01 Foundations

Shravan Vasishth

SMLP 2019

Preview: Steps in Bayesian analysis

The way we will conduct data analysis is as follows.

- Given data, specify a *likelihood function*.
- Specify *prior distributions* for model parameters.
- Using software, derive *marginal posterior distributions* for parameters given likelihood function and prior density.
- Simulate parameters to get *samples from posterior distributions* of parameters using some *Markov Chain Monte Carlo (MCMC) sampling algorithm*.
- Evaluate whether model makes sense, using *model convergence* diagnostics, fake-data simulation, *prior predictive* and *posterior predictive* checks, and (if you want to claim a discovery) calibrating true and false discovery rates.
- Summarize *posterior distributions* of parameter samples and carry out your scientific conclusion.

Bayes' rule

A and B are events. Conditional probability is defined as follows:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ where } P(B) > 0 \tag{1}$$

This means that $P(A,B) = P(A|B)P(B)$.

Since $P(B,A) = P(A,B)$, we can write:

$$P(B,A) = P(B|A)P(A) = P(A|B)P(B) = P(A,B). \tag{2}$$

Rearranging terms:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3}$$

This is Bayes' rule.

Random variable theory

A random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$.

S_X is all the x's (all the possible values of X, the support of X). I.e., $x \in S_X$. We can also sloppily write $X \in S_X$.

Good example: number of coin tosses till H

- $X : \omega \rightarrow x$
- ω: H, TH, TTH,... (infinite)
- $x = 0, 1, 2, \dots ; x \in S_X$

Random variable theory

Every discrete (continuous) random variable X has associated with it a **probability mass (distribution) function (pmf, pdf)**. I.e., PMF is used for discrete distributions and PDF for continuous. (I will sometimes use lower case for pdf and sometimes upper case. Some books use pdf for both discrete and continuous distributions.)

$$p_X : S_X \rightarrow [0,1] \tag{4}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{5}$$

Random variable theory

Probability density functions (continuous case) or probability mass functions (discrete case) are functions that assign probabilities or relative frequencies to all events in a sample space.

The expression

$$X \sim f(\cdot) \tag{6}$$

means that the random variable X has pdf/pmf $g(\cdot)$. For example, if we say that $X \sim N(\mu, \sigma^2)$, we are assuming that the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{7}$$

Random variable theory

We also need a **cumulative distribution function** or cdf because, in the continuous case, P(X=some point value) is zero and we therefore need a way to talk about P(X in a specific range). cdfs serve that purpose.

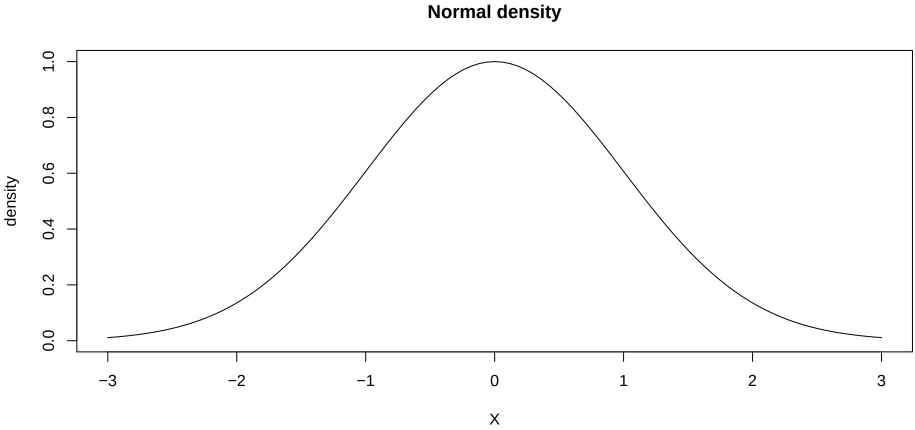
In the continuous case, the cdf or distribution function is defined as:

$$P(X < x) = F(X < x) = \int_{-\infty}^{X} f(x)\, dx \tag{8}$$

Random variable theory

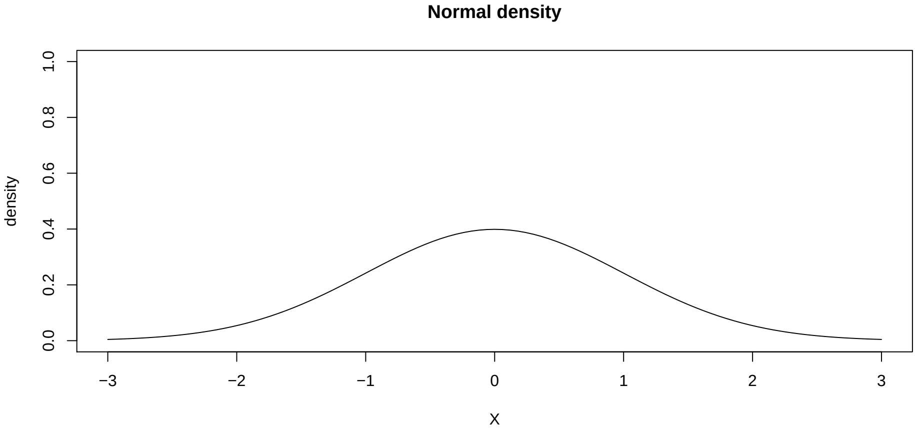
$$f(x) = \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{9}$$

This is the "kernel" of the normal pdf, and it doesn't sum to 1:



Random variable theory

Adding a normalizing constant makes the above kernel density a pdf.



Random variable theory

Recall that a random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$. S_X is all the x's (all the possible values of X, the support of X). I.e., $x \in S_X$.

X is a continuous random variable if there is a non-negative function f defined for all real $x \in (-\infty, \infty)$ having the property that for any set B of real numbers,

$$P\{X \in B\} = \int_B f(x)\, dx \tag{10}$$

Distributions

```
if ( !('devtools' %in%
       installed.packages()) )
  install.packages("devtools")

devtools::install_github("bearloga/tinydensR")
```

Then, run

```
library(tinydensR)
univariate_discrete_addin()
```

or

```
univariate_continuous_addin()
```

Binomial distribution

If we have x successes in n trials, given a success probability p for each trial. If $x \sim Bin(n, p)$.

$$P(x \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{11}$$

The mean is np and the variance $np(1-p)$.

```
dbinom(x, size, prob, log = FALSE)
### cdf:
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
### quantiles:
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
### pseudo-random generation of samples:
rbinom(n, size, prob)
```

The Poisson distribution

This is a distribution associated with “rare events”, for reasons which will become clear in a moment. The events might be:

- traffic accidents,
- typing errors, or
- customers arriving in a bank.

For psychology and linguistics, one application is in eye tracking: modeling number of fixations.

The Poisson distribution

Let λ be the average number of events in the time interval $[0, 1]$. Let the random variable X count the number of events occurring in the interval. Then:

$$f_X(x) = \mathbb{P}(X = x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{12}$$

Uniform distribution

A random variable (X) with the continuous uniform distribution on the interval (α, β) has PDF

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta, \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

The associated R function is dunif$(\mathtt{min} = a, \mathtt{max} = b)$. We write $X \sim$ unif$(\mathtt{min} = a, \mathtt{max} = b)$. Due to the particularly simple form of this PDF we can also write down explicitly a formula for the CDF F_X:

Uniform distribution

$$F_X(a) = \begin{cases} 0, & a < 0, \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha \leq t < \beta, \\ 1, & a \geq \beta. \end{cases} \tag{14}$$

$$E[X] = \frac{\beta + \alpha}{2} \text{ and } Var(X) = \frac{(\beta - \alpha)^2}{12} \tag{15}$$

```
dunif(x, min = 0, max = 1, log = FALSE)
punif(q, min = 0, max = 1, lower.tail = TRUE,
    log.p = FALSE)
qunif(p, min = 0, max = 1, lower.tail = TRUE,
    log.p = FALSE)
runif(n, min = 0, max = 1)
```

Normal distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \tag{16}$$

We write $X \sim \text{norm}(\texttt{mean} = \mu, \texttt{sd} = \sigma)$, and the associated R function is `dnorm(x, mean = 0, sd = 1)`.

