Lecture 10B
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 prior and posterior predictive probabilities.

Predicting new data

- In previous lectures, we looked at updating the probability of parameters (hypotheses) based on data.

- We have observed data (result of the experiment) $y \sim p(y \mid \theta)$, dependent on parameters θ.

- Suppose we have found the posterior distribution $p(\theta \mid y)$.

- **Question:** What is the probability distribution of new data x of a future experiment?

Predictive probabilities

- In this lecture, we are going to focus on predictive probabilities.
- Predictive probability means assigning a probability to each possible outcome of a future experiment.
- There are many examples where we want to make probabilistic prediction: weather forecasting,
  "Tomorrow it will rain with probability 60% "
- Other examples: medical treatment outcomes, climate change, sports betting etc

Predictive probabilities

Example: Three types of coins

There are three types of coins

- Type A coins are fair, with probability 0.5 of heads.
- Type B coins have probability 0.6 of heads.
- Type C coins have probability 0.9 of heads.

You have a drawer containing 4 coins: 2 of type A, 1 of type B, and 1 of type C.

You pick a coin at random.

Prior predictive probabilities

Example: Three types of coins

- Prior predictive probabilities. Before taking data, what is the probability that our chosen coin will land heads?
- Let $D_{1,H}$ be the event that the first toss lands heads.
- Let A be the event the chosen coin is of type A. Likewise for B and C. Then,

$$P(A) = 0.5, \quad P(B) = 0.25, \quad P(C) = 0.25.$$

Prior predictive probabilities

Example: Three types of coins

- By the **law of total probability**, the prior predictive probability that the coin lands heads is

$$P(D_{1,H}) = P(D_{1,H} \mid A)P(A) + P(D_{1,H} \mid B)P(B) \\ + P(D_{1,H} \mid C)P(C) = 0.625$$

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

Posterior predictive probabilities

Example: Three types of coins

- Take data: We flip the chosen coin once and it lands heads.
- We now have data, $D_{1,H}$ (first toss lands heads). Given the data $D_{1,H}$, we update the prior probabilities of the hypotheses to posterior probabilities.
- The Bayes updating table is

hypothesis	prior	likelihood	Bayes num.	posterior
H	$P(H)$	$P(D_{1,H}\|H)$	$P(D_{1,H}\|H)P(H)$	$P(H\|D_{1,H})$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
Total	1		$P(D_{1,H}) = 0.625$	1

- $P(D_{1,H}) = P(D_{1,H} \mid A)P(A) + P(D_{1,H} \mid B)P(B) + P(D_{1,H} \mid C)P(C) = 0.625 = P(\text{data})$.

Posterior predictive probabilities

Example: Three types of coins

- Posterior predictive probabilities. Given $D_{1,H}$ has happened (flipped the coin once and got heads), what is the probability that our chosen coin will land heads if flipped second time?

- Let $D_{2,H}$ the event "heads second time".

- We want to compute $P(D_{2,H} \mid D_{1,H})$, called the posterior probability that the next toss lands heads.

Posterior predictive probabilities

Example: Three types of coins

- We assume that $D_{1,H}$ and $D_{2,H}$ are independent **given** the chosen coin.
- By the law of total probability,

$$P(D_{2,H} \mid D_{1,H}) = P(D_{2,H} \mid A)P(A \mid D_{1,H}) + P(D_{2,H} \mid B)P(B \mid D_{1,H}) \\ + P(D_{2,H} \mid C)P(C \mid D_{1,H}) = 0.668$$

- We use the posterior probabilities $P(A \mid D_{1,H})$, $P(B \mid D_{1,H})$ and $P(C \mid_{1,H})$ as weights in place of the prior probabilities, $P(A)$, $P(B)$ and $P(C)$
- The heads on the first toss increases the probability of heads in the second toss.

Posterior predictive probabilities

- Posterior predictive probabilities give a prediction of a future outcome, after collecting data and updating prior to posterior.

- **Remember:**
    - Prior and posterior probabilities are for hypotheses/parameters.
    - Prior predictive and posterior predictive probabilities are for data.
    - Posterior predictive probabilities are used to predict future data when the experiment is performed again.

Predictive distributions: discrete prior, discrete data

- Discrete data: $y \sim p(y \mid \theta)$, with θ unknown
- Discrete likelihood: $p(y \mid \theta)$.
- Discrete hypothesis θ with values θ_1, θ_2, $\dots$ θ_K.
- Prior pmf $p(\theta_i)$ of θ, $p(\theta_i) = p(\theta = \theta_i)$, $i = 1, \dots, K$.
- Posterior pmf $p(\theta_i \mid y) = p(\theta = \theta_i \mid y)$, $i = 1, \dots, K$.

- Let x: future data of the same experiment. We assume that x and y are independent given θ_i.
- By, the **law of total probability**, the posterior predictive probability of x is

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

Board example

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.

Board example

- 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: $p(\theta|y)$
- 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 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 a new outcome x on the next Bernoulli trial.

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 probability 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

$$\theta_1, \theta_2, \dots, \theta_M$$

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

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

- Then

$$x_1, x_2, \dots, x_M$$

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

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

  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 $> c$.