MVE550 2020 Lecture 2.1
Tossing a biased coin

Petter Mostad

Chalmers University

November 4, 2020

Inference

- We want to use *stochastic models* to make probabilistic predictions about future observations based on previous observations (i.e., data).
- Simple example: Predict the range of an electric car. Use as model a normal distribution with *parameters* μ (expectation) and σ (standard deviation).
- Adapting a model to data is called *inference*.
- The *classical* or *frequentist* inference *paradigm*: Define a model in terms of *unknown parameters*, *estimate* these parameters using the data, and *predict* from the model with the estimated parameters plugged in.
- The *Bayesian* inference paradigm: Build a stochastic model (a probability distribution) with variables representing both observed data and the future data one would like to predict. Use for prediction the conditional distribution with data variables fixed to their observed values.

Throwing a dice

- If you are trowing a fair six-sided dice, your stochastic model would be that each outcome has probability $1/6$.
- New observations would be independent of old observations: To make predictions, you don't need data.
- Assume instead the dice may be biased in some way, but you don't know exactly how.
- A way to make predictions would be to first acquire data, i.e., record approximately how often each outcome occurs, and use that information when predicting. Outcomes would be *dependent*.
- Thus you use a more complex stochastic model that reasonably models the dependency.
- Given a sequence 1, 5, 6, 1, 3, 1, 1, 2, 1, 5, the probability for 1 in the next throw is then computed as

$$\Pr(1 \mid 1,5,6,1,3,1,1,2,1,5) = \frac{\Pr(1,5,6,1,3,1,1,2,1,5,1)}{\Pr(1,5,6,1,3,1,1,2,1,5)}$$

A biased coin example

You believe a coin is *biased*, and that the chance for heads is *either* 0.7 *or* 0.3. The probability for each of these possibilities is 0.5.

- Objective: Learn in which direction the bias goes by observing repeated throws of the coin. Then make predictions.
- The probability of observing y heads in n throws is

$$\Pr(y) = 0.5 \cdot \text{Binomial}(y; n, 0.7) + 0.5 \cdot \text{Binomial}(y; n, 0.3).$$

- The probability of observing a *specific sequence* with y heads in n throws is

$$\Pr(y) = 0.5 \cdot 0.7^y 0.3^{n-y} + 0.5 \cdot 0.3^y 0.7^{n-y}.$$

- One can compute the probability of observing any *sequence*. The prediction for observing H after observing for example $HTTHTTT$ can be computed as

