Lecture 11A
MTH6102: Bayesian Statistical Methods

Eftychia Solea

Queen Mary University of London

2023

Today's agenda

Today's lecture

- Learn how to use the law of total probability to compute posterior predictive probabilities.

Review: Predictive probabilities

- Posterior predictive probability describes how likely are different outcomes of a future experiment.
- We have observed data (result of the experiment) $y \sim p(y \mid \theta)$, dependent on parameters θ.
- Then we update our prior distribution for θ, $p(\theta)$, to the posterior distribution $p(\theta \mid y)$.

Posterior predictive probabilities

- Suppose we plan to perform the experiment again to observe new data x
- We want to compute the posterior predictive distribution $p(x \mid y)$ of x given the observed data y.
- Posterior predictive probabilities are used to predict future data x when the experiment is performed again, and they are computed after obsevring data y and updating prior to posterior.

Predictive distributions: discrete prior, discrete data

- Discrete observed data: $y \sim p(y \mid \theta)$, with θ unknown
- Discrete likelihood: $p(y \mid \theta)$.
- Discrete hypothesis θ with values $\theta_1, \theta_2, \ldots \theta_K$.
- Prior pmf $p(\theta_i)$ of θ, $p(\theta_i) = p(\theta = \theta_i)$, $i = 1, \ldots, K$.
- Posterior pmf $p(\theta_i \mid y) = \frac{p(y|\theta_i)p(\theta_i)}{p(y)}$, $i = 1, \ldots, K$.

Hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(y\|\theta)$	$p(y\|\theta)p(\theta)$	$p(\theta\|y)$
θ_1	$p(\theta_1)$	$p(y\|\theta_1)$	$p(y\|\theta_1)\ p(\theta_1)$	$p(\theta_1\|y)$
θ_2	$p(\theta_2)$	$p(y\|\theta_2)$	$p(y\|\theta_2)\ p(\theta_2)$	$p(\theta_2\|y)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
θ_K	$p(\theta_K)$	$p(y\|\theta_K)$	$p(y\|\theta_K)\ p(\theta_K)$	$p(\theta_K\|y)$
Total	1	NOT SUM TO 1	$p(y)$	1

Predictive distributions: discrete prior, discrete data

- By the Law of total probability,

$$p(y) = \sum_{i=1}^{K} p(y|\theta_i)p(\theta_i),$$

is called the prior predictive probability.

- Prior predictive probabilities. Assign a probability to an outcome of the experiment. They are computed **before we observe any data**.

Predictive distributions: discrete prior, discrete data

- Let x: future data from the same experiment. We assume that x and y are independent given θ.

- By, the **law of total probability**, the posterior predictive probability of x given the observed data y is

$$p(x|y) = \sum_{i=1}^{K} p(x|\theta_i)\, p(\theta_i|y).$$

Board example: Three type of coins

There are three type of coins in the drawer with probabilities 0.5, 0.6 and 0.9 of heads, respectively. Each coin is equally likely

Data: Pick one and toss 5 times. You get 1 head out of 5 tosses.

(a) Compute the posterior probabilities for the type of coin
(b) Compute the posterior predictive distributions of observing heads in a future toss.
(c) Compute the posterior predictive distributions of observing 2 heads in 5 future coin tosses.

Board example: Three type of coins

- Bayesian updating table

Hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(y\|\theta) \sim$ binomial$(5, \theta)$	$p(y\|\theta)p(\theta)$	$p(\theta\|y)$
$\theta_1 = 0.5$	$p(\theta_1) = 1/3$	$p(y = 1\|\theta_1) = 0.15625$	$p(y = 1\|\theta_1)\ p(\theta_1) = 0.0521$	$p(\theta_1\|y = 1) = 0.669$
$\theta_2 = 0.6$	$p(\theta_2) = 1/3$	$p(y = 1\|\theta_2) = 0.0768$	$p(y = 1\|\theta_2)\ p(\theta_2) = 0.0256$	$p(\theta_2\|y = 1) = 0.329$
$\theta_3 = 0.9$	$p(\theta_3) = 1/3$	$p(y = 1\|\theta_3) = 0.00045$	$p(y = 1\|\theta_3)\ p(\theta_3) = 0.00015$	$p(\theta_3\|y = 1) = 0.00193$
Total	1	NOT SUM TO 1	$p(y = 1) = 0.07785$	1

- Prior predictive probability: $p(y = 1) = p(y = 1|\theta_1)p(\theta_1) + p(y = 1|\theta_2)p(\theta_2) + p(y = 1|\theta_3)p(\theta_3) = 0.07785$

Board example: Three type of coins

- Does the order of the 1 head and 4 tails affect the posterior distribution of the coin type?
  (a) Yes
  (b) No.

