D})} \frac{\mathrm{P}(r_\star)}{\mathrm{P}(r)}.$$

This acceptance-step step is essentially identical to the Metropolis-Hastings acceptance-rejection step.

One final issue to address concerns how long the trajectories of the deterministic Hamiltonian dynamics ought to be. If the trajectories are too short, the sampler will move slowly through the posterior distribution. If the trajectory is longer, as we can see in Figure 12, the trajectory may loop back upon itself. The extent to which the trajectories will loop around will depend on the curvature of the $U(\theta)$ space: In flat regions, the loop arounds will not happen, but will happen sooner in curved regions. The *No-U-Turn* (NUTS) sampler by Hoffman and Gelman (2014) optimally tunes the trajectory lengths according to the local curvature. This allows optimal trajectory lengths without manual tuning or without tuning runs of the sampler. The probabilistic programming language Stan, to which Chapter 16 is devoted, is HMC sampler that uses NUTS.

As an illustration of a HMC sampler, in fact based on the Stan language, we will use the `brms` package (Bürkner 2018).

```
library(brms)
```

This package allows us to implement a very wide range of models using minimal and familiar R command syntax and creates, compiles, and executes the Stan sampler. We will use `brms` based models often in the remaining chapters. We can implement the logit normal prior based binomial model mentioned above as follows.

```
M_hmc <- brm(m | trials(n) ~ 1,
             data = tibble(n = n, m = m),
             prior = prior(normal(0, 0.5), class = Intercept),
             family = binomial)
```