$$\frac{\Pr(HTTHTTTH)}{\Pr(HTTHTTT)}.$$

Biased coin example

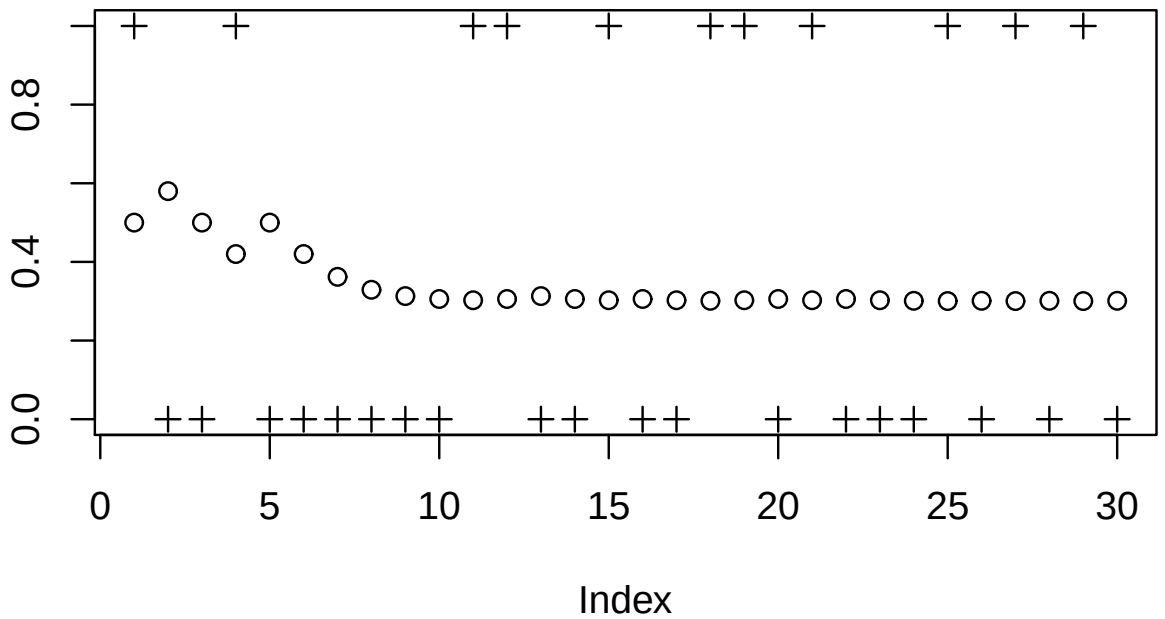

Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is that θ, the probability of heads, is either 0.7 or 0.5, with $\Pr(\theta = 0.7) = \Pr(\theta = 0.3) = 0.5$.

Reformulation using the underlying parameter θ

- A more common way to compute is to define the model in terms of an underlying unknown parameter θ, so that all observations are independent *given* θ.
- In our case, θ is a discrete random variable with possible values 0.7 and 0.3, with $\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5$. We then have, if y is the count of heads in the n first throws and y_{new} is the count of heads in the next throw

$$y \mid \theta \sim \text{Binomial}(n, \theta) \qquad \text{and} \qquad y_{new} \mid \theta \sim \text{Binomial}(1, \theta)$$

- We can use the general formulas

$$\pi(y_{new} \mid y) = \sum_{\theta} \pi(y_{new} \mid \theta)\pi(\theta \mid y) \text{ and } \pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}$$

- For example, $\pi(\theta = 0.3 \mid y) = \frac{\pi(y|\theta=0.3)\pi(\theta=0.3)}{\pi(y|\theta=0.7)\pi(\theta=0.7)+\pi(y|\theta=0.3)\pi(\theta=0.3)}$.
- We get exactly the same results as above. (Prove!)

General terminology

- The probability distribution for θ, $\pi(\theta)$, is called the *prior*.
- The probability distribution for the data y given θ, $\pi(y \mid \theta)$ is called the *likelihood*, when it is viewed as a function of θ.
- The probability distribution for θ given the value of the data y, $\pi(\theta \mid y)$ is called the *posterior*.

Generalizing the example: Learning about a proportion

- An experiment is performed n times. We assume there is a probability θ for "success" (or heads) each time, and that the outcomes are independent after θ has been fixed. Let y be the observed number of successes. We get $y \sim \text{Binomial}(n, \theta)$. Given y, what do we know about θ?
- For a Bayesian analysis, we need a stochastic model, i.e., a joint probability density (or probability mass function) $\pi(y, \theta)$. We have defined the likelihood $\pi(y \mid \theta)$. We need to define the prior $\pi(\theta)$.
- Let us first try with the prior $\theta \sim \text{Uniform}[0, 1]$.

Finding the posterior for θ

- The conditional model $\pi(\theta \mid y)$ (the posterior for θ) can be computed with Bayes formula. We get

$$\begin{aligned}\pi(\theta \mid y) &= \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} = \frac{\pi(y \mid \theta)\pi(\theta)}{\int_0^1 \pi(y \mid \theta)\pi(\theta)\, d\theta} \\ &= \frac{\text{Binomial}(y; n, \theta)}{\int_0^1 \text{Binomial}(y; n, \theta)\, d\theta} = \frac{\theta^y (1-\theta)^{n-y}}{\int_0^1 \theta^y (1-\theta)^{n-y}\, d\theta}.\end{aligned}$$

- Before we continue with computing the integral, we review the definition of the Beta distribution.

