TMS150 / MSG400 Stochastic data processing and simulation. Part I: Script and Lab 5. (Autumn 2023)

Moritz Schauer

October 7, 2024

1 Introduction

Bayesian inference means quantifying uncertainty with probability, that is considering unknown parameters as random variables.

We already use random variables to model

- physical random processes,
- random experiments, measurement errors, classification errors,

and the proposition is use them also

- to model uncertainty about values of unknown quantities relevant for our understanding of the world.

A physical random process is for example the time it takes that a radioactive isotope decays. Also the value of the card on top of a stack of poker cards after it has been shuffled is physically random. Measurement errors are inherent to all measurement processes. Uncertainty relates to lack of knowledge about quantities, which might be random or not: You don't know the number of coins in my pocket. It could be random, but it could also be that I might have removed all coins from my pocket intentionally before asking in which case their number is not very random, but still unknown to you.

Basis of Bayesian inference is to model unknown parameters as *random variables*. Equivalently, one assigns a *joint distribution* to all (unknown or observed) quantities. For example an unknown probability parameter $u \in [0, 1]$ of interest could be modelled as random variable $U \sim \mathrm{U}([0, 1])$, drawn from a uniform distribution on [0,1]. The statistical procedures are then all *derived* from basic probability theory which tells how to compute the conditional distribution of the unknowns given the information one has.

- Assume you have made an observation y, and you have variable x you would like

to make predictions about. You model both as random variables X and Y, for example you specify their joint probability density (or probability mass function) $f_{X,Y}(x, y)$.

- Then Bayesian inference uses posterior density, that is the conditional density $f_{Y|X=x}(y)$ of X given the observation $Y = y$. It is given by by *Bayes' theorem*,

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{1}$$

for predictions, where $f_Y(y)$ is the marginal density for y.

Comments:

1. Random variables, joint and marginal densities are reviewed in section 2.
2. Computing the marginal density typically requires us to *integrate* or *sum*.
3. Often the joint density can be written as product of a prior density $f_X(x)$ of x and a conditional (predictive) density (the density of Y given that $X = x$.)

$$f_{X,Y}(x, y) = f_{Y|X=x}(y) \cdot f_X(x).$$

4. Equation (1) a version of the elementary rule

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B)} \tag{2}$$

for random variables.

To sum up: based on reasonable assumptions, create a stochastic model containing random variables representing observed data and parameters or even future observations you would like to predict and use for prediction of unobserved random variables Bayes' theorem.

2 Random variables and (joint) distributions

Random variables are numeric quantities whose value depends on the outcome of a random experiment.

Events are subsets of the sample space Ω. An event A will happen with a probability p. p is a proportion, a number $0 \leq p \leq 1$. We write $\mathbf{P}$ for the function that gives probability p of an event A. For example $\mathbf{P}(A \cap B)$ is the probability that both events A and B happen.

To introduce random variables, consider the following example.

Hold a six-sided die in your hand. Trowing it on a table it constitutes a random experiment. The die will show a number of eyes $X \in \{1, 2, 3, 5, 6\}$ and X will depend on how the dice rolls. After the die is thrown, it will show one particular number $x \in \{1, 2, \ldots, 6\}$.

Random variables are often denoted by capital letters X, Y, Z etc. Different random variables can dependent on the outcome of the same experiment.

We can speak of the *events* that the random variables take certain values, for example the event A that the die shows a 6, which we write $A = \{X = 6\}$. Then $\mathbf{P}(X = 6)$ is the probability of the event $\{X = 6\}$. If the die is fair, any of those numbers will come up with equal probability, meaning that the events $\{X = 1\}$, ..., $\{X = 6\}$ have equal probability, or $\mathbf{P}(X = 1) = \cdots = \mathbf{P}(X = 6) = \frac{1}{6}$. For any other number $x \notin \{1, 2, 3, 4, 5, 6\}$ we have $\mathbf{P}(X = x) = 0$.

The *distribution* P_X of X is the function which assigns each set $B \subset \mathbb{R}$ the probability

$$B \mapsto P_X(B) = \mathbf{P}(X \in B).$$

The laws of probability hold for P_X, for example $P_X(A \cup B) = P_X(A) + P_X(B)$ for disjoint $A \cap B = \varnothing$.

One can also ask for the probability of two random variables X and Y taking values x, y at the same time, for example for the probability of the event $\{X = 3\} \cap \{Y = 4\}$, which is written $\mathbf{P}(X = 3, Y = 4)$. It may depend not only on the distributions of X and Y but also on the relationship of X and Y (which may depend on the same experiment). Given a pair of random variables X and Y, distributions P_X and P_Y are called *marginal distributions*.