By default, this will sample 4 independent chains, each with 1000 post-*warmup* samples. The warmup iterations are iterations prior to convergence on the typical set, and during this period, the HMC sampler is fine tuned. The summary of this model is as follows.

```
summary(M_hmc)
##  Family: binomial
##   Links: mu = logit
## Formula: m | trials(n) ~ 1
##    Data: tibble(n = n, m = m) (Number of observations: 1)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.23      0.13    -0.02     0.48 1.00     1463     2003
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The histogram of the 4000 posterior samples of θ is shown in Figure 13.

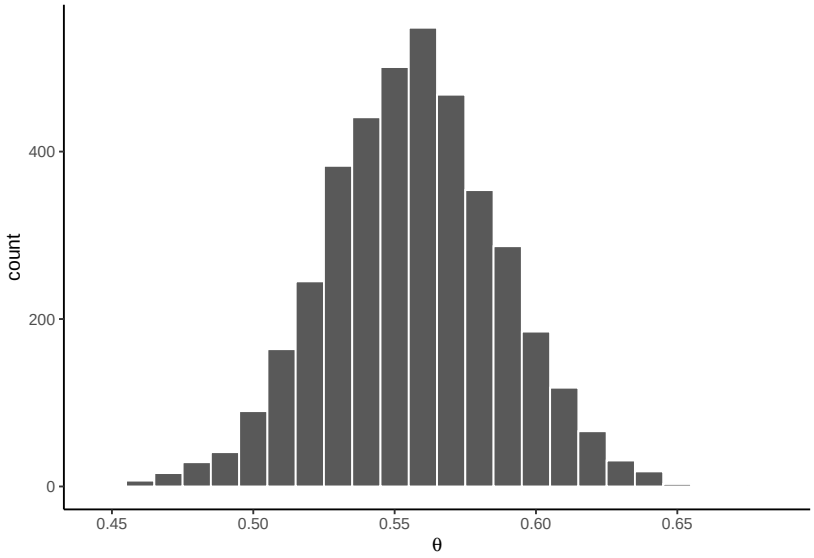
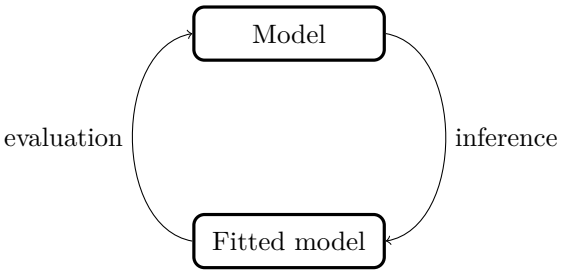


Figure 13: Histogram of the samples from the posterior distribution over θ from a HMC sampler of a *logit-normal* prior based binomial model.

Model evaluation

Let us return to the following diagram from the introduction to this chapter.



In our coverage of classical and then Bayesian methods of inference, we have dealt with the *inference* arc to right. This is where we begin with a statistical model that has variables or parameters whose values are unknown, and then use one general approach or another to effectively obtain estimates, as well as measures of uncertainty of these estimates, of these unknown variables. The result is a model that it fitted to the data. However, whether it is a good fit, and more importantly, whether the model is able to generalize to new data, remains to be seen. *Model evaluation* is the general term for this stage of the modelling process. Model evaluation is ultimately a large and multifaceted topic, involving various quantitative and graphical methods. Here, we will focus on the most common methods, and in particular, on *model comparison*, which is where we directly compare two or more models against one another.

Deviance and Log likelihood ratio tests

Let us first consider a widely used approach for model comparison for classical inference based models. We will call the model that we are evaluating $\mathcal{M}_1$, whose vector of unknown parameters is θ_1 and the maximum likelihood estimator of these values based on observed data $\mathcal{D}$ is $\hat{\theta}_1$. The probability of $\mathcal{D}$ according to $\mathcal{M}_1$ and $\hat{\theta}_1$ is

$$\mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1).$$

Note that this the value of the likelihood function for the model evaluated at its maximum value. In itself, this value is actually not very informative about how well $\mathcal{D}$ is predicted by $\mathcal{M}_1$ with parameters $\hat{\theta}_1$. It is almost always a very small value based on the fact that there is a very large set of possible values that the model could predict. However, it is more informative to compare the relative probabilities of the data according to two models that we wish to compare. If we name the comparison model $\mathcal{M}_0$, and name its parameter vector θ_0 and its estimator of these parameters $\hat{\theta}_0$, then the ratio of the probabilities of $\mathcal{D}$ according to these two models is

$$\frac{\mathrm{P}(\mathcal{D}|\mathcal{M}_0, \hat{\theta}_0)}{\mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1)}.$$

This ratio is usually referred to as the likelihood ratio. Note that it is specifically the ratio of the maximum values of the likelihood functions of the two models. The logarithm of this likelihood ratio is

$$\log\left(\frac{\mathrm{P}(\mathcal{D}|\mathcal{M}_0, \hat{\theta}_0)}{\mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1)}\right) = \log \mathrm{P}(\mathcal{D}|\mathcal{M}_0, \hat{\theta}_0) - \log \mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1).$$

For reasons that will become clear, we usually multiple this logarithm by -2 to obtain the following.

$$\begin{aligned} -2\log\left(\frac{\mathrm{P}(\mathcal{D}|\mathcal{M}_0, \hat{\theta}_0)}{\mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1)}\right) &= -2\log \mathrm{P}(\mathcal{D}|\mathcal{M}_0, \hat{\theta}_0) - -2\log \mathrm{P}(\mathcal{D}|\mathcal{M}_1, \hat{\theta}_1), \\ &= D_0 - D_1. \end{aligned}$$

Here, D_0 and D_1, which are simply -2 times the logarithm of the corresponding likelihoods, are known as the *deviance* of model $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. Note that because of the negative sign of the multiplier, the larger the deviance the lower the likelihood of the model. Thus, if $D_0 > D_1$, model $\mathcal{M}_1$ is better able to predict the data than model $\mathcal{M}_0$.

Differences of deviances, and so the log likelihood ratio, are particularly widely used when comparing *nested models*. Nested models are where one model's parameter space is subset of that of another. For example, if $\mathcal{M}_1$ is a normal distribution model with mean and variance parameters μ_1 and σ_1^2, both of which are unknown, and $\mathcal{M}_0$ is also a normal distribution but with fixed mean $\mu_0 = 0$ and only σ_0^2 being unknown, then $\mathcal{M}_0$ is nested in $\mathcal{M}_1$ because $\{\mu_0 = 0, \sigma_0^2\} \subseteq \{\mu_1, \sigma_1^2\}$. When comparing nested models, we can make use of *Wilks's theorem* that states that, of models $\mathcal{M}_0$ and $\mathcal{M}_1$ predict $\mathcal{D}$ equally well, then asymptotically (i.e., as sample size increases)

$$\Delta_D = D_0 - D_1 \sim \chi^2_{k_1 - k_0},$$

where k_1 and k_0 are the number of parameters in $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively. In other words, if $\mathcal{M}_0$ and $\mathcal{M}_1$ predict the data equally well, then the difference of their deviances, which would be due to sampling variation

alone, will be distributed as a χ^2 distribution whose degrees of freedom are equal to the difference of the number of parameters in the two models.

Wilks's theorem is widely used in regression models, as we will see in subsequent chapters. However, for now, we can illustrate with a simple example. Let us return to the `houseprices_df` data above. One model, $\mathcal{M}_1$, of the `price` variable could that its logarithm is normally distributed with some mean μ_1 and variance σ_1^2, both of which are unknown. A nested model, $\mathcal{M}_0$, could be that the logarithm of `price` is normally distributed with a fixed mean $\mu_0 = \log(60000)$ and unknown variance σ_0^2. Without providing too much details of the R commands, which we will cover in more detail in subsequent chapters, we can fit these two models as follows.