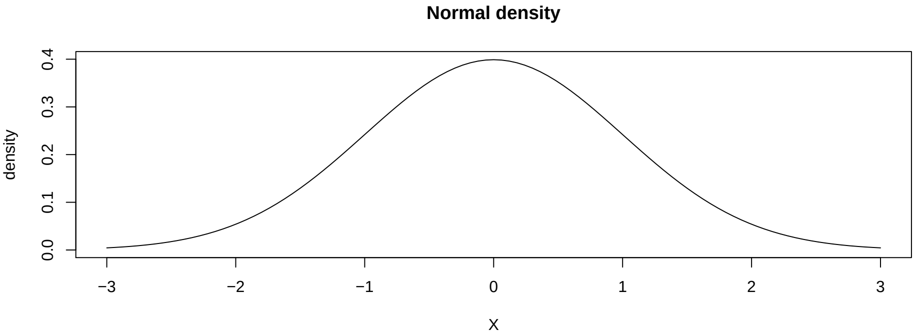


Figure 1: Normal distribution.

Normal distribution

If X is normally distributed with parameters μ and σ^2, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$.

Standard or unit normal random variable:

If X is normally distributed with parameters μ and σ^2, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters $0, 1$.

We conventionally write $\Phi(x)$ for the CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} \, dy \quad \text{where } y = (x - \mu)/\sigma \tag{17}$$

Normal distribution

The standardized version of a normal random variable X is used to compute specific probabilities relating to X .

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE,
                          log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE,
                          log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Beta distribution

This is a generalization of the continuous uniform distribution.

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$$

Beta distribution

We write $X \sim \text{beta}(\texttt{shape1} = \alpha, \texttt{shape2} = \beta)$. The associated R function is =dbeta(x, shape1, shape2)=.

The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{18}$$

t distribution

A random variable X with PDF

$$f_X(x) = \frac{\Gamma\left[(r+1)/2\right]}{\sqrt{r\pi}\,\Gamma(r/2)} \left(1 + \frac{x^2}{r}\right)^{-(r+1)/2}, \quad -\infty < x < \infty \tag{19}$$

is said to have Student's t distribution with r degrees of freedom, and we write $X \sim \mathtt{t}(\mathtt{df} = r)$. The associated R functions are dt, pt, qt, and rt, which give the PDF, CDF, quantile function, and simulate random variates, respectively.

We will just write:

$X \sim t(\mu, \sigma, r)$, where r is the degrees of freedom $(n-1)$, where n is sample size.

Jointly distributed random variables

Visualizing bivariate distributions

First, a visual of two uncorrelated normal RVs:

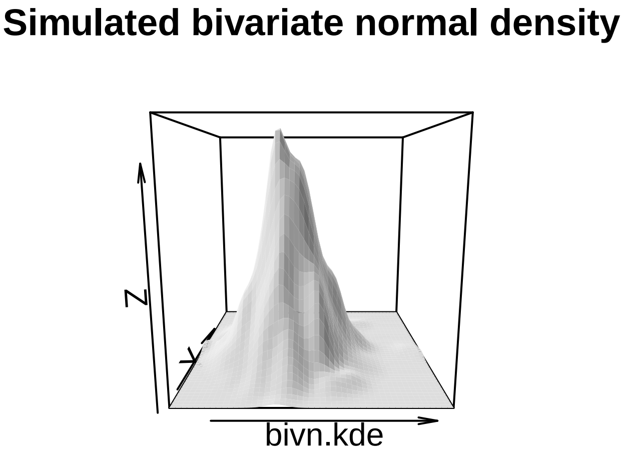


Figure 2: Visualization of two uncorrelated random variables.