If

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B) = P_X(A)P_Y(B) \text{ for all choices } A, B \subset \mathbb{R}$$

we call X and Y *independent*. This is the case if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all choices $A, B \subset \mathbb{R}$.

The first consequence of modelling uncertainty probabilistically that prediction are made using the rules of probability, in particular the transformation formulas for discrete and continuous random variables.

2.1 Discrete random variables

A random variable is called *discrete* if it is integer-valued or otherwise has only a finite or countable number of values. So X (the die) is integer valued, and therefore discrete, and $X/2$ is also discrete because it only takes on 6 different values, $\{1/2, 1, 3/2, 2, 5/2, 3\}$.

Discrete random variables can be described by their probability mass function, or *density* for short, a function $f\colon \mathbb{R} \to [0, \infty)$ such that such that

$$\mathbf{P}(X = u) = f(u).$$

Then

$$\mathbf{P}(X \in A) = P_X(A) = \sum_{\text{all } a \in A} f(a)$$

Here and below *all* means summing over all values a for which $f(a) > 0$. That means if X is discrete with probability mass function f and $A = \{1, 2, 3\}$, then $\mathbf{P}(X \in A) = f(1) + f(2) + f(3)$.

For the six-sided die

$$f(u) = \begin{cases} 1/6 & u \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

and $\mathbf{P}(X \in \{1, 2, 3\}) = f(1) + f(2) + f(3) = \frac{1}{2}$.

In `R` a random variable is modelled as function. Instead of throwing a die, we can call `sample(1:6, 1)`. Calling the function returns a numerical value which depends on an implicit random experiment. Note that calling `sample(1:6, 1)` twice returns independent random numbers (different random variables.)

> **Question 1:** For random variables, $X + X$ refers to two times the *same* random variable. This behaves different as `R`-code `sample(1:6, 1) + sample(1:6, 1)` which gives the sum of two independent copies. What are the possible values of $X + X$ can take? What values are possible when evaluating `sample(1:6, 1) + sample(1:6, 1)`?

If the two random variables X and Y take only a finite or countable number of values each

$$f_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y)$$

is the *joint density function* (or joint probability mass function in some texts) of X and Y with

$$\mathbf{P}(X \in A, Y \in B) = \sum_{\text{all } x \in A} \sum_{\text{all } y \in B} f_{X,Y}(x, y).$$

For discrete random variables the marginal densities can be obtained by summing out the other variable:

$$f_X(x) = \mathbf{P}(X = x) = \sum_{\text{all } y} f_{X,Y}(x, y), \quad f_Y(y) = \mathbf{P}(Y = y) = \sum_{\text{all } x} \int f_{X,Y}(x, y).$$

Finally,

$$f_{Y|X=x}(y) = \mathbf{P}(Y = y \mid X = x).$$

Question 2: a.) Throw a die, call the result X. Then, independently, throw a coin. If the coin shows heads, let $Y = X$. Else, let $Y = X + 1$.

Explain that the joint distribution model of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{12} & \text{if } y = x+1 \text{ and } x \in \{1,\dots,6\} \\ \frac{1}{12} & \text{if } y = x \text{ and } x \in \{1,\dots,6\} \\ 0 & \text{otherwise} \end{cases}$$

Assume $Y = 3$. Compute $f_{X|Y=y}(x)$ for $x \in \{1,\dots,6\}$.

b.) Show Bayes' rule (1) using the rules for conditional probabilities for discrete random variables.

If $Y = h(X)$ and X is discrete with density f then the *transformation formula* allows us to compute the mathematical expectation of the random variable Y

$$\mathbf{E}\left[h(X)\right] = \sum_{\text{all } x} h(x) f(x). \tag{4}$$

Using $h(x) = x$ this formula gives also the expectation of X.

Using $\mathbf{P}(X \in A) = \mathbf{E}\left[\mathbf{1}_A(X)\right]$ where $\mathbf{1}_A$ is the indicator function with $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise,

$$\mathbf{P}(Y \in A) = \sum_{\text{all } x} \mathbf{1}_A(h(x)) f(x).$$

If discrete random variables X and Y are independent with densities f_X and f_Y respective, their joint density is $f_{XY}(x,y) = f_X(x) f_Y(y)$ and

$$\mathbf{E}\left[h(X,Y)\right] = \sum_{\text{all } x} \sum_{\text{all } y} h(x,y) f_X(x) f_Y(y). \tag{5}$$

Question 3: Write a program that computes the expectation of the product of two independent six-sided dice throws X and Y using $h(x,y) = xy$ in (5).