```
M_1 <- lm(log(price) ~ 1, data = housing_df)
mu_0 <- rep(log(60000), nrow(housing_df))
M_0 <- lm(log(price) ~ 0 + offset(mu_0), data = housing_df)
```

The logarithms of the likelihoods of `M_1` and `M_0` at the maximum values are as follows.

```
logLik(M_1)
## 'log Lik.' -234.2995 (df=2)
logLik(M_0)
## 'log Lik.' -240.6163 (df=1)
```

The corresponding deviances are -2 times these values, which we may also obtain as follows.

```
(d_1 <- -2 * logLik(M_1))
## 'log Lik.' 468.599 (df=2)
(d_0 <- -2 * logLik(M_0))
## 'log Lik.' 481.2327 (df=1)
```

The difference of the deviances Δ_D is as follows.

```
delta_d <- as.numeric(d_0 - d_1)
delta_d
## [1] 12.6336
```

Assuming $\mathcal{M}_1$ and $\mathcal{M}_0$ predict the observed data equally well, this difference is distributed as a χ^2 distribution with 1 degree of freedom. This value for the degrees of freedom is based on the fact that one model has two paramters and the other has just one. We can calculate the p-value for Δ_D by calculating the probability of obtaining a result as or more extreme than Δ_D is a χ^2 distribution with 1 degree of freedom.

