
Tutorial

Introduction to Bayesian computing in R

D G Rossiter

April 5, 2026

Contents

1 Introduction	1
2 Bayes' Rule	2
2.1 What affects the posterior probability?	5
2.2 Bayes' Rule for multivariate outcomes	8
2.3 Incorporating more evidence	9
3 Introduction to Bayesian inference	10
3.1 Bayesian inference for the Binomial distribution	13
3.2 Likelihood ratio	16
3.3 Conjugate prior distribution	20
3.4 Parameterizing the prior distribution	21
3.5 From Binomial prior to Binomial posterior	25
3.6 Inferences from the posterior distribution	27
3.7 Prediction	32
3.8 Model checking	38
4 Hierarchical models	39
5 Multiparameter models: the Normal distribution	54
6 Numerical methods	58
6.1 Numerical analysis of a Bayesian multiple linear regression model	60
6.2 Bayesian linear regression model with informative priors .	67
6.3 Comparing models with the Bayes factor	69

Version 0.4 Copyright 2018 © D G Rossiter. All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (d.g.rossiter@cornell.edu).

6.4 The Markov chain	70
7 Spatial Bayesian analysis	73
7.1 The Gaussian linear model	75
7.2 Specifying priors	76
7.3 Bayesian parameter estimation	77
7.4 Bayesian prediction	77
References	78

1 Introduction

This tutorial is intended for applied earth or natural resources scientists who need an introduction to Bayesian concepts and computation. There are several excellent texts at different levels of mathematical sophistication [5, 6, 8, 9] and several tutorials on doing Bayesian computation [1, 11]; here we only assume familiarity with the exponential and logarithmic functions and the concepts of integration, and a basic knowledge of R. All concepts are illustrated with worked examples using R.

Bayesian computing is a general term for an approach to statistical inference based on the so-called *Bayesian interpretation* of probability: the probability of an outcome or a parameter of a statistical distribution is our belief in its truth, consistent with some axioms of probability and the evidence.

Note: The inspiration for this approach was Bayes' 1763 paper [2] in which he developed the theory of inverse probability in order to make probabilistic statements about the future. He examined the proposition "Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." We recognize this as an examination of the binomial probability distribution.

In this view, statistical inferences are based on a *posterior* ("after the fact") distribution of parameters of statistical distributions; these are updated versions of *prior* ("before the fact") beliefs based on data from experiments or observations. The updating depends on the *likelihood* of each possible value of the parameters, given the data actually observed.

Once the posterior distribution of parameters is known, they can be used to derive a distribution of predicted outcomes. The focus of Bayesian approaches is on **decision-making under uncertainty**, i.e., predictions about the future, based on what is known from the past. The aim is to express beliefs in the language of probability, so that rational decisions can be made based on these.

The Bayesian interpretation is contrasted to the "frequentist" (sometimes called the "classical") interpretation: the probability of an outcome is the proportion of experiments in which the outcome occurs, in some hypothetical repetitions of the experiment under the same conditions and with the same population. In a geosciences context, the "experiment" is usually a field sample from an area of interest. So the frequentist makes statements about what would be expected to be found upon repeated sampling under the same conditions.

The key difference between Bayesian and frequentist approach is that the Bayesian considers parameters of statistical distributions to be random variables, i.e., they also have their own statistical distributions, which in turn have parameters, often called *hyperparameters*. We approach knowledge of the full distribution of the parameters by experiment or observation, combined with *prior* estimates of the hyperparameters. By

contrast, the frequentist considers that the parameters of statistical distributions are fixed; we approach knowledge of these fixed “true” values by experiment or observation, and quantify our uncertainty about this knowledge due to sampling error. The Bayesian approach allows for a richer set of inferences.

In this tutorial we first consider Bayesian probability (Bayes’ Rule) (§2) and then Bayesian statistical inference (§3).

The R Project for Statistical Computing has a wide variety of contributed packages for Bayesian analysis; see the Task View¹.

2 Bayes’ Rule

One aspect of Bayesian computation is not controversial: Bayes’ Rule derived from the definition of conditional probability.

The definition of *joint probability* of two events A and B , i.e., that both occur, is $P(A \cap B)$, where $\cap$ indicates set intersection. This can be reformulated in terms of *conditional probability*, i.e., that one event occurs conditional on the other having occurred. This can be written in two ways:

$$P(A \cap B) = P(A | B) \cdot P(B) = P(B | A) \cdot P(A) \quad (1)$$

where $|$ indicates that the event on the left is conditional on the event on the right. In words, the probability of both events is the unconditional probability one (say, B), multiplied by the probability that, given that (i.e., *conditional on*) event, the other (say, A) occurs.

An example from remote sensing land-cover classification, developed further in §2: the probability that a pixel in the image covers an area of water (event A), and also that its reflectance value is below a certain threshold (event B), is equal to the probability of any pixel in the image covering water (i.e., proportion of water) $P(A)$ multiplied by the probability that, given that (i.e., conditional on the fact that) a pixel covers water, its reflectance is below the threshold $P(B|A)$. It is also equal to the probability of any pixel in the image having a low reflectance $P(B)$, multiplied by the probability that, given that a pixel has low reflectance, it covers water $P(A|B)$.

Equating the two right-hand sides and rearranging gives Bayes’ Rule. We chose one of the *unconditional* probabilities $P(A)$ as the *prior*, and modify this probability based on new information that an event B with prior probability $P(B)$ has occurred, and knowing that the *conditional* probability $P(B | A)$ of B given A , to a *posterior* conditional probability $P(A | B)$:

$$P(A | B) = P(A) \cdot \frac{P(B | A)}{P(B)} \quad (2)$$

¹ <https://cran.r-project.org/web/views/Bayesian.html>

In words, the probability of A knowing that B has occurred is the *prior* probability of A , multiplied by the *likelihood ratio* $P(B | A)/P(B)$, i.e., the ratio of the probability of event B in case A is observed, divided by the overall probability of event B .

The choice of left-hand side of Bayes' Rule depends on what conditional probability we wish to estimate, whereas the conditional probability on the right-hand side is the one we can measure. In the remote sensing example, we typically want to know the probability that a pixel is in a certain land cover class, e.g., water, given some information from the image, e.g., NDVI value. We can measure the inverse conditional probability from training observations: knowing that a pixel covers water, what is the probability of a low NDVI?

The denominator $P(B)$ can also be written as the sum of the two mutually-exclusive intersection probabilities, one if event A occurs $P(A)$ and one where it does not occur $P(\neg A)$:

$$P(B) = P(B | A) \cdot P(A) + P(B | \neg A) \cdot P(\neg A) \quad (3)$$

We rename the probabilities to correspond to the concept of an observed "event" E and an unobserved or unknowable event for which we want to estimate the probability H ("hypothesis"). Bayes' Rule for the binary case then can be written:

$$P(H | E) = P(H) \cdot \frac{P(E | H)}{P(E | H) \cdot P(H) + P(E | \neg H) \cdot P(\neg H)} \quad (4)$$

Where does the prior probability come from? One possibility is a Bayesian probability from previous experiments or observations from the same population. Another is the analyst's belief. Finally, it can represent ignorance, in some sense specific to the problem.

Example Consider production of a two-class land cover map by image analysis, for example, identifying the water bodies and non-water areas based on a multispectral remote sensed image. Our *hypothesis* H is that the area represented by a pixel is in fact mostly covered by water. The *event* E that we can observe is that a pixel's NDVI² value is less than some threshold, say 0.1. We want to know, if we observe this event (pixel NDVI is low) what is the probability that our hypothesis is true (the pixel covers mostly water).

To do this, we define $P(H)$ as the prior probability that a random pixel area mostly covers water. This is our prior; we need an estimate from some other source (say, a topographic map) or a visual assessment. If we have enough training data (evidence) this can be a very rough estimate.

² NDVI "Normalized Difference Vegetation Index" is defined as $(\text{near-infrared reflectance} - \text{red reflectance}) / (\text{near-infrared reflectance} + \text{red reflectance})$, e.g., Landsat 7 ETM+ $(B4 - B3) / (B4 + B3)$, and ranges from $-1 \dots +1$.

Knowing $P(H)$ we automatically know its complement $P(\neg H) = (1 - P(H))$. Notice that before we collect any evidence we could classify any pixel as water, with $P(H)$ probability that this is correct.

We then need to know $P(E | H)$, that is, if a pixel really does cover water, what is the conditional probability it will have this low NDVI. It might not, for example because the water body contains many aquatic plants, or because of specular reflection. We hope $P(E | H)$, will be large, say 0.8. That is, we hope for high *sensitivity*: if the hypothesis is true (water pixel is present), we want to identify it from the evidence, that is, we want to avoid *false negatives*, where we incorrectly reject a true hypothesis. These have probability $1 - P(E | H)$. We obtain $P(E | H)$ from training samples: field or visual image interpretation of a representative sample of known water pixels and the proportion on the corresponding image that have this low NDVI.

We also need to know the complement $P(E | \neg H)$: if a pixel is not covering water it still may have low NDVI, for example pixels covered mostly by roads or industry. Here we hope for high *specificity*: we only want to identify true water pixels by this method, and avoid *false positives*; $P(E | \neg H)$ is the probability of these. This is also estimated from the training sample.

Given all this, we can compute the *posterior* probability $P(H | E)$ that, given that we've observed a pixel with low NDVI, it is in fact water.

Let's look at this numerically:

```
# prior estimate 20% of the image covered by water
p.h <- 0.2
# sensitivity: 90% of water pixels have low NDVI
# (from training sample)
p.e.h <- 0.9
# false positive rate: 10% of non-water pixels have low NDVI
# (from training sample)
p.e.nh <- 0.1
# denominator of likelihood ratio:
# predicted overall proportion of low-NDVI pixels in the image
(p.e <- (p.e.h * p.h) + (p.e.nh * (1 - p.h)))

## [1] 0.26

# likelihood ratio: increase in probability of hypothesis given the evidence
(lr.h <- p.e.h/p.e)

## [1] 3.461538

# posterior probability
(p.h.e <- p.h * lr.h)

## [1] 0.6923077
```

The *likelihood ratio* 3.462 is the weight of evidence given by the observation that a pixel has low NDVI. This resulted in an increase from the prior 0.2, just based on our estimate of the proportion of water areas in the image, to a posterior 0.69, considerably higher.

We can check that the two posterior probabilities sum to 1, by computing the posterior probability that the pixel with low NDVI is *not* water, and

summing this to the the posterior probability that the pixel with low NDVI *is* water. This likelihood ratio is:

```
(lr.nh <- p.e.nh/p.e)
## [1] 0.3846154
```

Notice how the likelihood ratio 0.385 that the hypothesis is false is in this case less than 1, that is, the event of a low NDVI lowers the prior probability that the pixel is *not* water.

The two posterior probabilities are:

```
print(p.h.e); (p.nh.e <- (1-p.h) * lr.nh)
## [1] 0.6923077
## [1] 0.3076923
```

and these indeed sum to 1:

```
(p.h.e + p.nh.e == 1)
## [1] TRUE
```

2.1 What affects the posterior probability?

1. $P(H)$, the prior probability of the hypothesis. The higher the prior, the higher the posterior, other factors being equal. In the absence of any information in a two-class problem, we could set this to 0.5.
2. $P(E | H)$, the sensitivity of the hypothesis to the evidence. The higher this is, the more diagnostic is the NDVI; it is in the numerator of the likelihood ratio.
3. $P(E | \neg H)$, the false positive rate (complement of the specificity). The higher this is, the less diagnostic is the NDVI, since it is in the denominator.

Points (2) and (3) imply we should look for a sensitive, specific test.

We can see the effect of these in our small example. We write a function to compute the posterior probability:

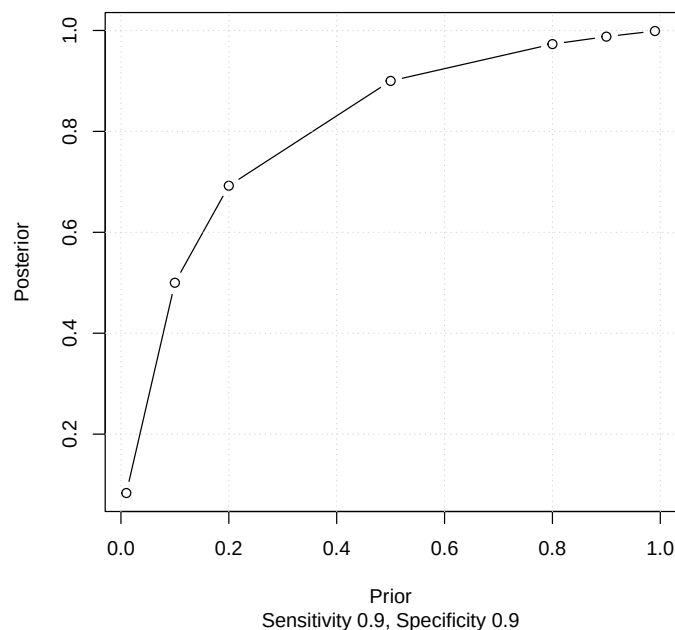
```
bayes.2 <- function(p.h, p.e.h, p.e.nh) {
  lr.h <- p.e.h/(p.e.h * p.h + p.e.nh * (1 - p.h))
  return(round(p.h * lr.h,4))
}
```

First, the effect of the prior, from very low to very high prior probability, with the same evidence:

```
# Effect of the prior
priors <- c(0.01, 0.1, 0.2, 0.5, 0.8, 0.9, 0.99)
(posterior <- bayes.2(priors, 0.9, 0.1))

## [1] 0.0833 0.5000 0.6923 0.9000 0.9730 0.9878 0.9989

plot(posterior ~ priors, type="b", xlab="Prior",
     ylab="Posterior", sub="Sensitivity 0.9, Specificity 0.9")
grid()
```



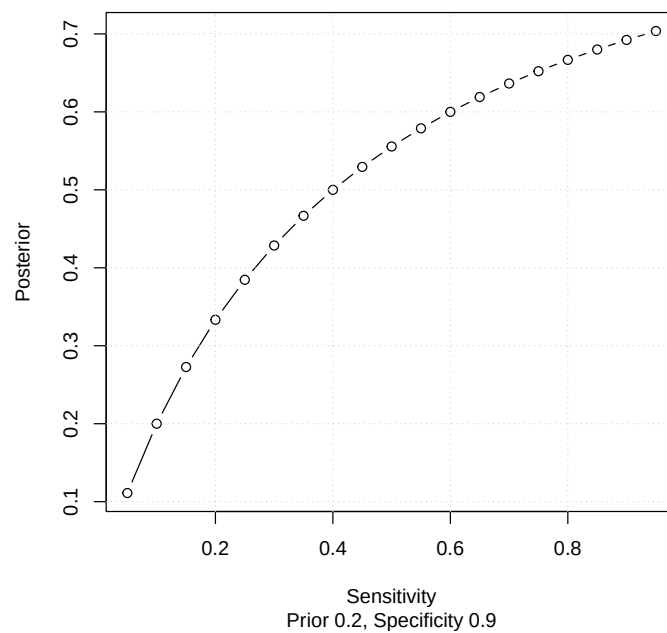
In all cases the posterior probability is increased by the likelihood ratio, but clearly the prior has an effect. If we had a map where we estimated only a very small amount of water overall, even this strong evidence does not raise the probability to more likely than not. On the other hand if we have already strong prior, e.g., a map that we think is largely water, such as a few islands in a large ocean, the evidence hardly makes any difference to our confidence that the pixel is in fact water.

Now the effect of the sensitivity of the evidence, with the same prior and specificity:

```
# Effect of the sensitivity
sens <- seq(0.05, 0.95, by=0.05)
(posterior <- bayes.2(0.2, sens, 0.1))

## [1] 0.1111 0.2000 0.2727 0.3333 0.3846 0.4286 0.4667 0.5000
## [9] 0.5294 0.5556 0.5789 0.6000 0.6190 0.6364 0.6522 0.6667
## [17] 0.6800 0.6923 0.7037

plot(posterior ~ sens, type="b", xlab="Sensitivity",
      ylab="Posterior", sub="Prior 0.2, Specificity 0.9")
grid()
```



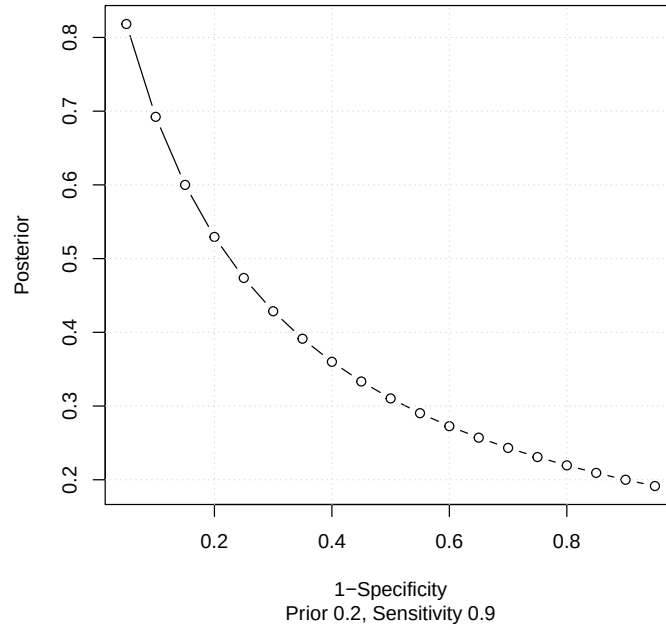
Obviously the more sensitive the test, the higher the posterior probability.

Finally the effect of the specificity, i.e., the false positive rate:

```
# Effect of the specificity
specs <- seq(0.05, 0.95, by=0.05)
(posterior <- bayes.2(0.2, 0.9, specs))

## [1] 0.8182 0.6923 0.6000 0.5294 0.4737 0.4286 0.3913 0.3600
## [9] 0.3333 0.3103 0.2903 0.2727 0.2571 0.2432 0.2308 0.2195
## [17] 0.2093 0.2000 0.1915

plot(posterior ~ specs, type="b", xlab="1-Specificity",
      ylab="Posterior", sub="Prior 0.2, Sensitivity 0.9")
grid()
```



More false positives lower the value of the evidence.

2.2 Bayes' Rule for multivariate outcomes

This can be generalized to a sequence of n mutually-exclusive hypotheses H_n , given some evidence E . The posterior probability of one of the hypotheses H_j is:

$$P(H_j | E) = P(H_j) \cdot \frac{P(E | H_j)}{P(E)} \quad (5)$$

where $P(E) = \sum_{j=1}^n P(E | H_j) \cdot P(H_j)$ is the overall probability of the event; this normalizes the conditional probability $P(H_j | E)$.

In words: The probability of a particular outcome, given the evidence, is proportional to the overall probability of the outcome without considering any evidence, multiplied by the probability of the evidence given that we know that outcome.

Example Continuing the land cover classification example, suppose we have H_1 water, H_2 grassland, H_3 forest, H_4 urban. From training samples we estimate $P(H_1 | E) = 0.9$ (as before), $P(H_2 | E) = 0.0$ (grassland never gives such low NDVI), $P(H_3 | E) = 0.1$ (there are dark areas of conifers or treefall), and $P(H_4 | E) = 0.8$ (most urban areas have a low NDVI). From prior knowledge we estimate the proportional cover to be $P(H_1) = 0.2$, $P(H_2) = 0.25$ (about one-quarter of the area is used for pasture), $P(H_3) = 0.5$ (about half the area is in forest), $P(H_4) = 0.05$ (a small urban area). Of course these must sum to 1. Then the overall $P(E)$ is:

```
# prior probabilities of each class, must sum to 1
(sum(p.h.m <- c(0.2, 0.25, 0.5, 0.05)) == 1)

## [1] TRUE

# probability of event for each class
p.e.h.m <- c(0.9, 0.0, 0.1, 0.8)
# total probability of event
(p.e.m <- sum(p.e.h.m * p.h.m))

## [1] 0.27
```

This is the posterior probability of the proportion of the image with low NDVI; it should match a direct calculation of this from the image.

Now we can compute the likelihood ratios for each class:

```
(lr.m <- p.e.h.m/p.e.m)

## [1] 3.333333 0.000000 0.3703704 2.9629630
```

In this case the likelihood ratios strongly favour water and urban areas; grassland will never be classified, and there is a lowered likelihood of forest; all these in the case that a low NDVI pixel is observed.

Now we compute the posterior probabilities, and check that they sum to 1:

```
(p.h.e.m <- p.h.m * lr.m)

## [1] 0.6666667 0.0000000 0.1851852 0.1481481

(sum(p.h.e.m) == 1)

## [1] TRUE
```

The evidence strongly favours water but also has some support for forest and urban areas. There are large changes in all the prior probabilities. The posterior probability for forest NA may seem surprisingly high, but that is because its prior probability was so high, i.e., we estimated that forest covered NA of the image area.

2.3 Incorporating more evidence

An important feature of Bayes' Rule is that it can be applied sequentially to incorporate new evidence. We just use the posterior probability from the first evidence as the prior probability when incorporating the new evidence. Returning to example 2, if we now use a different index to detect water pixels, e.g., NDWI³ values greater than 0.1, and we apply the test only to pixels tentatively identified as water by use of NDVI, we would now need to know:

$P(H)$: prior probability that a supposed water pixel really does covers water; we have this from the experiment with the NDVI, where we

³ NDWI "Normalized Difference Water Index" is defined as (near-infrared reflectance-short wave infrared)/(near-infrared reflectance + short wave infrared), e.g., Landsat 7 ETM+ (B4-B5)/(B4+B5), and ranges from $-1 \dots +1$.

estimated it to be 0.69. Knowing $P(H)$ we automatically know its complement $P(\neg H) = (1 - P(H))$.

$P(E | H)$: if a supposed water pixel really does cover water, what is the conditional probability it will have this high NDWI? (Sensitivity)

$P(E | \neg H)$: if a supposed water pixel really does not cover water, what is the conditional probability it will have this high NDWI? (Specificity)

The two conditional probabilities are estimated from the training sample, where these are a sample of pixels that in the first step (using NDVI) were identified as water pixels.

Given all this, we can compute the *posterior* probability $P(H | E)$ that, given that we've observed a supposed water pixel with high NDWI, it is in fact water.

Let's look at this numerically, supposing that this test is more sensitive but less specific, since it is being applied only to pixels that were tentatively classified as water.

```
# prior estimate is posterior from the estimate using NDVI
p.h.2 <- p.h.e
# sensitivity: 95% of supposed water pixels,
#   that are really water, have high NDWI
#   (from training sample)
p.e.h.2 <- 0.95
# false positive rate: 25% of supposed water pixels,
#   that are really not water pixels, have high NDWI
#   (from training sample)
p.e.nh.2 <- 0.25
# denominator of likelihood ratio:
#   predicted overall proportion of low-NDVI pixels in the image
(p.e.2 <- (p.e.h.2 * p.h.2) + (p.e.nh.2 * (1 - p.h.2)))

## [1] 0.7346154

# likelihood ratio: increase in probability of hypothesis given the evidence
(lr.h.2 <- p.e.h.2/p.e.2)

## [1] 1.293194

# posterior probability
(p.h.e.2 <- p.h.2 * lr.h.2)

## [1] 0.895288
```

The *likelihood ratio* 1.293 is the weight of evidence given by the observation than a supposed water pixel has high NDWI. This resulted in an increase from the prior 0.6923, just based on NDVI, to a posterior 0.8953, considerably higher.

This illustrates how evidence can be accumulated by successive applications of Bayes' Rule.

3 Introduction to Bayesian inference

The term “Bayesian” has been extended to a form of inference where we update a *prior probability distribution* (“**before** observations or experiments”) of model parameters with some evidence to obtain a *posterior*

probability distribution (“**after** observations or experiments”) of model parameters, based on the *likelihood* of the results of observations or experiments considering possible values of the parameters. This step is called *estimation* of the model parameters; we can then use these estimates for *prediction* of the target variable(s) (§3.7).

