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 θ_1, θ_2, ... θ_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. $p(x|y, \theta) = p(x|\theta)$
- By, the **law of total probability**, the posterior predictive probability of x given the observed data y is

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

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

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.

• new data $x=1$. The posterior predictive probability is

$$p(x=1|y=1) = \sum_{\theta \in \{0.5, 0.6, 0.9\}} p(x=1|\theta)\, p(\theta|y=1)$$

$$= p(x=1|\theta=0.5)\, p(\theta=0.5|y=1) + p(x=1|\theta=0.6)\, p(\theta=0.6|y=1)$$
$$+ p(x=1|\theta=0.9)\, p(\theta=0.9|y=1)$$
$$= 0.5\,(0.669) + 0.6\,(0.329) + 0.9\,(0.002)$$
$$= 0.46634 \approx 0.5$$

• $n=1$, $x=1 \sim \text{binomial}(1,\theta)$
$p(x=1|\theta) = \binom{1}{1} \theta^1 (1-\theta)^{1-1} = \theta$

• new data: $x=2$ heads out of 5 tosses.

$$p(x=2|y=1) = \sum_{\theta \in \{0.5, 0.6, 0.9\}} p(x=2|\theta)\, p(\theta|y=1)$$

$$= p(x=2|\theta=0.5)\, p(\theta=0.5|y=1) + p(x=2|\theta=0.6)\, p(\theta=0.6|y=1)$$
$$+ p(x=2|\theta=0.9)\, p(\theta=0.9|y=1)$$

binomial prob.

$$p(x=2|\theta) = \binom{5}{2} \theta^2 (1-\theta)^3 \qquad \theta \in \{0.5, 0.6, 0.9\}$$
$$= \binom{5}{2} 0.5^2 (0.5)^3 (0.669) + \binom{5}{2} 0.6^2 (0.4)^3 (0.329)$$
$$+ \binom{5}{2} 0.9^2 (0.1)^3 (0.002) = 0.28$$

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.

The likelihood in this case is

$$p(y=10|\theta) = \frac{(24\theta)^{10} e^{-24\theta}}{10!}, \quad \theta \in \{0.5, 0.25, \tfrac{1}{8}\}$$

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

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

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

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

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

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

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

$$p(x|y) = \int p(x|\theta, \cancel{y})\, p(\theta|y)\, d\theta$$

Predictive distributions: continuous prior, discrete data

Example

We have a coin with unknown probability θ of heads, $\theta \in [0,1]$.
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.

Solution

- Let y be the result of the first toss.

$$p(y=1) = \int_0^1 p(y=1|\theta)\, p(\theta)\, d\theta$$

$$= \int_0^1 \theta \cdot (2\theta)\, d\theta = \frac{2}{3}.$$

- Data $y=1$ (first flip was heads). First, we need to compute the posterior pdf, $p(\theta|y=1)$.

By Bayes' Theorem,

$$p(\theta|y=1) = \frac{p(\theta) \times p(y=1|\theta)}{p(y=1)} = \frac{(2\theta)\cdot\theta}{2/3} = \underline{3\theta^2}$$

Let x be the result of the second flip. Then,

$$p(x=1|y=1) = \int_0^1 p(x=1|\theta)\, p(\theta|y=1)\, d\theta$$

$$= \int_0^1 \theta\,(3\theta^2)\, d\theta = \frac{3}{4}$$

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.

Solution

First, the posterior pdf of q given the data x is

$$p(q|x) \sim \text{beta}(\alpha+x,\ \beta+n-x)$$

The posterior, predictive distribution of χ given x is

$$p(\chi|x) = \int_0^1 p(\chi|q)\, p(q|x)\, dq$$

Now,

$$p(\chi|q) = \binom{m}{\chi} q^{\chi} (1-q)^{m-\chi}$$

$$p(q|x) = \frac{q^{\alpha+x-1} (1-q)^{\beta+n-x-1}}{\text{Beta}(\alpha+x,\ \beta+n-x)}$$

Thus,

$$p(\chi|x) = \int_0^1 \binom{m}{\chi} q^{\chi} (1-q)^{m-\chi} \frac{q^{\alpha+x-1} (1-q)^{\beta+n-x-1}}{\text{Beta}(\alpha+x,\ \beta+n-x)}\, dq$$

$$= \binom{m}{\chi} \frac{1}{\text{Beta}(\alpha+x,\ \beta+n-x)} \int_0^1 \underbrace{q^{\chi+\alpha+x-1} (1-q)^{m-\chi+\beta+n-x-1}}\, dq$$

$X \sim beta(\alpha, \beta)$ the pdf is

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{Beta(\alpha,\beta)} \quad x \in [0,1]$$
$$\alpha > 0, \beta > 0$$

$$\int_0^1 f_X(x)\,dx = 1$$

So $$\int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{Beta(\alpha,\beta)}\,dx = 1$$

$$\Leftrightarrow \frac{1}{Beta(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = 1$$

$$\Leftrightarrow \boxed{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = Beta(\alpha,\beta)} \quad (*)$$

Use $\circledast$ with $a+x+\mathsf{x}$ instead of a and $\beta+m-x+n-\mathsf{x}$ instead of β, to find

$$p(x|\mathsf{x}) = \binom{m}{x}\frac{\mathrm{Beta}(x+a+\mathsf{x},\, \beta+m-x+n-\mathsf{x})}{\mathrm{Beta}(a+\mathsf{x},\, \beta+n-\mathsf{x})}.$$

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

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

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

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

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 $t_1, \ldots, t_n$. $t_i \sim \exp(\theta)$
- 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

Solution

Observed data $t = (t_1, \ldots, t_n)$, where each $t_i \sim \text{Exp}(\theta)$, $\theta > 0$.