- Does the order of the 1 head and 4 tails affect the posterior predictive distribution of the next flip?
  (a) Yes
  (b) No.

Board example

- Suppose that y is the number of expensive goods in a shop over 24 days. So $y \sim \text{Poisson}(24\theta)$ where $\theta = 1/2$, $\theta = 1/4$ or $\theta = 1/8$.
- Suppose the prior pmf is

$$p(\theta = 1/2) = p(1/2) = 0.2, \quad p(\theta = 1/4) = p(1/4) = 0.5,$$
$$p(\theta = 1/8) = p(1/8) = 0.3.$$

- We observe $y = 10$ expensive goods were sold in the last 24 days.

1. Compute the posterior pmf for θ.
2. Compute the posterior predictive distribution that $x = 10$ number of goods will be sold in the next 24 days.

Predictive distributions: continuous prior, discrete data

- Continuous parameter θ in the range $[a, b]$.
- Prior: $p(\theta)$, $\theta \in [a, b]$.
- Discrete data, y. Likelihood $p(y|\theta)$.

- By, the **law of total probability**, the prior predictive probability of y is

$$p(\text{data}) = p(y) = \int_a^b p(y|\theta)\, p(\theta)\, d\theta,$$

where the integral is computed over the entire range of θ.

- **Note:** $p(y)$ is a probability mass function, i.e., $p(y) = P(Y = y)$

Predictive distributions: continuous prior, discrete data

- Posterior: $p(\theta|y) = \frac{p(\theta)\times p(y|\theta)}{p(y)}$
- x: future data of the same experiment. We assume that x and y are independent given θ

- By, the **law of total probability**, the posterior predictive probability of x (given y) is

$$p(x|y) = \int_a^b p(x|\theta)\, p(\theta|y)\, d\theta.$$

Predictive distributions: continuous prior, discrete data

Example

We have a coin with unknown probability θ of heads.
Prior: $p(\theta) = 2\theta$, $\theta \in [0, 1]$.

- Find the prior predictive probability of throwing heads on the first toss.
- Suppose the first flip was heads. Find the posterior predictive probabilities of both heads and tails on the second flip.

Example: beta prior/ binomial data

- Data, $k \sim \text{binomial}(n, q)$
- Prior, $q \sim \text{beta}(\alpha, \beta)$.

  - Find the posterior predictive probability to observe success on the next Bernoulli trial.
  - Find the posterior predictive probability to observe new x successes on the next m Bernoulli trials.

Board example

Data: 10 patients have 6 successes. $\theta \sim \text{beta}(5, 5)$

- Find the posterior distribution of θ.
- Find the posterior predictive probability of success with the next patient.

Posterior predictive distribution: continuous prior, continuous data

- Continuous parameter θ in the range $[a, b]$.
- Prior pdf: $p(\theta)$, $\theta \in [a, b]$.
- Continuous data, y. Likelihood $p(y|\theta)$.
- The prior predictive pdf of y is

$$p(y) = \int_a^b p(y|\theta)\, p(\theta)\, d\theta,$$

  where the integral is computed over the entire range of θ.
- **Note:** $p(y)$ is a pdf.

Posterior predictive distribution: continuous prior, continuous data

- Posterior pdf: $p(\theta|y)$
- x: future data of the same experiment.

- The posterior predictive distribution of x is

$$p(x|y) = \int_a^b p(x|y,\theta)\, p(\theta|y)\, d\theta.$$

- As usual, we usually assume x and y are conditionally independent given θ. That is, $p(x|y,\theta) = p(x|\theta)$.
- In this case,

$$p(x|y) = \int_a^b p(x|\theta)\, p(\theta|y)\, d\theta.$$

Posterior predictive distribution

The posterior predictive distribution for x given the observed data y is

$$p(x \mid y) = \int p(x \mid \theta)\, p(\theta \mid y)\, d\theta$$

- This is the probability distribution for unobserved or future data x.
- This distribution includes two types of uncertainty:
  - the uncertainty remaining about θ after we have seen y;
  - the random variation in x.

