Bayesian Statistics, Chapter 11

Elizabeth Purdom

This document has last been compiled on Oct 14, 2024.

Contents

1 Bayesian Method **3**

1.1 Bayes Rule and Bayesian Statistics 3

1.2 Inference 6

1.3 Comparison of Frequentist and Bayesian 7

2 The Prior **8**

2.1 Bayesian Philosophy of Probability 10

2.2 Picking a Prior 11

2.3 Philosophies on Choosing the Prior 12

2.4 Objective or Non-informative priors 13

2.4.1 Equal Probability – Uniform prior 14

2.5 Jeffreys' Priors (Invariant priors) 16

2.6 Other Notions of Noninformative 19

3 Calculation of Posterior distribution **20**

3.1 Rejection Sampling 22

3.1.1 Apply to posterior distribution 26

3.1.2 Why this works 27

3.2 Importance Sampling 29

1 Bayesian Method

At a practical level, Bayesian methods consider parameters of interest θ to be random variables. So we have our data $X = (X_1, \ldots, X_n)$, as well as the parameter θ that are random, so that X, θ have a joint distribution. And since the whole point is that knowing the data tells us something about θ, and vice versa (knowing θ tells us what kind of data we will see) we certainly do not want to consider them independent.

We will observe X, and want to be able to say something about θ. For a Bayesian, this can be stated in probabilistic terms as determining

$$P(\theta|X),$$

the conditional distribution of θ given our observed data.

Definition 1.1 (Posterior Distribution)**.** The conditional distribution $P(\theta|X)$ is called the **Posterior Distribution** of θ

Once you know this conditional distribution, Bayesians answer questions like

- What is an estimate for θ? This is often interpretated as a likely value for θ, such as the average,
$$E(\theta|X)$$
- Is $\theta > \theta_0$? We could answer this question as what is the probability θ is greater than θ_0
$$P(\theta > \theta_0|X)$$
- Find an interval that θ has a 95% chance of being inside:
$$(L_n, U_n) : P(L_n < \theta < U_n|X) = 0.95$$
These are called **Bayesian Credible Intervals**

All Bayesian inference is solved by determining $P(\theta|X)$

1.1 Bayes Rule and Bayesian Statistics

Determining the Posterior Distribution for Bayesian statistics uses Bayes Theorem

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{f(x)}$$

where f above refers to the density or probability mass function

In terms of inference,

- $f(\theta)$ refers to the marginal distribution of the data, i.e. without any information about the data – the distribution of θ without looking at the data, or before the data is collected. It is thus called the **prior distribution**. Notice that the prior distribution is considered *known*.
- $f(x|\theta)$: this is the likelihood, $\mathcal{L}_n(\theta)$, but has of course a different interpretation for Bayesian analysis – joint density of the data for particular θ.
- $f(x)$: the normalizing constant that doesn't depend on θ– the marginal distribution for the data; can be hard to calculate
- $f(\theta|x)$: the posterior density – reflects knowledge of θ after seeing the data

Note that knowing $f(\theta)$ and $f(x|\theta)$ determines the entire joint distribution, so these two quantities are the decisions to be made to model the problem. And the marginal distribution is simply

$$f(x) = \int f(x|\theta)f(\theta)d\theta$$

Notation Notice that we are using f for all of the densities, even though clearly $f(x|\theta)$ versus $f(\theta|x)$ are entirely different functions. The idea is that you figure out, based on what are the arguments, which you are dealing with. Otherwise, we start having too many letters to keep track of. Similarly, we might write the joint likelihood of $i.i.d$ data as

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

so that f on the left-hand side refers to the joint density over n observations and f on the right-hand side refers to the univariate density for a single x_i.

Typical Bayesian Setup Clearly we need to assume we know some of these quantities. We want to determine $f(\theta|x)$, but what about the others? We will have the following two distributions are known/assumed

1. We will assume we know the distribution or density of X, *if we knew* θ. Only now, since we assume θ is random, we will write this as a conditional distribution:

   $$X \text{ distributed with density } f(x|\theta)$$

2. We also will assume we know the prior distribution of θ

   $$\theta \text{ distributed with density } f(\theta)$$

Note that with these two pieces of information, I can calculate $f(\theta|X)$ because

$$f(x) = \int f(x|\theta)f(\theta)d\theta$$

so I can calculate $f(\theta|X)$

Example: (Wasserman) Suppose

$$X_1, \ldots, X_n|\theta \overset{i.i.d}{\sim} N(\theta, \sigma^2), \theta \sim N(a, b^2)$$

where σ^2, and a, b are known. We need to calculate

$$f(\theta|X_n) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

Let's focus on the numerator and recall we want to focus on this as a function of θ,

$$\begin{aligned}
f(\theta|X_n) \propto f(X|\theta)f(\theta) &\propto \exp\{-\frac{1}{2\sigma^2}\sum_i (X_i-\theta)^2\}\exp\{-\frac{1}{2b^2}(\theta-a)^2\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i X_i^2 - 2n\theta\bar{X}_n + n\theta^2\right) + \frac{1}{2b^2}\left(\theta^2 - 2a\theta + a^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left((n/\sigma^2 + 1/b^2)\theta^2 - 2(\frac{n\bar{X}_n}{\sigma^2} + \frac{a}{b^2})\theta\right)\right\} \\
&= \exp\left\{-\frac{1}{2(\frac{\sigma^2b^2}{nb^2+\sigma^2})}\left(\theta^2 - 2(\frac{nb^2\bar{X}_n + \sigma^2 a}{nb^2+\sigma^2})\theta\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2(\frac{\sigma^2b^2}{nb^2+\sigma^2})}\left(\theta - \frac{nb^2\bar{X}_n + \sigma^2 a}{nb^2+\sigma^2}\right)^2\right\}
\end{aligned}$$

If we let

$$\mu = \frac{nb^2\bar{X}_n + \sigma^2 a}{nb^2+\sigma^2}$$
$$\tau^2 = \frac{\sigma^2b^2}{nb^2+\sigma^2}$$

We can see that the posterior distribution is in the form of a normal distribution,

$$\exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\}$$

Since we know the posterior density must integrate to 1 (over θ, our random variable), we don't need to work out what all the constants that don't involve θ are which we dropped along the way. They have to be equal to

$$\frac{1}{\sqrt{2\pi\tau^2}}.$$

Otherwise the density would not integrate to 1.

Tip If the posterior is something tractable (i.e. a standard distribution), then the key to calculation of the posterior is to not worry about any normalization factor, and only work with the part of the density that involves the parameters θ. If the posterior follows a known density, you can see this by the portion that matches θ; and in that case you *know* what form the normalizing constant has to take (you will see many examples of this).

Even if it *isn't* a standard distribution, the same principle holds true – but you will need to numerically figure out what that constant must be; but you still don't need the parts of the equation you dropped since they won't help you figure out that constant.

Exercise 1. Suppose $X_1, \ldots, X_n|\lambda \overset{iid}{\sim} Poisson(\lambda)$, and the prior is $\lambda \sim Gamma(a, b)$. Find the posterior distribution for λ.

Exercise 2. Suppose $X_1, \ldots, X_n|\sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-\alpha-1}\exp\{-\beta/\sigma^2\}$$

Find the posterior distribution for σ^2.

1.2 Inference

In Bayesian statistics, all inference is based on the posterior distribution. We can use the posterior to calculate quantities similar to those under frequentist statistics (point estimates and intervals), or we can examine the posterior probability of *any* event of interest.