Review of definition: The Beta distribution

θ has a Beta distribution on $[0, 1]$, with parameters α and β, if its density has the form

$$\pi(\theta \mid \alpha, \beta) = \frac{1}{\mathsf{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\mathsf{B}(\alpha, \beta)$ is the Beta *function* defined by

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

where $\Gamma(t)$ is the *Gamma function* defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx.$$

Recall that for positive integers, $\Gamma(n) = (n-1)! = 1 \cdot \cdots \cdot (n-1)$. See for example Wikipedia for more properties of the Beta distribution, and the Beta and Gamma functions. We write $\pi(\theta \mid \alpha, \beta) = \mathsf{Beta}(\theta; \alpha, \beta)$ for the Beta density; we then also write $\theta \sim \mathsf{Beta}(\alpha, \beta)$.

Example: Learning about a proportion, continued

- Comparing the computations above with the Beta density we can read off:
$$\theta \mid y \sim \text{Beta}\,(y+1, n-y+1)\,.$$
- To compute the probability given data y that a new observation y_{new} is a success, we can use general formulas
$$\pi(y_{new}=1 \mid y) = \int_0^1 \pi(y_{new}=1 \mid \theta)\pi(\theta \mid y)\,d\theta = \int_0^1 \theta\,\text{Beta}(\theta; y+1, n-y+1)\,d\theta.$$
Looking up the expectation of the Beta distribution, we get $\pi(y_{new}=1 \mid y) = \frac{y+1}{n+2}$.
- The posterior $\pi(\theta \mid y)$ can be computed most easily using proportionality notation. We get in our case
$$\pi(\theta \mid y) \propto_\theta \pi(y \mid \theta)\pi(\theta) = \pi(y \mid \theta) \propto_\theta \theta^y(1-\theta)^{n-y}.$$
- We can directly recognize this as a Beta density: $\theta \mid y \sim \text{Beta}(y+1, n-y+1)$

Biased coin example

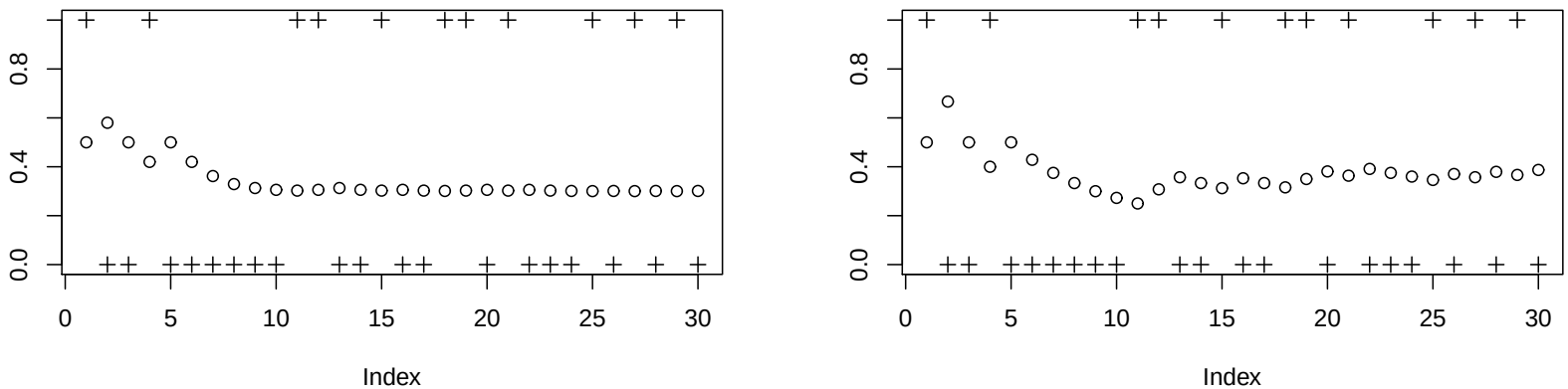

Figure: The probability of heads at each point in a sequence of observations, or the probability of "success", conditioning on the previous observations. The priors used are $\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5$ (left) and $\theta \sim \text{Uniform}(0, 1)$ (right).

Using a Beta distribution as prior

- Assume the prior is $\theta \sim \text{Beta}(\alpha, \beta)$.
- The posterior becomes (prove!)

$$\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$$

- The prediction becomes (prove!)

$$\pi(y_{new} = 1 \mid y) = \text{E}(\theta \mid y) = \frac{y + \alpha}{n + \alpha + \beta}.$$

- DEFINITION: Given a likelihood model $\pi(x \mid \theta)$. A *conjugate family of priors* to this likelihood is a parametric family of distributions for θ so that if the prior is in this family, the posterior $\theta \mid x$ is also in the family.