```
pchisq(delta_d, df = 1, lower.tail = F)
## [1] 0.0003788739
```

Simply put, this result tells us that a value of $\Delta_D = 12.63$ is not an expected result if models $\mathcal{M}_1$ and $\mathcal{M}_1$ predict the data equal well, as so we can conclude that the $\mathcal{M}_1$, which has the lower deviance, predicts the data significantly better than $\mathcal{M}_0$.

Cross validation and out-of-sample predictive performance

Although widely used and very useful, the deviance based model comparison just described is limited to both classical methods and to nested models. It is also limited in that it examines how well any pair of models can predict the observed data on which the model was fitted. A more important question is how the any model can generalize to new data that is from the same putative source as the original data. This is known as *out-of-sample* predictive performance. Any model may in fact predict the data on which it was fitted well, or even perfectly, and yet poorly generalize because it is too highly tuned to the pecularities or essentially random element of the data. This is known as *over-fitting* and it is a major problem, especially with complex models. To properly evaluate a model and identify any over-fitting we need to see how well the model can generalize to new data. Rather than waiting for new data to be collected, a simple solution is to remove

some data from the data that is used for model fitting, fit the model with the remaining data, and then test how well the fitted model predicts the reserved data. This is known as *cross-validation*.

One common approach to cross-validation is known as K-fold cross-validation. The original data set is divided randomly into K subsets. One of these subsets is randomly selected to be reserved for testing. The remaining $K-1$ are used for fitting and the generalization to the reserved data set is evaluated. This process is repeated for all K subsets, and overall cross validation performance is the average of the K repetitions. One extreme version of K-fold cross-validation is where $K = n$ where n is the size of the data-set. In this case, we remove each one of the observations, fit the model on the remaining set, and test on the held out observed. This is known as *leave one out* cross-validation.

For leave one out cross-validation, the procedure is as follows, with the procedure for any K-fold cross-validation being similarly defined. Assuming our data is $\mathcal{D} = y_1, y_2 \dots y_n$, we divide the data into n sets:

$$(y_1, y_{\neg 1}), (y_2, y_{\neg 2}), \dots (y_i, y_{\neg i}) \dots (y_n, y_{\neg n}),$$

where y_i is data point i and $y_{\neg i}$ is all the remaining data except for data point i. Then, for each i, we fit the model using $y_{\neg i}$ and test how well the fitted model can predict y_i. We then calculate the sum of the predictive performance over all data points. In classical inference based models, for each i, we calculate $\hat{\theta}^{\neg i}$, which is the maximum likelihood or other estimator of the parameter vector θ based on $y_{\neg i}$. We then calculate $\log \mathrm{P}(y_i|\hat{\theta}^{\neg i})$, which is the logarithm of the predicted probability of y_i based on the model with parameters $\hat{\theta}^{\neg i}$. This then leads to

$$\text{elpd} = \sum_{i=1}^{n} \log \mathrm{P}(y_i|\hat{\theta}^{\neg i})$$

as the overall measure of the model's out-of-sample predictive performance, which we refer to as *expected log predictive density* (ELPD). In Bayesian approaches, an analogous procedure is followed. For each i, we calculate the posterior distribution $\mathrm{P}(\theta|y_{\neg i})$, and the model's overall out-of-sample predictive performance is

$$\text{elpd} = \sum_{i=1}^{n} \log \int \mathrm{P}(y_i|\theta)\mathrm{P}(\theta|y_{\neg i}) d\theta.$$

This can be approximated by

$$\text{elpd} \approx \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} \mathrm{P}(y_i|\tilde{\theta}_s^{\neg i}) \right),$$

where $\tilde{\theta}_1^{\neg i}, \tilde{\theta}_2^{\neg i} \dots \tilde{\theta}_S^{\neg i}$ are S samples from $\mathrm{P}(\theta|y_{\neg i})$.

In order to illustrate this, let us use the `price` variable in the `housing_df` data. One model for this data that we can consider is, as described in the introduction to this chapter, is as follows $i \in 1 \dots n$, $y_i \sim \text{logN}(\mu, \sigma^2)$, where $\text{logN}(\mu, \sigma^2)$ is a log-normal distribution whose parameters are μ and σ^2. Using a classical inference based model, this can be implemented in R as follows.

```
m1 <- lm(log(price) ~ 1, data = housing_df)
```

We can remove any $i \in 1 \dots n$ and fit the model with the remaining data as follows, using $i = 42$ as an example.

```
i <- 42
m1_not_i <- lm(log(price) ~ 1,
               data = slice(housing_df, -i)
)
```

The logarithm of the probability of y_i is logarithm of the normal density for $\log y_i$ with the mean and standard deviation based on their maximum likelihood estimators in the fited model. We can extract the maximum likelihood estimators of the mean and standard deviation from `m1_not_i` as follows.

```
mu <- coef(m1_not_i)
stdev <- sigma(m1_not_i)
```

Then , the logarithm of normal density of $\log y_i$ based on these estimators is

```
y_i <- slice(housing_df, i) %>% pull(price)
dnorm(log(y_i), mean = mu, sd = stdev, log = T)
## [1] 0.03483869
```

We can create a function to calculate the logarithm of the prediction for any $\log y_i$ as follows.

```
logprediction_m1 <- function(i){
  m1_not_i <- lm(log(price) ~ 1,
                 data = slice(housing_df, -i)
  )
  mu <- coef(m1_not_i)
  stdev <- sigma(m1_not_i)
  y_i <- slice(housing_df, i) %>% pull(price)
  dnorm(log(y_i), mean = mu, sd = stdev, log = T)

}
```

We can then apply this to all data points and sum the result.

```
n <- nrow(housing_df)
map_dbl(seq(n), logprediction_m1) %>%
  sum()
## [1] -236.2388
```

Now, let us compare this log-normal model's performance to a normal model, i.e., where we assume that for $i \in 1 \ldots n$, $y_i \sim \mathrm{N}(\mu, \sigma^2)$. The log of the prediction of the held out data point is calculated in almost an identical manner, except that we use the values of the `price` variable directly and not their logarithm.