In the Bayesian view, a *statistical model* has the following general form, using the notation $[\cdot]$ to indicate a probability distribution:

$$[Y, S \mid \theta] \quad (6)$$

where Y is the joint distribution of some variable(s) for given values of model parameter(s) θ , the values of the variables determined by some unobservable process S , which we call the *signal*, as opposed to the *noise*, i.e., random variations not accounted for by the process. We can decompose this as:

$$[Y, S \mid \theta] = [S \mid \theta] [Y \mid S, \theta] \quad (7)$$

showing that the process depends on the model parameters, and given these, the joint probability distribution of the variables is determined.

We observe some of the Y , we use these to *estimate* θ , and then we can use the statistical model to *predict* other values produced by the process.

$$[S \mid Y] = \int_{\theta} [S \mid Y, \theta] [\theta \mid Y] d\theta \quad (8)$$

The decomposition shows that we are predicting the future observations from a process based on integrating the predictions from the entire posterior distribution of the estimated parameters θ . Knowing S , we can also predict any deterministic function of it $T = f(S)$.

For example, in simple linear regression S is a presumably linear process relating a stimulus to a response, producing observations $Y = (\mathbf{x}, \mathbf{y})$, i.e., ordered pairs of a predictor $\mathbf{x}$ and predictand $\mathbf{y}$, and $\theta = (\beta_0, \beta_1)$, i.e., the intercept and slope of the linear process. Then prediction uses this parameterized process $[S \mid \theta]$.

In the following development we omit the signal S , which is implicit, and work directly with the variables.

In Bayesian inference we assume that the true values of model parameters θ are *random variables*, and therefore have a *joint* probability distribution with the observations:

$$[Y, \theta] = [Y \mid \theta] [\theta] \quad (9)$$

The term $[\theta]$ is the *marginal* distribution for θ , i.e., before any data is known; therefore it is called the *prior* distribution of θ .

Inference is then based on sampling from the posterior distributions of the different model parameters. For example, in linear regression the Bayesian approach provides a full (joint) probability distribution of the parameters of the regression model; these include the coefficients of the

linear regression model, their errors, and their inter-correlation. This allows a rich set of inferences for decision making. We can find the most likely value, in some sense, but also use the full distribution for simulating possible scenarios.

Note: In the so-called *frequentist* or *classical* approach, parameters of statistical models are considered to be fixed, but unknowable by finite experiment. We conduct more experiments, collecting more evidence, to come closer to the “true” value as a point estimate using the most likely value, based on the results of experiments, and their assumed error distribution.

Bayesian statistical inference follows these steps:

1. We select a *statistical model form* to represent some process S of the real world; these have *model parameters* which must be fit to apply the model form to the particular situation. For example:
 - The relation between two different laboratory methods to measure the same chemical is linear with constant bias; the model is a linear regression model, the parameters are the slope and bias of the linear regression line and their standard errors.
 - The average number of successes in an experiment with two outcomes (a Bernoulli trial); the model is a binomial distribution with a single parameter, the proportion of successes.
 - The number of events in a given time period; the model is a Poisson distribution with a single parameter, the average number of events in the time period.
2. We assume some *functional form* for a probability distribution for the model parameters, based on external knowledge, prior experience, or assumed ignorance.
3. We estimate the *prior* distribution of these parameters from prior knowledge. We represent our uncertainty about the prior by a *probability distribution* which reflects our subjective prior opinion about plausible values of the parameters. Thus the distribution of parameters also has parameters! These are called *hyperparameters* – once they are known, either as fixed values or their distribution, the distribution of the parameters can be specified.
4. We apply the evidence by a modified Bayes’ Rule to compute the *posterior* distribution of the model parameters:

$$[\theta | Y] = \frac{[Y | \theta] [\theta]}{[Y]} \quad (10)$$

where $[Y] = \int [Y | \theta] [\theta] d\theta$, the *marginal* distribution of Y from the specified model, over all possible values of the model parameters. ,

Note: $[\theta | Y]$ is also called the *evidence* that the observed data values were generated by the chosen statistical model, no matter

what parameter values are used. For example, observations from a non-linear bivariate relation will not be well-fit by a linear model, no matter what slope and intercept are chosen.

5. We then examine the posterior distribution to make statistical *inferences*. For example, we may want to know the most probable value of the parameters (their modes), or their *credible intervals* in the parameters are found which a certain probability, or some probability of extreme values of the parameters. These inferences are directly usable for *decision making*.
6. We can also compute a *predictive distribution* for future events based on the the prior or posterior probability distributions of the parameters. For example, the distribution of predicted values of a chemical by a second lab. method given a value of the chemical by a first method; or, the number of successes in a future Bernoulli trial; or, the distribution of the number of events in a time period. These are also useful for *decision-making*, since they predict future occurrences.
7. We evaluate the model by comparing the predictive distribution with the observations: is it a reasonable functional form? Is the prior distribution appropriate?

3.1 Bayesian inference for the Binomial distribution

We introduce Bayesian inference with a simple example: inference about the Binomial distribution. This is a *continuous* probability distribution, with one parameter. It models any real-world situation where a number of independent observations n is made, each with one of two mutually-exclusive outcomes. The process S is thus some process that only gives one of these outcomes for each observation. The classic example is flipping a coin (not necessarily a fair coin!), which must land either heads or tails. Another example is answering a multiple-choice question on an examination, when only one of the options is correct.

Note: This was the distribution studied by Bayes [2]: “to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.”

It is defined as follows: The number k of “successes” in n independent, exchangeable⁴ Bernoulli trials, i.e., with two mutually-exclusive possible outcomes conventionally referred to as “successes” and “failures”, follows the Binomial distribution with parameter $\theta \in [0 \dots 1]$, the probability of success of any one trial:

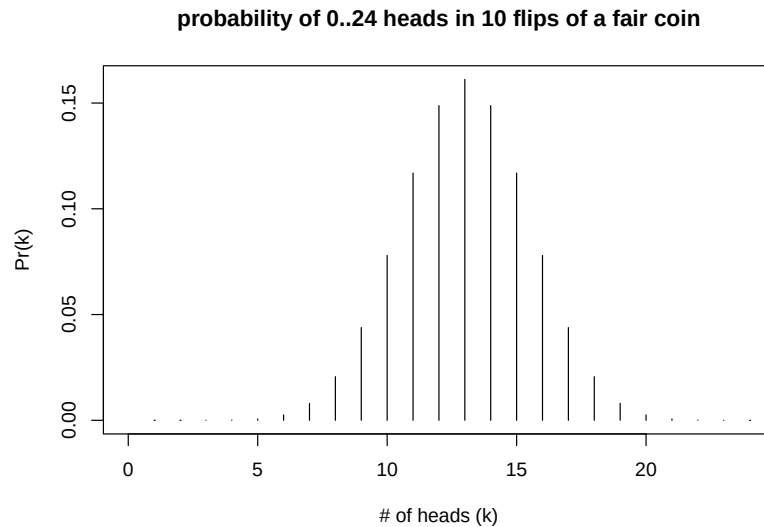
⁴ i.e., their order does not matter

$$p(k, n) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad (11)$$

A typical example is a set of flips of a coin (if fair, $\theta = 0.5$), where “heads” is counted as a success.

TASK 1 : Plot the probability of 0...24 heads in 24 flips of a fair coin with the `dbinom` “binomial density” function as a histogram. Compute the probability of exactly 10 heads in 24 flips. •

```
plot(dbinom(0:24, size=24, prob=0.5), type="h",
     xlim=c(0,24),
     xlab="# of heads (k)", ylab="Pr(k)",
     main="probability of 0..24 heads in 10 flips of a fair coin")
dbinom(10, size=24, prob=0.5)
## [1] 0.1169
```



The probability of exactly 10 heads in 24 flips of a fair coin is $\binom{24}{10} 0.5^{10} (1 - 0.5)^{24-10} = 0.1169$.

Looking at this distribution from the opposite perspective, we see that if we observe any number 0...24 heads in 24 trials, this is *evidence* of different strength for $\theta = 0.5$. For example, observing more than 16 or fewer than 8 heads in 24 flips means that, based on this evidence, the probability we would get this result if $\theta = 0.5$ is very low, 0.0639.

TASK 2 : Compute this probability using the `dbinom` function. •

```
1 - sum(dbinom(8:16, size=24, prob=0.5))
## [1] 0.06391466
```

This is the conditional probability $p(\theta = 0.5 \mid (k \leq 8 \wedge k \geq 16), n = 24)$. The key point here is that a given result from an experiment (flipping

the coin 24 times and obtaining fewer than 8 or more than 16 heads) provides very little evidence that the coin is fair, i.e., it provides evidence that the coin is biased towards tails.

But we can go further. The aim of Bayesian inference is to have a *full probability distribution* for a parameter, here θ of the Binomial distribution. That is, we do not want to determine a single most probable value for θ , instead we want to determine the probability of any value, or that the value is within a certain range, or that the value exceeds a certain number. For this we need a distribution for θ , parameterized by one or more *hyperparameters*.

In this section we will use the example of 10 successes in 24 Bernoulli trials.

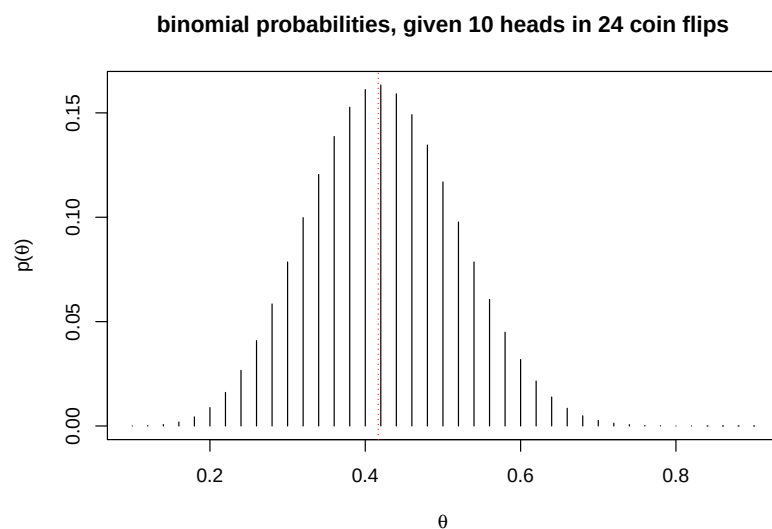
TASK 3 : Assign these to the variables k and n , so that the following code can easily be re-used with other values. •

```
k <- 10; n <- 24
```

Based on the evidence of 10 heads in 24 flips, we can compute $P(\theta | k = 10, n = 24)$ for any θ .

TASK 4 : Plot the probabilities of $\theta = 0.1, 0.12, \dots, 0.88, 0.9$ as a histogram, with the maximum likelihood value $10/24 = 0.4167$ shown as a red vertical bar. •

```
theta <- seq(0.1, 0.9, by=0.02)
plot(dbinom(k, size=n, prob=theta) ~ theta, type="h",
     xlab=expression(theta),
     ylab=expression(paste(plain("p("), theta, plain(")"))),
     main="binomial probabilities, given 10 heads in 24 coin flips")
abline(v=k/n, col="red", lty=3)
```



Any set of independent, exchangeable trials with a binary outcome can

be modelled in this way. A continuous outcome can be made binary with some thresholding, for example, concentration of a heavy metal in a soil sample can be divided into contaminated (“success”) / not contaminated at a regulatory threshold value.

Our interest is in estimating this parameter from some data, for example, the proportion of contaminated soils at a site, based on random sampling. Obviously the maximum likelihood estimate from one set of n observations, where k are above the threshold, is $\theta = k/n$, because the expected value $k = n \cdot \theta$. The variance then is $n \cdot \theta(1 - \theta)$; this increases with n because it is a count.

However, this estimate has two problems:

1. It does not give a probability distribution for θ , only the most likely value and its variance;
2. It does not take into account prior information.

The Bayesian approach solves both of these.

Note: In the frequentist approach the mean and variance are used to estimate intervals, usually with the Normal approximation interval; however with small sample sizes or where the proportion is close to the limits 0 or 1 this is quite poor. Various other intervals have been proposed and have better statistical properties.

3.2 Likelihood ratio

We extend Bayes’ Rule to full distributions of a parameter:

$$p(\theta | k, n) = p(\theta) \cdot \frac{p(k, n | \theta)}{p(k, n)} \quad (12)$$

In words, the *posterior* probability of any proportion of successes θ , given that we observe k successes in n trials, is the *prior* probability distribution of $\theta \in [0 \dots 1]$ from previous evidence or knowledge, multiplied by the *likelihood ratio*:

likelihood
ratio

$$\frac{p(k, n | \theta)}{p(k, n)} \quad (13)$$

This is the ratio of the probability of finding a given number k success in n trials for a known value of θ , divided by the probability of finding this number of successes, no matter what value of θ . The denominator is an integral over all possible values of θ , which reduces to a very simple

form:

$$\begin{aligned}
p(k, n) &= \int_{\theta=0}^1 p(k, n|\theta) d\theta \\
&= \binom{n}{k} \cdot \text{Beta}(k+1, (n-k)+1) \\
&= \binom{n}{k} \cdot \frac{\Gamma(k+1)\Gamma((n-k)+1)}{\Gamma(n+2)} \\
&= \binom{n}{k} \cdot \frac{k!(n-k)!}{(n+1)!} \\
&= \frac{n!}{k!(n-k)!} \cdot \frac{k!(n-k)!}{(n+1)!} \\
&= \frac{1}{n+1}
\end{aligned} \tag{14}$$

which does *not* depend on the number of successes k . That is, *a priori*, all $n+1$ possible number of successes in n trials $0, 1, \dots, n$ are equally likely, if the parameter θ is unknown.

For $k = 10, n = 24$ is 0.04. That is, in 24 Bernoulli trials, the probability of 10 successes, no matter the probability θ of success in one trial, is 4%, the same as the probability of any number of successes.

Note: Here we know the outcome is binary, and we selected the Binomial distribution, with some unknown (to us) value of θ , as having produced the observed outcome, i.e., 10 successes in 24 trials. But there are other models that produce binary outcomes. For example, a model with θ that increases or decreases during the trial according to some second parameter λ and some change function. These would also provide evidence for the observed number of successes having been produced by that model.

Of course, this probability is different depending on the true values of θ ; this is the numerator in Eqn. 13. In the previous section we graphed these for the $n = 24$ case; here are some values of $p(10, 24 | \theta)$:

```
dbinom(k, size=n, prob=theta)

## [1] 4.486725e-05 2.028169e-04 6.867410e-04 1.877777e-03
## [5] 4.351643e-03 8.832712e-03 1.607258e-02 2.667086e-02
## [9] 4.088009e-02 5.844799e-02 7.854515e-02 9.980650e-02
## [13] 1.204865e-01 1.387008e-01 1.527052e-01 1.611579e-01
## [17] 1.633124e-01 1.591085e-01 1.491507e-01 1.345840e-01
## [21] 1.169000e-01 9.771136e-02 7.853831e-02 6.063900e-02
## [25] 4.490599e-02 3.183367e-02 2.154872e-02 1.388574e-02
## [29] 8.485520e-03 4.894667e-03 2.649795e-03 1.336816e-03
## [33] 6.229861e-04 2.652330e-04 1.017180e-04 3.450278e-05
## [37] 1.010387e-05 2.471763e-06 4.822934e-07 7.012913e-08
## [41] 6.838477e-09
```

To convert these to a likelihood ratio we need to divide by the normalizing constant $p(k, n)$. In the discrete case we compute it directly as the mean of all the discrete probabilities, given the set of possible $\theta_1, \theta_2, \dots, \theta_m$:

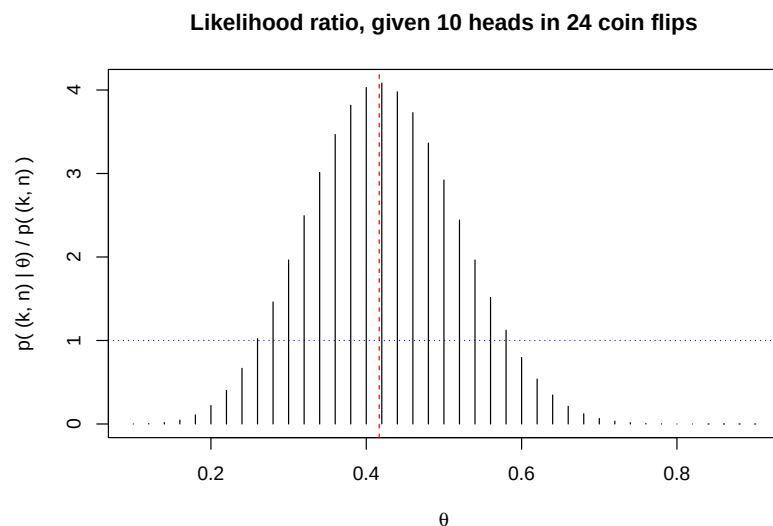
$$p(k, n) = \frac{1}{m} \sum_{i=1}^m p(k, n|\theta_i) \tag{15}$$

TASK 5 : Plot the likelihood ratio for $\theta = 0.1, 0.12, 0.14 \dots 0.9$ as a histogram, with a vertical line at the maximum value of the ratio, and a horizontal line at likelihood = 1. Likelihoods above this line show that this value of θ is more probable than the prior; below less probable. •

```
(p.k.n <- 1/(n+1))

## [1] 0.04

theta <- seq(0.1, 0.9, by=0.02)
plot(dbinom(k, size=n, prob=theta)/p.k.n ~ theta, type="h",
     xlab=expression(theta),
     ylab=expression(paste(plain("p( k, n ) | ",
                               theta, plain(") / p( k, n )")))),
     main="Likelihood ratio, given 10 heads in 24 coin flips")
abline(v=k/n, col="red", lty=2)
abline(h=1, col="blue", lty=3)
```



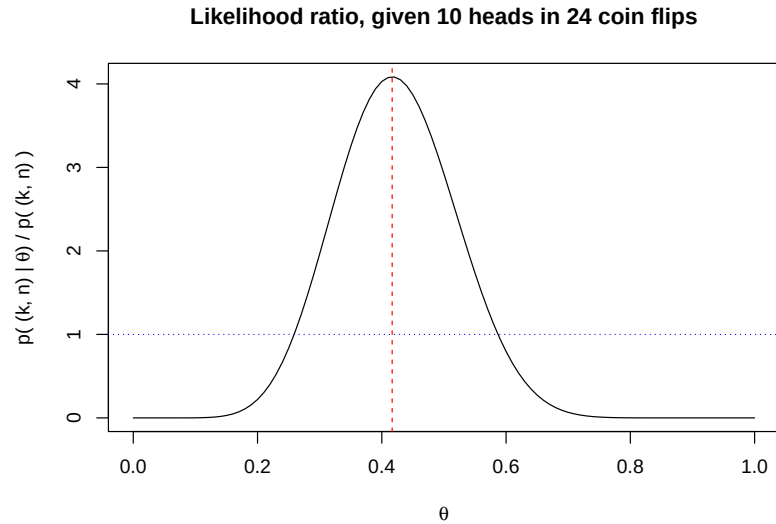
Note: The `expression` function allows for $\text{\LaTeX}$ -like mathematical notation in base graphics plots.

In this graph the horizontal blue line at 1 separates the θ for which the observation (10, 24) provides more evidence (> 1) for the prior probability of that θ from those for which the observation reduces the probability (< 1). For example, if we observe 10 successes in 24 trials, this strengthens the prior evidence for $\theta = 0.4$ by a ratio of 4.029; this is of course close to the maximum likelihood estimate for θ given the data, $10/24 = 0.417$. But it also provides some evidence for any value; for example $\theta = 0.7$ becomes less likely, with a likelihood ratio of 0.066.

However, θ can have any value between $[0 \dots 1]$, so the summation of Equation 15 must be replaced with the integral of Equation 14.

TASK 6 : Plot the continuous distribution of the likelihood with the curve function. •

```
(p.k.n <- 1/(n+1))
## [1] 0.04
curve(dbinom(k, size=n, prob=x)/p.k.n,
      xlab=expression(theta),
      ylab=expression(paste(plain("p( k, n ) | "),
                             theta, plain("/") / p( k, n ) "))),
      main="Likelihood ratio, given 10 heads in 24 coin flips")
abline(v=k/n, col="red", lty=2)
abline(h=1, col="blue", lty=3)
```



The ratio of Equation 13 can also be written with the reverse functional relation, i.e., θ as a function of k, n :

$$\ell(\theta | k, n) = p(k, n | \theta) \quad (16)$$

where the ℓ function is read as “the likelihood of”. This is another way of thinking about the relation between the observations and the parameter: the likelihood that the parameter has a certain value, knowing the observations, i.e., considering the data as fixed. The likelihood function contains all the information about the parameter that can be inferred from the data. It is also called the *sampling density*, because it depends on having taken a sample, i.e., having made a trial.

sampling
density

Once we have the prior probability distribution and the likelihood function, we compute the (un-normalized) posterior probability distribution by a modification of Bayes’ Rule, applying to distributions:

$$p(\theta | x) \propto p(\theta) \cdot \ell(\theta | x) \quad (17)$$

Equation 17 is the fundamental equation of Bayesian inference.

Note the use of $\propto$ “proportional to” instead of $=$ “equals”. This will be explained in more detail below; in short, it gives an *unnormalized* posterior distribution (the probability does not integrate to 1), and simplifies calculations.

3.3 Conjugate prior distribution

Since θ can take any value from $[0 \dots 1]$, we need to find a *probability distribution* form for it; this distribution will be parametrized by one or more *hyperparameters*, i.e. “parameters of the probability distribution of the parameters of the distribution we are interested in”. This probability distribution must assign a probability to all values on $[0 \dots 1]$, summing to 1. In the example just above we could find the probability of any θ but we did not have a continuous distribution. In other words, we need a function with domain $\mathbb{R} \in [0 \dots 1]$ (possible values of θ) and range $[0 \dots 1]$ (their probability).

The functional form for a parameter must have certain properties that match our intuition. The distribution for Binomial θ must:

1. Have one mode;
2. be monotonically non-increasing on both sides of the mode, unless the mode is at $\theta \in (0, 1)$;
3. have a functional form that will be the same after it has been multiplied by the likelihood.

Many distributions satisfy conditions (1) and (2); condition (3) is the most difficult. It is desirable because we may want to later use the posterior distribution as a prior in further work.

A form with property (3) is called a *conjugate prior* for the particular likelihood. This is defined a prior distribution such that when multiplied by the likelihood it has the same form. That is, in Equation 17 $p(\theta | x)$ and $p(\theta)$ must have the same form, knowing the form of the likelihood $\ell(\theta | x)$ that will multiply the prior $p(\theta)$ to obtain the posterior $p(\theta | x)$.

In the case of the parameter for the Binomial distribution, the likelihood $\ell(\theta | x)$ is an unnormalized binomial distribution with constants k, n :

$$\theta^k \cdot (1 - \theta)^{n-k} \quad (18)$$

Note the exponents on the proportion and its complement.

The usual conjugate prior distribution for binomial θ is the Beta distribution with two hyperparameters α and β :

$$\text{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad (19)$$

where the first term is a normalizing constant to ensure that the total probability integrates to 1, using the Beta function:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (20)$$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, the generalization to the real numbers of the factorial. For integer x , $\Gamma(x + 1) = x!$. So, the normalizing constant is:

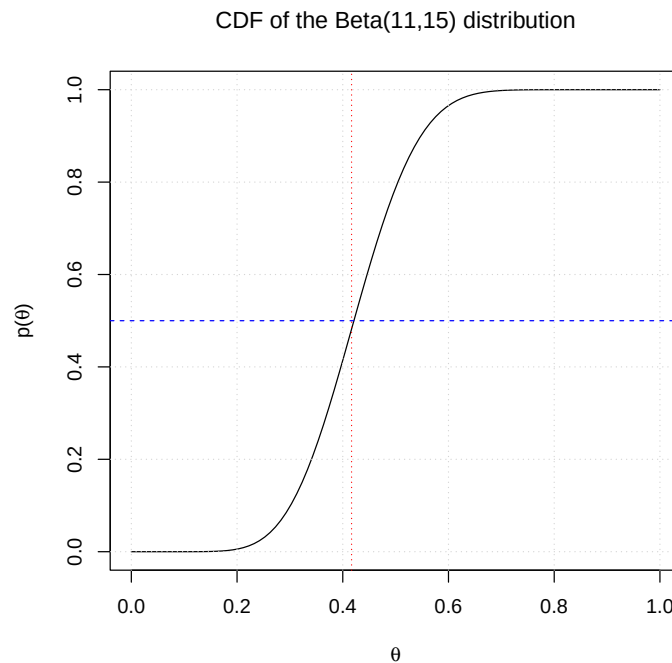
$$1/B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad (21)$$

Notice the second term of Equation 19 with θ and $(1 - \theta)$ to different powers; this resembles the binomial distribution if we set $\alpha = k + 1$ and $\beta = (n - k) + 1$, with k the number of successes in n trials.