Point Estimator The **posterior mean** is a commonly used point estimator:

$$E[\theta|X_1, \ldots, X_n] = \int \theta f(\theta|X_1, \ldots, X_n)d\theta$$

It can often be written as a weighted average of the prior mean and the MLE. For example, in the second example on the previous page,

$$E[\theta|X_1, \ldots, X_n] = \mu = \frac{nb^2}{nb^2 + \sigma^2}\,\bar{X}_n + \frac{\sigma^2}{nb^2 + \sigma^2}\,a$$

A less common Bayesian point estimator is the **maximum a posteriori probability (MAP)** estimate, which is just the mode of the density. This is not usually the preferred estimator, but are often easier to find numerically (using the same gradient methods we used in the MLE) since the normalizing constant is not needed to find the maximum of $f(\theta|x)$ with respect to θ.

Intervals A $1-\alpha$ **credible interval** for θ (also called a posterior interval) is an interval C_n satisfying

$$P(\theta \in C_n | X_1, \ldots, X_n) = 1 - \alpha$$

Note a couple of differences compared to a confidence interval:

- The probability statement is about θ, not C_n. C_n is a function of $X_1, \ldots, X_n$, which we are conditioning on in the probability statement.
- The intervals constructed this way may or may not have good frequentist coverage rates.

Note that C_n is not uniquely defined . There are several popular methods for finding such intervals.

- A $1-\alpha$ **equal-tail credible interval** is an interval (a, b) such that

$$\int_{-\infty}^{a} f(\theta|x)d\theta = \int_{b}^{\infty} f(\theta|x)d\theta = \alpha/2$$

- A $1-\alpha$ **highest posterior density (HPD)** region R_n is defined such that
  1. $P(\theta \in R_n|x) = 1 - \alpha$
  2. $R_n = \{\theta : f(\theta|x) > k\}$ for some k.

  When $f(\theta|x)$ is unimodal, R_n is an interval.

Often, if θ is real-valued, it is more informative to plot $f(\theta|x)$ than it is to report an interval.

1.3 Comparison of Frequentist and Bayesian

In terms of *estimation*, the MLE frequentist estimators and Bayesian techniques often converge to the same solution, assuming we have the underlying regularity conditions we generally assume for the MLE.

Theorem 1. *Let $\hat{\theta}_n$ be the MLE with $\hat{se} = 1/\sqrt{I_n(\theta)}$. Then the posterior is asymptotically normal with mean $\hat{\theta}$ and standard deviation $\hat{se}$.*

We can see this in our Normal example:

$$E[\theta|X_1,\ldots,X_n] = \mu = \frac{nb^2}{nb^2+\sigma^2}\,\bar{X}_n + \frac{\sigma^2}{nb^2+\sigma^2}\,a$$

Our Bayesian estimator is a trade-off between $\bar{X}_n$ and our prior mean a. As n gets larger, $\bar{X}$ "takes over" and the prior assumptions become less important. However, in finite samples, the differences can be important (and we will see when we do decision theory these differences can improve the finite sample properties of the estimator.)

In fact we can make similar statements about confidence intervals and credible intervals. However we should note that they have very different interpretations. Bayesian credible intervals are a statement about the probability of θ being in a certain range. Frequentist confidence intervals are statements about the probability of the CI *method* to cover the true θ in the long-run – but say nothing about the actual interval (the actual interval either does or doesn't cover the true θ). So Bayesian statements are much closer to a natural interpretation that a lay-person would give to the interval.

Finally, we note that hypothesis testing, which we will get to later, is quite different and often *doesn't* lead to the same answer. Indeed, there are many "paradoxes" in statistics involving the divergence of these two approaches.

2 The Prior

While we initially started the discussion as a joint distribution of θ and our data X, generally Bayesian analysis is thought of as starting with a prior distribution on θ, i.e. the distribution describing likely values of θ without seeing any data. This is provided entirely by the experimenter/analyst.

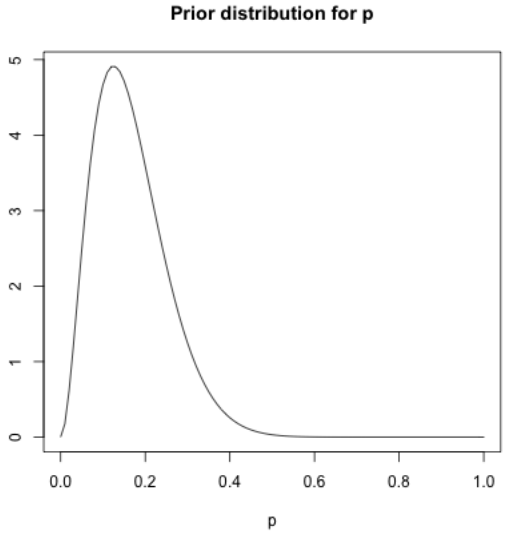
Then once we see the data, we "update" our prior to get a new distribution of likely values for θ now that we have seen actual data – this is our posterior distribution of θ. Again, as we've seen, this update often results in a distribution that can be viewed as an accomodation between the prior beliefs and the data we have seen, with larger amounts of data resulting in an "update" that is determined more and more by our data than our prior.

Role of Prior: Baseball Example Suppose you track a major league baseball player, and you've seen the player bat 30 times. Each time the player is "at bat", they either hit it or they don't (they get multiple attempts, but we only care if they were ultimately successful or not). So we observe data $X_1,\ldots,X_{30}$ Bernoullis with a parameter p giving the parameter of the Bernoulli.

Let's say the player was successful 16/30 times. Our standard frequentist/MLE estimator is $\hat{p}_{MLE} = 16/30 = 0.533$ with a 0.95 CI based on asymptotic normality of (0.35,0.70).

But if you know anything about baseball–even though you don't know anything about this player–you would be unlikely to think that this player's probability of hitting a ball is really 0.533 and even the lower end of the CI of 0.35 is remarkable. Why? Because only an excellent batter will reach more than 0.3 over a season, and over 0.4 basically doesn't happen anymore in major leagues. So in fact 0.533 is not realistic at all, nor is most of the range of the CI.

A Bayesian would instead try to quantify this prior information into a quantifiable prior. This could include past data, but could also be somewhat convenient specification of a vague sense. So I could say I think my prior on p was a $Beta(3, 15)$



For a $Beta(\alpha, \beta)$ distribution for $\alpha > 0, \beta > 0$ we have pdf,

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

and mean $\alpha/(\alpha + \beta)$. With $\alpha = 3$ and $\beta = 15$, we get a prior mean of 0.16 (so I'm maybe putting my prior a bit too pessimistic, since the yearly average in major leagues is closer to 0.24, depending on the year)

Then we can show that our posterior distribution is

$$p|X \sim Beta(\sum_i X_i + \alpha, n - \sum_i X_i + \beta),$$

and

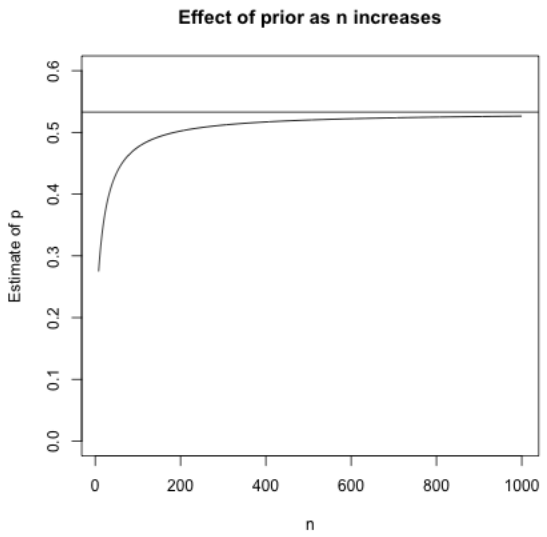
$$\hat{p}_{Bayes} = E(p|X) = \frac{\sum_i X_i + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta}\hat{p}_{MLE} + \frac{\alpha + \beta}{n + \alpha + \beta}p_0$$

where $p_0 = \frac{\alpha}{\alpha+\beta}$, is the expected value under the prior distribution. In our case this gives us

$$\hat{p}_{Bayes} = 0.62\hat{p}_{MLE} + 0.375p_0 = 0.39$$

So we have an explicit tradeoff between the data we have so far, and our prior information. This would still suggest a phenominal player, but a much more reasonable estimate of their probability to hit any particular future ball.