```
logprediction_m0 <- function(i){
  m0_not_i <- lm(price ~ 1,
                 data = slice(housing_df, -i)
  )
  mu <- coef(m0_not_i)
  stdev <- sigma(m0_not_i)
  y_i <- slice(housing_df, i) %>% pull(price)
  dnorm(y_i, mean = mu, sd = stdev, log = T)

}

map_dbl(seq(n), logprediction_m0) %>%
  sum()
## [1] -6342.375
```

It is common to multiply these sums of log predictions by -2 to put the values on a deviance scale. This is sometimes known as the leave-one-out-information-criterion (LOOIC). Thus, our measure of the out-of-sample predictive deviance of the log-normal model is 472.48, and for the normal model, it is 12684.75. The log-normal model is clearly the better model with a much lower deviance score. In the next section, we will consider how best to interpret differences in LOOIC and related quantities.

To perform leave one out cross validation using Stan models, we can make use of an efficient implementation of this based on *Pareto-smoothed importance sampling* (Vehtari, Gelman, and Gabry 2017). This means that we do not have to manually divide the data set into subsets, as we did in the previous example. In fact,

this efficient method allows us to effectively implement the leave one out cross validation method without repeatedly re-running the HMC models, which would be extremely computationally burdensome. All we need do is implement the Stan model, which we can do easily using `brm` as follows.

```
m1_bayes <- brm(log(price) ~ 1, data = housing_df)
m0_bayes <- brm(price ~ 1, data = housing_df)
```

We may then get the ELPD and the LOOIC using the command `loo`.

```
loo(m1_bayes)$estimates
##             Estimate         SE
## elpd_loo -236.334568 15.9128481
## p_loo       1.983721  0.1665893
## looic     472.669135 31.8256962
loo(m0_bayes)$estimates
##             Estimate         SE
## elpd_loo -6341.91351 23.4794215
## p_loo        3.02761  0.6239585
## looic    12683.82701 46.9588429
```

Leaving aside `p_loo` (a measure of effective number of parameters) and the standard errors (a measure of uncertainity of the estimates, as defined above), we see here that the `elpd_loo` and the `looic` estimates from the Bayesian models are very similar to those calculated using the classical inference based models above.

AIC

Cross-validation is an excellent method of model evaluation because it addresses the central issue of out-of-sample generalization, rather than fit to the data, and can be applied to any models, regardless of whether these models are based on classical or Bayesian methods of inference. On the other hand, traditionally, cross validation has been seen as too computationally demanding to be used in all data analysis situations. This is becoming less of a concern now, both because of the computational power and the development of efficient implementation such as the Pareto smoothed importance sampler method mentioned above. Nonetheless, one still widely used model evaluation model, the Akaike Information Criterion (AIC), originally proposed by Akaike (1973), can be justified as a very easily computed approximation to leave one out cross validation (see Stone 1977; Fang 2011). AIC is defined as follows.

$$\begin{aligned} \text{AIC} &= 2k - 2\log \text{P}(\mathcal{D}|\hat{\theta}), \\ &= 2k + \text{Deviance}, \end{aligned}$$

where k is the number of parameters in the model. Obviously, for models where the log of the likelihood is available, and where the number of parameters of the model is straightforward to count (which is not always the case in complex models), the AIC is simple to calculate. For example, for the log normal and normal models that we evaluated above using cross-validation, both of which have $k = 2$ parameters, the AIC is calculated as follows.

```
m1 <- lm(log(price) ~ 1, data = housing_df)
m0 <- lm(price ~ 1, data = housing_df)
k <- 2
aic_1 <- as.numeric(2 * k - 2 * logLik(m1))
aic_0 <- as.numeric(2 * k - 2 * logLik(m0))