For example, here is the cumulative distribution for $\alpha = (10 + 1)$, $\beta = (14 + 1)$; the crossover at $p(\theta) = 0.5$ is at the estimate $(\alpha - 1)/(\alpha + \beta - 2) = 10/24 = 0.41\bar{6}$ obtained from the parameters $(\alpha = 11, \beta = 15)$ of the Beta distribution. The `pbeta` function computes the probability density function (PDF) for the Beta distribution.

TASK 7 : Plot the CDF of the Beta distribution on a sequence of 100 possible values of θ , specified with the `seq` function. Show the value of θ at the crossover $p(\theta) = 0.5$ as a vertical bar.

```
theta <- seq(0, 1, length=100)
plot(pbeta(theta, 1+k, 1+(n-k)) ~ theta, type="l", xlab=expression(theta),
     ylab=expression(paste(plain("p("), theta, plain(")"))),
     main=expression(paste("CDF of the Beta(11,15) distribution")))
grid()
abline(v=k/n, lty=3, col="red")
abline(h=0.5, lty=2, col="blue")
```



3.4 Parameterizing the prior distribution

Now we know the form of the conjugate prior density, but it must be parameterized, i.e., the two hyperparameters α and β must be selected.

There are two ways to do this, referred to as *informative* and *non-informative* priors, respectively.

Informative prior If we have prior information, presumably from similar situations, or even from expert judgement, we express this as the parameters of the prior Beta distribution. This is especially important if the evidence we now will collect is sparse, compared to the evidence from other studies or physical principles. An example of sparse data is estimating θ for the proportion of contaminated soils in a study area, where we only have budget for a small number of (expensive) observations (soil samples). An example of common sense is the proportion of hydric soils⁵ in a low plain in a humid region; proportions less than, say, 0.20, are highly unlikely from what we know about geomorphology and soil hydrology.

One possibility is to use the parameterized Beta distribution of θ for similar areas as the prior, i.e., use the hyperparameters of the posterior probability distribution from other studies as the prior hyperparameters.

If this is not possible (no similar study) or desirable, values of the hyperparameters can be computed from estimated mean (expected value) and variance among trials, as follows.

We use the fact that the expected value of a Beta-distributed θ is:

$$\begin{aligned} \mathbf{E}\theta &= \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \theta d\theta \\ &= \alpha / (\alpha + \beta) \end{aligned} \quad (22)$$

which can be interpreted as the prior belief of the proportion of success.

The denominator $(\alpha + \beta)$ in Equation 22 can be interpreted as the number of trials (or the sample size) on which your belief is based; it resembles the sum of success and failures in the binomial distribution, i.e., the total number of trials. Because of the -1 in the exponents of the Beta distribution, this number is better given as $(\alpha + \beta + 2)$, and the numerator as $(\alpha + 1)$. These are not actual trials, the idea is that increasing this number (and therefore α to keep the proportion the same) represents your strength of evidence.

TASK 8: Compute the parameterization of the Beta prior, if we consider that our prior belief is based on 5, 10, ... 30 trials, each with an expected proportion of 0.4 successes. •

```
e.prop <- 0.4
s <- seq(5, 30, by=5)
(alpha <- (e.prop * s) + 1)

## [1] 3 5 7 9 11 13

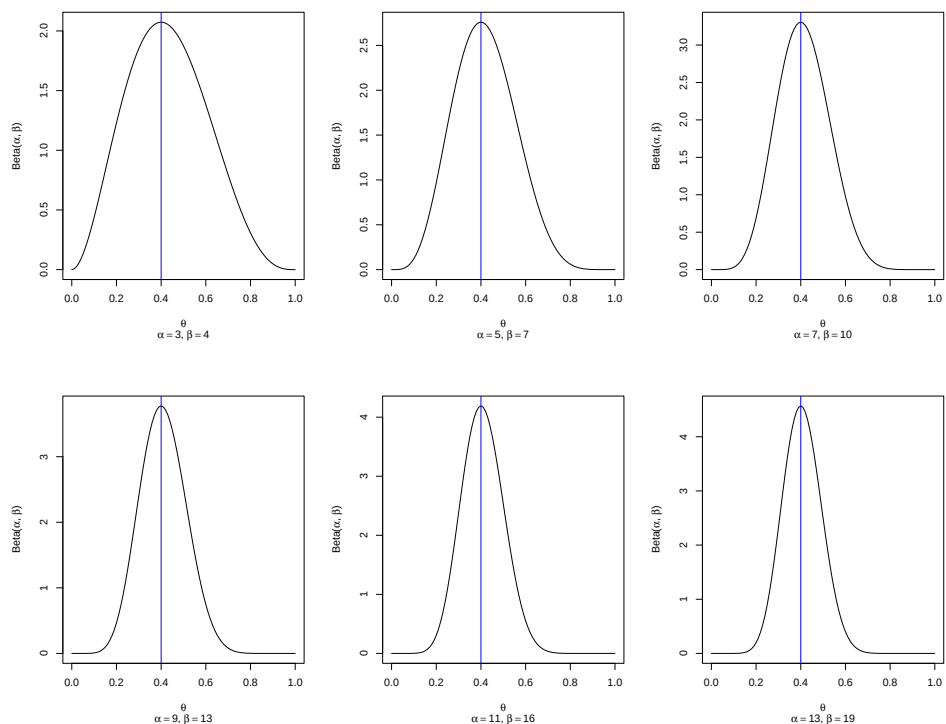
(beta <- (s - alpha) + 2)

## [1] 4 7 10 13 16 19
```

⁵ “a soil that formed under conditions of saturation, flooding or ponding long enough during the growing season to develop anaerobic conditions in the upper part” – https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/hydric/?cid=nrcs142p2_053961

TASK 9 : Graph all these, using the curve function applied to the dbeta function to evaluate it at some number (default 101) of points in the range. This shows how the form of the Beta distribution becomes more concentrated with more trials, with the same proportion of successes •

```
par(mfrow=c(2,3))
for (i in 1:length(s)) {
  curve(dbeta(x, alpha[i], beta[i]), xlab=expression(theta),
        from=0, to=1,
        ylab=expression(paste(plain("Beta("), alpha,
                              plain(", "), beta, plain(")"))),
        sub=substitute(paste(plain(alpha) == x, ", ",
                              plain(beta) == y),
                      list(x=alpha[i],y=beta[i])),
        )
  abline(v=0.4, col="blue")
}
par(mfrow=c(1,1))
```



The mode is at the expected proportion, $(\alpha - 1)/(\alpha + \beta - 2)$.

The Beta densities become more concentrated, with the same expected value 0.4 as mode, as the number of trials increases. Viewing the spread of these densities may help in selecting a “reasonable” informative prior, based on expert opinion. The more concentrated the densities, the more effect the prior will have on the posterior.

It is also possible to use the expected mean, as above (Equation 22), and the variance to solve two equations in two unknowns to obtain the two

hyperparameters:

$$\mathbf{E}\theta = \frac{\alpha}{(\alpha + \beta)} \quad (23)$$

$$\mathbf{Var}\theta = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (24)$$

But this requires an expert judgement of a variance, which is not as intuitive as a mean and sample size.

Non-informative prior The idea here is to parameterize the Beta distribution such that all values of θ are *a priori* equally likely, and all inferences about the distribution of θ come from the data.

Note: “Non-informative” is not really good terminology, as even absence of information is information. The idea is to represent in some sense the least amount of information, i.e., maximum *a priori* ignorance, consistent with the form of the prior distribution.

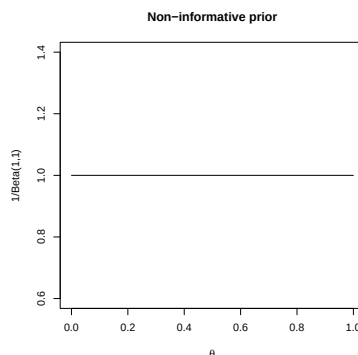
One choice⁶ is $\alpha = \beta = 1$:

$$\text{Beta}(x; 1, 1) = \frac{1}{B(1, 1)} x^{1-1} (1-x)^{1-1} = \frac{1}{B(1, 1)} = 1 \quad (25)$$

which is uniform on $[0 \dots 1]$ and does not depend on the Binomial parameter – hence it is uninformative about θ , which we are trying to estimate. So, all information about θ will come from the data.

TASK 10 : Visualize the uniformity by graphing $1/B(1, 1)$ against θ . •

```
curve(1/dbeta(x, 1, 1), xlab="expression(theta)", ylab="1/Beta(1,1)",  
      from=0, to=1,  
      main="Non-informative prior")
```



This is a boring graph! but proves the point.

Note: There is quite some controversy about the choice of non-informative prior for the Beta distribution, and even about the meaning of “non-informative”. Other choices, besides Beta(1,1), with good arguments are Beta(0, 0) and Beta(0.5, 0.5). Since these exponents are small, all ≤ 1 , these different choices have very little practical difference in the presence of any reasonable amount of data, as will be shown in the next section by the updating rule.

⁶ used by Bayes in his Essay

3.5 From Binomial prior to Binomial posterior

We've decided to use the Beta distribution to represent the probability of each θ . The prior distribution is updated by the likelihood of each θ , given the observation of k successes in n trials:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (26)$$

$$\ell(k, n | \theta) \propto \theta^k(1-\theta)^{n-k} \quad (27)$$

$$p(\theta | k, n) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+(n-k)-1} \quad (28)$$

The step from Equations 26 and 27 to 28 follows from law of exponents, applying the fundamental Bayesian inference of Equation 17 to update the hyperparameters of the Beta distribution, which in turn determines the distribution of the binomial parameter. Now it should be clear why the Beta distribution is a good representation of the probability distribution of the binomial parameter θ .

In other words, if prior $p(\theta) \sim \text{Beta}(\alpha, \beta)$, and the number of successes k in n trials follows the binomial distribution with parameter θ , then the posterior becomes $p(\theta | k, n) \sim \text{Beta}(\alpha + k, \beta + (n - k))$. This simple updating formula allows us to modify a prior Beta distribution to posterior Beta distribution that takes into account the data. Note that the larger the n , the less important are the prior values of the hyperparameters. This also shows why there is not much difference between Beta(0,0), Beta(1,1) and Beta(0.5,0.5) priors in practice, since $+k$ and $+(n - k)$ will dominate them in the posterior as n increases.

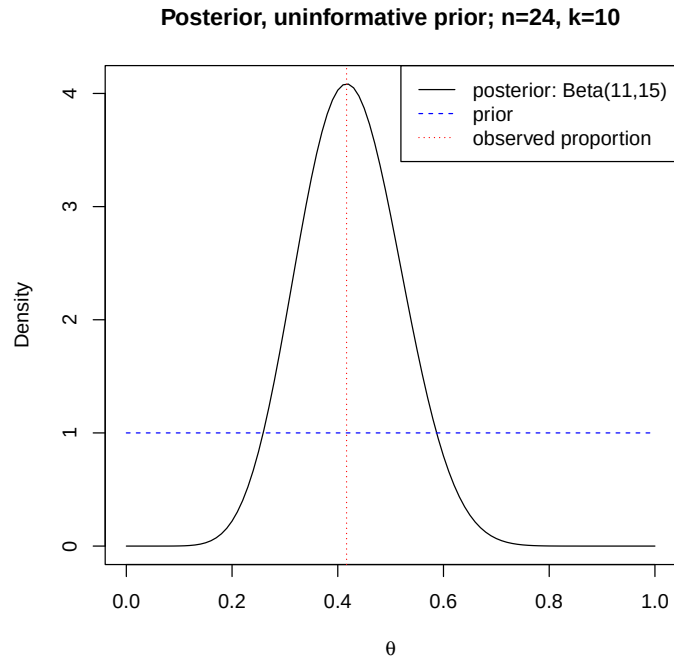
The use of $\propto$ "proportional to" instead of $=$ "equals" simplifies most calculations. The usual method in Bayesian computation is to:

1. simplify the form of the prior distribution by dropping any multiplicative constants that do not contain the parameter(s) of interest;
2. multiply this by the numerator of the likelihood ratio to obtain the non-normalized posterior distribution of the parameter(s);
3. make some inferences directly from the non-normalized posterior, e.g., credible intervals;
4. if necessary, normalize by the denominator of the likelihood ratio and any multiplicative constants that were removed. In the case of the Beta distributions everything is already normalized.

TASK 11 : Plot the posterior probability density of θ based on inference from the non-informative prior. Also show the non-informative prior distribution (in this case, a constant value). •

```
curve(dbeta(x, 1+k, 1+(n-k)), xlab=expression(theta),  
      ylab="Density",  
      from=0, to=1,  
      main="Posterior, uninformative prior; n=24, k=10")
```

```
curve(dbeta(x, 1, 1), from=0, to=1, add=TRUE, lty=2, col="blue")
abline(v=k/n, col="red", lty=3)
legend("topright", legend=c("posterior: Beta(11,15)", "prior", "observed proportion"),
      col=c("black", "blue", "red"), lty=1:3)
```



We see that even this fairly small amount of evidence has concentrated the posterior probability, with mode being the maximum likelihood value; this is the result of using a non-informative prior. This is an extreme example of *shrinkage*: the posterior distribution is more concentrated than the prior.

We contrast this with an example with a strongly informative prior for expected proportion of successes which is quite different from what is observed. We choose a proportion of successes 0.6 and a supposed number of trials 30, i.e., 18 supposed successes, which gives a prior of Beta(19,13). Multiplying this with the likelihood gives a posterior Beta{19 + (n - k) = 33, 13 + (n - k) = 27} distribution for θ :

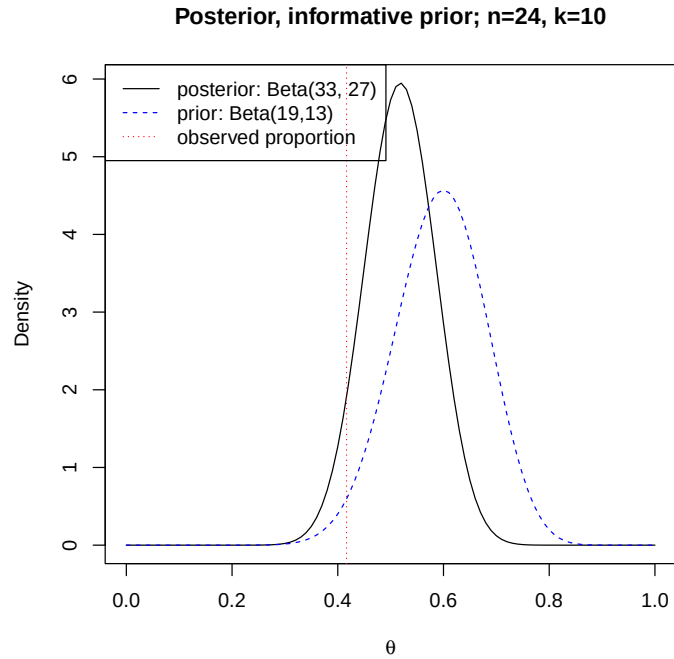
We assign the informative priors to variables, so that the code can easily be re-run with other values:

```
prior.a <- 19; prior.b <- 13
```

TASK 12 : Plot the posterior probability density of θ based on inference from the informative prior. Also show the informative prior distribution.

```
curve(dbeta(x, prior.a+k, prior.b+(n-k)), xlab=expression(theta),
      ylab="Density",
      from=0, to=1,
      main="Posterior, informative prior; n=24, k=10")
curve(dbeta(x, prior.a, prior.b), lty=2, from=0, to=1, add=TRUE, col="blue")
```

```
abline(v=k/n, col="red", lty=3)
legend("topleft", legend=c("posterior: Beta(33, 27)", "prior: Beta(19,13)", "observed proportion"),
      col=c("black", "blue", "red"), lty=1:3)
```



Notice how in this case the posterior from the informative prior is more concentrated, because the data increases the amount of information over what was assumed for the prior. The mode of the posterior distribution starting from the informative prior is between the observed proportion and the mode of the prior distribution. Here also there is shrinkage.

3.6 Inferences from the posterior distribution

Given a probability distribution for a parameter, we can make many inferences based on it, for example means, variances, and posterior intervals, by making random draws from the probability distribution. This is a major advantage of the Bayesian approach.

Posterior mean and mode First, the posterior mean and mode can be directly computed from the properties of the Beta distribution, knowing the posterior values of the hyperparameters α and β : These are in some sense the most likely values of the binomial parameter.

$$E(\theta | k, n) = \alpha / (\alpha + \beta) \quad (29)$$

$$\text{Mode}(\theta | k, n) = (\alpha - 1) / (\alpha + \beta - 2) \quad (30)$$

TASK 13 : Compute the posterior mean and mode, based on inference

from the informative prior. •

```
alpha <- prior.a+k; beta <- prior.b+(n-k)
alpha/(alpha + beta)

## [1] 0.5178571

(alpha-1)/(alpha + beta - 2)

## [1] 0.5185185
```

The mean and mode are near 0.52; these are compromises between the prior with mode $19/32 = 0.5938$ and the maximum likelihood point estimate of the observations $10/24 = 0.4167$. Note that the more observations, the closer the posterior mode will come to the maximum likelihood point estimate.

TASK 14 : Compute the posterior mean and mode, based on inference from the non-informative prior. •

```
alpha <- 1+k; beta <- 1+(n-k)
alpha/(alpha + beta)

## [1] 0.4230769

(alpha-1)/(alpha + beta - 2)

## [1] 0.4166667
```

The mean and mode are near 0.42, in fact the mode is exactly the maximum likelihood point estimate 0.4167; all the information about the posterior came from the observations.

Credible intervals A single most probable value is not so useful. In the Bayesian approach we can also compute a *credible interval* within which we believe, with some probability, the parameter lies. Recall that in the Bayesian approach we assume that the true value is not fixed but unknowable by finite experiment (that is the frequentist view), rather, that is a random variable. That is why we needed to define and parameterized a prior probability distribution for the parameter – even without experiment we needed to characterize the parameter by a distribution. So after the experiment, and updating by the likelihood, we also have a distribution, now posterior to the experiment. The word “credible” is linked to “belief”, that is, our probabilistic belief that the random variable is in some interval.

We obtain credible intervals from the quantiles of the distribution, prior or posterior. To do this, we find the upper limit c of the definite integral of the distribution, such that it equals the desired quantiles q , for example $q = 0.05$ and $q = 0.95$ for the 90% credible interval.

$$\int_0^c p(\theta|k, n) d\theta = q \quad (31)$$

The `qbeta` “quantile of the Beta distribution” function makes this computation.

TASK 15 : Compute the 90% credible interval for the prior and posterior distributions of the binomial parameter for the informative and non-informative priors. •

```
# informative
(cred.inf.pre <- qbeta(c(0.05, 0.95),
                      shape1=prior.a, shape2=prior.b))

## [1] 0.4495566 0.7311737

diff(cred.inf.pre)

## [1] 0.2816171

(cred.inf.post <- qbeta(c(0.05, 0.95),
                      shape1=prior.a+k, shape2=prior.b+(n-k)))

## [1] 0.4085025 0.6264798

diff(cred.inf.post)

## [1] 0.2179773

# non-informative
(cred.non.inf.pre <- qbeta(c(0.05, 0.95),
                          shape1=1, shape2=1))

## [1] 0.05 0.95

diff(cred.non.inf.pre)

## [1] 0.9

(cred.non.inf.post <- qbeta(c(0.05, 0.95),
                          shape1=1+k, shape2=1+(n-k)))

## [1] 0.2698531 0.5831620

diff(cred.non.inf.post)

## [1] 0.3133089
```

Notice how the interval is much wider in the case of the non-informative prior, with the same observations.

We can also compute this by simulation, drawing samples with the `rbeta` “random value from the beta distribution” function and then finding the quantiles of the simulated draw with the `quantile` function. Although in the present example we are able to compute posterior statistics analytically, with more complicated models that will not be possible, and simulation must be used. So we introduce simulation here to show how it can be used, and then discuss the more difficult case in §6.

TASK 16 : Compute the 90% credible interval for θ by simulation from the posterior distribution, based on the informative prior. Plot a histogram of the simulated parameter, with vertical lines showing the val-

ues at the limits of the credible interval.

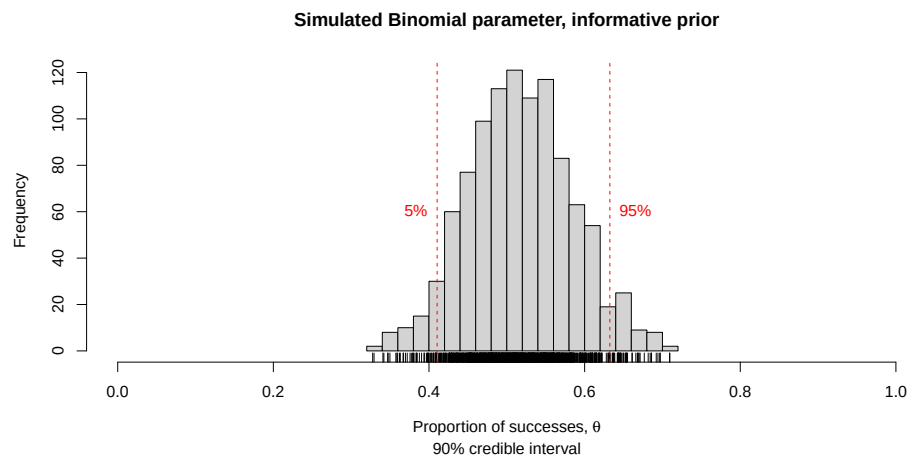
```
sim <- rbeta(n=1024, shape1=prior.a+k, shape2=prior.b+(n-k))
hist(sim, main="Simulated Binomial parameter, informative prior", breaks=24,
      xlim=c(0,1),
      xlab=expression(paste(plain("Proportion of successes, "),
                             theta)),
      sub="90% credible interval")
rug(sim)
(cred.theta <- quantile(sim, c(0.05, 0.95)))

##          5%          95%
## 0.4104378 0.6324019

(cred.inf <- qbeta(c(0.05, 0.95),
                  shape1=prior.a+k, shape2=prior.b+(n-k)))

## [1] 0.4085025 0.6264798

abline(v=cred.theta[1], col="red", lty=2)
abline(v=cred.theta[2], col="red", lty=2)
text(cred.theta[1], 60, labels="5%", col="red", pos=2)
text(cred.theta[2], 60, labels="95%", col="red", pos=4)
```



TASK 17 : Compute the 90% credible interval for θ by simulation from the posterior distribution, based on the non-informative prior. Plot a histogram of the simulated parameter, with vertical lines showing the values at the limits of the credible interval.

```
sim <- rbeta(n=1024, shape1=1+k, shape2=1+(n-k))
hist(sim, main="Simulated Binomial parameter, non-informative prior", breaks=24,
      xlim=c(0,1),
      xlab=expression(paste(plain("Proportion of successes, "),
                             theta)),
      sub="90% credible interval")
rug(sim)
(cred.theta <- quantile(sim, c(0.05, 0.95)))

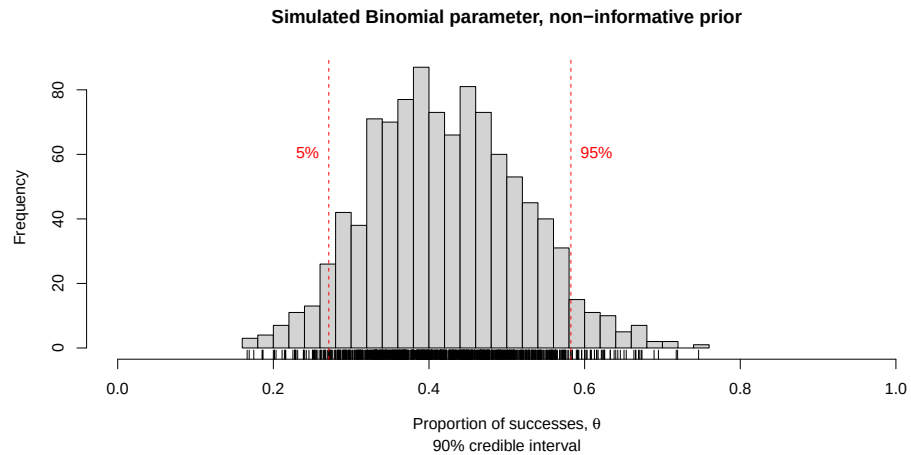
##          5%          95%
## 0.2712176 0.5823391

(cred.non.inf <- qbeta(c(0.05, 0.95),
                      shape1=1+k, shape2=1+(n-k)))

## [1] 0.2698531 0.5831620

abline(v=cred.theta[1], col="red", lty=2)
```

```
abline(v=cred.theta[2], col="red", lty=2)
text(cred.theta[1], 60, labels="5%", col="red", pos=2)
text(cred.theta[2], 60, labels="95%", col="red", pos=4)
```



The simulated intervals are quite close to the exact intervals. Increasing the number of simulations will make the match even closer.