Biased coin example

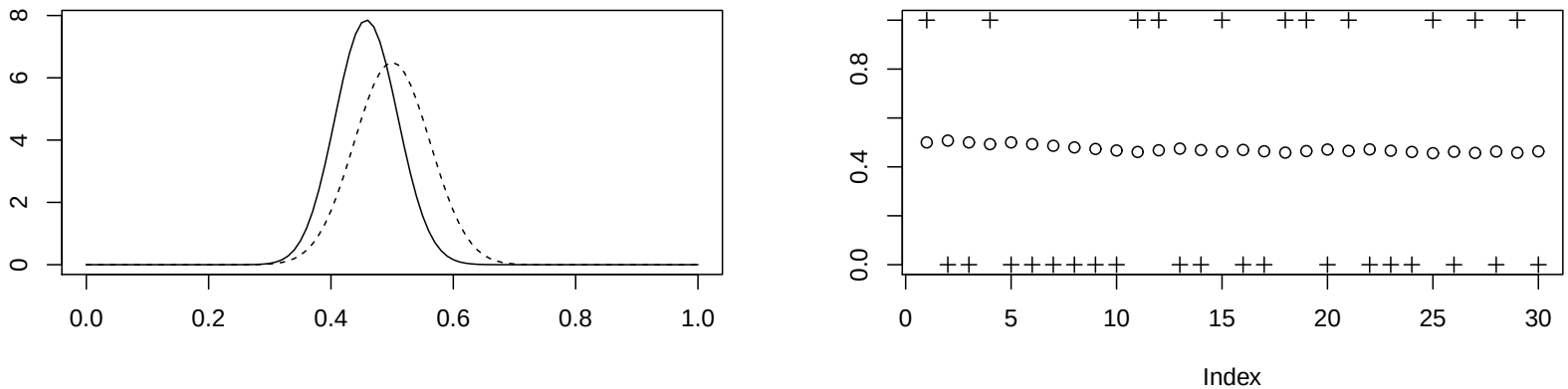

Figure: Left: The prior Beta$(33.4, 33.4)$ and the posterior Beta$(33.4 + 11, 33.4 + 19)$ for θ. Right: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails, using the shown prior.

MVE550 2020 Lecture 2.2
Conjugacy. Prediction. Discretizations

Petter Mostad

Chalmers University

November 4, 2020

Review / overview

Outline:

- DEFINITION: Given a likelihood model $\pi(x \mid \theta)$. A *conjugate family of priors* to this likelihood is a parametric family of distributions for θ so that if the prior is in this family, the posterior $\theta \mid x$ is also in the family.
- Second example: The Poisson Gamma conjugacy.
- Computations for predictive distributions.
- Bayesian inference using discretization or numerical integration.

Example: The Poisson-Gamma conjugacy

- Assume $\pi(x \mid \theta) = \text{Poisson}(x; \theta)$, i.e., that

$$\pi(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!}$$

- Then $\pi(\theta \mid \alpha, \beta) = \text{Gamma}(\theta; \alpha, \beta)$ where α, β are positive parameters, is a conjugate family. Recall that

$$\text{Gamma}(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$

- Specifically, we have the posterior

$$\pi(\theta \mid x) = \text{Gamma}\left(\theta; \alpha + x, \beta + 1\right).$$

Poisson-Gamma example

- We make repeated observations of a Poisson(θ) distributed variable for some $\theta > 0$. The observed values are $x_1 = 20$, $x_2 = 24$, and $x_3 = 23$. What is the posterior distribution for θ given this data?
- We first must decide on a prior for θ. In this example we use $\pi(\theta) \propto_\theta \frac{1}{\theta}$.
- Note that this is an *improper* prior; it is a "density" that does not integrate to 1! However, using such improper priors is possible in Bayesian statistics.
- We get the posterior after observing x_1:

$$\theta \mid x_1 \sim \text{Gamma}(20, 1)$$

- Using this as prior, we get after also observing x_2:

$$\theta \mid x_1, x_2 \sim \text{Gamma}(20 + 24, 1 + 1)$$

and similar for the last observation x_3.