We can consider how the data influences our estimate for different sizes of data observed. If we assume that the proportion hit stays the same, but is based on different size n, we see that eventually the data will "win out".



Exercise 3. Do the above calculations for yourself to validate the answers. What would be a credible interval for p?

Exercise 4. Of course, I could have picked a different prior, for example a prior that didn't allow any value of p greater than 0.4 (e.g. $U(0, .4)$). What's the problem here? Would it work for $U(0, 1)$, and if so, what is your estimate and credible interval for p?

2.1 Bayesian Philosophy of Probability

The entire concept of a posterior as an update on our prior is built on the Bayesian philosophy of probability.

What we have discussed so far is very much mechanical. But ultimately, Bayesian statistics has a fundamental different interpretation of the analysis and exactly what we mean when we make probabilistic statements about a system under study.

So far, we have been using the classical concept of probability based on a long sequence of repetitions of a concept (hence the term "frequentist"). Some times this makes a great deal of sense (flips of the coin); other times, it's rather artificial (what do we mean when we discuss the probability of a weather event or a stock market event – alternative universes where we rerun the system?)

Bayesian probability is built upon interpreting probability as a subjective measure of personal belief. What exactly is meant by "subjective" is a source of controversy. It can mean simply that probability statements are judgements made by the practitioner. We use personal ideas of probability all the time, like what is the chance a horse will win in a race (e.g. for the purposes gambling, where we might express it in odds). It allows us to express levels of uncertainty beyond just yes/no.

One way of comparing the difference is to consider two different kinds of uncertainty about a problem:

- *aleatory uncertainty*: due to inherent randomness in a system or observations of the system
- *epistemic uncertainty*: due to our own incomplete understanding of the system; the point of scientific inquiry is to reduce this

Bayesian statistician uses the language of probability to reflect both of these kinds of uncertainty, while frequentists use probability to only refer to aleatory uncertainty.

2.2 Picking a Prior

The first thing to note is that there are often two choices to make in choosing a prior: the distributional form of the prior, and the value of parameters of the prior distribution. So in our baseball example, we used a common choice of a prior distribution for a proportion to be a $Beta(\alpha, \beta)$. But that's not enough for a prior – you also have pick the specific α and β, which we did based on my (limited) baseball knowledge.

There are often standard or convenient choices of distributions that people will make for certain types of parameters. So far, we've only looked at examples where we can easily calculated the posterior by hand, so that the final posterior distribution follows a known, standard distribution. This has a lot of conveniences, in terms of being able to work with it to understand your problem.

A special class of problems of convenient priors are **conjugate prior distributions**. This is a prior distribution for θ where $f(\theta)$ and $f(\theta|x)$ belong to the same parametric family. Both examples we have looked at so far were examples of such

a setting: the normal prior distribution for the mean θ of our normally distributed data resulted in a normal posterior. And a Beta prior for our probability for the p Bernoulli resulted in a Beta posterior distribution for p. (Note that in the normal case, the likelihood $f(x|\theta)$ was also normal, but that isn't required for a conjugate prior. In our Baseball example, the data was distributed Bernouilli – it was only the prior and posterior that were the same distribution).

Again, this doesn't solve the problem of what parameters you choose for the prior.

2.3 Philosophies on Choosing the Prior

How should we choose a prior distribution? Several schools of thought answer this question differently:

- Subjective Bayesianism: The prior should reflect in as much detail as possible the researcher's prior knowledge of and uncertainties about the problem. These should be determined through *prior elicitiation*.

  A problem with subjective approach is that the prior information may not be precise enough to define a specific distribution. For example, suppose prior information says that the distribution of θ should be symmetric, with median zero and quartiles of ± 1. While that is important and pretty specific prior information, that doesn't define a single distribution and analysis with different such distributions might result in very different results. (There are strategies for trying to further quantify a subjective probability, which we won't go into.)

  Alternatively, you might pick a convenient distributional form for a prior, and then use this type of information to guide your choice of the parameters of this distribution. This is basically what I did on the baseball example, but it has real limitations since you are stuck with what that distribution can encode.

- Objective Bayesianism: This is the idea that the prior should not incorporate the subjective beliefs of the researcher. They should reflect ignorance of the parameter, and have minimal downstream influence on our inference. Priors with this property are often known as *non-informative* or *objective*. What it means to be non-informative is tricky, and there's not a single agreement on what properties a non-informative prior should have. Because objective priors can be used "off-the-shelf", we will discuss objective priors in more detail below.

- Robust Bayesianism: This is the idea that reasonable people may hold different priors, and it is difficult to precisely express even one person's prior; we should therefore consider how our inference depends on our prior.

  A simple way to do this is to try different reasonable priors and compare the results, often called *sensitivity analysis*.

Another way is to start at the end, and start with posterior result, perhaps developed under a specific choice of prior, like $P(\theta > 1|X) > 0.70$. Then work backward and figure out what priors are compatible with this result. This can help see what kind of prior assumptions must be made to *not* get such a result; if they seem more unreasonable than those priors that do imply such a result, then this gives greater confidence to this result.

A Bayesian analysis will often incorporate more than one of these ideas.

2.4 Objective or Non-informative priors

The idea of being objective is that there should be a structural rule for how to create a prior, based on the particular data generation method $f(x|\theta)$ that is assumed.[1] This is in contrast with the idea of subjective Bayes which tries to determine the researchers' prior knowledge and construct a reasonable prior that reflects this information.

The question becomes how to create a rule to create a prior? The goal is often stated as wanting to be "non-informative", but there are not clear definitions of what this mean. Non-informative could mean that the prior reflects ignorance of the parameter, or that it have minimal influence on our inference. What it means to be non-informative is tricky, and there's not a single definition on what it means to be non-informative. Moreover, the study of objective priors over many decades has resulting in the realization that being completely "non-informative" is not really achievable – a prior of any form shapes the posterior. These types of priors can be better thought of as objective priors, which simply means there is a clear rule that can be used across many settings for how to choose the prior (differentiating it from subjective priors, which use problem-specific information to choose an appropriate prior). This allows the existance of priors used by convention, creating a "standard of reference",[2] allowing the results of one analyst to be reasonably compared with that of another. Therefore, non-informative is often a poor choice of words, except in the fact that using one demonstrates you as an investigator are declaring that you have ignorance of specific information about the problem that should be used.

There are different strategies for picking priors depending on how one defines the end goal.[3]

[1] "Generally it is standard to describe as "objective" any statistical analysis which only depends on the model assumed [i.e. $f(x|\theta)$] and the data observed. Logically, the prior distribution should come before the data model, but, in practice, priors are often chosen with reference to a likelihood function." Gelman (2017, *Entropy*)

[2] This concept is due to Jeffreys', who reinvigorated the idea of objective priors, which before that had been mainly the flat, uniform prior of Laplace/Bayes.

[3] A good review of these ideas is Kass and Wasserman, "The Selection of Prior Distributions by Formal Rules"

2.4.1 Equal Probability – Uniform prior

The simplest way to define what it means to be non-informative is to give equal probability to all possible values of the parameter – i.e. place a uniform distribution on θ. When the range of θ is bounded, this prior gives a valid PDF, since it integrates to 1.