2.2 Continuous random variables

A random variable X is *continuous* if $\mathbf{P}(X = u) = 0$ for all $u \in \mathbb{R}$ and there is a probability density function $f\colon \mathbb{R} \to [0,\infty)$, that is a function

$$\mathbf{P}(a \le X \le b) = \int_a^b f(x)\mathrm{d}x$$

Because $\mathbf{P}(X = u) = 0$ one can also replace $\le$ by $<$ in the definition.

The probability density function is also called *density* for short.

The *distribution function* of X,

$$F_X(u) = \mathbf{P}(X \leq u) = P_X((-\infty, u])$$

can be used to compute the probability of events $\{a < X \leq b\}$:

$$\mathbf{P}(a < X \leq b) = F_X(b) - F_X(a)$$

If X has density f_X, F_X is a antiderivative of the density function with $F_X(a) \to 1$ for $a \to \infty$.

Example: We say that X has a normal distribution (or Gaussian distribution) with parameters mean μ and variance $\sigma^2 > 0$, in symbols

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

to indicate that X is a random variable with density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Later we denote the Gaussian density with parameters μ, σ^2 by

$$\phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{6}$$

The program `rnorm(n, mean = m, sd = s)` returns a $\mathrm{N}(m, s^2)$ distributed random variable (mind the square.)

Example: Uniform random variable. The program `runif(1)` corresponds to a random variable giving uniformly distributed random numbers in the interval $[0, 1]$.

We say that X has a uniform distribution with parameters a and b in symbols

$$X \sim \mathrm{U}(a, b)$$

to indicate that X is a random variable with distribution function

$$F(u) = \begin{cases} 1 & u > b \\ \dfrac{u-a}{b-a} & u \in [a, b] \\ 0 & u < a. \end{cases} \tag{7}$$

Let $a = 0$ and $b = 1$ and F the function from (7). Then the statements "$X \sim \mathrm{U}(0, 1)$" and "$F(u) = \mathbf{P}(X \leq u)$ for all real numbers u" are synonymous.

For example `runif(1)` has this distribution function with parameters $a = 0$ and $b = 1$ and `runif(1, min = a, max = b)` gives uniformly distributed random numbers for

other parameters.

A pair of random variables X, Y is *jointly continuous* if $\mathbf{P}(X = x, Y = y) = 0$ for all $x, y \in \mathbb{R}$, and there is a joint probability density $f_{X,Y} \colon \mathbb{R} \times \mathbb{R} \to [0, \infty)$ such that

$$\mathbf{P}(a \le X \le b, c \le Y \le d) = \int_{y=c}^{d} \int_{x=a}^{b} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \int_{x=a}^{b} \int_{y=c}^{d} f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x$$

For jointly continuous random variables the marginal densities can be obtained by integrating out the other variable:

$$f_X(x) = \int f_{X,Y}(x,y)\mathrm{d}y, \quad f_Y(y) = \int f_{X,Y}(x,y)\mathrm{d}x.$$

Example: Two jointly normal random variables X and Y with correlation ρ, means μ_X, μ_Y and variances σ_X^2, σ_Y^2 have joint density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho^2)}\right),$$

where

$$z = \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}.$$

and for example

$$f_X(x) = \int f_{X,Y}(x,y)\mathrm{d}y = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right),$$

is the marginal density of X, that is $X \sim N(\mu_X, \sigma_X^2)$.

If continuous random variables X and Y are independent with densities f_X and f_Y respective, their joint density is $f_{XY}(x,y) = f_X(x)f_Y(y)$ and

$$\mathbf{E}\left[h(X,Y)\right] = \int\int h(x,y)f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y = \int\int h(x,y)f_X(x)f_Y(y)\mathrm{d}y\mathrm{d}x \qquad (8)$$

You can choose the order of inner and outer integral, take whatever makes integration easier.

This is the last transformation formula for expectations (and for probabilities, by choosing $h(X,Y)$ as indicator function) we cover. There is also transformation formula for densities. They are concerned with the question: If $Y = h(X)$, and X has density f_X, what about the density f_Y of Y? We don't cover it here.

3 Integration

3.1 Riemann approximations

Equation (8) expresses expectations of continuous random variables as integral. Let us recall.