Why simulate when we can compute the exact value? This is because in more complicated situations there will not be an analytical form of the posterior distribution, and simulation will be the only way to estimate credible intervals or other statistics.

Probability of exceedence Another question of interest to decision-makers is the probability that the parameter θ exceeds or is less than some threshold value θ_0 . For example, if the proportion of polluted soil in a brownfield site is above some threshold, the decision can be to fence it in and abandon it, rather than reclaim it by decontaminating the soil. This probability can be computed from the posterior distribution as:

$$p(\theta > \theta_0) = \int_{\theta_0}^1 p(\theta|k, n) d\theta = 1 - \left(\int_0^{\theta_0} p(\theta|k, n) d\theta \right) \quad (32)$$

This can be computed with the `pbeta` “Beta distribution CDF” function.

TASK 18 : Compute the probability of exceedence, supposing that the threshold $\theta_0 = 0.4$, i.e., 40% ‘successes’, based on both the informative and non-informative priors. •

```
1 - pbeta(0.4, shape1=prior.a+k, shape2=prior.b+(n-k)) # informative prior
## [1] 0.9620749

1 - pbeta(0.4, shape1=1+k, shape2=1+(n-k)) # non-informative prior
## [1] 0.585775
```

In this case the posterior probability of exceedence using the informative prior leads us to a more conservative conclusion, i.e., more likely that the

threshold is exceeded. This is because the informative prior was centred at a higher value of θ than this threshold, i.e., we had a prior belief of a high proportion of contaminated soils, while the non-informative prior had no mode. Here is a case where prior belief has a large influence on decision-making.

3.7 Prediction

Once we have a distribution for the parameter θ (either a prior or posterior), we can use it to predict the values of future observations, e.g., number of successes in a binomial trial. Recall that the general form of a predictive distribution is given by Eqn. 8, which we repeat here:

$$[S | Y] = \int_{\theta} [S | Y, \theta] [\theta | Y] d\theta \quad (8)$$

Here the process is the set of Bernoulli trials, and we want to predict results of a future set, based on the set we've seen (Y) and the inferred distribution of the parameters of the process (θ).

The *predictive* density for the Binomial distribution is the number of successes $\tilde{y}$ in a future experiment of n trials. The tilde $\sim$ over a variable name indicates a future observation.

For example, the number of soil samples, out of n , that will exceed a contamination threshold in a future sampling campaign in the same study area as was used to establish the posterior distribution of θ .

The prediction is obtained from the distribution of the parameter θ by integrating the predictions from each value of the parameter based on its probability:

$$f(\tilde{y}) = \int f(\tilde{y} | \theta) g(\theta) d\theta \quad (33)$$

where $g(\theta)$ is the density for θ . This can be either a prior or posterior; it represents our belief about the distribution of the parameter. The term $f(\tilde{y} | \theta)$ is the density of predicted outcomes for a specific value of θ .

For predictions post-experiment, this is formally a prediction of the future $\tilde{y}$ based on the past observations y :

$$p(\tilde{y} | y) = \int p(\tilde{y} | \theta) p(\theta | y) d\theta \quad (34)$$

where $g(\theta)$ is conditioned on the experiment, i.e., $p(\theta | y)$. Note the similarity with the conceptual predictive Eqn. 8.

An important note; this formula assumes that there is no dependence of future observations on past, that is, making the past observations did not change the context in which future observations will be made.

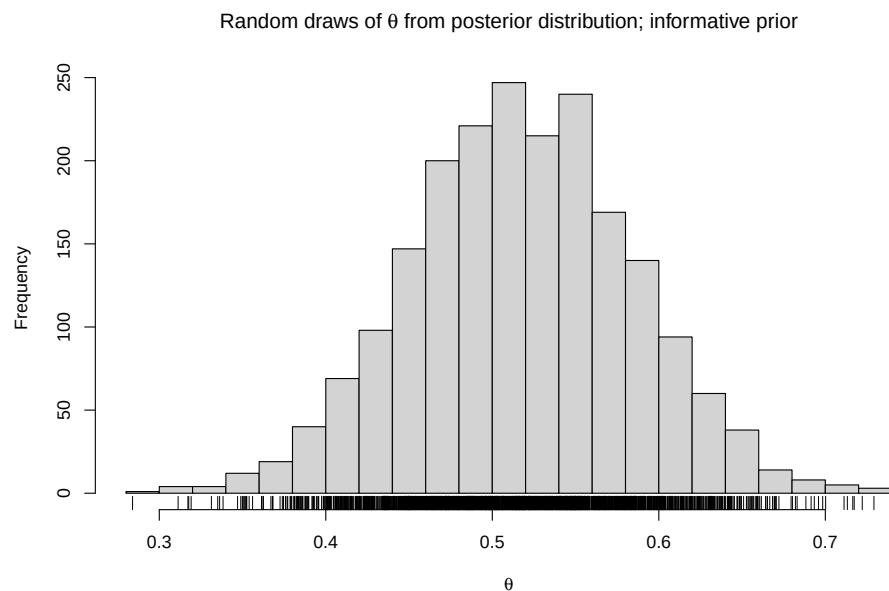
A general way to compute the predictive distribution is by *simulation*

1. Draw random values of the parameter θ from the posterior distribution for it, i.e., the Beta distribution with the posterior values of its hyperparameters;

2. Draw random values of the number of successes from Binomial distributions parameterized by these random values of its parameter.

TASK 19 : Draw 2048 random values from the posterior distribution of θ , using the `rbeta` “random value from the Beta distribution” function, based on the *informative* prior $\text{Beta}(19, 13)$ and the experiment with outcome $k = 10, n = 24$. •

```
n.sim <- 2048
alpha <- prior.a+k; beta=prior.b+(n-k)
g <- rbeta(n.sim, shape1=alpha, shape2=beta)
hist(g, xlab=expression(theta),
     main=expression(paste(plain("Random draws of "),
                             theta,
                             plain(" from posterior distribution; informative prior"))),
     breaks=24)
rug(g)
```

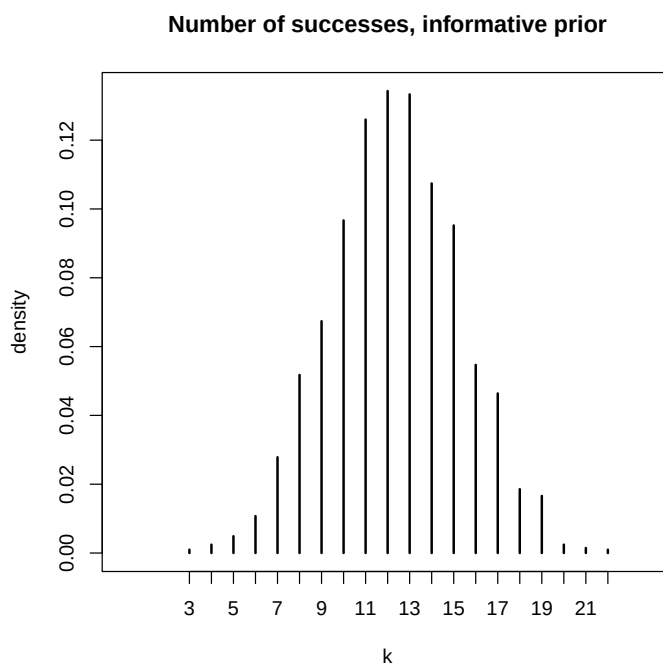


TASK 20 : Based on the random values of θ , draw random numbers of successes from the Binomial distribution of the number of successes k in n trials, using the `rbinom` “random draw from the binomial distribution” function; summarize these with the `table` function. •

```
(table.k <- table(pred.k <- rbinom(n.sim, size=n, prob=g)))

##
##  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
##  2  5 10 22 57 106 138 198 258 275 273 220 195 112 95 38
## 19 20 21 22
## 34  5  3  2

plot(table.k/n.sim, xlab="k", ylab="density", xlim=c(0,24),
     main="Number of successes, informative prior")
```



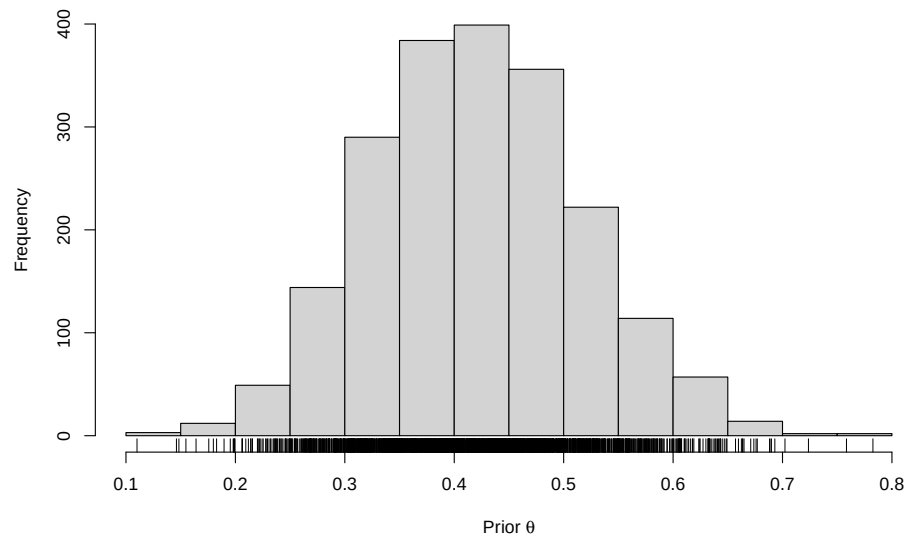
Note: Because the argument `g` to `rbinom` is the same length as the argument for the number of simulations `n.sim`, the function operates on equal-length vectors, thus each draw is from a Binomial distribution has its own value of θ , as drawn from the posterior Beta distribution.

These results are based on two random draws: (1) the parameters of the Binomial distribution, and then (2) the number of successes in 24 trials, given that parameter. In this case the resulting mode is $k = 12$, a compromise between the informative prior (mode $k = 18$, $n = 30$) and the likelihood function (mode $k = 10$, $n = 24$).

TASK 21 : Repeat the above two draws with the posterior distribution obtained from the *non-informative* prior and the experiment with outcome $k = 10$, $n = 24$. •

```
alpha <- 1+k; beta=1+(n-k)
g <- rbeta(n.sim, shape1=alpha, shape2=beta)
hist(g, xlab=expression(paste(plain("Prior ")), theta)),
     main=expression(paste(plain("Random draws of "),
                             theta,
                             plain(" from posterior distribution; non-informative prior"))),
     breaks=24)
rug(g)
```

Random draws of θ from posterior distribution; non-informative prior

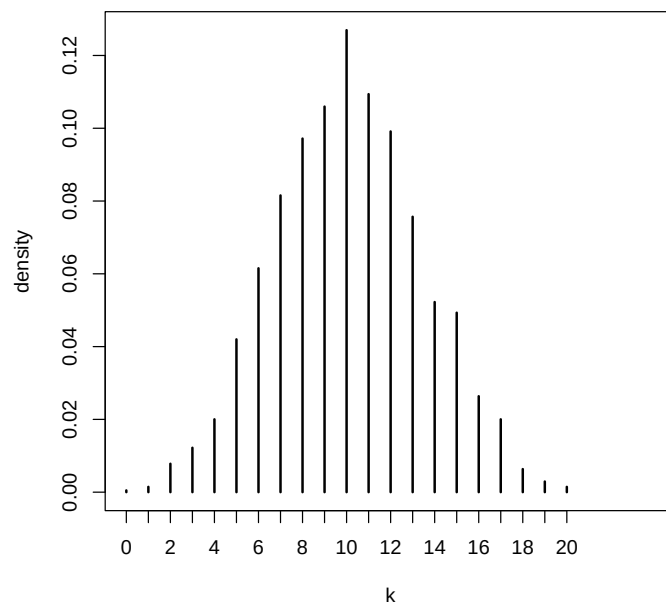


```
(table.k <- table(pred.k <- rbinom(n.sim, size=n, prob=g)))

##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
##  1  3 16 25 41 86 126 167 199 217 260 224 203 155 107 101
## 16 17 18 19 20
## 54 41 13  6  3

plot(table.k/n.sim, xlab="k", ylab="density", xlim=c(0,24),
      main="Number of successes, non-informative prior")
```

Number of successes, non-informative prior



The mode should be $k = 10$ but because of the simulation it may be displaced. The larger the number of simulations, the closer to the theo-

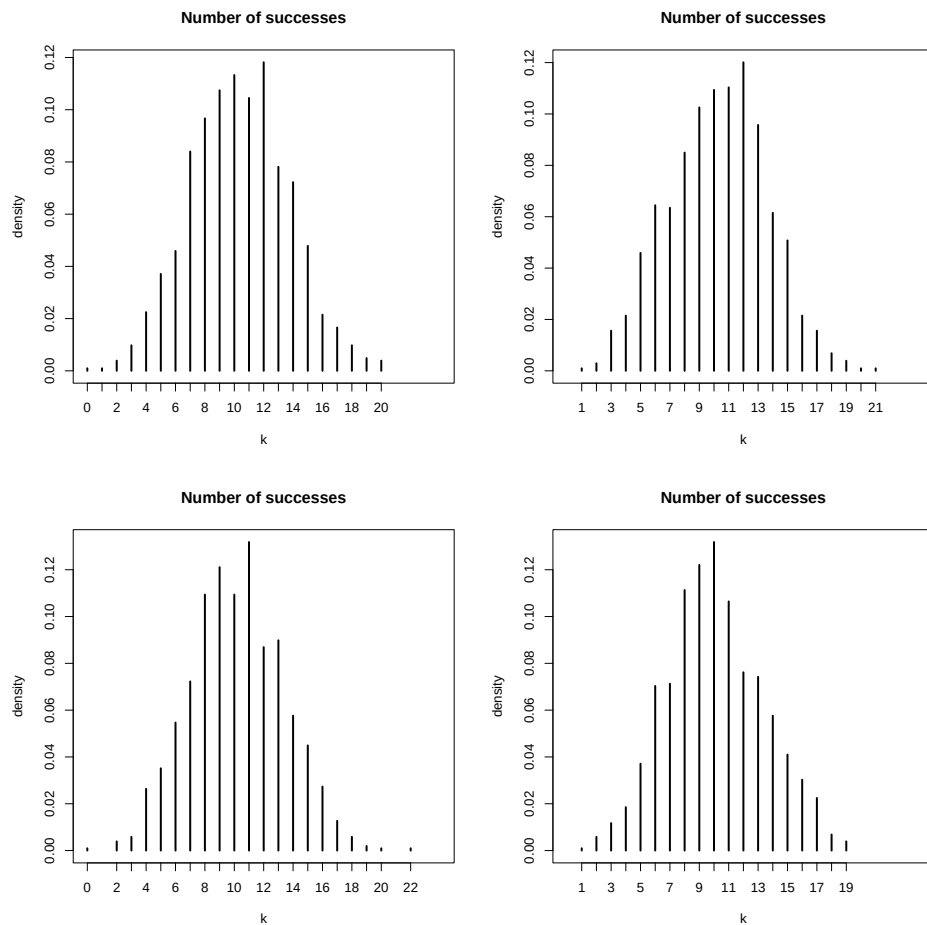
retical value.

As expected, there is a large difference between the predictions based on the posterior distribution of the parameters.

We can see the effect of simulation by comparing results with the same number of simulations and same distributions.

TASK 22 : Compute four versions of the predictive distribution of the number of successes k in 24 trials, from the posterior distribution of θ computed from 1024 simulations of the (10, 24) trial and the non-informative prior distribution for θ . Plot them in a 2x2 display. •

```
n.sim <- 1024
par(mfrow=c(2,2))
alpha <- 1+k; beta=1+(n-k)
for (i in 1:4) {
  g <- rbeta(n.sim, shape1=alpha, shape2=beta)
  table.k <- table(pred.k <- rbinom(n.sim, size=n, prob=g))
  plot(table.k/n.sim, xlab="k", ylab="density",
       main="Number of successes",
       xlim=c(0,24))
}
par(mfrow=c(1,1))
```



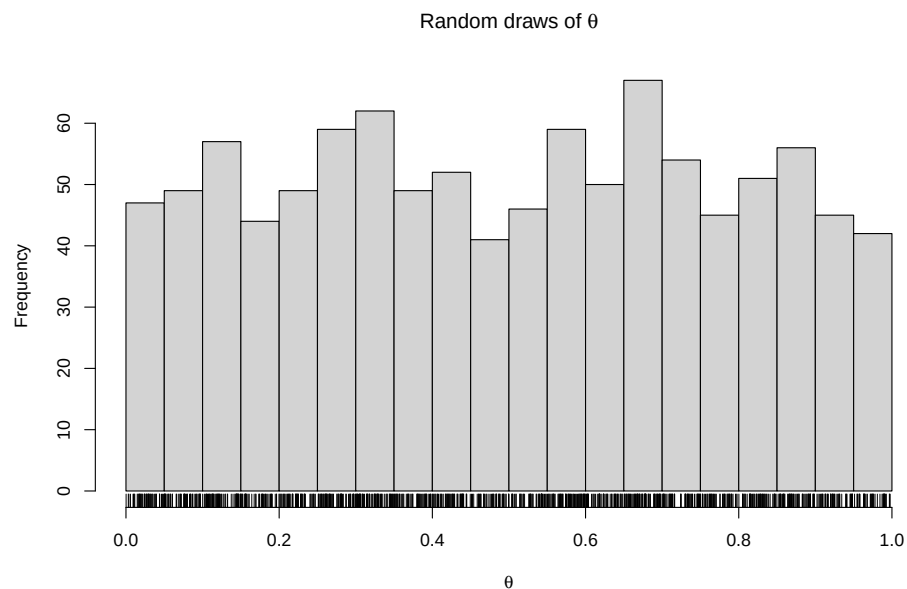
Challenge: Repeat the predictions with more or fewer simulations, each

several times, to see the effect of the number of simulations on the consistency of the predictions.

It's interesting to see the predictions based on the non-informative prior alone; this is a *prior* prediction, i.e., what we expect to see before we've conducted the experiment (for example, taking soil samples).

TASK 23 : Draw 1024 random values from the posterior distribution of θ based on the *reference* prior Beta(1, 1); plot these as a histogram. •

```
alpha <- 1; beta=1
g <- rbeta(n.sim, alpha, beta)
hist(g, xlab=expression(theta), freq=T,
     main=expression(paste(plain("Random draws of "),
                             theta)),
     breaks=24)
rug(g)
```



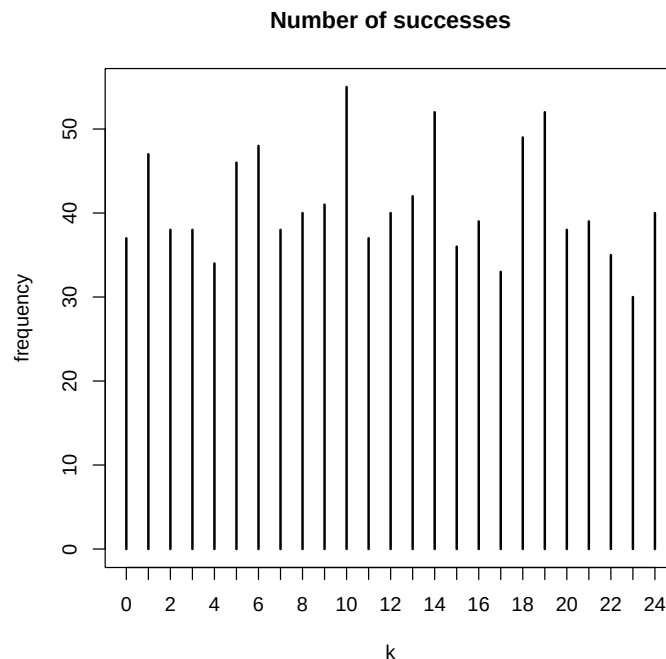
This should be a uniform distribution (see the graph of Beta(1, 1) above), but because of random sampling it has some fluctuations.

TASK 24 : Use these simulated values of the binomial parameter to predict the number of successes in 24 trials, again using 1024 simulated trials. •

```
(table.k <- table(pred.k <- rbinom(n.sim, size=n, prob=g)))

##
## 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
## 37 47 38 38 34 46 48 38 40 41 55 37 40 42 52 36 39 33 49 52 38 39
## 22 23 24
## 35 30 40

plot(table.k, xlab="k", ylab="frequency",
     main="Number of successes")
```



Again, this should be uniform, but deviates from a uniform distribution due to random sampling. All values from 0 to 24 are equally likely, because all values of θ are equally likely. As expected there is no clear mode. This is our Bayesian prediction from a non-informative prior before any data is collected.

Challenge: Repeat the predictions with more or fewer simulations, each several times, to see the effect of the number of simulations on the consistency of the predictions.

3.8 Model checking

We use the posterior density to perform inference about the parameter of interest (e.g., find the mode or credible intervals), but this depends on the Bayesian model selected to represent the problem, i.e., the sampling density, and the assumptions made in parameterizing the prior density. How can we have confidence that these choices are valid?

One method is to inspect the *predictive* density, but applied to the data we have actually observed. If the observed data values are consistent with the predictive density, this is evidence that the model is appropriate. But if the observed data are quite unusual in terms of the predictive density, i.e., they are in the tails of the density, then something is likely wrong. Either we have selected an inappropriate model, or we have mis-specified the sampling or prior densities.

For example, using the informative prior centred on 0.6 with 30 supposed trials, i.e., the Beta(19, 13) prior, the posterior predictive distribution estimates that 10 successes of 24 trials (what we observed in the

experiment), would be found in about 10% of the 1024 simulated predictive draws (see §3.7 first example), and 10 or less in about 42% of the draws. These are not excessively small, so we would not, on this basis, reject either our model form nor our prior probability distribution for θ .

4 Hierarchical models

A *hierarchical* model, also called a *multilevel* model, is one where several posterior distributions must be estimated, with some depending on others.

The example we will develop is a *multinomial mixture* of *binomial distributions*. These arise when the population of interest is divided into m separate groups, each with its own separate binomial distribution with a parameter ‘proportion of successes’ θ_j , resulting in a probability $p(k_j)$ of exactly k_j successes in n_j trials:

$$p(k_j) = \binom{n_j}{k_j} \theta_j^{k_j} (1 - \theta_j)^{n_j - k_j} \quad (35)$$

Further, the division of the population into groups is also probabilistic and represented by a *multinomial* distribution, the multivariate extension of the binomial distribution, with the probability ψ_j of being in a given group j .

Note: As before we use θ for the parameter of a binomial distribution, so we introduce ψ (‘psi’) for the parameter of the multinomial distribution.