This was the notion of non-informative posited in the 1700s as an objective strategy for how to pick a prior and this prior is still very prevalent; this is also known as the principle of insufficient reason. Bayesian analysis using uniform priors was the statistical analysis that existed until Fisher and Neyman's introduction of frequentist and MLE theory in the 1920s, which was seen as a more objective strategy and quickly took over.

Example (Bernoulli, Wasserman p.178) Suppose $X_1, \ldots, X_n|p \overset{i.i.d}{\sim} Bernoulli(p)$ and $p \sim U(0,1)$. Then we have

$$\begin{aligned} f(p|X) &\propto f(p)\mathcal{L}(p) \\ &= p^{\sum_i X_i}(1-p)^{n-\sum_i X_i} \\ &= p^{(S+1)-1}(1-p)^{n-(S+1)-1} \end{aligned}$$

where $S = \sum_i X_i$. We write it this way to see the similarity to the $Beta(\alpha, \beta)$ distribution for $\alpha > 0, \beta > 0$ with pdf,

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

and mean $\alpha/(\alpha+\beta)$. So we can see that

$$p|X \sim Beta(\sum_i X_i + 1, n - \sum_i X_i + 1),$$

and

$$\hat{p} = E(p|X) = \frac{\sum_i X_i + 1}{n+2} = \frac{n}{n+2}\bar{X}_n + (1 - \frac{n}{n+2})p_0$$

where $p_0 = 1/2 = E(p)$, the expected value under the prior distribution. Again we see a trade-off between the signal in the data ($\bar{X}_n$) and the prior (p_0).[4]

Improper Prior It is also possible at times to assign a uniform prior when the range of θ is not bounded. For example, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. We could

[4]Indeed, the uniform is a $Beta(1,1)$ distribution, so we've actually already done this calculation in our baseball example.

take $f(\theta) \propto 1$. This prior is called **improper** since

$$\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$$

so it does not define a proper distribution.

However, we can still apply the Bayesian machinery to get

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \exp\left\{-\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n]\right\} \\ &\propto \exp\left\{-\frac{n}{2}[\theta^2 - 2\theta\bar{X}_n + \bar{X}_n^2]\right\} \end{aligned}$$

which is the form of a $N(\bar{X}_n, 1/n)$ distribution for θ. Therefore in this case we still have a proper posterior, i.e. the posterior is a well-defined distribution.

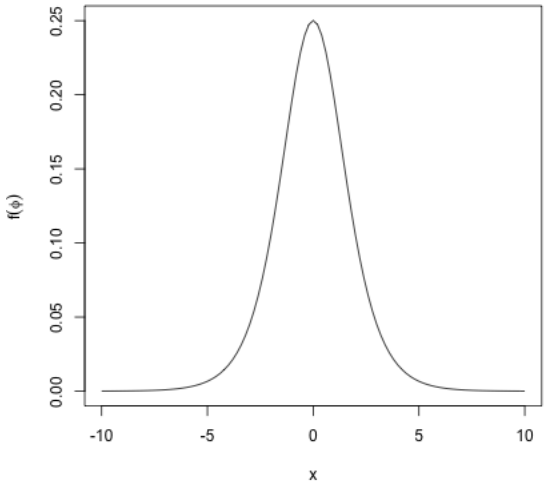
Notice that this improper prior gives a posterior distribution for θ which has mean $\bar{X}_n$, i.e. an estimate of $\hat{\theta} = \bar{X}_n$, which is the frequentist estimate. You could also think of the improper prior as the limit of taking our previous (proper) prior $\theta \sim N(0, b^2)$ with $b^2 \to \infty$.

Issue with "Non-informative": parameterizations An issue with deciding that "non-informative" means giving equal probability to all values is that this depends on the parameterization of the problem. The prior may give equal probability for one parameterization of the likelihood, but not for another.

Notice that the flat prior above for the Bernoulli has this problem. Let $\phi = \log(\frac{p}{1-p})$. This is a common alternative way to parameterize a Bernoulli problem, like in logistic regression. Then if the prior distribution for p is $U(0,1)$, this implies that the prior distribution of ϕ is given by

$$f(\phi) = \frac{e^{\phi}}{(1+e^{\phi})^2}$$

This is not at all giving equal probability to possible values of ϕ:



So this "uninformative" prior about p gives us an informative prior about ϕ, which is counter intuitive.

Flat priors also do not generalize well to high dimensions. If you are in higher dimensions (i.e. multivariate priors), and your parameters are each bounded (lies in a hyper-rectangle), then if you put a uniform prior on each parameter, this results in saying you think the vector of the parameters lies near the surface of the hyper-rectangle. Generally all intuitions about geometry that we have from 3D space break down in higher dimensions.

Parameterizations and the behavior in high dimensions are a frequent problem in trying to come to a definition of non-informative. Trying to find a rule for making non-informative priors that holds in high dimensions and for alternative parameterizations is basically impossible. This leads to looking for alternative notions of what we want out of a strategy for creating non-informative/objective prior.

2.5 Jeffreys' Priors (Invariant priors)

Invariant priors One property we might want a strategy for creating priors to possess is that it be **transformation invariant**. For example, if θ represents a distance, our inference shouldn't depend on whether θ is expressed in miles or kilometers. Or more generally a monotone function $\phi = g(\theta)$. Such a requirement does not require the selection of any specific parameterization, which could in many problems be rather arbitrary

If we want a rule for choosing a prior that is transformation invariant, this means if you you apply the rule for creating a prior $f_\theta(\theta)$ for θ, and if the same rule is applied

to $\phi = g(\theta)$ to obtain $f_\phi(\phi)$ then both priors (f_θ and f_ϕ) should imply the same prior distributions for θ and ϕ: f_ϕ should give the same prior to ϕ that you would have gotten if you instead first obtained f_θ and then found what prior it implied for ϕ by the change of variable rule:

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|.$$

So for an invariant procedure, it doesn't matter whether you apply the rule to θ and then find the distribution of ϕ or apply the rule directly to ϕ (and then find the distribution θ), you wind up with the same priors for θ and ϕ, respectively. (But that doesn't mean that $f_\theta(\theta) = f_\phi(\phi)$ – we wouldn't want that since θ and ϕ are on completely different scales!)

Clearly the flat prior strategy does not have this property, as we demonstrated above.

Jeffreys' Prior The **Jeffreys' prior** satisfies this property.[5] For 1-dimensional θ, the Jeffreys' prior is

$$f(\theta) \propto I_n(\theta)^{1/2},$$

where $I_n(\theta)$ is the Fisher information for θ based on the likelihood $\mathcal{L}(\theta)$. The Jeffreys prior can be improper.

Jeffreys' method also has the advantage that it can easily generalize to multivariate θ , though it can have some unfavorable properties in higher dimensions.

Jeffreys' Prior for Bernoulli In our above example,

$$I(p) = \frac{1}{p(1-p)}$$

so that the Jeffreys' prior is

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}$$

This is the shape of a $Beta(1/2, 1/2)$, so the Jeffreys' prior is a Beta distribution,

$$f(p) = \frac{\Gamma(1)}{\Gamma(1/2)\Gamma(1/2)} p^{-1/2}(1-p)^{-1/2} = \frac{1}{\pi} p^{-1/2}(1-p)^{-1/2}$$

And our posterior distribution is

$$f(p|X) \propto p^{\sum X_i - 1/2}(1-p)^{n - \sum X_i - 1/2}$$

[5]It is not the only prior that is invariant however. There are other proposals that are also invariant.

which is again a Beta distribution, so that

$$f(p|X) = Beta(\sum_i X_i + 1/2, n - \sum_i X_i + 1/2)$$

Setting the prior to be $f(p) = Beta(1/2, 1/2)$ implies a prior distribution for ϕ. With $g^{-1}(\phi) = \frac{e^\phi}{1+e^\phi}$ we have

$$\begin{aligned} f_\phi(\phi) &= \frac{1}{\pi}\left(\frac{e^\phi}{1+e^\phi}\right)^{-1/2}\left(\frac{1}{1+e^\phi}\right)^{-1/2}\left|\frac{e^\phi}{(1+e^\phi)^2}\right| \\ &= \frac{1}{\pi}\frac{e^{\phi/2}}{1+e^\phi} \end{aligned}$$

Notice that it doesn't matter for i.i.d data if I use $I(p)$ or $I_n(p)$ since we are only defining the prior up to a constant.