c(aic_1, aic_0)
## [1]   472.599 12682.711
```

Clearly, these are very similar to the cross-validation ELPD for these two models.

Like ELPD and other measures, a model's AIC value is of little value in itself, and so we only interpret differences in AIC between models. Conventional standards (see, for example, Burnham and Anderson 2003,

Chapter 2) hold that AIC differences of greater than between 4 or 7 indicate clear superiority of the predictive power of the model with the lower AIC, while differences of 10 or more indicate that the model with the higher value has essentially no predictive power relative to the model with the lower value. We can appreciate why these thresholds are followed by considering the concept of *Akaike weights* (see Burnham and Anderson 2003, 75). Akaike weights provide probabilities for each of a set of K models that are being compared with one another. They are defined as follows:

$$\rho_k = \frac{e^{-\frac{1}{2}\Delta \text{AIC}_k}}{\sum_{k=1}^{K} e^{-\frac{1}{2}\Delta \text{AIC}_k}},$$

where ΔAIC_k is the difference between the AIC of model k and the lowest AIC value in the set of K models. These probabilities can be interpreted as the probabilities that any model has better predictive performance than the others. If we have just two models, with model $k = 1$ being the one with the lower AIC, then $\Delta\text{AIC}_1 = 0$ and ΔAIC_2 will be some value positive quantity δ. Using Akaike weights, the probability that the model with the lower AIC has the better predictive performance is

$$\rho = \frac{1}{1 + e^{-\delta/2}}.$$

For δ values of 4, 6, 9, the corresponding probabilities are 0.88, 0.95, 0.99. These values provide a justification for the thresholds proposed by Burnham and Anderson (2003).

WAIC

AIC is appealing because of its simplicity. As we have seen, despite its simplicity, it can be a highly accurate approximation to cross-validation ELPD. However, this approximation will not hold in all models, especially large and complex ones. Addition, in some situations calculating the maximum of the likelihood function is not straightforward, and nor is defining the number of free parameters in the model. A more widely applicable version of the AIC is the *Watanabe Akaike Information Criterion* (WAIC). WAIC was introduced in (Watanabe 2010) by the name *Widely Applicable Information Criterion* rather than *Watanabe Akaike Information Criterion*. It has been shown that WAIC is more generally or widely applicable than AIC, and is a close approximation to Bayesian leave-one-outcross-validation, yet can be calculated easily from a model's posterior samples. WAIC is calculated as follows:

$$\text{waic} = -2\left(\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} \text{P}(y_i|\theta^s)\right) - \sum_{i=1}^{n} V_{s=1}^{S}\left(\log \text{P}\left(y_i|\theta^s\right)\right)\right),$$

where y_i, θ^s etc are as they are defined above for the case of Bayesian ELPD. The term $V_{s=1}^{S}(\cdot)$ signifies the variance of its arguments.

Using Stan based models, WAIC is easily calculated using the `loo` package.

```
waic(m1_bayes)$estimates
##              Estimate         SE
## elpd_waic -236.332645 15.9127202
## p_waic       1.981799  0.1664625
## waic       472.665291 31.8254403
waic(m0_bayes)$estimates
##               Estimate         SE
## elpd_waic -6341.908993 23.4784786
## p_waic        3.023098  0.6228416
## waic      12683.817986 46.9569573
```

As we can see, this is very similiar to the Bayesian cross-validation ELPD, but can be calculated directly from posterior samples.

References

Akaike, Hirotogu. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." Edited by F Petrov B. N AND Caski.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo."

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411.

Burnham, Kenneth P, and David R Anderson. 2003. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Clopper, Charles J, and Egon S Pearson. 1934. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." *Biometrika* 26 (4): 404–13.

Diaconis, Persi, Susan Holmes, and Richard Montgomery. 2007. "Dynamical Bias in the Coin Toss." *SIAM Review* 49 (2): 211–35.

Fang, Yixin. 2011. "Asymptotic Equivalence Between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models." *Journal of Data Science* 9 (1): 15–21.

Gelman, Andrew, and Deborah Nolan. 2002. "You Can Load a Die, but You Can't Bias a Coin." *The American Statistician* 56 (4): 308–11.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy* 19 (10): 555.

Hoffman, Matthew D, and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.

Kerrich, J. E. 1946. *An Experimental Introduction to the Theory of Probability*. E. Munksgaard.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.

Metropolis, Nicholas, and S. Ulam. 1949. "The Monte Carlo Method." *Journal of the American Statistical Association* 44 (247): 335–41.

Neal, Radford. 2011. "MCMC Using Hamiltonian Dynamics." In *Handbook of Markov Chain Monte Carlo*, edited by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. CRC press.

Stone, M. 1977. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 44–47.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic." *Statistics and Computing* 27 (5): 1413–32.

Watanabe, Sumio. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research* 11 (Dec): 3571–94.