The probability mass function of the multivariate model is:

$$\begin{aligned} f(n_1, n_2, \dots, n_m; n; \psi_1, \psi_2, \dots, \psi_m) &= \Pr(X_1 = n_1, X_2 = n_2, \dots, X_m = n_m) \\ &= \frac{n!}{n_1! n_2! \dots n_m!} \psi_1^{n_1} \psi_2^{n_2} \dots \psi_m^{n_m} \end{aligned} \quad (36)$$

This is the probability of exactly n_1 items of the n total being of type 1, n_2 items being of type 2, etc.; these then become the number of observations for each of the separate binomial distributions. These numbers depend on the total number of observations n and the model parameters ψ_j .

Note: If $m = 2$ this reduces to the binomial distribution of Equation 11.

TASK 25 : Use the `dmultinom` “multinomial density” function to compute the probability of observing 10 of each of three types in 30 observations, assuming $\psi_1 = 0.2, \psi_2 = 0.5, \psi_3 = 0.3$. •

```
dmultinom(c(10,10,10), prob=c(.2,.5,.3))
```

```
## [1] 0.003277808
```

TASK 26 : Visualize a random sample of 128 draws of 100 items from this distribution, created with the `rmultinom` “random sample from the multinomial distribution” function, on a *ternary diagram*, created with the `coord_tern` function of the `ggtern` package, applied to a `ggplot` object of the `ggplot2` graphics package.

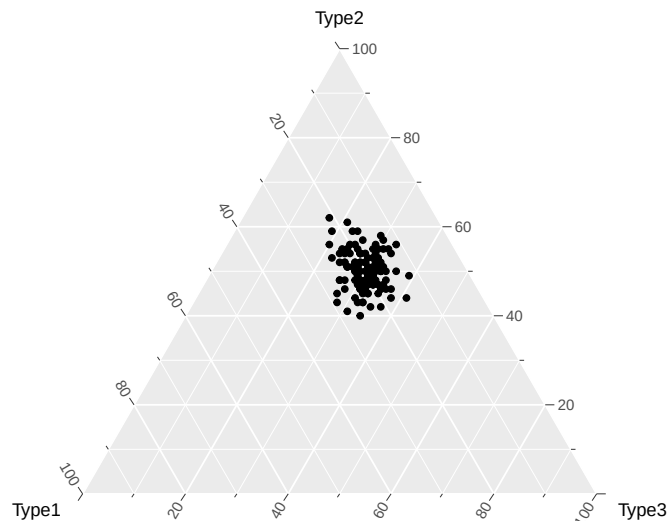
```
require(ggtern)
samp <- t(rmultinom(n=128, size=100, prob=c(.2,.5,.3)))
samp <- as.data.frame(samp)
names(samp) <- c("Type1", "Type2", "Type3")
summary(samp)
```

	Type1	Type2	Type3
## Min.	:11.00	Min. :40.00	Min. :17.00
## 1st Qu.:	18.00	1st Qu.:47.00	1st Qu.:27.00
## Median :	20.00	Median :50.00	Median :30.00
## Mean :	19.95	Mean :50.29	Mean :29.76
## 3rd Qu.:	22.00	3rd Qu.:53.00	3rd Qu.:32.00
## Max. :	29.00	Max. :62.00	Max. :41.00

```
summary(apply(samp, 1, "sum")) # all draws sum to 100, i.e., total items
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	100	100	100	100	100	100

```
ggplot(data=samp, mapping= aes(Type1, Type2, Type3)) +
  coord_tern() +
  geom_point()
```



These are scattered around the most likely value, (20, 50, 30).

Suppose we design a soil sampling campaign where we will make a fixed number n of spatially-random observations, constrained by the budget, to determine soils that require remedial applications of some fertilizer (lime, P, ...) based on some user-defined threshold, e.g., a critical level

determined by previous field experiments. Each of these observations can be of one of several soil types; we recognize this as a realization of the multinomial distribution. The number of observations in soil type j is n_j , and $\sum_i n_i = n$. Within each soil type there is a proportion of soils θ_j that will fall below the threshold, so that k_j of the n_j samples will require remedial applications. These are expected to be different for the different soil types. We recognize these as separate realizations, per soil type, of the binomial distribution treated in §3.

Prior to sampling we do not know the distribution of observations into soil types: this will be assessed in the field. Nor do we know the proportion of each that will fall below the threshold: this will be determined in the laboratory. At the end of the procedure we want to know:

1. the proportion of soil types in the study area;
2. the conditional probability of the need for remedial action for each soil type;
3. the number of sites which require remediation;
4. the probability that a new random observation under the same conditions will require remediation.

All of these are probability distributions, not single values. Of course we can compute the single values of (1-3) from the actual observations (the *sample*), but we are interested in these values for the *population*, i.e., all the soils in this study area. Question (4) can only be answered with a *predictive* distribution, not simply from the proportions in the sample.

This is a *multilevel model*, with the following levels:

- Level 1 : $k_j \mid \theta_j, n_j \sim \text{Binomial}(\theta_j, n_j)$, the number of observations of the total n_j in soil type j requiring remediation;
- Level 2 : $\theta_j \mid \alpha_j, \beta_j \sim \text{Beta}(\alpha_j, \beta_j)$, the distribution for the binomial parameter θ_j in soil type j ;
- Level 3 : $n_j \mid \psi_1, \psi_2, \dots, \psi_m, n \sim \text{Multinomial}(\psi_1, \psi_2, \dots, \psi_m, n)$, the number of observations of soil type j , out of the total number of observation n , for each of the m possible soil types;
- Level 4 : $\psi_j \mid \alpha_1, \alpha_2 \dots \alpha_m \sim \text{Dirichlet}(\alpha_1, \alpha_2 \dots \alpha_m)$, the distribution of the m multinomial parameters.

where the Dirichlet distribution is the multivariate analogue of the Beta distribution:

$$D(\alpha) = \frac{1}{\mathbf{B}(\alpha)} \prod_{j=1}^m x_j^{\alpha_j-1} \quad (37)$$

with m shape parameters α_j , and with the requirement that the proportion x_j form an $m - 1$ simplex, i.e., $\sum_{j=1}^m x_j = 1$.

The denominator of the normalizing constant is:

$$\mathbf{B}(\alpha) = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^m \alpha_j)} \quad (38)$$

If $m = 2$, this reduces to the Beta distribution of Equation 19 with , $\alpha_1 = \alpha$, $\alpha_2 = \beta$.

The Dirichlet distribution is the conjugate prior for the multinomial distribution. This is the multivariate extension of the Beta distribution, as Multinomial is the multivariate extension of the Binomial. It gives us the distribution of the number of observations in each of the m classes. We then select a separate Beta distribution as the conjugate prior for each of the m binomial distributions This gives us the distribution of the number of soils needing remediation in each of the classes.

Suppose there are three major soil types in the study area:

1. alluvial soils along rivers;
2. well-drained upland soils; and
3. thin, rocky soils on steep hillsides.

Suppose further than in our random field survey of 128 soil samples, field observation classifies them as 24 alluvial, 64 upland, and 40 hillside. Of these, 12, 20 and 10, respectively were found to be above the action threshold.

TASK 27 : Assign these numbers to R variables, so that they can be changed one time (here) and the subsequent code can be run with the changed values. •

```
n.class <- c(24, 64, 40)
n.action <- c(12, 20, 10)
n <- sum(n.class)
```

Hierarchical analysis with non-informative priors We start with the reference *non-informative* prior for the Dirichlet distribution for three classes: $\alpha_1 = \alpha_2 = \alpha_3 = 1$; this is by analogy to the reference non-informative prior for the Beta distribution:

$$\frac{\Gamma(1 + 1 + 1)}{\Gamma(1)\Gamma(1)\Gamma(1)} x_1^0 x_2^0 x_3^0 = \frac{3!}{1 \cdot 1 \cdot 1} 1 \cdot 1 \cdot 1 = 2 \quad (39)$$

since $\Gamma(3) \equiv (3 - 1)! = 2! = 2$, $\Gamma(1) = (1 - 1)! = 0! \equiv 1$ and $x^0 \equiv 1$. This is uniform, i.e., does not depend on the counts x_i .

TASK 28 : To visualize this, draw a sample from this non-informative prior distribution, to get possible realizations of the proportions, i.e., the parameters of the multinomial distribution, using the `rdirichlet` function of the `MCMCpack` package. Display on a ternary diagram. Save

the ternary diagram as a graphics object for later comparison with the informative case.

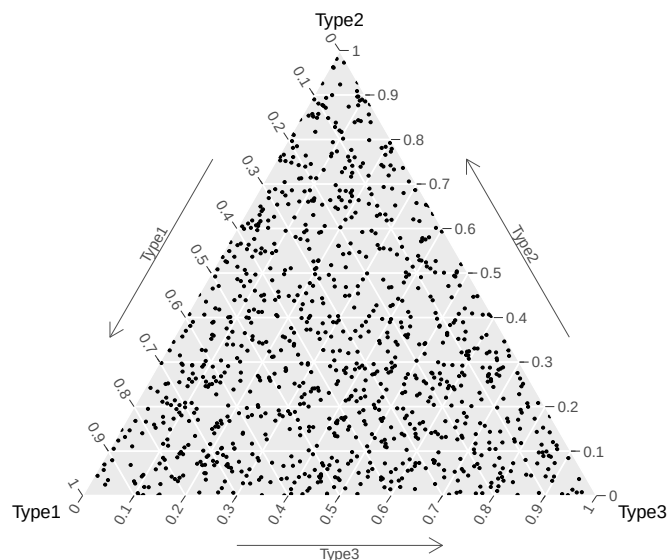
The parameters of the Dirichlet distribution are (1,1,1), i.e., noninformative for the three classes; this can be written more simply using the rep “repeat” function as rep(1,3).

```
require(MCMCpack)
n.sim <- 1024
samp.prior <- as.data.frame(rdirichlet(n=n.sim, alpha=rep(1,3)))
names(samp.prior) <- c("Type1", "Type2", "Type3")
summary(samp.prior)

##      Type1      Type2      Type3
## Min. :0.0009021 Min. :0.0001057 Min. :0.0007329
## 1st Qu.:0.1439104 1st Qu.:0.1338795 1st Qu.:0.1312275
## Median :0.3069527 Median :0.2780973 Median :0.2960234
## Mean :0.3370223 Mean :0.3282380 Mean :0.3347396
## 3rd Qu.:0.4984405 3rd Qu.:0.4945408 3rd Qu.:0.5046303
## Max. :0.9486909 Max. :0.9811721 Max. :0.9673029

axis <- seq(0, 1, 0.1)
g.dir.noninf.prior <- ggplot(data=samp.prior,
  mapping= aes(Type1, Type2, Type3)) +
  coord_tern() +
  theme_showarrows() +
  scale_T_continuous(breaks = axis, labels = axis) +
  scale_L_continuous(breaks = axis, labels = axis) +
  scale_R_continuous(breaks = axis, labels = axis) +
  geom_point(size=0.5) +
  ggtitle("Non-informative prior")
plot(g.dir.noninf.prior)
```

Non-informative prior



This figure shows clearly that the prior is uniform for all three classes, with no correlation between them.

We now update to the posterior Dirichlet distribution, using the evidence

from the field survey:

$$p(x_1, x_2, x_3) \propto x_1^{25} x_2^{65} x_3^{41} \quad (40)$$

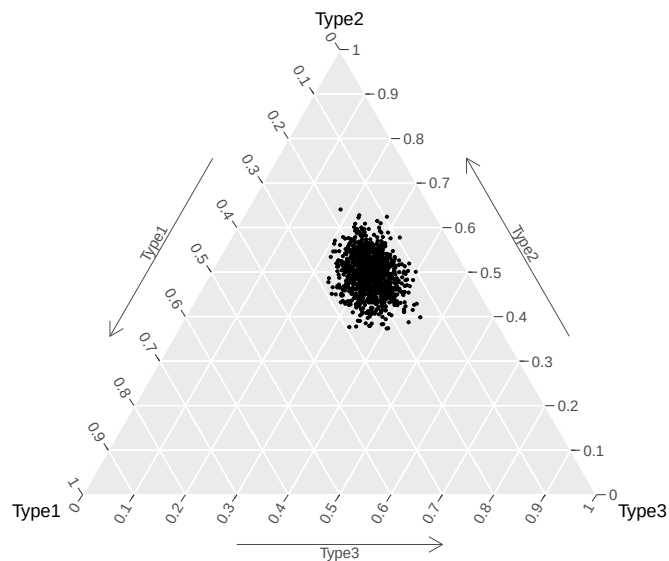
TASK 29 : Draw a sample from this posterior distribution, to get possible realizations of the proportions. Display on a ternary diagram. Save the ternary diagram as a graphics object for later comparison with the informative case. •

```
n.sim <- 1024
samp.post <- as.data.frame(rdirichlet(n.sim, n.class+rep(1,3)))
names(samp.post) <- c("Type1", "Type2", "Type3")
summary(samp.post)

##      Type1      Type2      Type3
## Min.   :0.09403   Min.   :0.3734   Min.   :0.1815
## 1st Qu.:0.16756   1st Qu.:0.4640   1st Qu.:0.2821
## Median :0.18979   Median :0.4976   Median :0.3094
## Mean   :0.19127   Mean    :0.4968   Mean    :0.3120
## 3rd Qu.:0.21355   3rd Qu.:0.5301   3rd Qu.:0.3403
## Max.    :0.29275   Max.    :0.6408   Max.    :0.4577

axis <- seq(0, 1, 0.1)
g.dir.noninf.post <- ggplot(data=samp.post,
  mapping= aes(Type1, Type2, Type3)) +
  coord_tern() +
  theme_showarrows() +
  scale_T_continuous(breaks = axis, labels = axis) +
  scale_L_continuous(breaks = axis, labels = axis) +
  scale_R_continuous(breaks = axis, labels = axis) +
  geom_point(size=0.5) +
  ggtitle("Posterior from non-informative prior")
plot(g.dir.noninf.post)
```

Posterior from non-informative prior



The evidence has concentrated the posterior distribution into one compact area of the ternary diagram, i.e., parameter space.

TASK 30 : Multiply these predicted proportions by the total sample size to obtain realizations of the multinomial distribution; Use the `round` function to ensure integer numbers of per-class observations. Display the results numerically with the `summary` function. Also, summarize the sum of the rows obtained with the `apply` “apply a function row- or column-wise” function. •

```
samp.n <- round(n*samp.post)
dim(samp.n)

## [1] 1024      3

summary(samp.n)

##      Type1      Type2      Type3
## Min.   :12.00  Min.   :48.00  Min.   :23.00
## 1st Qu.:21.00  1st Qu.:59.00  1st Qu.:36.00
## Median :24.00  Median :64.00  Median :40.00
## Mean   :24.48  Mean   :63.58  Mean   :39.94
## 3rd Qu.:27.00  3rd Qu.:68.00  3rd Qu.:44.00
## Max.   :37.00  Max.   :82.00  Max.   :59.00

summary(apply(samp.n, 1, sum)) # close to 128

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      127    128    128     128    128     129
```

Because of the rounding, the total number of samples per simulation varies a bit, generally 128 ± 1 .

Now we need to analyze the next level of the model: the proportion of observations above the action threshold for each soil type. This is the Binomial case, for which the prior distribution is the Beta distribution, as examined in §3.3. The non-informative values for the hyperparameters are $\alpha = \beta = 1$; these are updated with the evidence. These are per-class, taking advantage of R’s vectorized operations:

TASK 31 : Update the hyperparameters of the Binomial distribution, based on the evidence. •

```
(alpha.n <- 1 + n.action)

## [1] 13 21 11

(beta.n <- 1 + (n.class - n.action))

## [1] 13 45 31
```

TASK 32 : Use these values of the hyperparameters to simulate draws from the Beta distribution for each class. Summarize the results; compare to the observed proportions in the sample of each class. •

```
# class 1
sim.1.p <- rbeta(n=n.sim, shape1=alpha.n[1], shape2=beta.n[1])
```

```

sim.2.p <- rbeta(n=n.sim, shape1=alpha.n[2], shape2=beta.n[2])
sim.3.p <- rbeta(n=n.sim, shape1=alpha.n[3], shape2=beta.n[3])
summary(sim.1.p); summary(sim.2.p); summary(sim.3.p)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1783  0.4328  0.4972  0.4992  0.5675  0.7874
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1553  0.2753  0.3137  0.3154  0.3512  0.4981
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06369 0.21078 0.25258 0.25682 0.29500 0.52723

(prop.obs <- n.action/n.class)

## [1] 0.5000 0.3125 0.2500

```

The medians and means of the simulations are close to the observed proportions of the experiment: 0.5, 0.3125, 0.25; however with a large number of simulations some will be quite far from these.

Now we can simulate how many observations are over the action threshold.

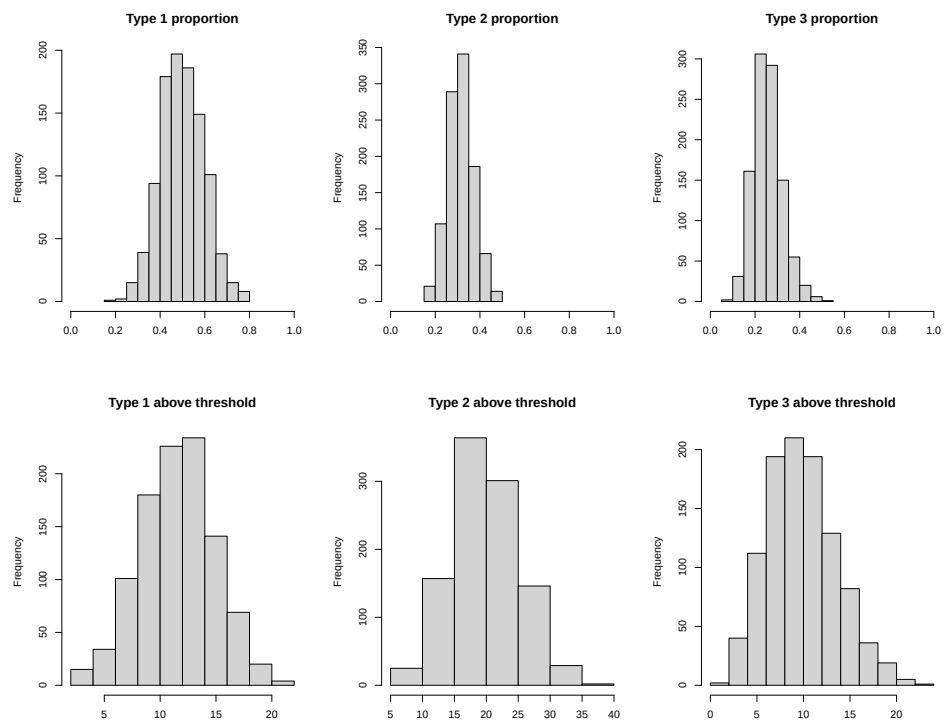
TASK 33 : Simulate the number of observations over the action threshold, for each simulated distribution of observations within classes. Display as per-class histograms of proportions and numbers above the threshold.

The most efficient way to do this is to match each binomial draw from a simulated number of observations with a single simulated parameter from the previous step. Another approach would be to use all simulated proportions for each binomial draw, but this would result in $10^{10} \cdot 10^{10} = 10^{20}$ draws.

```

sim.1.n <- rbinom(n=n.sim, size=n.class[1], prob=sim.1.p)
sim.2.n <- rbinom(n=n.sim, size=n.class[2], prob=sim.2.p)
sim.3.n <- rbinom(n=n.sim, size=n.class[3], prob=sim.3.p)
par(mfrow=c(2,3))
hist(sim.1.p, xlim=c(0,1), xlab="", main="Type 1 proportion")
hist(sim.2.p, xlim=c(0,1), xlab="", main="Type 2 proportion")
hist(sim.3.p, xlim=c(0,1), xlab="", main="Type 3 proportion")
hist(sim.1.n, xlab="", main="Type 1 above threshold")
hist(sim.2.n, xlab="", main="Type 2 above threshold")
hist(sim.3.n, xlab="", main="Type 3 above threshold")
par(mfrow=c(1,1))

```

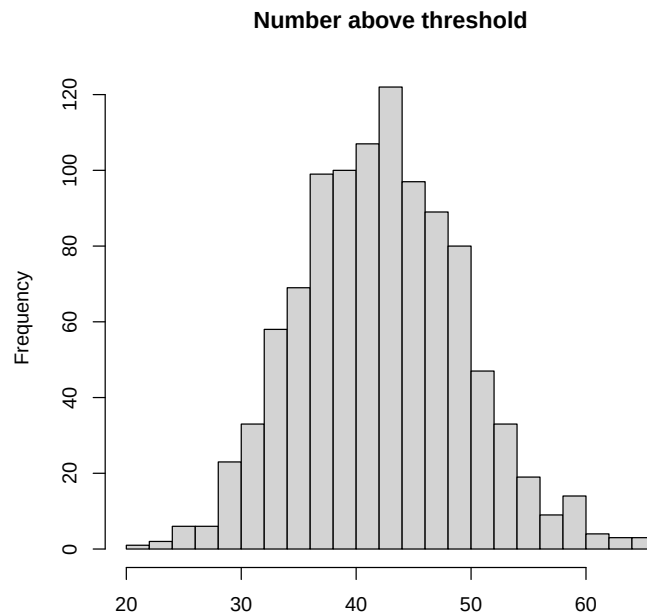


TASK 34 : Sum the numbers above the threshold for each class, to obtain the simulated numbers of the total above the threshold. Display as a histogram. •

```
summary(sim.n <- sim.1.n + sim.2.n + sim.3.n)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      21.0   38.0   43.0   42.7   47.0   66.0

hist(sim.n, main="Number above threshold",
      xlab="", breaks=24)
```



TASK 35 : Compute the 90% credible interval for the number of points above the threshold. Round to the nearest integer with the round function, because only integer counts of soil samples are possible. •

```
round(quantile(sim.n, probs=c(0.05, 0.95)))
```

```
## 5% 95%
## 31 55
```

The analysis shows that in a completely random sample in this study area with three soil types, based on the one field campaign, if we repeat the campaign with the same number of observations, our belief should be that, with 90% probability, we will find between 31 and 55 of the 128 above the action threshold.

Notice that we have a hierarchy of uncertainty

1. the true proportion of the soil types in the area;
2. sampling uncertainty because of random allocation of observations across the study area;
3. the true proportion of soils with values above the action threshold for each soil type;
4. sampling uncertainty because of random allocation of observations within each soil type.

We used the single field trial to update our beliefs in (1) and (3), leading to distributions for (2) and (4).

Hierarchical analysis with informative priors We now repeat the analysis of the previous section, but with an *informative* prior.

At the top level of the analysis, suppose that by expert knowledge or previous survey we estimate the proportions of the three soil types in the area as 0.2, 0.5, 0.3, respectively. We estimate this evidence to be equivalent to 20 total “pseudobservations”, so there are 4, 10, and 6 “pseudocounts”, respectively. This corresponds to a prior Dirichlet distribution:

$$\frac{\Gamma(20)}{\Gamma(4)\Gamma(10)\Gamma(6)} x_1^{4-1} x_2^{10-1} x_3^{6-1} = \frac{20!}{3! \cdot 9! \cdot 5!} \cdot x_1^3 x_2^9 x_3^5 \quad (41)$$

TASK 36 : Assign these parameter values to an R vector, so they can be changed and the subsequent analysis be rerun. •

```
alpha <- c(4, 10, 6)
```

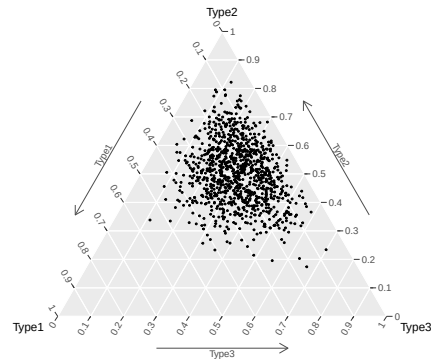
TASK 37 : Draw a sample from this informative prior distribution, to get possible realizations of the proportions. This updates the prior with the observations. Display on a ternary diagram, side-by-side with the diagram based on the non-informative prior computed above. •

