

Introduction to Bayesian Modeling

In the frequentist approach, estimated probabilities are viewed as one of an infinite sequence of possible data of the same experiment, while Bayesian analysis is based on expressing uncertainty about unknown quantities as formal probability distributions.

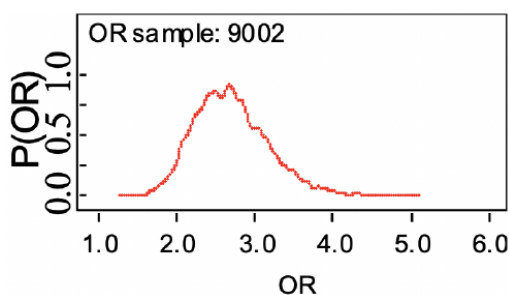
Bayes Problem: Given the number of times in which an unknown event has happened and failed - Requires the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

Bayesian statistics used probability as the only way to describe uncertainty in both data and parameter. Everything that is not known for certain are modeled with distributions and treated as random variables. Bayesian inference focuses on the uncertainty in value of parameter and decisions which is conditional on sample data.

	Exposed	Unexposed	
Diseased	a	b	m1
Not Diseased	c	d	m0
	n1	n0	n

Recall the odds ratio (OR) is the probability of having an outcome (disease) compared to the probability of not having the outcome: ad / bc

Bayesian Statistics takes the point of view that OR and RR are uncertain unobservable quantities, which we can model with a probability distribution.



Prior probability: Model/probability distribution using existing data/knowledge

Posterior probability: Updated model with new and prior data; These are used in inference

Bayes theorem allows us to learn from experience and turn a prior distribution into a posterior distribution.

Bayes Theorem

Bayes Theorem

$$P(\theta|y_1, \dots, y_n) = \frac{P(y_1, \dots, y_n|\theta) \times P(\theta)}{P(y_1, \dots, y_n)} \propto P(y_1, \dots, y_n|\theta) \times P(\theta)$$

The denominator in Bayes Theorem is a normalizing constant. If a conjugate of the family of distribution then the normalizing constant can be derived.

Applying Bayes' Theorem for inference in a diagnosis problem

- **Data:** Paul has hoarse voice (D)
- Three hypotheses (un-observable):
 - ① Paul has a Cold (H_1)
 - ② Paul has Thyroiditis (H_2)
 - ③ Paul has a stomach Flu (H_3)
- **Prior:** $P(H)$ would favor (H_1) and (H_2) over (H_3)
- **Likelihood:** $P(D|H)$ strongly favors (H_1) and (H_2) over (H_3)
- **Posterior:** $P(H|D)$ favors (H_1) and (H_2) over (H_3)

$P(H)$ = Probability of H (Prior Density)

$P(H | \text{data}) = (P(\text{data} | H) * P(H)) / P(\text{Data})$ = Posterior probability of H

Consider a generic probability distribution $p(\theta)$ for a single parameter θ . The probability distributions are defined as:

Distribution function: $F(\theta^*) = \Pr(\theta < \theta^*)$, sometimes referred to as the “tail area.”

Expectation: $E[\theta] = \int \theta p(\theta) d\theta$, where the integral is replaced by a summation for discrete θ .

Variance, standard deviation and precision:

$Var[\theta] = \int (\theta - E[\theta])^2 p(\theta) d\theta = E[\theta^2] - E^2[\theta]$; standard deviation = $\sqrt{\text{variance}}$; precision = $1/\text{variance}$.

Percentiles: the 100 q th percentile is the value θ_q such that $F(\theta_q) = q$, in particular the median is the 50th percentile $\theta_{0.5}$.

% interval: A subset of values of θ with specified total probability: generally a 100 q % interval will be (θ_1, θ_2) such that $F(\theta_2) - F(\theta_1) = q$. Such an interval might be “equi-tailed,” in that $F(\theta_2) = 1 - q/2$, $F(\theta_1) = q/2$, although for asymmetric distributions narrower intervals will be possible. The narrowest interval available is known as the Highest Posterior Density (HPD) interval: see below for an example.

Mode: the value of θ that maximises $p(\theta)$.

Conjugate Analysis: Beta-Binomial

In conjugate analysis the prior density for theta has the same functional form as the likelihood function.

Likelihood function

$$P(y|\theta) = \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$
$$\propto \theta^y (1-\theta)^{n-y}$$

Conjugate Prior

$$P(\theta|\alpha_1, \alpha_2) = \frac{\theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)}$$
$$\propto \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$$

Credible Interval

A **Confidence Interval** represents uncertainty we have about the method used in generating an interval that captures the true value in repeated sampling. The parameter is fixed and the interval is random.

A **Credible Interval** is a statement about the probability that a parameter lies in a fixed interval. The uncertainty is related to the value of the true parameter and the interval is fixed.