Take a function f defined on the interval $[a, b]$. The definite integral of a function f from a to b, denoted the symbol

$$\int_a^b f(x)\mathrm{d}x$$

is defined by the following limit:

Divide the area under the curve in n boxes, each of the same width $\Delta x = (b-a)/n$ and with height $f(x_i^*)$, where i runs from 1 to n and x_i^* are points between the start $(i-1)\Delta x$ and the end $i\Delta x$ of each box on the x-axis. Then compute the Riemann sum (which depends on the choice of x_i^*)

$$R_n = \sum_{i=1}^{n} \underbrace{f(x_i^*)}_{\text{height}} \cdot \underbrace{\Delta x}_{\text{width}}$$

and let n go to infinity:

$$\int_a^b f(x)\mathrm{d}x = \lim_{n\to\infty} R_n$$

We only define the integral for functions where the limit on the right hand side does not change if we choose $x_i^* = a+(i-1)\Delta x$ (the left endpoint of a box), or $x_i^* = a+(i-1)\Delta x$ the right endpoint of a box, or any other points $x_i^* \in [a + (i-1)\Delta, a + i\Delta x]$. So when computing or approximating $\int_a^b f(x)\mathrm{d}x$ for an integrable function f we can choose whatever suits as best!

Note, that when $f(x_i^*)$ is negative, then the terms $f(x_i^*)\cdot\Delta x$ in the sum become negative too.

Computing these limits is difficult. We can use a left Riemann sum

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{i=1}^{n} f(a+(i-1)\Delta x)\cdot\Delta x, \quad \Delta x = (b-a)/n$$

for large n to approximate the integral:

Question 4: Let $X \sim \mathrm{N}(0,1)$. Use a left Riemann sum to approximate

$$\mathbf{P}(-1 \le X \le 1) = \int_{-1}^{1} \varphi(x)\mathrm{d}x$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is the standard normal density. Compare with integration using R's `integrate`:

```
> integrate(dnorm, -1, 1)
0.6826895 with absolute error < 7.6e-15
```

Life is easier, if we know an antiderivative F of f,

$$F' = f.$$

Then the fundamental theorem of calculus says

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a) = [F]_a^b .$$

That's why we like to know the distribution function F if possible.

We can also approximate double integrals by (left) Riemann sums:

$$\int_{y=c}^{d}\int_{x=a}^{b} g(x,y)\mathrm{d}x\mathrm{d}y \approx \sum_{i=1}^{n}\sum_{j=1}^{n} g(a+(i-1)\Delta x, c+(j-1)\Delta y)\Delta y\Delta x \tag{9}$$

where $\Delta x = (b-a)/n, \Delta y = (d-c)/n$.

Question 5:

Consider the ordinary differential equation

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = -Yu(t), \quad u(0) = X$$

with random parameter Y and random starting value X, quantified by

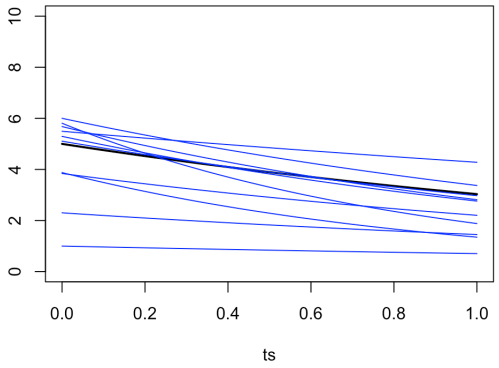
$$X \sim \mathrm{N}(5, 1.5^2)$$
$$Y \sim \mathrm{N}(0.5, 0.3^2).$$

with X and Y *independent*.

Solving the ODE up to time t gives the functional form of the value of the random value of the trajectory of u at time t as

$$\exp(-tY)X.$$

a.) Sample from X and Y and plot some (20) random trajectories u on the time interval $t = 0$ to $t = 1$.



b.) Compute

$$\mathbf{E}\left[\exp(-Y)X\right]$$

the expectation of the random value of the trajectory u at time $t = 1$. Use (8) and approximate the integrals by Riemann sums (9) choosing $n = 500$ and $a = 0$, $b = 10$, $c = -1, d = 2$.

3.2 Grid prior

Instead of approximating an integral over probability densities with a Riemann sum, one may directly approximate a continuous 1D or 2D density with a discrete probability distribution: If for example a real variable X with $X \in [a, b]$ has a continuous density $f(x)$, we may approximate it by choosing equally spaced values $x_1, \ldots, x_n$ in the interval $[a, b]$ (for example $x_i = a + \frac{i}{n}(b - a)$), $i \in \{1, \ldots, n\}$.)

Setting a probability for each such value to

$$\pi(x_i) = \frac{f(x_i)}{\sum_{j=1}^{n} f(x_j)},$$

we obtain a probability mass function for a discrete approximation of the random variable X.