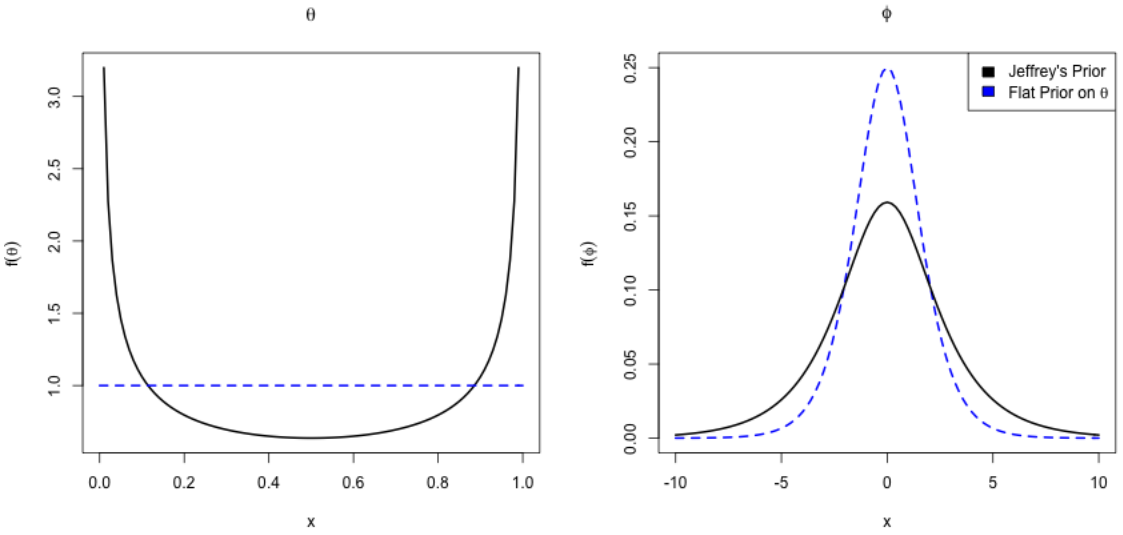
If instead I apply the Jeffrey's rule directly to ϕ to determine a prior for ϕ, this means the I would use a prior density $\propto \sqrt{I_n(\phi)}$. The log-likelihood with respect to ϕ is

$$\ell(\phi) = \phi s - n\log(1+e^\phi)$$

so that

$$\frac{\partial^2}{\partial\phi^2}\ell(\phi) = \frac{-ne^\phi}{(1+e^\phi)^2}$$

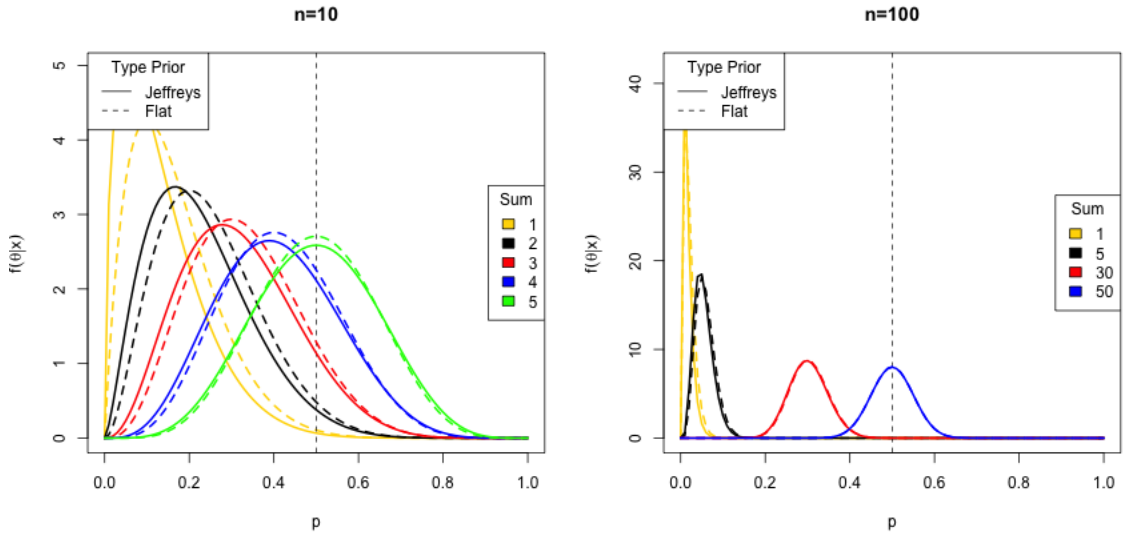
Therefore $I_n(\phi) = \frac{ne^\phi}{(1+e^\phi)^2}$, so that a prior with density $\propto \sqrt{I_n(\phi)}$ is the same as that implied by the Jeffrey's prior on p found above.



Notice that this gives us an objective rule for picking the prior, and that rule for how to pick the prior is invariant to the parameterization of the problem. It

doesn't mean that the priors give equal probability to the parameters nor that they are equally "ignorant" of the parameters, in the sense of the probabilities across the parameter range.

We can also consider its effect on the *posterior* distribution for various observed values of $\sum_i X_i$ and n. This is usually the question of largest concern in picking a prior – how does it influence my inference?



As is frequently the case with (reasonable) priors, the choice of priors has the greates effect when there is a small amount of data (relative to the number of parameters in the data model).

Exercise 5. Find the Jeffreys prior for λ when $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$.

Exercise 6. Is the Jeffreys prior proper?

Exercise 7. Find the implied prior distribution for $\phi = \log \lambda$.

Exercise 8. Show the prior in 3 is the same as the Jeffreys prior for ϕ.

2.6 Other Notions of Noninformative

There are many other objective strategies for creating priors, satisfying different criteria. A common goal for objective priors is to want that the prior knowledge doesn't influence the analysis – it is dominated by the information provided by the data. Then the concept of how a prior should be chosen is relative to the information provided by the data, i.e. that provided by the likelihood. We don't see the data before picking our prior (that would severely complicate our probability statements), but we

can consider the average amount of information the data will be expected to bring, and choose our prior based on this. This is the flavor of the Jeffreys' prior that uses the Fisher Information

Reference priors are determined by explicitly choosing a prior so that the amount of information "added" by using the data is maximized. This is formalized by considering how far $f(\theta|X)$ is from $f(\theta)$ using the Kullbeck-Leibner divergence.[6] We don't know the posterior distribution in advance, but like with the Fisher information, we can consider the expected distance between the posterior and the prior, and set the prior to be the one furthest, on average, from the posterior. By maximizing the divergence, we allow the data to have the maximum effect on the posterior estimates.

This method leads to the Jeffreys' prior in the case of 1-dimensional θ. But it has better properties for higher-dimensional θ.

Other strategies for objective priors are based on the goal the inference results will align asymptotically with frequentist results for large enough sample sizes, ensuring that the information in data is guaranteed to "dominate" the analysis.