```
n.sim <- 1024
samp.prior <- as.data.frame(rdirichlet(n.sim, alpha))
names(samp.prior) <- c("Type1", "Type2", "Type3")
summary(samp.prior)
```

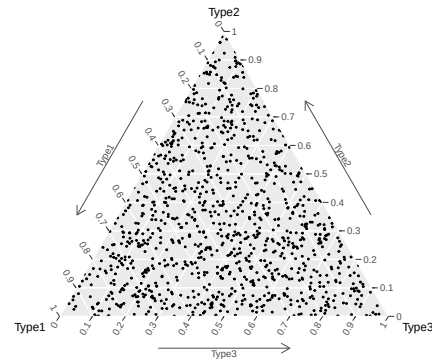
	Type1	Type2	Type3
## Min.	:0.02159	Min. :0.1739	Min. :0.06031
## 1st Qu.:	:0.13735	1st Qu.:0.4247	1st Qu.:0.21872
## Median :	:0.19354	Median :0.5038	Median :0.29207
## Mean :	:0.20157	Mean :0.5019	Mean :0.29657
## 3rd Qu.:	:0.25955	3rd Qu.:0.5762	3rd Qu.:0.36746
## Max.	:0.55039	Max. :0.8217	Max. :0.69896

```
axis <- seq(0, 1, 0.1)
g.dir.inf.prior <- ggplot(data=samp.prior,
  mapping= aes(Type1, Type2, Type3)) +
  coord_tern() +
  theme_showarrows() +
  scale_T_continuous(breaks = axis, labels = axis) +
  scale_L_continuous(breaks = axis, labels = axis) +
  scale_R_continuous(breaks = axis, labels = axis) +
  geom_point(size=0.5) +
  ggtitle("Informative prior")
plot(g.dir.inf.prior)
plot(g.dir.noninf.prior)
```

Informative prior



Non-informative prior



The informative prior distribution is somewhat diffuse, because of the low number of pseudocounts, but centred on the assumed proportion (0.2, 0.5, 0.3). Unlike the uninformative prior, there are some areas of the ternary diagram with no simulated points, i.e., these combinations of proportions are highly improbable.

Again we update by the field observations:

$$p(x_1, x_2, x_3) \propto x_1^{3+24} x_2^{9+64} x_3^{5+40} \quad (42)$$

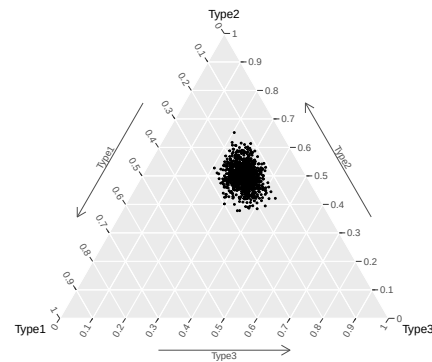
TASK 38 : Draw a sample from this posterior distribution, to get possible realizations of the proportions. This updates the prior with the observations. Display on a ternary diagram, side-by-side with the diagram based on the non-informative prior computed above. •

```
n.sim <- 1024
samp.post.inf <- as.data.frame(rdirichlet(n.sim, n.class+alpha))
names(samp.post.inf) <- c("Type1", "Type2", "Type3")
summary(samp.post.inf)
```

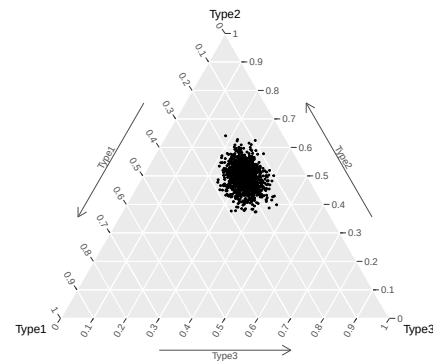
	Type1	Type2	Type3
## Min.	:0.1075	Min. :0.3778	Min. :0.2044
## 1st Qu.:	:0.1675	1st Qu.:0.4726	1st Qu.:0.2844
## Median :	:0.1896	Median :0.4999	Median :0.3080
## Mean :	:0.1905	Mean :0.4999	Mean :0.3097
## 3rd Qu.:	:0.2124	3rd Qu.:0.5278	3rd Qu.:0.3352
## Max.	:0.2984	Max. :0.6521	Max. :0.4448

```
axis <- seq(0, 1, 0.1)
g.dir.inf.post <- ggplot(data=samp.post.inf,
  mapping= aes(Type1, Type2, Type3)) +
  coord_tern() +
  theme_showarrows() +
  scale_T_continuous(breaks = axis, labels = axis) +
  scale_L_continuous(breaks = axis, labels = axis) +
  scale_R_continuous(breaks = axis, labels = axis) +
  geom_point(size=0.5) +
  ggtitle("Posterior from informative prior")
plot(g.dir.inf.post)
plot(g.dir.noninf.post)
```

Posterior from informative prior



Posterior from non-informative prior



There is not much difference between these posterior distributions. The ternary plots show a bit more compactness (fewer points scattered outside the “cloud”) for the informative prior.

TASK 39 : Summarize the percentage difference between proportions for the three soil types.

```
summary(100*(sort(samp.post.inf$Type1) - sort(samp.post$Type1)))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.72219 -0.11989 -0.02421 -0.07890  0.03901  1.34308

summary(100*(sort(samp.post.inf$Type2) - sort(samp.post$Type2)))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.0426 -0.1912  0.1737  0.3092  0.8444  1.5723

summary(100*(sort(samp.post.inf$Type3) - sort(samp.post$Type3)))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.3993 -0.4991 -0.1717 -0.2303  0.1295  2.2916
```

TASK 40 : Multiply these by the total sample size to obtain realizations of the multinomial distribution.

```
samp.n.inf <- round(n*samp.post.inf)
summary(samp.n.inf)

##      Type1      Type2      Type3
## Min.   :14.00 Min.   :48.00 Min.   :26.00
## 1st Qu.:21.00 1st Qu.:60.00 1st Qu.:36.00
## Median :24.00 Median :64.00 Median :39.00
## Mean   :24.37 Mean   :64.01 Mean   :39.62
## 3rd Qu.:27.00 3rd Qu.:68.00 3rd Qu.:43.00
## Max.   :38.00 Max.   :83.00 Max.   :57.00
```

Again we analyze the next level of the model: the proportion of observations above the action threshold for each soil type. To illustrate the effect of an informative prior, we will assume that previous studies on these soil types in other areas has shown the proportion of samples above the action threshold to be 0.7 (river valleys), 0.2 (uplands), 0.1 (hillslopes). Due to the size of the previous studies we consider these to be based on

100 pseudocounts each.

Then the informative values for the hyperparameters are $\alpha = 70, \beta = 30$ (Type 1), $\alpha = 20, \beta = 80$ (Type 2), $\alpha = 10, \beta = 90$ (Type 3).

TASK 41 : Update the hyperparameters with the evidence from the field trial. •

```
(alpha.n <- c(70, 20, 10) + n.action)

## [1] 82 40 20

(beta.n <- c(30, 80, 90) + (n.class - n.action))

## [1] 42 124 120
```

TASK 42 : Use these values of the hyperparameters to simulate draws from the Beta distribution for each class. •

```
# class 1
sim.1.p <- rbeta(n=n.sim, shape1=alpha.n[1], shape2=beta.n[1])
sim.2.p <- rbeta(n=n.sim, shape1=alpha.n[2], shape2=beta.n[2])
sim.3.p <- rbeta(n=n.sim, shape1=alpha.n[3], shape2=beta.n[3])
summary(sim.1.p); summary(sim.2.p); summary(sim.3.p)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.5408  0.6340  0.6636  0.6619  0.6907  0.7669
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1228  0.2210  0.2446  0.2446  0.2663  0.3740
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06078 0.12028 0.14118 0.14280 0.16155 0.27312

(prop.obs <- n.action/n.class)

## [1] 0.5000 0.3125 0.2500
```

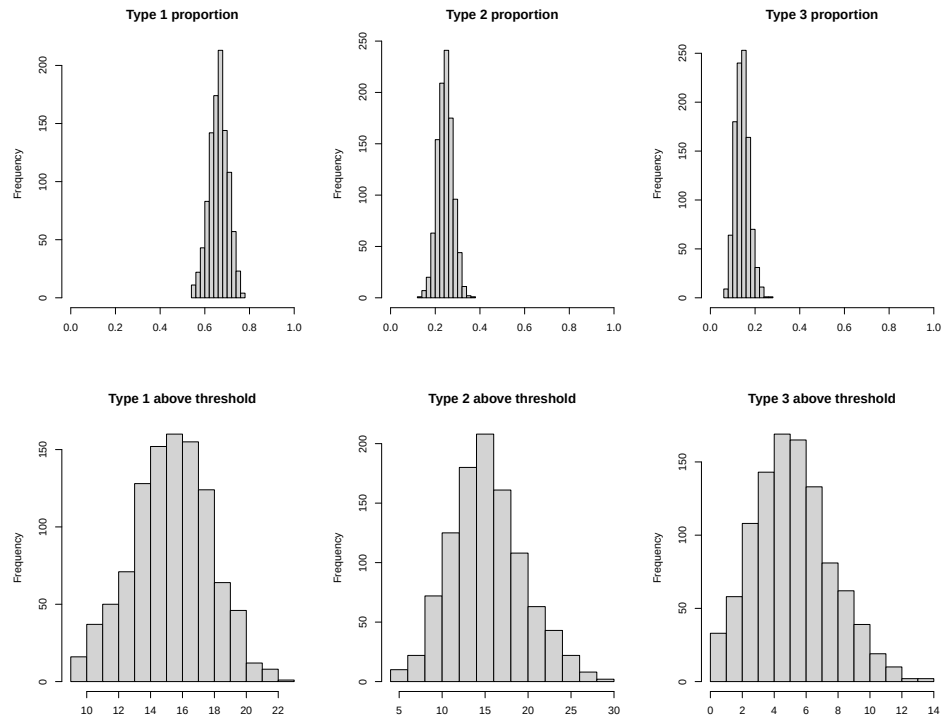
The medians and means of the simulations are between to the observed proportions of the experiment: 0.5, 0.3125, 0.25 and the prior proportions 0.7, 0.2, 0.1, i.e., there has been *shrinkage* of the prior distribution towards the evidence.

TASK 43 : Use these simulated proportions to simulate the number of observations over the action threshold, for each simulated distribution of observations within classes. Display as per-class histograms of proportions and numbers above the threshold. •

We match each binomial draw from a simulated number of observations with a single simulated parameter from the previous step.

```
sim.1.n <- rbinom(n=n.sim, size=n.class[1], prob=sim.1.p)
sim.2.n <- rbinom(n=n.sim, size=n.class[2], prob=sim.2.p)
sim.3.n <- rbinom(n=n.sim, size=n.class[3], prob=sim.3.p)
par(mfrow=c(2,3))
hist(sim.1.p, xlim=c(0,1), xlab="", main="Type 1 proportion")
hist(sim.2.p, xlim=c(0,1), xlab="", main="Type 2 proportion")
hist(sim.3.p, xlim=c(0,1), xlab="", main="Type 3 proportion")
hist(sim.1.n, xlab="", main="Type 1 above threshold")
hist(sim.2.n, xlab="", main="Type 2 above threshold")
hist(sim.3.n, xlab="", main="Type 3 above threshold")
```

```
par(mfrow=c(1,1))
```



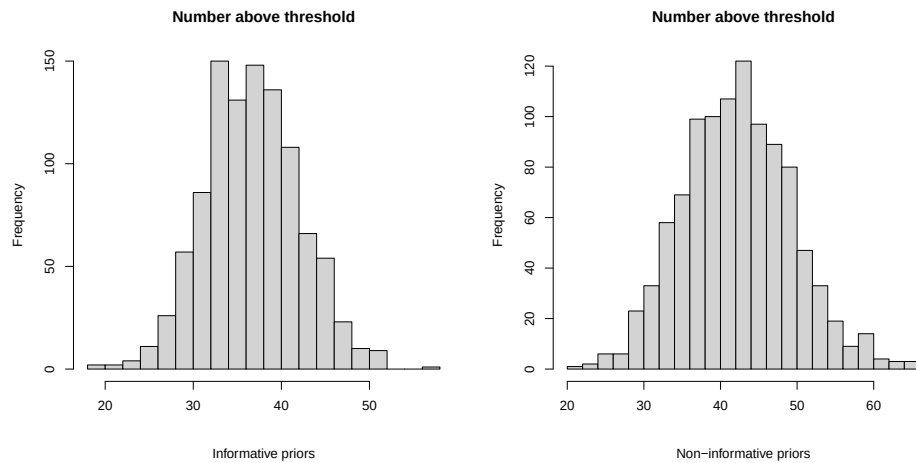
The simulated proportions of each class cover a much narrower range (more concentrated) in this case where we specified an informative prior for the Dirichlet distribution of their proportions, compared to the non-informative prior. Therefore the predicted counts above the threshold also cover a narrower range.

TASK 44 : Compute the simulated numbers of the total above the threshold. Display as a histogram, side-by-side with the histogram from the non-informative prior. •

```
summary(sim.n.i <- sim.1.n + sim.2.n + sim.3.n)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      18.00   33.00   37.00   37.19   41.00   58.00

par(mfrow=c(1,2))
hist(sim.n.i, main="Number above threshold", sub="Informative priors",
     xlab="", breaks=24)
hist(sim.n, main="Number above threshold", sub="Non-informative priors",
     xlab="", breaks=24)
par(mfrow=c(1,1))
```



TASK 45 : Compute the 90% credible interval for the informative case. Compare to the non-informative case. •

```
round(quantile(sim.n.i, probs=c(0.05, 0.95)))

## 5% 95%
## 29 46

round(quantile(sim.n, probs=c(0.05, 0.95)))

## 5% 95%
## 31 55
```

The analysis shows that in a completely random sample in this study area with three soil types, based on the one field campaign, if we repeat the campaign with the same number of observations, our belief should be that, with 90% probability, we will find between 29 and 46 of the 128 above the action threshold.

This range is narrower and with lower counts than that obtained with the non-informative prior. The narrower range is because the prior information was equivalent to a large number of pseudo-observations. The lower values is because of the prior proportions both of the proportion of observations in each class, and the proportion of each class above the action threshold.

5 Multiparameter models: the Normal distribution

In the simple example of the previous section we only had to determine the prior distribution, likelihood function, and posterior distribution for one parameter of the target distribution. Other distributions have multiple parameters. The most common multiparameter distribution is the *normal* probability distribution (also called the Gaussian distribution), which has two parameters:

1. the *location* μ , also called the *mean*;
2. the *dispersion* σ^2 , also called the *variance*. It is sometimes more

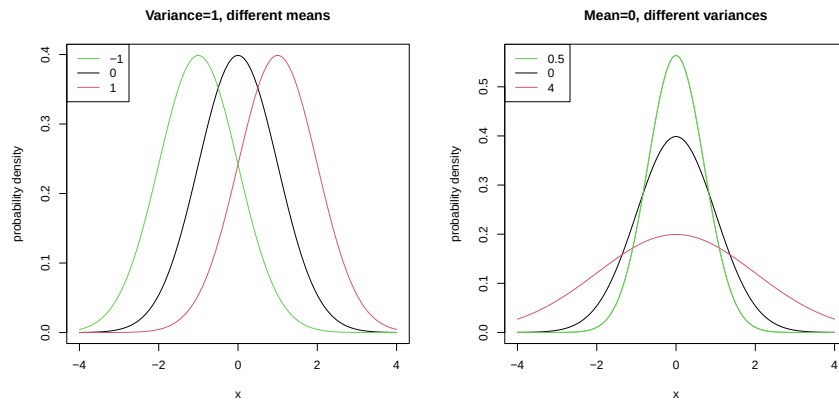
convenient to work with the inverse $1/\sigma^2$, called the *precision*, written as τ .

The density function is:

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\} \quad (43)$$

The following figure shows the Normal distribution for (1) same variance, several mean values (location centres); (2) same mean, several variances (dispersions):

```
curve(dnorm(x, mean=0, sd=1), from=-4, to=4,
      main="Variance=1, different means",
      ylab="probability density")
curve(dnorm(x, mean=1, sd=1), add=T, col=2)
curve(dnorm(x, mean=-1, sd=1), add=T, col=3)
legend("topleft", legend=c(-1, 0, 1), lty=1, col=c(3,1,2))
curve(dnorm(x, mean=0, sd=sqrt(0.5)), from=-4, to=4, col=3,
      main="Mean=0, different variances",
      ylab="probability density")
curve(dnorm(x, mean=0, sd=1), add=T)
curve(dnorm(x, mean=0, sd=sqrt(4)), add=T, col=2)
curve(dnorm(x, mean=0, sd=sqrt(0.5)), add=T, col=3)
legend("topleft", legend=c(0.5, 0, 4), lty=1, col=c(3,1,2))
```



The estimation of these parameters is based on the *likelihood* of observing a set $\mathbf{x} = [x_1, x_2, \dots, x_n]$ of independently and identically distributed (i.i.d.) variates whose the distribution is assumed to be $\mathcal{N} \sim (\mu, \sigma)$, i.e., with the parameters to be estimated. The likelihood is:

$$p(\mathbf{x} | \mu, \sigma^2) = \prod_{i=1}^n p(x_i | \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \quad (44)$$

As the parameters μ and σ^2 change, so does the likelihood of having observed the values $\mathbf{x}$.

In some situations this can be reduced to a single-parameter model:

- The location μ may be known from physical principles or the design of an experiment;

- The dispersion σ^2 may be known from previous experiments, for example calibration of a measuring instrument by a large number of repeated measurements on the same object.

But in general both parameters must be estimated together.

In the frequentist approach, these are estimated by maximum likelihood from the observations, as the sample average $\bar{x}$ and squared deviation from the sample average s^2 . Confidence intervals for these can then be constructed under the theory of repeated experiments.

However, in the Bayesian approach, both parameters have prior and posterior *distributions*, with hyperparameters. That is, we do not have a “best estimate” in the frequentist sense; rather, we have a full distribution of the probability of the values of the two parameters.

The most common distributions for these are:

- For μ , another Normal distribution, with hyperparameters (μ_0, σ_0^2) ;
- For σ^2 , the inverse χ^2 distribution with the hyperparameter ν , i.e., the degrees of freedom:

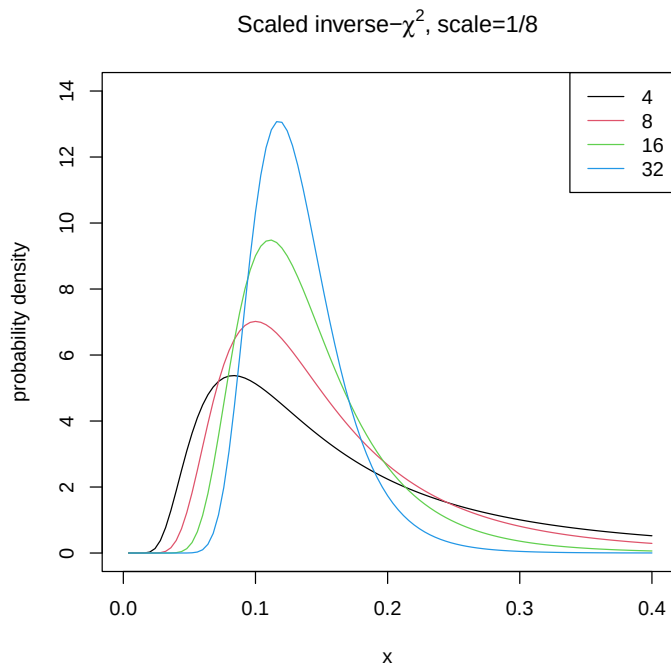
$$\chi_{\nu}^{-2}(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2)-1} e^{-1/(2x)} \quad (45)$$

The more the degrees of freedom, the more of the distribution approaches 0, i.e., the more probable that the variance σ^2 is small. The mean of this distribution is $1/(\nu - 2)$.

This is usually *scaled*, with an additional parameter $\tau^2 = 1/\sigma^2$, which is interpreted as the inverse of the variance of the presumed process. This is because the scaled inverse chi-squared distribution with ν degrees of freedom is the distribution for $x = 1/s^2$ where s^2 is a sample mean of the squares of ν independent random variables $\sim \mathcal{N}(0, 1/\sigma^2)$. That is, given that the process we want to model produces Normally-distributed random variates, this distribution gives its precision.

The following figure shows the scaled inverse χ^2 for several degrees of freedom with a constant scale:

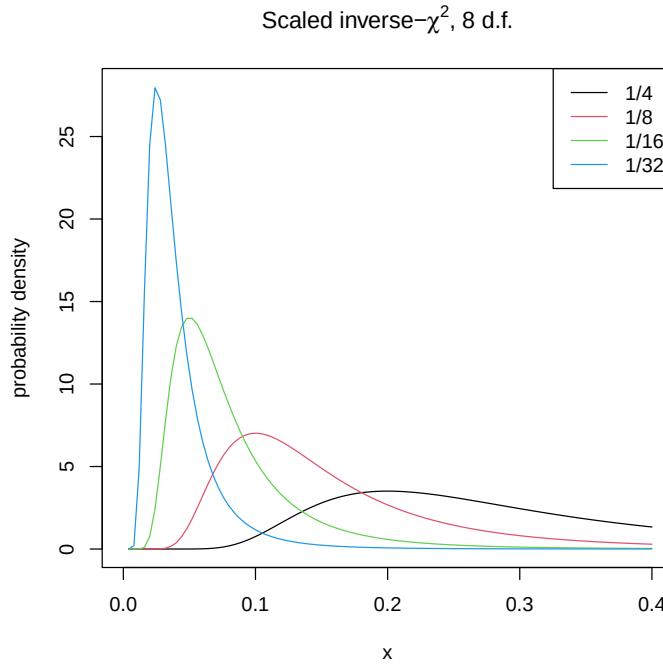
```
require(ggplot2)
curve(dinvchisq(x, df=4, scale=1/8), ylim=c(0,14),
      xlim=c(0,0.4), ylab="probability density",
      main=expression(paste("Scaled inverse-", chi^2, ", scale=1/8")))
curve(dinvchisq(x, df=8, scale=1/8), col=2, add=T)
curve(dinvchisq(x, df=16, scale=1/8), col=3, add=T)
curve(dinvchisq(x, df=32, scale=1/8), col=4, add=T)
legend("topright", legend=c("4", "8", "16", "32"), col=1:4, lty=1)
```



These have modes $(\nu\tau^2)/(\nu + 2)$, in this case 0.083, 0.1, 0.117, 0.1176471. That is, with the same scale, as the degrees of freedom (independent sampling variances of s^2) increases, the distribution shifts slightly to the right, but becomes more concentrated.

The following figure shows the scaled inverse χ^2 for eight degrees of freedom with a variable scale:

```
curve(dinvchisq(x, df=8, scale=1/4), ylim=c(0,28),
      xlim=c(0,0.4), ylab="probability density",
      main=expression(paste("Scaled inverse-", chi^2, ", 8 d.f.")))
curve(dinvchisq(x, df=8, scale=1/8), col=2, add=T)
curve(dinvchisq(x, df=8, scale=1/16), col=3, add=T)
curve(dinvchisq(x, df=8, scale=1/32), col=4, add=T)
legend("topright", legend=c("1/4", "1/8", "1/16", "1/32"), col=1:4, lty=1)
```



Here we can see how the scale influences the concentration: smaller variance of the process leads to a more concentrated density, closer to zero.

Note: Some analysts use the closely-related inverse- Γ distribution for the variance.

The Bayesian analysis updates the hyperparameters based on the observations, resulting in posterior distributions of the parameters of the Normal distribution.

6 Numerical methods

The real interest in Bayesian computation is for models which can not be reduced to analytical forms, as we could do with the Binomial example of §3. That is, the posterior distribution can not be computed as a closed form; this is often because the denominator (proportionality constant) in the fundamental Bayesian inference formula 10 has no closed form.

This constant is the integral

$$\int p(\theta) \cdot p(Y | \theta) d\theta$$

This is usually the case for complex models (multi-parameter, hierarchical...). In these cases the required integration over the parameter space must be done numerically. These models require *numerical* simulations of the posterior distribution. This requires substantial computer power and some mathematical tricks.

The most common method to simulate posterior distributions is the Markov chain Monte Carlo⁷ (MCMC) method. This is an algorithm for sampling from a (multivariate) probability distribution that can not be expressed as a closed form, based on constructing a Markov chain that has the desired distribution, e.g., posterior or predictive, as its equilibrium distribution.

A *Markov chain* is a sequence of steps of computations, each one depending on one or more previous steps. It is called *ergodic* if under certain conditions it reaches an equilibrium, so that samples from it represent the joint posterior distribution. This allows us to compute any statistic of the complete posterior distribution.

One MCMC method is the Gibbs sampler [3, 14]. It repeatedly samples from the full *conditional* distribution of each of the k parameters in the posterior distribution, one parameter i at a time:

$$p(\theta_i \mid \theta_{j \neq i}, i = 1, 2, \dots, k) \quad (46)$$

It uses the most recent draw of the remaining parameters to condition the draw for the selected parameter. Generally, these conditional distributions can be expressed analytically and simulated as we did above for posterior distributions in closed form.

To see this more explicitly [14, §3.2], suppose we have k parameters whose distribution we want to determine

1. Pick arbitrary starting values $x^0 = (x_1^0, \dots, x_k^0)$ This does not depend (yet) on the observations Y .
2. Make a *random* drawing from the full conditional distribution $\pi(x_i \mid x_{-i}, i = 1, \dots, k)$, as follows:

$$\begin{aligned} x_1^1 &\text{ from } \pi(x_1 \mid x_{-1}^0 \mid Y) \\ x_2^1 &\text{ from } \pi(x_2 \mid x_1^1, x_3^0, \dots, x_k^0 \mid Y) \\ x_3^1 &\text{ from } \pi(x_3 \mid x_1^1, x_2^1, x_4^0, \dots, x_k^0 \mid Y) \\ &\dots \\ x_k^1 &\text{ from } \pi(x_k \mid x_{-k}^1 \mid Y) \end{aligned}$$

This results in the next full conditional distribution $x^1 = (x_1^1, \dots, x_k^1 \mid Y)$. Note how at each step we sample from the full conditional distribution, but then update the next draw with the sampled value. Under certain conditions this converges to a steady-state distribution.

The MCMCpack R package⁸ allows the user to perform Bayesian inference via simulation from the posterior distributions of several popular models, including linear regression with Gaussian errors.

⁷ Just a fancy name for “random”

⁸ <http://mcmcpack.berkeley.edu/>

6.1 Numerical analysis of a Bayesian multiple linear regression model

The `MCMCregress` function of the `MCMCpack` package [10] generates a sample from the posterior distribution of a (multiple) linear regression model with Gaussian errors, using the Gibbs sampler. The prior distribution for the β vector (regressors) must be multivariate Gaussian, and that for the error variance an inverse- Γ prior. The returned sample from the posterior distribution can be analyzed with functions provided in the coda “Convergence Diagnosis and Output Analysis and Diagnostics for MCMC” package [12].

This model has the well-known form: $y_i = (X_i)^T \beta + \varepsilon_i$, with i.i.d. Gaussian errors: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The priors are *semi-conjugate* and *a priori* independent:

$$\beta \sim \text{MVN}(b_0, B_0^{-1}) \quad (47)$$

$$1/\sigma^2 = \tau \sim \Gamma(c_0/2, d_0/2) \quad (48)$$

The “semi-” in “semi-conjugate” means that we assume *a priori* (without evidence) that the distribution of the β vector is independent of the distribution of the $1/\sigma^2 = \tau$. In general this is a reasonable assumption: the amount of noise in the regression does not influence the values of the linear model coefficients. The hyperparameter b_0 is a vector of the means of the prior distributions of the β . Their variance-covariance matrix is B_0 ; this implies that the coefficients can be correlated.

Because the two distributions are independent, the Gibbs sampler can be used, alternating between sampling the two distributions.

The MCMC approaches the *posterior* distribution of the regression parameters of the linear model, and the posterior standard error. Each step updates the posterior probability according to the likelihood of the observed data, using Bayes’ rule.

The conditional posterior probability for the β vector is multivariate-normal (MVN):

$$p(\beta \mid \sigma^2, \mathbf{y}, \mathbf{X}) \sim (\text{MVN}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}', \sigma^2(\mathbf{X}'\mathbf{X})^{-1})) \quad (49)$$

which is the OLS formulation. Note how the variance-covariance matrix of the regression parameters depends on the residual variance of the regression.

However, we also need the marginal posterior distribution for each of the β_m , which requires integrating out the variance:

$$p(\beta_m \mid \mathbf{y}, \mathbf{X}) = \int_0^{+\omega} p(\beta_m \mid \sigma^2, \mathbf{y}, \mathbf{X}) d\sigma^2 \quad (50)$$

Similarly, for the marginal posterior distribution of the regression variance σ^2 , we need to integrate out the regression coefficients.

To compute these integrals we use the ergodic Markov chain that converges to a stable distribution, as computed by the Gibbs sampler.

Another way to write the conditional distributions, considering the evidence Y are:

$$P(\beta | Y, \sigma^2) = \mathcal{N} \left((X^T X)^{-1} X^T Y, \sigma^2 (X^T X)^{-1} \right) \quad (51)$$

$$P(\sigma^2 | Y, \beta) = \mathcal{IG} \left(a + n/2, b + \frac{1}{2} (Y - X\beta)^T (Y - X\beta) \right) \quad (52)$$

Notice how the distribution of each parameter depends on the other, as well as on the data.

To illustrate Bayesian linear regression we use the `meuse` Meuse River heavy metals dataset from the `sp` package. The regression is the $\log_{10}$ Zn concentration modelled as a linear function of position, distance from the river, and elevation. Here is the conventional linear model, with frequentist uncertainty estimates:

```
require(sp)
data(meuse)
m <- lm(log10(zinc) ~ x + y + dist.m + elev, data=meuse)
summary(m)

##
## Call:
## lm(formula = log10(zinc) ~ x + y + dist.m + elev, data = meuse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4107 -0.1193 -0.0007  0.1049  0.4912
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.654e+00  4.962e+00  -1.139   0.2563
## x           -6.143e-05  4.589e-05  -1.339   0.1827
## y             6.157e-05  3.244e-05   1.898   0.0596 .
## dist.m       -6.804e-04  8.525e-05  -7.982 3.47e-13 ***
## elev        -1.168e-01  1.655e-02  -7.060 5.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1802 on 150 degrees of freedom
## Multiple R-squared:  0.6783, Adjusted R-squared:  0.6698
## F-statistic: 79.08 on 4 and 150 DF, p-value: < 2.2e-16
```

In the Bayesian analysis, we also need to specify prior values of the hyperparameters.

b0 : a vector of the mean prior values of β ;

B0 : a matrix of the prior precisions of each β ; this can be a single scalar (same precision for each), a diagonal matrix (independent but different precisions per-predictor), or a full matrix (precisions of different predictors are correlated).

- c_0 : $c_0/2$ is the *shape* parameter of the inverse- Γ prior for σ^2 ; the amount of information represents c_0 pseudo-observations;
- d_0 : $d_0/2$ is the *scale* parameter of the inverse- Γ prior for σ^2 ; it represents the sum of squared errors of the c_0 pseudo-observations;

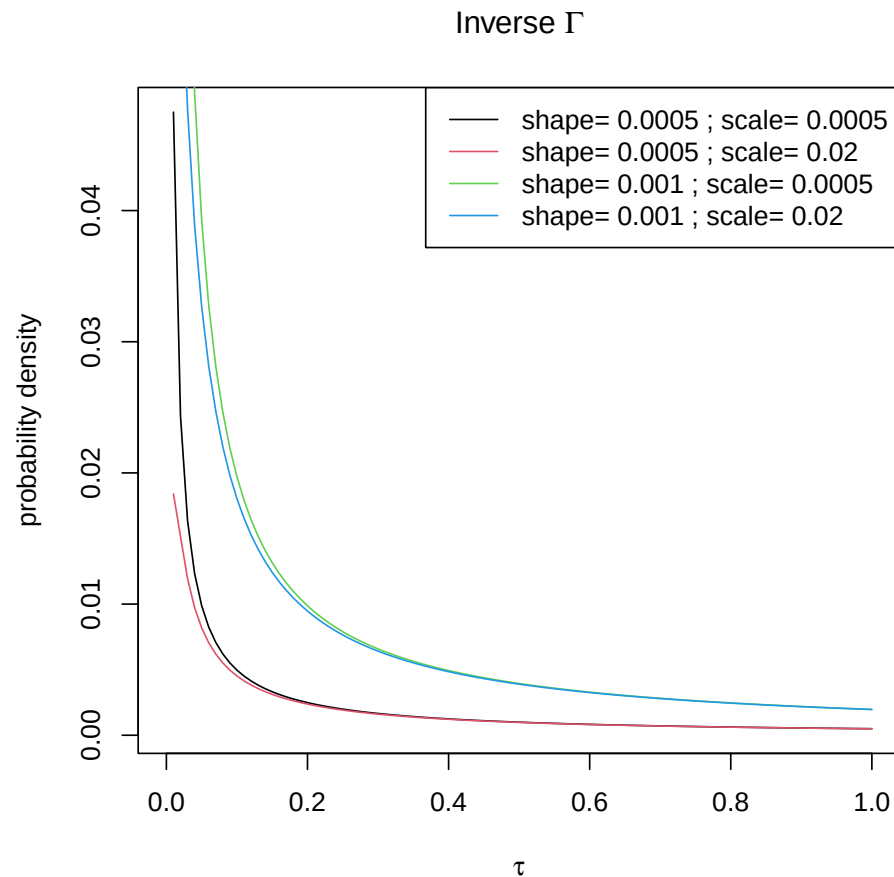
The inverse- Γ distribution is:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \quad (53)$$

Here the shape α is the parameter $c_0/2$, and the scale β is $d_0/2$.

The *non-informative* defaults are b_0 , $B_0 = 0$, c_0 , $d_0 = 0.001$. That is, all regression coefficients are 0 (no effect on the predictand) with zero precision, i.e., infinite variance. The parameters of the Γ distribution for the precision are very small values, to concentrate the prior at very low precision, i.e., high residual variance. The following Figure show this distribution of the prior probability density for τ ; for comparison the same distribution with higher precision (larger scale parameter) and more information (larger shape parameter) are also shown.

```
curve(dinvgamma(x, shape=0.0005, scale=0.0005), ylab="probability density",
      xlab=expression(tau))
curve(dinvgamma(x, shape=0.0005, scale=0.01), add=T, col=2)
curve(dinvgamma(x, shape=0.002, scale=0.0005), add=T, col=3)
curve(dinvgamma(x, shape=0.002, scale=0.01), add=T, col=4)
legend("topright", legend=paste("shape=", c("0.0005",
      "0.0005", "0.001", "0.001"), ";",
      "scale=", c("0.0005",
      "0.02", "0.0005", "0.02")),
      lty=1, col=1:4)
title(expression(paste("Inverse ", Gamma)))
```



We now run the same regression in a Bayesian context, using the Gibbs sampler for MCMC, and the default non-informative priors. There are two control arguments:

burnin : the number of *burn-in* iterations, i.e., before statistics are collected for the posterior distribution; default 1000;

mcmc : The number of MCMC iterations after burn-in; default 10000.

```
require(MCMCpack)
lm.posterior <- MCMCregress(formula(m),
                             data=meuse)
summary(lm.posterior)

##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##              Mean      SD Naive SE Time-series SE
## (Intercept) -5.680e+00 5.014e+00 5.014e-02    4.904e-02
## x           -6.125e-05 4.656e-05 4.656e-07    4.837e-07
## y            6.156e-05 3.287e-05 3.287e-07    3.287e-07
## dist.m       -6.799e-04 8.567e-05 8.567e-07    8.567e-07
## elev        -1.171e-01 1.663e-02 1.663e-04    1.663e-04
```

```
## sigma2      3.295e-02 3.884e-03 3.884e-05      3.974e-05
##
## 2. Quantiles for each variable:
##
##           2.5%      25%      50%      75%
## (Intercept) -1.555e+01 -9.035e+00 -5.681e+00 -2.328e+00
## x           -1.542e-04 -9.213e-05 -6.129e-05 -3.039e-05
## y           -2.515e-06 3.950e-05 6.126e-05 8.345e-05
## dist.m      -8.455e-04 -7.388e-04 -6.795e-04 -6.218e-04
## elev        -1.498e-01 -1.283e-01 -1.170e-01 -1.057e-01
## sigma2      2.611e-02 3.019e-02 3.268e-02 3.528e-02
##           97.5%
## (Intercept) 4.101e+00
## x           2.996e-05
## y           1.287e-04
## dist.m      -5.131e-04
## elev        -8.451e-02
## sigma2      4.132e-02
```

Comparing the mean values of the Bayesian estimates of the parameters to the OLS estimates:

```
round(coefficients(m)[2:5],6)

##           x           y      dist.m      elev
## -0.000061  0.000062 -0.000680 -0.116840

round(summary(lm.posterior)$statistics[2:5,"Mean"],6)

##           x           y      dist.m      elev
## -0.000061  0.000062 -0.000680 -0.117053
```

Comparing the variances of the random error:

```
round(summary(m)$sigma^2,4)

## [1] 0.0325

round(summary(lm.posterior)$statistics["sigma2","Mean"],4)

## [1] 0.0329
```

We see that the mean estimates are similar but not identical.

Recall that the prior for the precision of the residuals was $1/0.001 = 1000$; here the posterior distribution of the precision is centred on $1/0.0329 = 30$, i.e., much more precise.

Comparing the confidence limits from OLS with the credible intervals for Bayesian regression:

```
# lower limit of the 95% credible interval
round(summary(lm.posterior)$quantiles[2:5,"2.5%"],6)

##           x           y      dist.m      elev
## -0.000154 -0.000003 -0.000845 -0.149779

# half-width of the 95% confidence interval
ci.halfwidth <- summary(m)$coefficients[,"Std. Error"]*1.96
# lower limit of the 95% confidence interval
round(summary(m)$coefficients[, "Estimate"] - ci.halfwidth,6)

## (Intercept)           x           y      dist.m      elev
## -15.380394    -0.000151    -0.000002    -0.000848    -0.149279

# upper limit of the 95% credible interval
round(summary(lm.posterior)$quantiles[2:5,"97.5%"],6)
```

```
##           x           y      dist.m      elev
## 0.000030 0.000129 -0.000513 -0.084513

# upper limit of the 95% confidence interval
round(summary(m)$coefficients[, "Estimate"] + ci.halfwidth, 6)

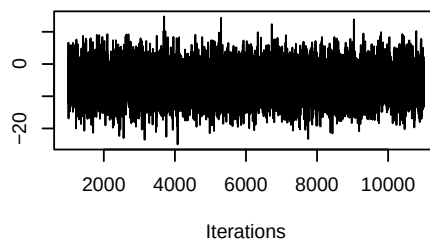
## (Intercept)           x           y      dist.m      elev
## 4.071938      0.000029  0.000125  -0.000513  -0.084401
```

These also do not differ much; we will see a larger effect with the informative prior.

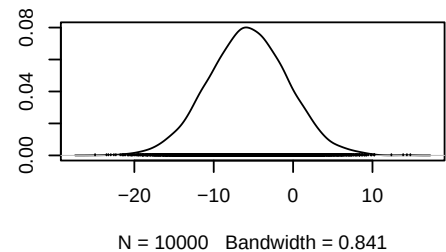
There are several diagnostic plots; these show the trace of the Markov chain, and the probability density, for each model parameter.

```
plot(lm.posterior, ask=FALSE)
```

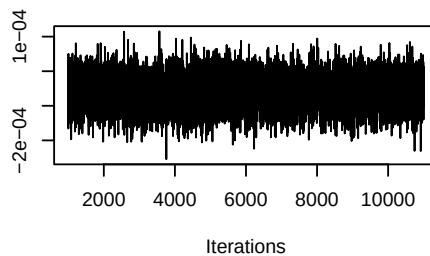
Trace of (Intercept)



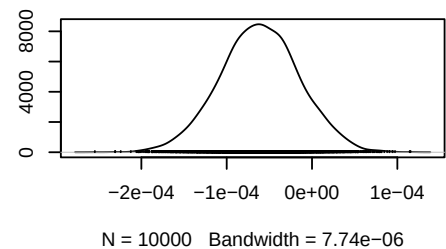
Density of (Intercept)



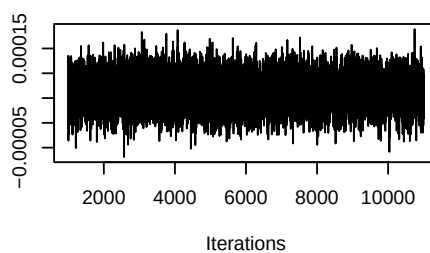
Trace of x



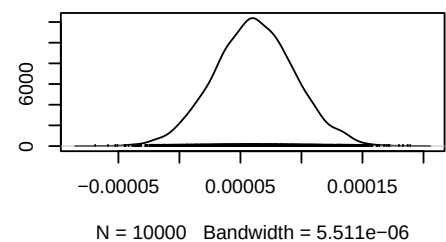
Density of x

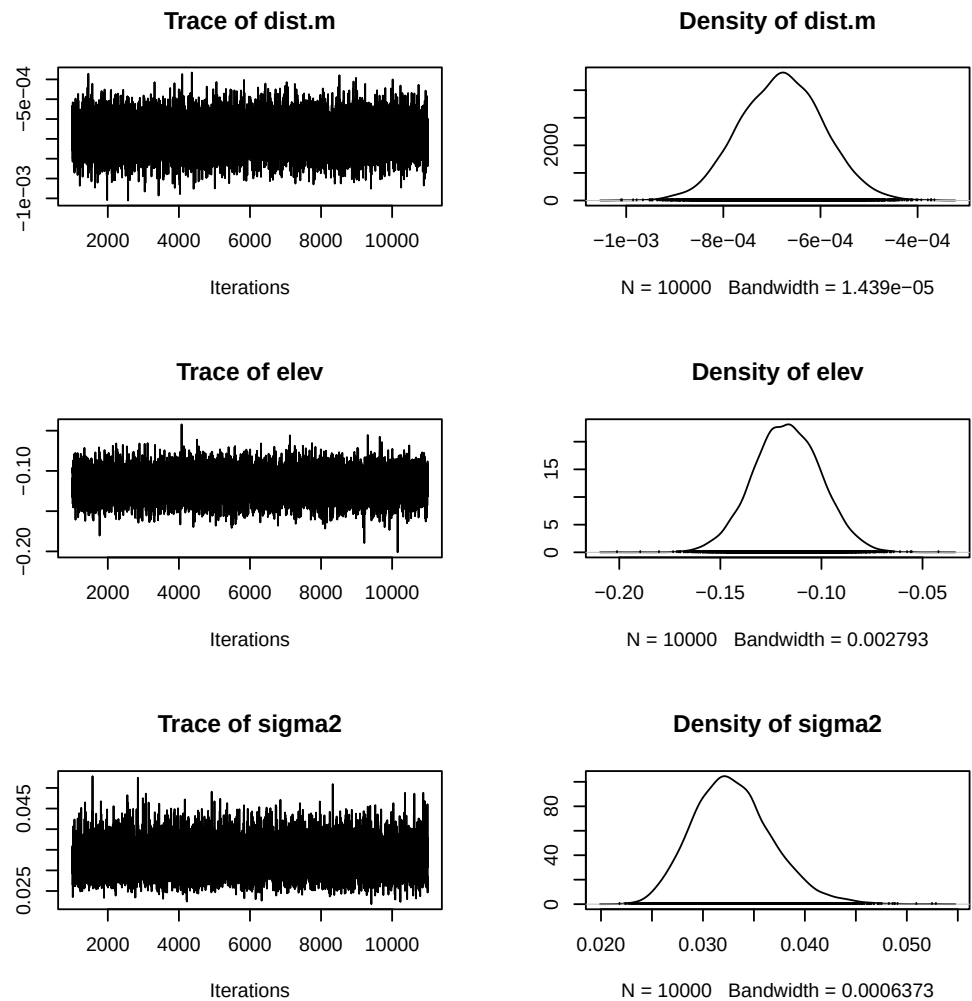


Trace of y



Density of y





The left-hand plots (traces) show that, after the burn-in period, the Markov chain is stationary, so a long chain gives a good estimate of the distribution of the parameters. The right-hand plots (densities) shows the posterior distribution of each parameter.

Several diagnostics are available. The Raftery and Lewis diagnostic shows the number of iterations to estimate a certain quantile of the posterior distribution of a parameter to a specified accuracy, by default 0.005. Suppose we want to estimate the 0.975 quantile, with 95% probability:

```
raftery.diag(lm.posterior, q=0.975)
```

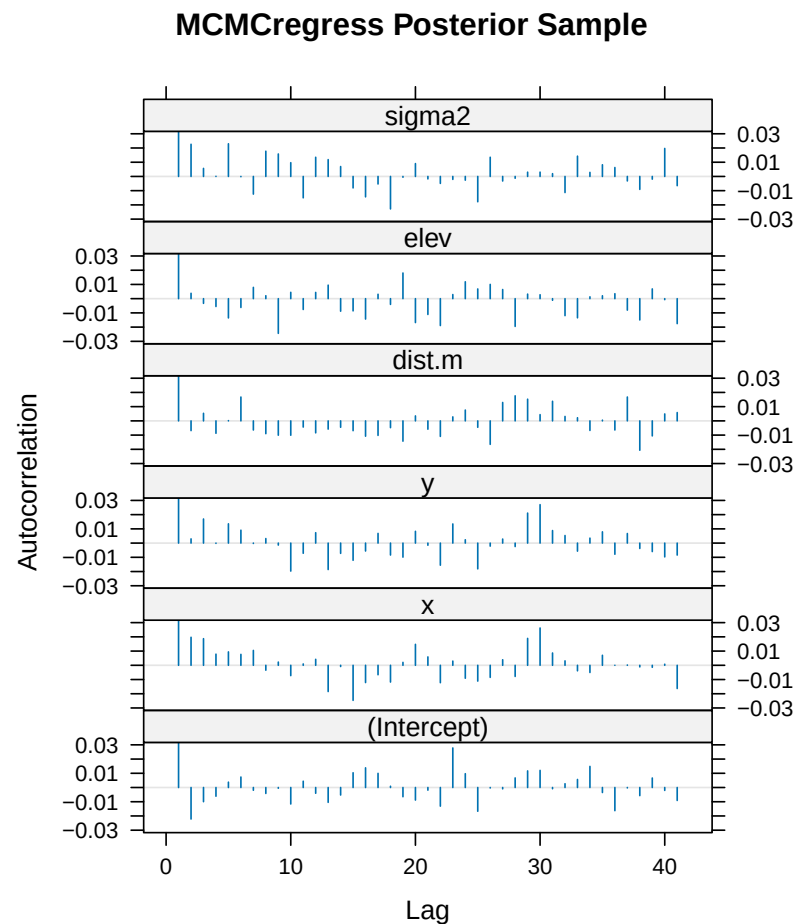
```
##
## Quantile (q) = 0.975
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##          Burn-in Total Lower bound Dependence
##          (M)      (N)  (Nmin)      factor (I)
## (Intercept) 2      3741  3746      0.999
## x            2      3865  3746      1.030
## y            2      3710  3746      0.990
## dist.m       2      3771  3746      1.010
## elev         2      3741  3746      0.999
```

```
## sigma2      2      3834 3746      1.020
```

The required sample size is somewhat less than 4000 with only 2 “burn-in” steps; the default length of the chain in `MCMCregress` is 1000 burn-in and 1000 iterations after that; so in this case we can be confident in the estimates.

The `coda` package provides more diagnostics. The `acfplot` plot shows the autocorrelation at various lags between estimates of the parameters during the Markov chain:

```
require(coda)
acfplot(lm.posterior)
```



We see no pattern in the time-evolution of the sample, except for positive auto-correlation at the first lag, i.e., the next estimate.

6.2 Bayesian linear regression model with informative priors

The real strength of the Bayesian approach comes if we have prior information about the coefficients, their standard error, and the precision of the residuals. For example, the precision could be based on repeated measurements of the same sample, as in laboratory quality control. The

prior coefficients could be based on previous studies or theoretical values. For example, two laboratory methods for measuring the concentration of the same thing (e.g., pH, electrical conductivity) could be expected to give identical results, or could have a theoretical basis (e.g., pH in various dilution of soil:water). It is inefficient to discard this information.

We continue with the Meuse example, with some assumed priors:

b0 : The prior mean vector of β , from previous knowledge.

B0 : The prior precision of β ; we keep the default of 0, an improper uniform prior.

c0 : $c_0/2$ (*shape* parameter of the inverse- Γ prior for σ^2)

d0 : $d_0/2$ (*scale* parameter of the inverse- Γ prior for σ^2); recall this represents the sum of squared errors of the c_0 pseudo-observations;

Suppose we have no evidence for the effect of the coördinates, but a prior estimate of the decrease in Zn for each 100m from river is 200 mg kg⁻¹, i.e., the coefficient for log10(Zn) per m is log 102 = 0.3, and the estimate for the decrease per 1 m elevation is also 200 mg kg⁻¹. Supposing we have good evidence for these two predictors, we also increase the precision of the prior coefficient estimates. We also specify less precision than the default, with substantial evidence of this.