More generally, if X and Y are random variables, $X \in [a, b]$ and $Y \in [c, d]$ with joint density $f(x, y)$, then by choosing a grid of points (x_i, y_j), with $i = 1, \ldots, n$ and $j = 1, \ldots, m$ for example $(x_i, y_j) = (a + \frac{i}{n}(b - a), c + \frac{j}{m}(d - c))$, and setting

$$\pi(x_i, y_j) = \frac{f(x_i, y_j)}{\sum_{k=1}^{n} \sum_{s=1}^{m} f(x_k, y_s)}.$$

we obtain a joint probability mass function for a discrete approximation of the random variables X and Y.

3.3 Approximating expectations by averaging over samples

Generally, if X is a random variable, the expected value of $f(X)$ can be approximated by taking a sample $X_1, \ldots, X_n$ of independent copies of X and computing the average

$$\mathbf{E}[f(X)] \approx \frac{1}{n} \sum_{j=1}^{n} f(X_j)$$

That this works is essentially the Law of Large Numbers, and how fast the approximation becomes how good is generally governed by the Central Limit Theorem. Roughly, quadrupling the sample size doubles the accuracy. However, the accuracy of an optimisation using an approximate sum as above is not always easy to assess.

4 Bayesian inference

4.1 Conditional probability

The next consequence of modelling uncertainty probabilistically is observing data or in fact anything modelled as random variable changes the *conditional* distribution of unobserved variables by the rules of conditional probability. For example if X is a parameter of interest modelled as random variable, and any event B was observed (for example the event that another observable random variable Y took the value 5, $B = \{Y = 5\}$, the probability of the event $A = \{X \in (a, b]\}$ for $a, b \in \mathbb{R}$ changes to the conditional probability $\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$.

If $\mathbf{P}(B) > 0$, the *conditional probability* given B is defined by

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Let us think about this formula: The conditional probability $\mathbf{P}_{|B}(\cdot) := \mathbf{P}(\cdot \mid B)$ assigns probability 0 to the complement of B,

$$\mathbf{P}_{|B}(C) = 0 \text{ for } C \subset B^c$$

and proportional probability to sets $A_1, A_2 \subset B$,

$$\frac{\mathbf{P}_{|B}(A_1)}{\mathbf{P}_{|B}(A_2)} = \frac{\mathbf{P}(A_1)}{\mathbf{P}(A_2)}.$$

such that $\mathbf{P}_{|B}(C) = 1$ for $C \supset B$.

We have

$$\mathbf{P}(A \cap B) = \mathbf{P}(B \mid A)\mathbf{P}(A) = \mathbf{P}(A \mid B)\mathbf{P}(B)$$

and therefore if $\mathbf{P}(B) > 0$, *Bayes' rule* for probabilities,

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B)}$$

Here $\mathbf{P}(A)$ is the *prior* probability of the event A, a proposition or hypothesis of interest and *data* B has been observed/obtained. The formula computes $\mathbf{P}(A \mid B)$, the *posterior* probability of A incorporating information about the data B. $\mathbf{P}(B \mid A)$ represents the probability of observing B given A or the *likelihood* of A having seen B (note the inversion).

Similarly, if $\mathbf{P}(X \in B) > 0$, we can consider distribution of the random variable X under the conditional probability $\mathbf{P}(\cdot \mid X \in B)$. Then

$$f_{X|X\in B}(x) := \frac{f_X(x)\mathbf{1}_{x\in B}}{\mathbf{P}(X \in B)}$$

is the *conditional density of X given B.*

Example: If $X \in \{1, \ldots, 3\}$ and $B = \{1, 3\}$ and f_X given below, then $f_{X|B}$ is found by setting $f_{X|B}(k) = 0$ for cases $k \notin B$ and renormalising:

k	1	2	3
$f_X(k)$	$\frac{1}{3}$	$\frac{5}{12}$	$\frac{1}{4}$
$f_{X\|X\in B}(k)$	$\frac{1}{3}/C$	0	$\frac{1}{4}/C$

where $C = \mathbf{P}(X \in B) = \frac{1}{3} + \frac{1}{4}$.

Question 6: You have two coins: one fair coin with head (H) on one side and tails (T) on the other, and one unfair coin having heads on both sides. You pick a coin randomly with equal probability and throw it a first time. What is $\mathbf{P}(X_1 = H)$? You throw the *same* coin again. What is $\mathbf{P}(X_2 = H \mid X_1 = H)$? Perhaps draw a probability tree. (Here X_1 is the result of the first throw, X_2 the result of the second.)