3 Calculation of Posterior distribution

Notice, that in Bayesian analysis, there's not any big question about what estimator to use, which test to use, etc. Once you have a model for these components, the conceptual answer to the problem arrives immediately.

However, calculation of these quantities *can* be quite difficult, and indeed consideration of the calculation of these quantities is a major part of Bayesian thought.

In particular the inferential quantities we are interested in are generally integrals, which are harder to numerically evaluate when θ is multivariate than maximization problems.[7]

Conjugate Priors First consider a special class of problems in which the calculations can be done in closed form. A **conjugate prior distribution** for θ is one for which $f(\theta)$ and $f(\theta|x)$ belong to the same parametric family. We just saw an example of such a setting, above, with the normal distribution with a normal prior

[6] See parametric model for a definition of KL divergence between distributions; the KL divergence $KL(f|g)$ is loosely defined as gain in information in f relative to g.

[7] MAP estimates only require finding the maximum of the posterior density, and can be solved much more easily than integrals, but those are 1) less desirable and 2) still leaves the problem of creating interval estimates

for the mean. (Note that in that case the likelihood $f(x|\theta)$ was also normal, but that isn't required for a conjugate prior).

Numerical Calculation Many common Bayesian estimators and methods were traditionally based on conjugate priors because you can determine a solution. However, this is rather contrived, and there's not a great reason that conjugate priors are a good choice, other than for ease of calculation.

So it is typically the case that the posterior distribution can't be calculated in closed form. This difficulty was a major roadblock for Bayesian statistics until the last 30 years or so. However, with computers, Monte Carlo sampling from the posterior became widespread.

For example, suppose we can sample $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta|X)$. Then basic Monte Carlo approximation to the posterior mean of any function $q(\theta)$ is

$$\begin{aligned} E[q(\theta)|X] &= \int q(\theta) f(\theta|X) d\theta \\ &\approx \frac{1}{B} \sum_{b=1}^{B} q(\theta_b) \end{aligned}$$

This is broader than it might seem at first glance. For example, q could be an indicator function, giving us a way of approximating the posterior probability of any event.

We should note that sampling from a distribution when you only know its density (or its density up to a constant) is tricky, and there are a lot of Bayesian methods for doing so.[8]

Methods for calculating $E(h(\theta)|X)$ We'll consider two basic methods for sampling from the posterior:

- Importance sampling: Uses sampling from an alternative distribution and then corrects the Monte-Carlo step so that can provides correct approximation to $E(h(\theta)|X)$.

[8] *If* you know the CDF F of a distribution, a standard form of sampling is *Inverse transform sampling* or *inversion sampling*. You sample $U \sim U(0,1)$ and let $Z = F^{-1}(U)$. Then the distribution of your random sample Z will come from a distribution with CDF F. However, this assumes you know the CDF F and can invert it, which is generally not the case in Bayesian settings. Usually inversion sampling is used for known, well-characterized distributions.

- Rejection sampling: Uses sampling from an alternative distribution that is feasible to sample from, and filters that sample appropriately to produce an exact, iid sample, which can be used for Monte-Carlo approximation.
- (Not discussed) Markov Chain Monte Carlo (MCMC) methods construct a Markov chain that has the posterior distribution as the chain's stationary distribution. These are very flexible and probably the most commonly used but are beyond the scope of this course. The most well-known MCMC method for doing this is the Metropolis–Hastings algorithm (of which the Gibbs sampler is an example)

3.1 Rejection Sampling

Setup: The basic idea is that we want to sample from a distribution with pdf $g(z)$, but we don't know how. But under some conditions, we can sample from another distribution with density $p(z)$ that is easy to sample from, and manipulate our sampling so that the sample actually comes from the distribution g.

Suppose that:

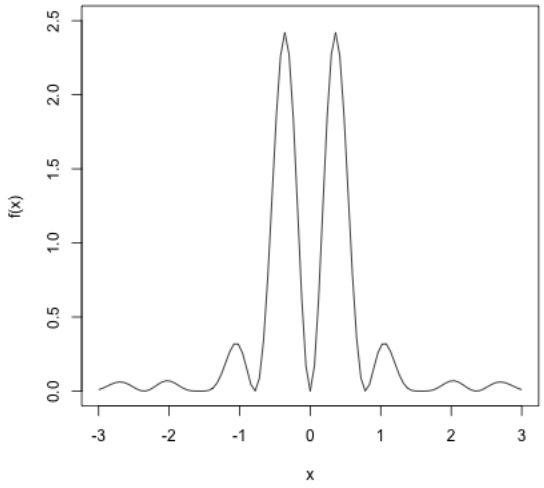
- We want to sample from a target distribution with density $g(z)$.
- We can easily sample from some density $p(z)$
- The density of p is similar to g with respect to location and spread.
- Moreover, suppose that we can find $c > 0$ such that

$$g(z) \leq cp(z) \; \forall z \quad \text{(envelope condition)}$$

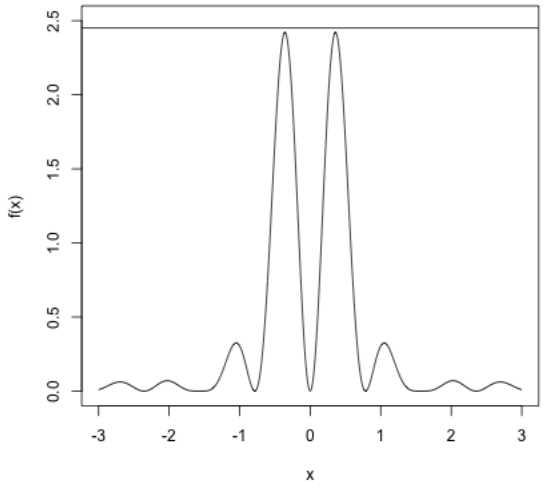
Example Consider example of a density on [-3,3],[9]

$$g(x) \propto e^{-x^2/2}(3cos^2(x)sin^2(4 * x))$$

[9]From `https://towardsdatascience.com/what-is-rejection-sampling-1f6aff92330d`



Then if I take the $U(-3,3)$, its density is $p(x) = \frac{1}{6}$ and I can easily sample from it. I can also clearly find a c so that $g(x) \leq cp(x)$ by setting (for example), c=14.7:



Of course if I sample from $U(-3,3)$, I'm going to get far too many points in the tails, and not enough in the [-1,1] range. The idea is to sample from $U(-3,3)$, but throw away some of the samples, so in the end we get a sample that matches our complicated $g(z)$. For each possible sample, we evaluate it's density under $U(-3,3)$ compared to $g(z)$.

Algorithm Then the following algorithm produces B *iid* draws from a distribution with density $g(z)$.

1. Draw $Z^{cand} \sim p(z)$.
2. Generate $U \sim Unif(0,1)$.
3. If $U \leq \frac{g(Z^{cand})}{cp(Z^{cand})}$, accept Z^{cand},
$$Z_b = Z^{cand}$$
otherwise reject Z^{cand} and repeat above until get Z_b.