```
lm.i.posterior <- MCMCregress(formula(m), data=meuse,
                             b0=c(0, 0, 0, 0.3, 0.3),
                             B0=c(1e-6, .01, .01, .0001, .0001),
                             c0=10, d0=10)

summary(lm.i.posterior)

##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##              Mean          SD Naive SE Time-series SE
## (Intercept) -5.695e+00 8.477e+00 8.477e-02 8.294e-02
## x           -6.107e-05 7.864e-05 7.864e-07 8.189e-07
## y            6.151e-05 5.554e-05 5.554e-07 5.554e-07
## dist.m      -6.793e-04 1.448e-04 1.448e-06 1.448e-06
## elev        -1.172e-01 2.815e-02 2.815e-04 2.815e-04
## sigma2       9.424e-02 1.076e-02 1.076e-04 1.098e-04
##
## 2. Quantiles for each variable:
##
##              2.5%          25%          50%          75%
## (Intercept) -2.244e+01 -1.138e+01 -5.691e+00 -2.074e-02
## x           -2.160e-04 -1.135e-04 -6.123e-05 -8.766e-06
## y           -4.744e-05 2.435e-05 6.086e-05 9.848e-05
## dist.m      -9.578e-04 -7.783e-04 -6.787e-04 -5.819e-04
## elev        -1.725e-01 -1.362e-01 -1.172e-01 -9.802e-02
## sigma2       7.519e-02 8.662e-02 9.347e-02 1.007e-01
##              97.5%
## (Intercept) 1.080e+01
## x           9.375e-05
## y           1.750e-04
## dist.m      -3.975e-04
```

```
## elev      -6.217e-02
## sigma2    1.178e-01
```

There is not much difference in the coefficients, but the estimate of the residual error is much larger, due to the prior.

```
round(summary(lm.posterior)$statistics[2:5, "Mean"], 6)

##          x          y      dist.m      elev
## -0.000061  0.000062 -0.000680 -0.117053

round(summary(lm.i.posterior)$statistics[2:5, "Mean"], 6)

##          x          y      dist.m      elev
## -0.000061  0.000062 -0.000679 -0.117231

round(summary(lm.posterior)$statistics["sigma2", "Mean"], 4)

## [1] 0.0329

round(summary(lm.i.posterior)$statistics["sigma2", "Mean"], 4)

## [1] 0.0942
```

6.3 Comparing models with the Bayes factor

We see from the wide credible intervals for the coördinates x , y that they may not be important predictors. So, we refit the model with just distance to river and elevation, and then compare them with the *Bayes factor*. This is the ratio of posterior likelihoods of the data, given the fitted models:

$$BF = \frac{p(y | X, m_a)}{p(y | X, m_b)} \quad (54)$$

where m_a, m_b are two models to compare, X is the design matrix, and y the observed data. The Bayes factor quantifies the support from the data for one model compared to another.

Jeffreys [7] suggested a subjective scale for how strongly the evidence favours m_a :

factor	ln(factor)	strength of evidence for m_a
$< 10^0$	< 0	negative, supports m_b
$10^0 \dots 10^{0.5}$	$0 \dots \approx 1.5$	barely worth mentioning
$10^{0.5} \dots 10^1$	$\approx 1.5 \dots \approx 2.3$	substantial
$10^1 \dots 10^{3/2}$	$\approx 2.3 \dots \approx 3.5$	strong
$10^{3/2} \dots 10^2$	$\approx 3.5 \dots \approx 4.6$	very strong
$> 10^2$	$> \approx 4.6$	decisive

In order to compute these, we need to compute the likelihoods $p(y | X, m)$ for each model to be compared, and then compute their ratio.

Here we use the default priors, except that we need to specify a proper prior (one that intergrates to 1 over its range) for the regression coefficient standard errors.

```

lm.1.posterior <- MCMCregress(formula(m), data=meuse,
                             B0=c(1e-6, .01, .01, .01, .01), marginal.likelihood="Chib95")
lm.2.posterior <- MCMCregress(update(formula(m), . ~ . -x -y),
                             data=meuse,
                             B0=c(1e-6, .01, .01), marginal.likelihood="Chib95")
round(summary(lm.1.posterior)$statistics[2:5,"Mean"],6)

##           x           y      dist.m      elev
## -0.000061  0.000062 -0.000680 -0.117053

round(summary(lm.2.posterior)$statistics[2:3,"Mean"],6)

##      dist.m      elev
## -0.000762 -0.114598

# summary(lm.1.posterior)
# summary(lm.2.posterior)
(bf.1.2 <- BayesFactor(lm.1.posterior, lm.2.posterior))

## The matrix of Bayes Factors is:
##           lm.1.posterior lm.2.posterior
## lm.1.posterior      1.00e+00      5.58e-11
## lm.2.posterior      1.79e+10      1.00e+00
##
## The matrix of the natural log Bayes Factors is:
##           lm.1.posterior lm.2.posterior
## lm.1.posterior          0.0          -23.6
## lm.2.posterior         23.6           0.0
##
## lm.1.posterior :
##   call =
## MCMCregress(formula = formula(m), data = meuse, B0 = c(1e-06,
##   0.01, 0.01, 0.01, 0.01), marginal.likelihood = "Chib95")
##
##   log marginal likelihood = -15.92415
##
##
## lm.2.posterior :
##   call =
## MCMCregress(formula = update(formula(m), . ~ . - x - y), data = meuse,
##   B0 = c(1e-06, 0.01, 0.01), marginal.likelihood = "Chib95")
##
##   log marginal likelihood = 7.685869

```

The coefficients for the two retained variables change somewhat, as expected. The log-Bayes factor for the simpler model, compared to the more complex model, is 23.6, which is decisive evidence in its favour.

6.4 The Markov chain

To see how the Markov chain converges, it's instructive to see the beginning of the chain with no burn-in and only a short chain. We can also compare the posterior density distributions.

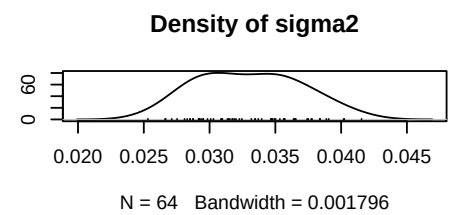
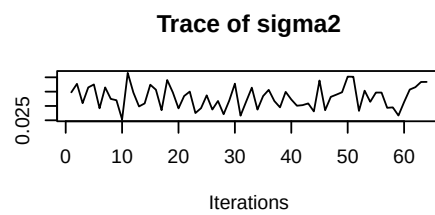
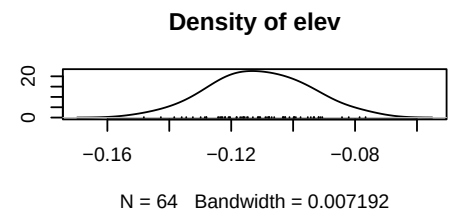
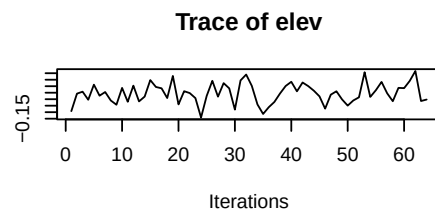
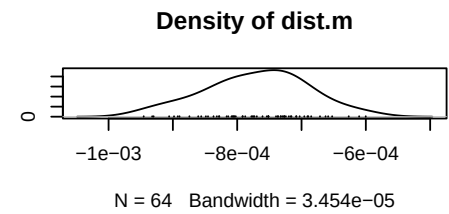
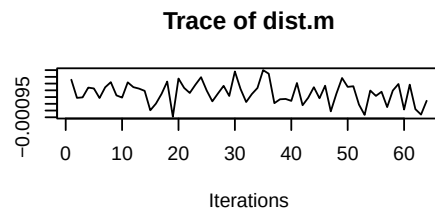
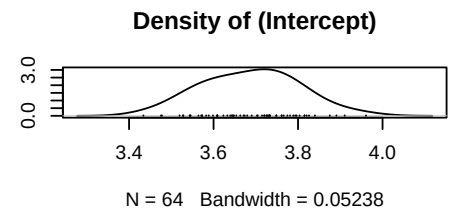
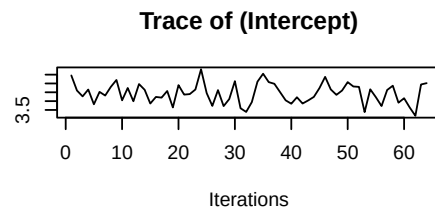
```

lm.2.short.posterior <- MCMCregress(update(formula(m), . ~ . -x -y),
                                    data=meuse,
                                    B0=c(1e-6, .01, .01), burnin=0, mcmc=64, verbose=16)

##
##
## MCMCregress iteration 1 of 64
## beta =
##   3.89081
##  -0.00067
##  -0.13840

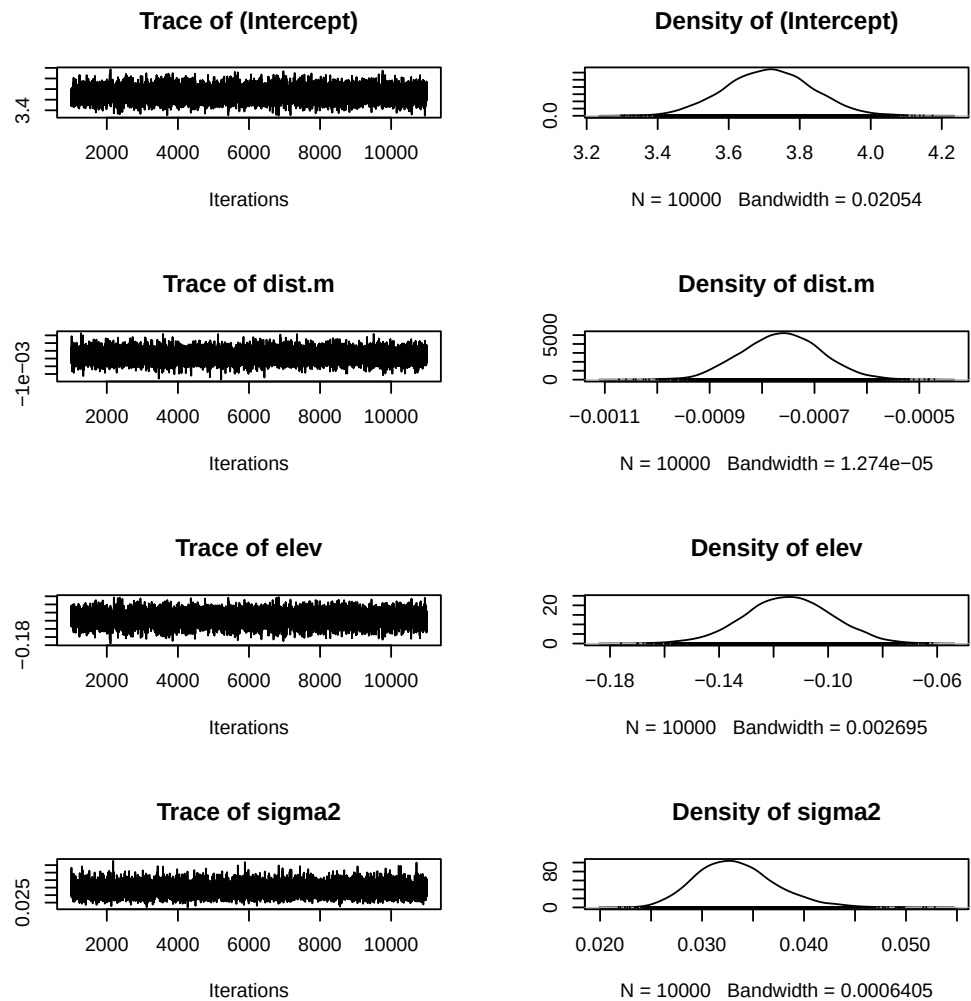
```

```
## sigma2 =    0.03476
##
##
## MCMCregress iteration 17 of 64
## beta =
##    3.63949
##   -0.00078
##   -0.10323
## sigma2 =    0.02852
##
##
## MCMCregress iteration 33 of 64
## beta =
##    3.58856
##   -0.00078
##   -0.09953
## sigma2 =    0.03639
##
##
## MCMCregress iteration 49 of 64
## beta =
##    3.72005
##   -0.00066
##   -0.12036
## sigma2 =    0.03479
plot(lm.2.short.posterior, ask=FALSE)
```



Compare this to the distribution and trace for the longer chain:

```
plot(lm.2.posterior, ask=FALSE)
```



7 Spatial Bayesian analysis

The same kind of reasoning for non-spatial models applies to spatial models: we have a model form, which in the spatial case includes a model of spatial dependence, and we consider now the parameters of the model to be random variables. There are several R packages that take a Bayesian approach to spatial modelling, including `spBayes` [4], `geoR` (for geostatistics) [13], `geoGLM` (for generalised Linear Spatial Models). Here we use `geoR`, which includes Bayesian methods for point geostatistics, analogous to the `gstat`, `spatial` and `fields` packages, among others, that take a frequentist approach to geostatistical inference, i.e., model estimation and prediction by kriging.

```
require(geoR)
```

This package provides a `krige.bayes` function to perform kriging, allowing specifications of different levels of uncertainty in the model parameters, i.e., they can have prior distributions. It then computes the posterior distributions for the model parameters and on the predictive

distributions for prediction locations.

The authors justify this approach as follows:

“[E]stimation and prediction are usually operationally separate in geostatistical practice, meaning that we first formulate our model [of spatial dependence] and estimate its parameters, then plug the estimated parameter values into theoretical prediction equations [i.e., kriging equations] as if they were the true values. An obvious concern with this two-phase approach is that ignoring uncertainty in the parameter estimates may lead to optimistic assessments of predictive accuracy.”

In their opinion, the most consistent way to account for parameter uncertainty is to describe the parameters as probability distributions, using the Bayesian approach to update prior estimates with the data. In the geostatistical setting, the parameters are the variogram or spatial correlation function parameters.

The unobserved (unknowable) spatial process that gives rise to the data is symbolized as $\{S(x) : x \in \mathbb{R}^2\}$; Ribeiro and Diggle [13] call this the “signal”, i.e., reality. But this is only observed through data collected at n locations, symbolized as Y . The data we see is obviously conditioned on the signal – a different signal would result in different observations. If we symbolize the model parameters as a vector θ this relation is:

$$[Y, S | \theta] = [S | \theta] [Y | S, \theta] \quad (55)$$

That is, the distribution of the observed data Y produced by signal S with parameters θ is the product of the distribution of the signal S given known values of the model parameters θ , and the distribution of the observed data conditional on the signal and model parameters.

The **predictive** distribution, i.e., what we would like to map, is then $[S | Y, \theta]$. That is, given the data Y and estimates of θ , we want to predict the signal, for example by ordinary kriging. But of course the actual signal is not observable, so instead we have to map some **target** of S , symbolized as T . This introduces an additional level in the hierarchy. In the special case of the linear Gaussian model, $T = S(x)$, i.e., we want the best linear unbiased prediction of the value of the signal at each prediction location.

In practice we only see Y , and both S and θ are unobservable random variables, thus they have probability distributions. This results in a three-level hierarchy:

$$[Y, S, \theta] = [\theta] [S | \theta] [Y | S, \theta] \quad (56)$$

where $[\theta]$ is the prior distribution for θ (the parameters of the spatial dependence model), that will be updated during the Bayesian analysis,

$[S | \theta]$ is the distribution of the spatial field given values for its parameters, and $[Y | S, \theta]$ is the distribution of the observed values, given the parameterized spatial field. So starting from the prior $[\theta]$, this updates the model of the spatial field, which is then updated by the likelihood of the observed Y given that parameterized field.

Then the predictive distribution for the spatial field is:

$$[S | Y] = \int [S | Y, \theta] [\theta | Y] d\theta \quad (57)$$

That is, the predictive distribution is a weighted average (i.e., the integral) of plug-in predictive distributions, where the weights are based on the posterior uncertainty about the values of the model parameters θ , which were estimated in a previous step in the hierarchy; this is $[\theta | Y]$.

7.1 The Gaussian linear model

As with any Bayesian analysis, we need to specify a model form to be parameterized. The most common form for geostatistics is the Gaussian linear model, based on the assumption of a Gaussian spatial process $\{S(x) : x \in \mathbb{R}^2\}$ such that for any set of locations in $\mathbb{R}^2$ their joint distribution is multivariate Gaussian, so that the process can be specified by just two functions:

mean : $\mu(x) = E[S(x)]$

covariance : $\gamma(x, x') = \text{Cov}\{S(x), S(x')\}$

The key point is that S follows a multivariate Gaussian distribution with a mean vector μ_S and a positive-definite covariance matrix G between the points in the field.

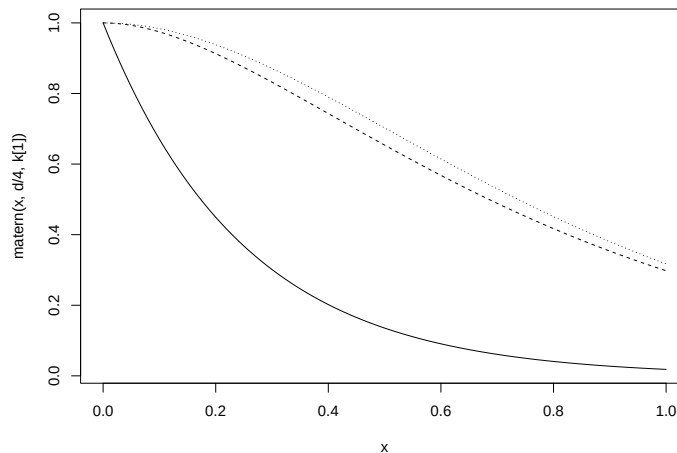
The covariance between points depends only on their spatial separation and some function of that – this is what needs to be determined in order to predict using geostatistics. Many functions can be used, but they should be such that correlation decreases with separation. Further, not only the range till essentially no correlation must be specified (this is a parameter to be estimated), but also the *smoothness* of the decay with distance. Common models such as exponential and spherical have a defined smoothness; more general is to use the Matérn correlation function, which includes smoothness as an additional parameter:

$$\rho(u) = \frac{2^{\kappa-1} \Gamma(\kappa)}{(u/\phi)^\kappa K_\kappa(u/\phi)} \quad (58)$$

where K_κ is a modified Bessel function of order κ , $\phi > 0$ is the scale parameter (“range”), and the order κ determines the smoothness. For $\kappa = 0.5$ the Matérn function reduces to the familiar exponential $\rho(u) = \exp(-u/\phi)$ (least smooth); as $\kappa \rightarrow \omega$, $\rho(u) \rightarrow \exp\{-(u/\phi)^2\}$, sometimes called the Gaussian correlation function (smoothest). This function is popular in Bayesian geostatistics, because it avoids the problem of specifying the smoothness of the spatial correlation (except as a prior).

Here is a graph of the Matérn function with various degrees of smoothness. The range parameter is adjusted to give the same effective range for all smoothnesses.

```
d <- 1 # xlim
x <- seq(0,d, by=0.01) # plot at these separations
k <- c(0.5, 1.5, 2.5) # smoothness
# practical range, as the distance
# at which the correlation is 0.05.
a <- 2*sqrt(k) # adjust phi to effective range
plot(matern(x,d/4,k[1]) ~ x, type="l")
lines(matern(x,d/a[2],k[2]) ~ x, type="l", lty=2)
lines(matern(x,d/a[3],k[3]) ~ x, type="l", lty=3)
```



7.2 Specifying priors

A major difference between frequentist and Bayesian approaches is that the Bayesian must specify a prior distribution for all parameters to be estimated. These are updated based on the likelihood of parameter values, given the model and data. Ribeiro and Diggle [13] explain the problem in the geostatistical context:

“In the absence of clearly articulated **prior knowledge priors should be diffuse**, in which case their effects are swamped by the effect of the likelihood provided that the sample size is sufficiently large. However, in the geostatistical setting where data are often highly correlated, intuitive ideas of what constitutes a ‘large’ sample may be misleading. Our experience has been that with data-sets of size several hundred, **apparently diffuse priors can still have a noticeable influence on the inferences**. It seems to be a general feature of geostatistical problems that the **models are poorly identified, in the sense that widely different combinations of parameter values lead to very similar fits**. This may not matter if parameter estimation is not of direct interest but is only a means towards the goal of prediction. Even so, it remains a lingering concern because the prior does potentially influence the

predictive distribution which we report for any target.”

7.3 Bayesian parameter estimation

Recall the general Bayesian formulation for updating a prior distribution $\pi(\theta)$ with a likelihood function $\ell(\theta; \mathcal{Y})$ based on the data:

$$p(\theta \mid \mathcal{Y}) = \frac{\ell(\theta; \mathcal{Y})\pi(\theta)}{\int \ell(\theta; \mathcal{Y})\pi(\theta)d\theta} \quad (59)$$

7.4 Bayesian prediction

This is the main focus of geostatistical Bayesian analysis.

References

- [1] Jim Albert. *Bayesian Computation with R*. Springer-Verlag New York, 2nd edition, 2009. ISBN 978-0-387-92298-0. URL <http://site.ebrary.com/id/10294526>.
- [2] G. A. Barnard and Thomas Bayes. Studies in the history of probability and statistics: IX. Thomas Bayes's 'Essay Towards Solving a Problem in the Doctrine of Chances'. *Biometrika*, 45(3/4):293315, 1958. doi: 10.2307/2333180.
- [3] George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- [4] Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4), 2007. doi: 10.18637/jss.v019.i04. URL <http://www.jstatsoft.org/v19/i04/>.
- [5] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 3 edition edition, Nov 2013. ISBN 978-1-4398-4095-5.
- [6] Peter D Hoff. *A first course in Bayesian statistical methods*. Springer Verlag, 2009. ISBN 978-0-387-92407-6. URL <https://link.springer.com/openurl?genre=book&isbn=978-0-387-92299-7>.
- [7] Harold Jeffreys. *Theory of probability*. Clarendon Press, 3d ed. edition, 1961.
- [8] Peter M Lee. *Bayesian statistics: an introduction*. Arnold, 2004. ISBN 0 340 81405 5. URL <http://www-users.york.ac.uk/~pm11/bayes/book.htm>.
- [9] Jean-Michel Marin and Christian P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer New York, 2007. ISBN 978-0-387-38979-0. doi: 10.1007/978-0-387-38983-7.
- [10] Andrew D. Martin and Kevin M. Quinn. Applied Bayesian inference in R using MCMCpack. *R News*, 6(1):27, 2006.
- [11] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Dec 2015. ISBN 978-1-4822-5344-3.
- [12] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):711, 2006.
- [13] Jr. Ribeiro and Peter J. Diggle. geoR: A package for geostatistical analysis. *R News*, 1(2):1418, 2001.

- [14] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3–23, Jan 1993. doi: 10.2307/2346063.