4.2 Conditional densities

The density of the conditional distribution of X after observation that Y took value y is (both for continuous or discrete)

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

for $f_Y(y) > 0$ (this is meaningful for continuous random variables even though $\mathbf{P}(Y = y) = 0$ and therefore $\mathbf{P}(\cdot \mid Y = y)$ is not defined.) Note that $f_{X|Y=y}(x)$ depends on

y.

Thus $f_{X|Y=y}(x)$ is proportional to the joint density and the denominator makes sure that $f_{X|Y=y}(x)$ integrates or sums to 1. If $f_{X|Y=y}(x)$ is continuous for example,

$$\int f_{X|Y=y}(x)\mathrm{d}x = \int \frac{f_{X,Y}(x,y)}{f_Y(y)}\mathrm{d}x = \frac{\int f_{X,Y}(x,y)\mathrm{d}x}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1$$

Example: Two jointly normal random variables X and Y with correlation ρ, means μ_X, μ_Y and variances σ_X^2, σ_Y^2. If $Y = y$ was observed, the conditional density of X given $Y = y$ is the density of the Gaussian distribution $\mathrm{N}(\mu_{X|Y=y}, \sigma^2_{X|Y=y})$ with mean

$$\mu_{X|Y=y} = \mu_X + \rho\sigma_X \frac{y - \mu_Y}{\sigma_Y}$$

and variance

$$\sigma^2_{X|Y=y} = (1 - \rho^2)\sigma_X^2,$$

that is, see (6),

$$f_{X|Y=y}(x) = \phi(x, \mu_{X|Y=y}, \sigma^2_{X|Y=y})$$

4.3 Bayes' rule

We have again *Bayes' rule*

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}.$$

That is all we need to do inference, or, more precisely, to formulate answers to statistical problems in a principled way that can be solved by integration.

In the statistical setting, where X is a parameter of interest with prior density f_{prior} $(= f_X)$, and Y with the conditional density $f_{Y|X=x}(y; x)$ is the model for the data given $X = x$.

The *likelihood* of x

$$L(x; y) = f_{Y|X=x}(y)$$

is obtained from fixing y in the conditional density and considering it as a function of x. The likelihood assigns to each x a measure how well x explains the observation.

Bayes formula then reads

$$f_{\text{post}}(x) = \frac{L(x; y)f_{\text{prior}}(x)}{C}, \tag{10}$$

where $f_{\text{post}}(x) = f_{X|Y=y}(x)$ is the posterior density of X given the data $Y = y$ and

$$C = \int L(x; y)f_{\text{prior}}(x)\mathrm{d}x$$

if X continuous and

$$C = \sum_{\text{all } x} L(x; y) f_{\text{prior}}(x)$$

if X discrete.

If $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$ is a vector of n independent observations, each modelled with same conditional/predictive density $f_{\text{pred}}(y \mid x)$ given $X = x$, the likelihood in (10) takes product form

$$L(x; \mathbf{y}^{(n)}) = \prod_{i=1}^{n} f_{\text{pred}}(y_i \mid x).$$

Example: Noisy observation. Hypothetical `data.csv` contains n measurements $y_1, \ldots, y_n$, of an unknown quantity x with error. The measurement process can be modelled as realisation of the random variables $Y_1, \ldots, Y_n$,

$$Y_i = x + Z_i, \quad Z_i \sim \mathrm{N}(0, \sigma_\varepsilon^2),$$

for $i \in \{1, \ldots, n\}$ and known noise level $\sigma_\varepsilon > 0$ and the Z_i are independent. The likelihood of the parameter x is by (6)

$$L(x; y_1, \ldots, y_n) = \prod_{i=1}^{n} \varphi(y_i - x, 0, \sigma_\varepsilon^2).$$

Previous studies have found $x = 3.2 \pm 0.2$. Modelling x correspondingly as random variable X

$$X \sim \mathrm{N}(\mu = 3.2, \sigma^2 = 0.2^2)$$

which implies that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has conditional density $f_{\mathbf{Y}|X=x} = L(x; y_1, \ldots, y_n)$ and that the data $\mathbf{y} = (y_1, \ldots, y_n)$ determine the posterior density of $X \mid \mathbf{Y} = \mathbf{y}$ by Bayes' rule

$$f_{X|\mathbf{Y}=\mathbf{y}}(x) = \frac{f_{\mathbf{Y}|X=x}(y) f_X(x)}{\int f_{\mathbf{Y}|X=x}(y) f_X(x) \mathrm{d}x} = \frac{\varphi(x, \mu, \sigma^2) \prod_{i=1}^{n} \varphi(y_i - x, 0, \sigma_\varepsilon^2)}{\int \varphi(x, \mu, \sigma^2) \prod_{i=1}^{n} \varphi(y_i - x, 0, \sigma_\varepsilon^2) \mathrm{d}x}.$$