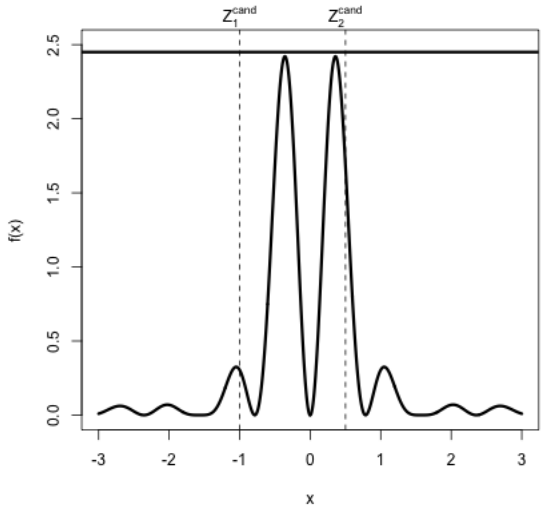
Repeat this for $b = 1, \ldots, B$ (i.e. until B values of Z^{cand} have been accepted). The sample $Z_1, \ldots, Z_B$ is an i.i.d. from the distribution with density g. [10]

Back to Example So I draw both a candidate $Z^{cand} \sim U(-3,3)$ and a uniform $U \sim U(0,1)$. (Don't be confused that in this case my sampling distribution is also a uniform!)

Suppose I draw my Z^{cand} and I get
$$Z_1^{cand} = -1$$
Then
$$\frac{g(Z_1^{cand})}{14.7p(Z_1^{cand})} = \frac{0.3}{2.45} = 0.12$$
Clearly this is a much less likely option under my target distribution g than under my sampling distribution. If I do a random $U(0,1)$ I will have only a 0.12 chance of getting a value less than this. So most of the time I will be rejecting such a draw.

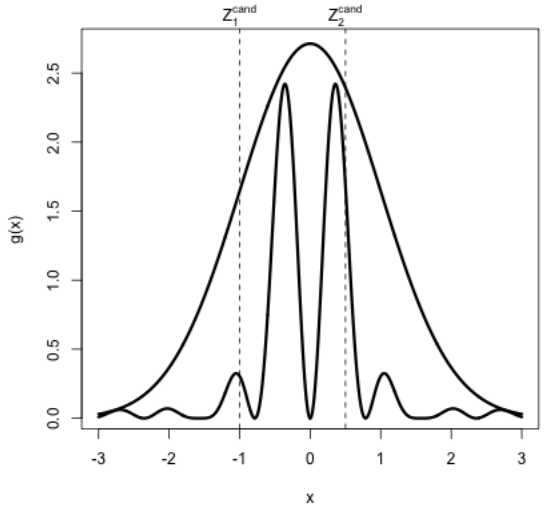
On the other hand, if
$$Z_2^{cand} = 0.4,$$
this is a much more likely value and
$$\frac{g(Z_2^{cand})}{14.7p(Z_2^{cand})} = \frac{2.35}{2.45} = 0.95$$
So my random $U(0,1)$ has a probability of 0.95 chance of being under this value.

[10] If you are familiar with the Metropolis-Hastings Algorithm, it uses rejection criteria as well, but this is in the context of MCMC



Notice how intuitively this will even out our sampling probabilities, so that those areas with much greater probability in our sampling distribution than our target distribution will be rejected and vice versa.

Choice of sampling distribution Also notice that we will be wasting a lot of our draws, because there is such a gap between $U(-3, 3)$ and our target density. We get to choose a convenient sampling distribution, and we could be a lot more efficient if we chose another distribution. For example, for a $N(0, 1)$, if I let $c = 6.8$, then I have:



Returning to my previous draws,

$$\frac{g(Z_1^{cand})}{6.8p(Z_1^{cand})} = \frac{0.3}{1.64} = 0.18$$

And

$$\frac{g(Z_2^{cand})}{6.8p(Z_2^{cand})} = \frac{2.35}{2.39} = 0.98$$

My probability of keeping these value went up from 0.12 to 0.18 and 0.95 to 0.98. It's even more dramatic if I consider

$$Z_3^{cand} = -2.5$$

where the ratio is 0.25 under the $N(0,1)$ and only 0.01 under $U(-3,3)$.

However, the really important difference is not just the rejection probability is less, but also that under the $U(-3,3)$ I'm equally likely to draw Z_3^{cand} as Z_1^{cand}; so by choosing a normal sampling distribution, I will be drawing more draws to begin with from regions with lower rejection regions. Notice that there are two probabilities going on here – the probability of drawing Z^{cand} and then the probability that $U <\leq \frac{g(Z^{cand})}{cp(Z^{cand})}$ *conditional* on knowing Z^{cand}.

Only need to know g up-to constant For our Bayesian analysis, we will generally not know the density, but only up to a constant. But that is okay – in fact the function I used above to describe the density is missing a constant. We can apply this algorithm, if we only know the target density $g(z)$ up to some proportionality constant.

Specifically, suppose the actual density is

$$g(z) = k(z)/\int k(u)du$$

where $k(z)$ is a function we know but we are missing the normalizing constant. Then we only need an envelope condition on k so that

$$k(z) \leq cp(z)\ \forall z$$

And we can do the above algorithm based on $k(z)$ instead of $g(z)$

3.1.1 Apply to posterior distribution

Now consider rejection sampling where g is the posterior distribution, with

$$g(\theta) \propto k(\theta) = f(x|\theta)f(\theta).$$

We want to generate a sample of values $\theta_1, \ldots, \theta_B$ with this posterior distribution (remember θ is a random variable now).

Choice of Sampling distribution We can choose any distribution p that satisfies our conditions. One such choice is the prior (assuming we can sample from it), so that

$$p(\theta) = f(\theta).$$

It's a reasonable assumption that we can sample from our prior – our posterior may be difficult, but usually our priors will be built from standard distributions.

Note that by definition,

$$\frac{k(\theta)}{p(\theta)} = \frac{f(x|\theta)f(\theta)}{f(\theta)} = f(x|\theta) \leq f(x|\hat{\theta}_n) \equiv c$$

where $\hat{\theta}_n$ is the MLE. So we find the MLE, and we calculate the likelihood to get our c.

Algorithm In this case, the rejection sampling algorithm becomes

1. Draw $\theta^{cand} \sim f(\theta)$.
2. Generate $U \sim Unif(0, 1)$.
3. If
$$U \leq \frac{f(x|\theta^{cand})}{f(x|\hat{\theta}_n)f(\theta)},$$
accept θ^{cand},
$$\theta_b = \theta^{cand}$$
otherwise reject θ^{cand} and repeat above until get θ_b.

Repeat 1-3 until B values of θ_b have been generated.

3.1.2 Why this works

Let's look at why this gives us the sample we want.

First of all, consider the probability at any step that we will accept our candidate Z^{cand},

$$P(U \leq \frac{g(Z^{cand})}{cp(Z^{cand})})$$

We have two random variables here, U and Z^{cand}, so we will condition to calculate this

$$\begin{aligned} P(U \leq \frac{g(Z^{cand})}{cp(Z^{cand})}) &= \int P(U \leq \frac{g(t)}{cp(t)} | Z^{cand} = t) p(t) dt \\ &= \int \frac{g(t)}{cp(t)} p(t) dt \\ &= \frac{1}{c} \int g(t) dt = \frac{1}{c} \end{aligned}$$

So the overall (unconditional) probability that we accept any particular candidate is $1/c$.[11]

But this doesn't tell us why is Z_b a draw from $g(z)$? We need to prove that the distribution of the Z_b we get out is what we claimed. To do that, we need to prove that

$$P(Z_b < t) = \int_{-\infty}^{t} g(z) dz$$

Z_b is the result of the interaction of 2 random variables: Z^{cand} and U. Specifically the distribution of Z_b is the distribution of only the "successful" draws of Z^{cand}, so the distribution of Z_b is the distribution of

$$Z^{cand} | U < \frac{g(Z^{cand})}{cp(Z^{cand})}$$

So

$$\begin{aligned} P(Z_b < t) &= P(Z^{cand} < t | U < \frac{g(Z^{cand})}{cp(Z^{cand})}) \\ &= \frac{P(Z^{cand} < t, U < \frac{g(Z^{cand})}{cp(Z^{cand})})}{P(U < \frac{g(Z^{cand})}{cp(Z^{cand})})} \\ &= cP(Z^{cand} < t, U < \frac{g(Z^{cand})}{cp(Z^{cand})}) \end{aligned}$$

[11] Notice that if we only know $k(z)$, i.e. the density up to a constant, then we get $P(U \leq \frac{g(Z^{cand})}{cp(Z^{cand})}) = \frac{\int k(t)dt}{c}$. But the calculation that follows will now work out to give us

$$P(Z_b < t) = \frac{\int_{-\infty}^{t} k(z) dz}{\int k(z) dz} = \int_{-\infty}^{t} g(z) dz$$

So it doesn't change the result. But it does emphasize that calculating the unconditional probability as $1/c$ does rely on knowing g, and not just up to a constant.