Bivariate normal distributions

And here is an example with a negative correlation:

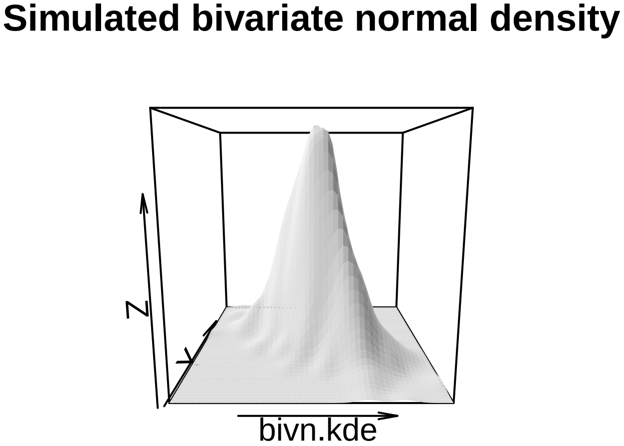


Figure 4: Visualization of two negatively correlated random variables.

Bivariate normal distributions

Visualizing conditional distributions

You can run the following code to get a visualization of what a conditional distribution looks like when we take “slices” from the conditioning random variable:

```r
for(i in 1:50){
  plot(bivn.kde$z[i,1:50],type="l",ylim=c(0,0.1))
  Sys.sleep(.5)
}
```

Maximum likelihood estimation

Discrete case

Suppose the observed sample values are $x_1, x_2, \dots, x_n$. The probability of getting them is

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta) \tag{20}$$

i.e., the function f is the value of the joint probability **distribution** of the random variables $X_1, \dots, X_n$ at $X_1 = x_1, \dots, X_n = x_n$. Since the sample values have been observed and are fixed, $f(x_1, \dots, x_n; \theta)$ is a function of θ. The function f is called a **likelihood function**.

Maximum likelihood estimation

Continuous case

Here, f is the joint probability **density**, the rest is the same as above.

Definition

If $x_1, x_2, \dots, x_n$ are the values of a random sample from a population with parameter θ, the **likelihood function** of the sample is given by

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) \tag{21}$$

for values of θ within a given domain. Here, $f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$ is the joint probability distribution or density of the random variables $X_1, \dots, X_n$ at $X_1 = x_1, \dots, X_n = x_n$.

So, the method of maximum likelihood consists of finding the maximum point in the likelihood function with respect to θ.
The value of θ that maximizes the likelihood function is the **MLE** (maximum likelihood estimate) of θ.

Finding maximum likelihood estimates

For simplicity consider the case where $X \sim N(\mu = 0, \sigma = 1)$.

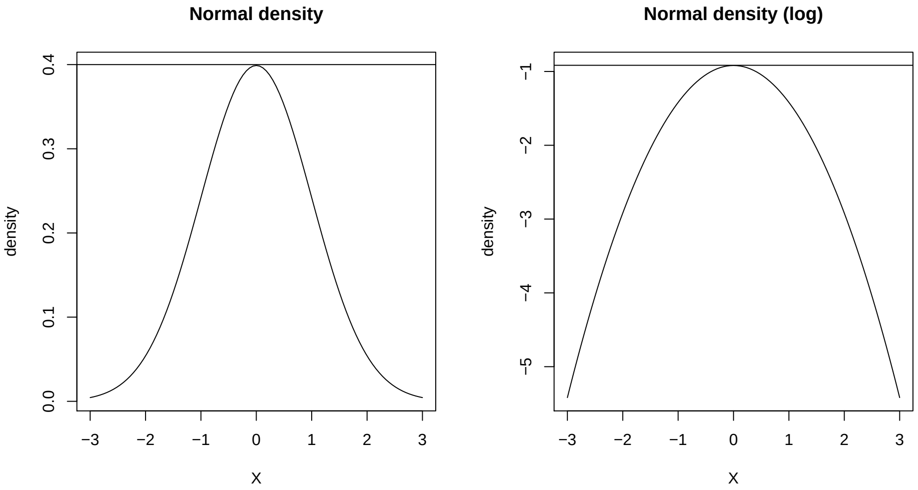


Figure 5: Maximum likelihood and log likelihood.

Finding maximum likelihood estimates

Practical implication

Suppose you sample 10 data points:
The sample mean gives you the MLE of μ, and the sample variance gives you the MLE of σ^2:

```
mean(x)
```

```
## [1] 0.14775
```

```
var(x)
```

```
## [1] 1.3952
```

Because the samples will randomly vary from one experiment to another, this does not mean the the above sample means and variances reflect the true μ and σ^2!