This can be computed exactly by integration ($f_{X|\mathbf{Y}=\mathbf{y}}$ is Gaussian again), or we can approximate quantities of interest like the mean of the posterior

$$\mathbf{E}[X \mid \mathbf{Y} = \mathbf{y}] = \int x f_{X|\mathbf{Y}=\mathbf{y}}(x) \mathrm{d}x$$

by numerical integration. Note that when working with a product of densities, it can be required to take logarithms for numerical accuracy,

$$\prod_{i=1}^{n} f_{Y_i|X=x}(y_i) = \exp\left(\sum_{i=1}^{n} \log f_{Y_i|X=x}(y_i) \right)$$

4.4 Bayesian convention

Often the subscripts of the densities are dropped and densities are just called p.

Say X, Y have joint bivariate density, denoted p instead of $f_{X,Y}$. Then the convention is to use the name of the variable to indicate which (marginal or condition) density is meant:

$p(x)$	$p(y)$	$p(x, y)$	$p(x \mid y)$	$p(y \mid x)$
$f_X(x)$	$f_Y(y)$	$f_{X,Y}(x, y)$	$f_{X\mid Y=y}(x)$	$f_{Y\mid X=x}(y)$

This can be extended to three or more variables.

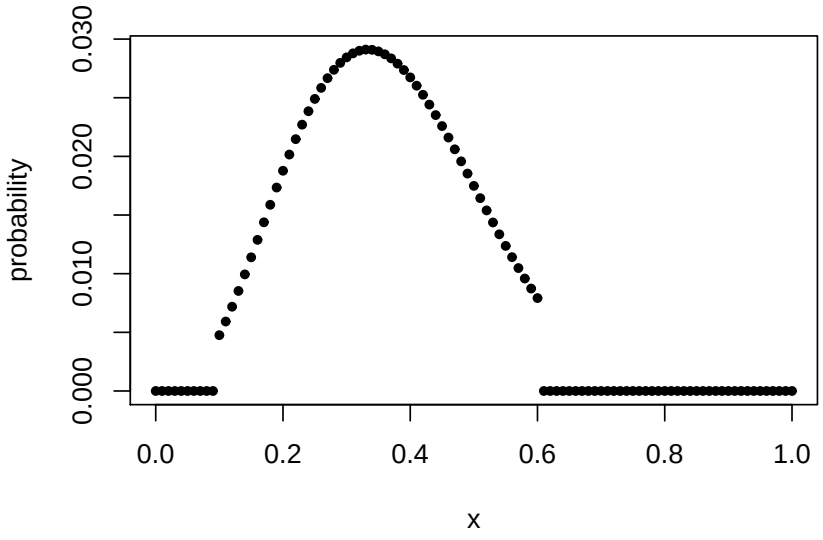


Figure 1: Prior for X for example ($\star$): The plot shows the discrete prior density $f_{\text{prior}}(x)$ for $x \in (0.00, 0.01, \ldots, 0.99, 1.0)$.

Example ($\star$): Binomial likelihood. As an example, assume Y given $X = x$ has a Binomial distribution with parameters n and x (with $0 < x < 1$) written $Y \mid X = x \sim \text{Binomial}(n, x)$, so that y is the number of "successes" among n independent trials when the probability of "success" in each trial is x. Then (using Bayesian convention) we have the likelihood

$$p(y \mid x) = \binom{n}{y} x^y (1-x)^{n-y}. \tag{11}$$

Assume y is observed and we want to use that observation to learn about x. Assume,

for example, we have the discrete prior for x which is illustrated in figure 1.

We can compute the posterior $p(x \mid y)$ by computing the prior probability $p(x)$ for each of the x for which it is non-zero (see figure 1), and multiply with the likelihood $p(y \mid x)$ computed from equation 11 at the same values. We do not need to explicitly compute $p(y)$, as we can instead normalise the vector of products so that it sums to 1. Assuming $y = 16$ and $n = 31$, figure 2 shows the prior, the likelihood $p(y \mid x)$ and the posterior.