Poisson-Gamma example

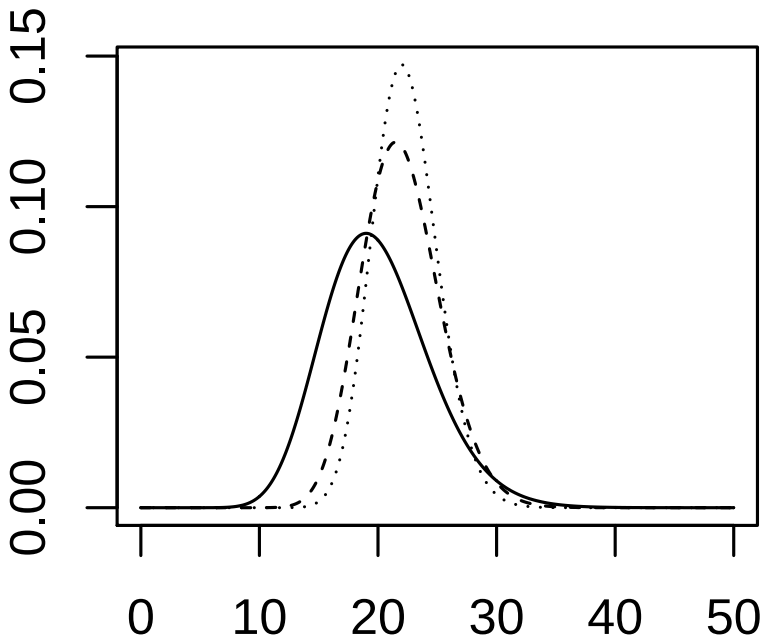

Figure: The posteriors after one, two, and three observations, where $x_1 = 20, x_2 = 24$, and $x_3 = 23$. Note how increasing amounts of data leads to a narrower posterior density.

Prediction

- In the Bayesian paradigm, we create a joint probability model for the parameters θ, the observed data x, and data we would like to predict x_{new}. Often on the form $\pi(\theta, x, x_{new}) = \pi(\theta)\pi(x \mid \theta)\pi(x_{new} \mid \theta)$.
- The distribution for x_{new} is given by conditioning on the observed x and marginalizing out θ:

$$\begin{aligned} \pi(x_{new} \mid x) &= \int_\theta \pi(\theta, x_{new} \mid x)\, d\theta = \int_\theta \pi(x_{new} \mid \theta, x)\pi(\theta \mid x)\, d\theta \\ &= \int_\theta \pi(x_{new} \mid \theta)\pi(\theta \mid x)\, d\theta \end{aligned}$$

  This is called the *posterior predictive distribution*.
- It is also possible to look at the predictive distribution for x before it has been observed. This is called the *prior predictive distribution*:

$$\pi(x) = \int_\theta \pi(x, \theta)\, d\theta = \int_\theta \pi(x \mid \theta)\pi(\theta)\, d\theta$$

Predictive distributions when using conjugate priors

- When using a conjugate prior, not only do we have an analytic expression for the posterior density for θ, we also have analytic expressions for the prior predictive density and the posterior predictive density.
- To see this for the prior predictive density, use this formula derived from Bayes formula:

$$\pi(x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)}$$

  The prior predictive density is on the left and all expressions on the right have analytic formulas.
- Note that, when using the right hand side for computing, θ will necessarily eventually disappear.
- As the posterior predictive distribution is on the same form as the prior predictive, we also get an analytic formula for it. Specifically, we can write

$$\pi(x_{new} \mid x) = \frac{\pi(x_{new} \mid \theta)\pi(\theta \mid x)}{\pi(\theta \mid x_{new}, x)}.$$

Predictive distribution for the Poisson Gamma conjugacy

- We have seen: If $k \mid \theta \sim \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$ then $\theta \mid k \sim \text{Gamma}(\alpha + k, \beta + 1)$.
- Direct computation gives the prior predictive distribution

$$\pi(k) = \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(\theta \mid k)} = \frac{\beta^{\alpha}\Gamma(\alpha + k)}{(\beta + 1)^{\alpha+k}\Gamma(\alpha)k!}$$

- Note that the positive integer x has a Negative Binomial distribution if its probability mass function is

$$\pi(x \mid r, p) = \binom{x + r - 1}{x} \cdot (1 - p)^x p^r = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)}(1 - p)^x p^r$$

- We get that the prior predictive is Negative-Binomial$(\alpha, \beta/(1 + \beta))$.
- Note that we can get the posterior predictive by simply replacing the α and β of the prior with the corresponding $\alpha + k$ and $\beta + 1$ of the posterior.