Board example: Exponential data/Gamma prior

- The time until failure for a type of light bulb is exponentially distributed with parameter $\theta > 0$, where θ is unknown.
- We observe n bulbs, with failure times $t_1, \ldots, t_n$.
- We assume a Gamma(α, β) prior distribution for θ, where $\alpha > 0$ and $\beta > 0$ are known.

1. Determine the predictive posterior distribution for future data x

Finding the posterior predictive distribution

$$p(x \mid y) = \int p(x \mid \theta)\, p(\theta \mid y)\, d\theta$$

- In conjugate examples, one can usually derive $p(x \mid y)$.
- It is generally easier to find the mean and variance of $p(x \mid y)$ than deriving the full distribution.

Conditional mean and variance in general

- Suppose that X and W are general random variables.
- Then

$$E(X) = E(E(X \mid W)) \quad \text{law of iterated expectation}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(E(X \mid W)) + E(\mathrm{Var}(X \mid W)) \quad \text{law of total variance}$$

- In Bayesian inference, we replace W with parameters and X with the new data we would like to predict.

Mean and variance of posterior predictive distribution

- For new data x and parameter(s) θ

$$E(x) = E(E(x \mid \theta))$$

$$Var(x) = Var(E(x \mid \theta)) + E(Var(x \mid \theta))$$

Mean and variance of posterior predictive distribution

- Add conditioning on observed data y, since we want *posterior* predictions

$$E(x \mid y) = E(E(x \mid \theta, y)) \quad \text{law of iterated expectation}$$

$$Var(x \mid y) = Var(E(x \mid \theta, y)) + E(Var(x \mid \theta, y)) \quad \text{law of total variance}$$

- These are the posterior predictive mean and posterior predictive variance of x, respectively.

Example: beta prior, binomial data

- Data, $k \sim \text{binomial}(n, q)$
- Prior, $q \sim \text{beta}(\alpha, \beta)$.
- New data, $x \sim \text{binomial}(m, q)$, m is known.

(1) Find the posterior predictive mean and variance of x

Using simulation (Monte Carlo)

- Suppose we know the posterior distribution $p(\theta \mid y)$, or we have a sample from it.
- Then it is easy to use simulation to generate a sample from the posterior predictive distribution of a new data-point x.
- Because we know the distribution of x for any given value of θ: it's the same as the distribution of the original data y.

Simulating the posterior predictive distribution

- Suppose that we have a sample from the posterior distribution

$$\theta_1, \theta_2, \ldots, \theta_M$$

- We can simulate the posterior predictive distribution $p(x \mid y)$.
- We just generate

$$x_j \text{ from } p(x \mid \theta_j, y) = p(x \mid \theta_j),\ j = 1, 2, \ldots, M$$

- Then

$$x_1, x_2, \ldots, x_M$$

  is a sample from the posterior predictive distribution $p(x \mid y)$.
- (Since

$$(x_1, \theta_1), (x_2, \theta_2), \ldots, (x_M, \theta_M)$$

  is a sample from $p(x, \theta \mid y) = p(\theta \mid y)\, p(x \mid \theta, y)$).

Simulating the posterior predictive distribution

- When do we have a sample from $p(\theta \mid y)$?
- Almost always, because we use MCMC to make inferences about θ.
- Or in simpler conjugate cases, we can directly generate an independent sample from $p(\theta \mid y)$.
- The latter is an example of simple Monte Carlo.

Using the the posterior predictive sample

- Suppose we have generated a sample from the posterior predictive distribution $x_1, x_2, \ldots, x_M$.
- We can summarize the sample for whatever interests us:
  - Posterior predictive mean, median, variance - just summarize sample $x_1, x_2, \ldots, x_M$
  - Prediction intervals, e.g. with 95% probability, x will be in some interval- just take the 0.025 and 0.975 sample quantiles of the sample $x_1, x_2, \ldots, x_M$.
  - Posterior predictive probability that $x = 0$ - just count what proportion of sample are 0.
  - Posterior predictive probability that $x > c$, for some c - count what proportion of sample are $> c$.