Markov Chain Monte Carlo (MCMC)

Simulation

There are several ways to calculate the properties of probability distributions for unknown parameters. We will be focusing on using simulation of the known random variables and estimate the property of interest.

Note that there are entire classes on algorithms used for simulation, we will not go into great depth on the underlying theory.

Monte Carlo algorithms are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.

Suppose we have a random variable X which has a probability distribution $p(x)$ and we have an algorithm for generating a large number of independent trials $x^{(1)}, x^{(2)}, \dots, x^{(T)}$. Then:

$$E(X) = \int xp(x) dx \approx \frac{1}{T} \sum_{t=1}^T x^{(t)}$$

By the law of large numbers the approximation becomes exact as T approaches infinity.

Additionally, if $I(l < X < u)$ is an indicator function where the value is 1 when X lies between l and u and 0 otherwise, the probability can be estimated with Monte Carlo integration by taking the sample average for each realization $x^{(t)}$

$$\Pr(l < X < u) \approx \frac{\text{number of realisations } x^{(t)} \in (l, u)}{T}$$

Note that, in general, MCMC methods generate a dependent sample from the joint posterior of interest, since each realisation depends directly on its predecessor.

In R:

```
# Step 1: Generate samples
x <- rbinom(1000, 8, 0.5)
# Represent histogram
hist(x, main = "")
# Step 2: Estimate P(X<=2) as
# Proportion of samples <= 2
sum(x <= 2)/1000
## 0.142
```

Gibbs Sampling

Gibbs Sampling is MCMC algorithm that reduces high dimensional simulation to lower dimensional simulations. Used in BUGS (Bayesian Inference Using Gibbs Sampling) and JAGS as the core algorithm for sampling from posterior distributions.

The algorithm generates a multi-dimensional Markov chain by splitting the vector of random variables θ into subvectors (often scalars) and sampling each subvector in turn, conditional on the most recent values of all other elements of θ . I will not go into great detail on the process because the computer does this for us.

With Gibbs sampling we can give a value to $P(X > .2)$ or $P(Y > 5)$, but not $P(Y > 5 | X = .2)$

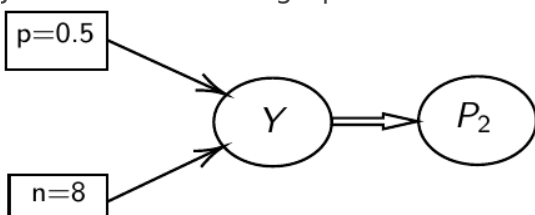
JAGS

JAGS is functionally for the MCMC. It is not a procedural language like R; it can be called from R. You can write complex model code in any order using a very limited syntax

The below code does the same thing as the previous R snippet but using JAGS:

```
library(rjags)
### model is defined as a string
model11.bug <- "model {
Y ~ dbin(0.5, 8)
P2 <- step(2 - Y)
}"
writeLines(model11.bug, "model11.txt")
# Now run the Gibbs sampling for 1000 iterations
### (Step 1) Compile BUGS model
M11 <- jags.model("model11.txt", n.chains = 1, n.adapt = 1000)
### (Step 2) Generate 1000 samples and discard
mcmc_11 <- update(M11, n.iter = 1000)
### (Step 3) Generate 10000 samples and retain for
### inference
test_11 <- coda.samples(M11, variable.names = c("P2", "Y"), n.iter = 10000)
```

JAGS uses directed graphical models for internal representation



- Nodes represent variables
- Directed arrows represent conditional probability distributions
- Double arrows are mathematical functions

The joint probability distribution can be derived using Markov properties of marginal and conditional independence (next section).

Beta Distribution

Beta is a family of probability distributions with support between 0 and 1, it is the typical choice for working with probability parameters. The density function depends on two hyper-parameters α_1 and α_2 .

$\alpha_1 + \alpha_2$ = effective sample size with α_1 events (more on this in a later section)

$$B(\alpha_1, \alpha_2) = \int_0^1 \theta^{\alpha_1-1} (1 - \theta)^{\alpha_2-1} d\theta$$

$$P(\theta|\alpha_1, \alpha_2) = \frac{\theta^{\alpha_1-1} (1 - \theta)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)}$$

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_2}; \quad \alpha_1 + \alpha_2 : \text{Prior precision}$$

$$\begin{aligned} V(\theta) &= \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)} \\ &= \frac{E(\theta)(1 - E(\theta))}{\alpha_1 + \alpha_2 + 1} \end{aligned}$$

Functions `rbeta()`, `dbeta()`, `qbeta()`, `pbeta()` available in R for working with Beta distributions. It's also implemented in JAGS as function `dbeta(,)`

The prior hyper-parameters determine the expected Y events. The posterior mean thus be written as:

$$E(Y) = E(E(Y|\theta)) = E(n\theta) = n \times \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

Revision #15

Created 20 January 2023 19:07:37 by Elkip

Updated 3 March 2023 21:06:17 by Elkip