To calculate this probability we are going to condition

$$\begin{aligned}
P(Z^{cand} < t, U < \frac{g(Z^{cand})}{cp(Z^{cand})}) &= E\left(I\{Z^{cand} < t\} I\{U < \frac{g(Z^{cand})}{cp(Z^{cand})}\}\right) \\
&= E\left(E\left\{I\{Z^{cand} < t\} I\{U < \frac{g(Z^{cand})}{cp(Z^{cand})}\} | Z^{cand}\right\}\right) \\
&= E\left(I\{Z^{cand} < t\} \frac{g(Z^{cand})}{cp(Z^{cand})}\}\right) \\
&= \frac{1}{c}\int I\{z < t\} \frac{g(z)}{p(z)} p(z) dz \\
&= \frac{1}{c}\int_{-\infty}^{t} g(z) dz
\end{aligned}$$

So this gives us that

$$P(Z_b < t) = \int_{-\infty}^{t} g(z) dz$$

The role of c The overall (unconditional) probability that we accept any particular candidate is $1/c$, indicating the importance of the value of c.[12]

If c has to be very large to satisfy the envelope condition, then we will have to sample and reject many candidates before we are successful. In our previous example, when we chose $U(-3, 3)$, we had $c = 14.7$, while for $N(0, 1)$ we had $c = 6.8$ of accepting a candidate. We only had the density proportional to a constant, but so we can't translate those into individual probabilities, but we do know that

$$\frac{P(\text{reject } Z^{cand} | U[-3, 3])}{P(\text{reject } Z^{cand} | N(0, 1)} = \frac{14.7}{6.8} = 2.16$$

So we are twice as likely to reject candidates under $U(-3, 3)$ as $N(0, 1)$ as our choice of sampling distribution $p(x)$

3.2 Importance Sampling

Importance sampling is an adaptation to the usual Monte Carlo integration that allows us to sample from an "importance" distribution defined by density p rather

[12]Notice that before in our example we were calculating conditional probabilities, i.e. if we knew Z^{cand}, what is the probability of keeping it. Here we are also accounting for what is the probability that we draw Z^{cand} as well as the probability we accept it.

than the desired density g. Note that

$$\begin{aligned} E_g[h(Z)] &= \int h(z)g(z)dz \\ &= \int h(z)\frac{g(z)}{p(z)}p(z)dz \\ &= E_p\left[h(Z)\frac{g(Z)}{p(Z)}\right] \end{aligned}$$

In other words, we can change an expectation of $h(Z)$ assuming Z came from distribution g (which we can't sample from) into an expectation of $h(Z)\frac{g(Z)}{p(Z)}$ under the assumption that Z came from distribution p (which we can sample from).

If we had a sample $Z_1, \ldots, Z_B$ i.i.d from distribution with density p, we could approximate the expection $E_p[h(Z)\frac{g(Z)}{p(Z)}]$, as

$$\begin{aligned} E_g[h(Z)] = E_p\left[h(Z)\frac{g(Z)}{p(Z)}\right] &\approx \frac{1}{B}\sum_{b=1}^{B} h(Z_b)\frac{g(Z_b)}{p(Z_b)} \\ &= \frac{1}{B}\sum_{b=1}^{B} h(Z_b)w(Z_b). \end{aligned}$$

Notice that our final approximation is a average of our $h(Z_b)$, weighted with

$$w(Z_b) = \frac{g(Z_b)}{p(Z_b)}$$

(note these "weights" don't sum to 1, so not technically a weighted average)

Only need to know g up-to constant Again, if we only know g upto a constant,

$$g(z) = k(z)/\int k(u)du$$

then we have a problem immediately applying the above, since we would get

$$w(Z_b) = \frac{k(Z)}{p(Z_b)\int k(u)du}$$

and we don't know the integral in the denominator.

But we can rework the expectation calculation to apply the same trick to both

the numerator and denominator,

$$\begin{aligned} E_g[h(Z)] &= \frac{\int h(z)k(z)dz}{\int k(z)dz} \\ &= \frac{\int h(z)\frac{k(z)}{p(z)}p(z)dz}{\int \frac{k(z)}{p(z)}p(z)dz} \\ &= \frac{E_p\left[h(Z)\frac{k(Z)}{p(Z)}\right]}{E_p\left[\frac{k(Z)}{p(Z)}\right]} \end{aligned}$$

With our sample $Z_1, \ldots, Z_B$ i.i.d from distribution with density p, we have

$$\begin{aligned} E_g[h(Z)] = \frac{E_p\left[h(Z)\frac{k(Z)}{p(Z)}\right]}{E_p\left[\frac{k(Z)}{p(Z)}\right]} &\approx \frac{\frac{1}{B}\sum_{b=1}^{B} h(Z_b)\frac{k(Z_b)}{p(Z_b)}}{\frac{1}{B}\sum_{b=1}^{B}\frac{k(Z_b)}{p(Z_b)}} \\ &= \frac{\frac{1}{B}\sum_{b=1}^{B} h(Z_b)w(Z_b)}{\frac{1}{B}\sum_{b=1}^{B} w(Z_b)} \\ &= \frac{1}{B}\sum_{b=1}^{B} h(Z_b)w^*(Z_b) \end{aligned}$$

where $w(Z) = \frac{k(Z)}{p(Z)}$ and $w^*(Z)$ is those values normalized to sum to 1, so we indeed get a weighted average.

Apply to posterior distribution We can apply the idea of importance sampling to calculate the expectation of any $h(\theta)$ under our posterior, i.e.

$$E[h(\theta)|x].$$

In the case of the posterior mean,

$$f(\theta|x) \propto k(\theta) = \mathcal{L}(\theta)f(\theta)$$

and we have

$$w(\theta_b) = \frac{\mathcal{L}(\theta)f(\theta)}{p(\theta_b)}$$

Again, we can use any appropriate distribution p. If we pick $p(\theta) = f(\theta)$, our prior, we sample from the prior: $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta)$, then our weights simplify to

$$w^*(\theta) = \frac{\mathcal{L}_n(\theta)}{\sum_{b=1}^{B}\mathcal{L}_n(\theta_b)}$$

and

$$E[h(\theta)|x] \approx \sum_{b=1}^{B} h(\theta_b)w^*(\theta_b).$$

Other distributions In the above discourse, we have picked p to be the prior distribution, because we can guarantee that it satisfies our conditions and is an appropriate distribution for θ. . Other choices are to pick common distributions that often work for a range of problems, like the t distribution. This has the advantage of being able to preprogram the algorithms for general problems as well as being useful if the prior distribution is not easy to sample from.

Comparison with Rejection sampling Notice that Rejection sampling gives you a sample exactly from the distribution with density g, unlike Importance Sampling. On the other hand, to get a sample of size B we must generate, on average, $c \times B$ candidates from p and reject $(c-1) \times B$ of them. Importance sampling uses every one of the B samples that are created.