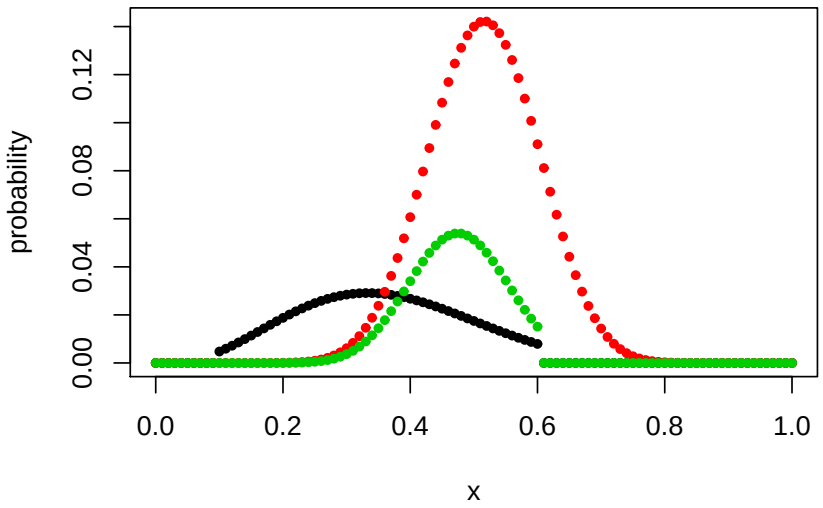


Figure 2: The prior for x from Figure 1 is shown in black; the likelihood $p(y \mid x)$ from equation 11 with $y = 16$ and $n = 31$ is shown in red, and the posterior for x is shown in green.

4.5 Predictive density

Bayes formula also gives a prediction for the next observation y_{n+1} given a vector $(y_1, \ldots, y_n)$ of n independent observations realisations, each modelled with same conditional/predictive density $p(y_i \mid x) \equiv p(y \mid x)$ (using Bayesian convention). This in form of the predictive density

$$p(y_{n+1} \mid y_1, \ldots, y_n) = \int p(y_{n+1} \mid x) p(x \mid y_1, \ldots y_n) \mathrm{d}x \tag{12}$$

if x is continuous and

$$p(y_{n+1} \mid y_1, \ldots, y_n) = \sum_{\text{all } x} p(y_{n+1} \mid x) p(x \mid y_1, \ldots y_n) \tag{13}$$

if x is discrete. In earlier notation the predictive density is $f_{Y_{n+1}|\mathbf{Y}^{(n)}=\mathbf{y}^{(n)}}$.

Example: Continuing example ($\star$), assume we would like to predict the number of successes in 10 new trials. In a classical analysis one would first use maximum likelihood to estimate $\hat{x} = 16/31 = 0.5161$ from the data. Using the Binomial distribution with parameters 10 and 0.5161 one would compute the predictive probabilities shown as black dots in figure 3. In a Bayesian analysis, one would instead use the posterior obtained above, and equation 13, replacing $p(y_{\text{new}} \mid x)$ with the Binomial distribution with parameters 10 and x, to obtain the predictive probabilities shown in red triangles in figure 3.

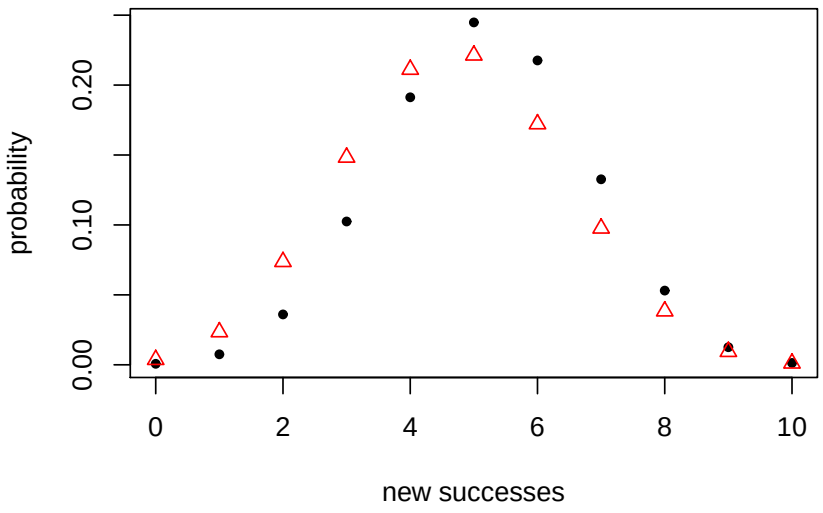


Figure 3: In black dots: The predictions for the number of "successes" in 10 new trials in a classical analysis. In red triangles are the predictions using a Bayesian analysis.