Poisson-Gamma example

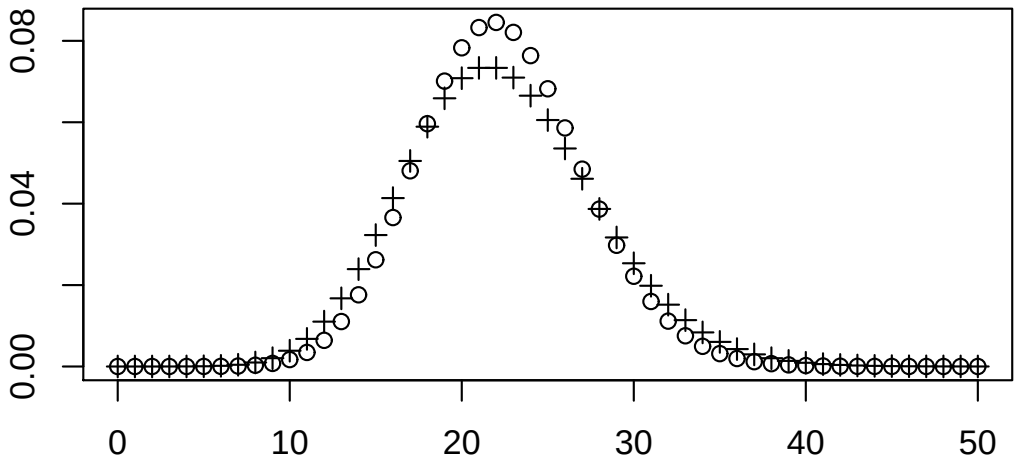

Figure: Two different ways of predicting the values of k_4, given the observations $k_1 = 20, k_2 = 24, k_3 = 23$. The pluses represent the Bayesian predictions using the posterior predictive; the circles represent the Frequentist predictions, using the Poisson distribution with parameter $(20 + 24 + 23)/3 = 22.33$.

Bayesian inference using discretization

If the sample space of θ is finite, Bayesian inference is quite easy:

- The prior distribution $\pi(\theta)$ is represented by a vector.
- The posterior distribution $\pi(\theta \mid y)$ is obtained by termwise multiplication of the vectors $\pi(y \mid \theta)$ and $\pi(\theta)$ and normalizing so the result sums to 1.
- The prediction $\pi(y_{new} \mid y) = \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta$ simplifies to taking the sum of the termwise product of the vectors $\pi(y_{new} \mid \theta)$ and $\pi(\theta \mid y)$.
- USAGE: Approximate a 1D (and 2D) prior $\pi(\theta)$ by finding $\theta_1, \ldots, \theta_k$ equally spaced in the definition area for θ, compute $\pi(\theta_i)$ and normalize these values so that they sum to 1.
- Check out the R code in the example of Section 1.5 of the Compendium!

Bayesian inference using numerical integration

- The prediction we want to make can be expressed as a quotient of integrals:

$$\begin{aligned}
\pi(y_{new} \mid y) &= \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta \\
&= \int_\theta \pi(y_{new} \mid \theta) \frac{\pi(y \mid \theta)\pi(\theta)}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta}\, d\theta \\
&= \frac{\int_\theta \pi(y_{new} \mid \theta)\pi(y \mid \theta)\pi(\theta)\, d\theta}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta}
\end{aligned}$$

- One idea: Compute these integrals using numerical integration.
- Can work well as long as the dimension of θ is low (max 2 or 3?) and the functions are well-behaved.
- Check out the R code in the example of Section 1.6 of the Compendium!