Since $\text{Gamma}(a, \beta)$ is conjugate to the exponential likelihood, the posterior of θ given the data t, is

$$p(\theta \mid t) \sim \text{Gamma}(a+n, \beta + S), \text{ where } S = \sum_{i=1}^{n} t_i$$

$\tilde{a} = a + n$

$\tilde{\beta} = \beta + S$

Future data $x \sim \text{Exp}(\theta)$

The posterior predictive distribution of x given t is

$$p(x \mid t) = \int_0^\infty p(x \mid \theta)\, p(\theta \mid t)\, d\theta$$

$$= \int_0^\infty \theta e^{-\theta x} \frac{(\tilde{\beta})^{\tilde{a}} \theta^{\tilde{a}-1} e^{-\tilde{\beta}\theta}}{\Gamma(\tilde{a})}\, d\theta$$

$$= \frac{(\tilde{\beta})^{\tilde{a}}}{\Gamma(\tilde{a})} \int_0^\infty \underbrace{\theta^{(\tilde{a}+1)-1} \exp\left(-(x+\tilde{\beta})\theta\right)}\, d\theta$$

$X \sim \text{Gamma}(a, \beta)$ then

$$\int_0^\infty f_X(x)\, dx = 1 \iff \int_0^\infty \frac{\beta^a}{\Gamma(a)} x^{a-1} e^{-\beta x}\, dx = 1$$

$$\iff \frac{\beta^a}{\Gamma(a)} \int_0^\infty x^{a-1} e^{-\beta x}\, dx = 1 \iff$$

$$\Leftarrow \boxed{\int_0^\infty x^{a-1} e^{-\beta x}\,dx = \frac{\Gamma(a)}{\beta^a} \quad (*)}$$

Use (*) but substitute in $\tilde{a}+1$ instead of a and $x+\tilde{\beta}$ instead of β to get

$$p(x|t) = \frac{(\tilde{\beta})^{\tilde{a}}}{\Gamma(\tilde{a})} \frac{\Gamma(\tilde{a}+1)}{(x+\tilde{\beta})^{\tilde{a}+1}}$$

$$\Gamma(x+1) = x\,\Gamma(x)$$

$$p(x|t) = \frac{(\tilde{\beta})^{\tilde{a}}\,\tilde{a}\,\cancel{\Gamma(\tilde{a})}}{\cancel{\Gamma(\tilde{a})}\,(x+\tilde{\beta})^{\tilde{a}+1}}$$

$$= \frac{(\tilde{\beta})^{\tilde{a}}\,\tilde{a}}{(x+\tilde{\beta})^{\tilde{a}+1}}$$

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.

$\mathbb{E}(X|W) = g(W)$ $\Rightarrow$ random variable

if $W = w$, $w \in \mathbb{R}$

$\mathbb{E}(X|W=w) = g(w)$ non-random

- $\mathbb{E}(g(W)) = \mathbb{E}(\mathbb{E}(X|W)) = \mathbb{E}(X)$

Mean and variance of posterior predictive distribution

- For new data x and parameter(s) θ

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

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

$$\mathbb{E}(x \mid y) = \mathbb{E}(\mathbb{E}(x \mid \theta, y))$$

Mean and variance of posterior predictive distribution

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

$$E(x \mid y) = E(E(x \mid \theta, y))$$ law of iterated expectation

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

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

Because X and y are independent given θ

$$f_{X|Y,\theta}(x|\cancel{y},\theta) = f_{X|\theta}(x|\theta)$$

$$\mathbb{E}(x|\cancel{y},\theta) = \mathbb{E}(x|\theta)$$

$$\mathrm{var}(x|\cancel{y},\theta) = \mathrm{var}(x|\theta)$$

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

Solution

By the law of iterated expectation,

$$\mathbb{E}(x|x) = \mathbb{E}\left(\mathbb{E}(x|q,\cancel{x})\right)$$

$$= \mathbb{E}\left(\mathbb{E}(x|q)\right)$$

Now given q, $x \sim$ binomial(m, q)

So $\mathbb{E}(x|q) = m \cdot q$

$$\underbrace{\mathbb{E}(x|x)}_{\text{predictive mean}} = \mathbb{E}(m \cdot q) = m\,\underbrace{\mathbb{E}(q)}_{\text{prior expectation}} = m \cdot \underbrace{\frac{a}{a+b}}_{\mu} = m \cdot \mu$$

$$\operatorname{Var}(x|x) = \mathbb{E}\left(\operatorname{var}(x|x,q)\right) + \operatorname{var}\left(\mathbb{E}(x|x,q)\right)$$

$$\mathbb{E}\left(\operatorname{var}(x|\cancel{x},q)\right) = \mathbb{E}\left(\operatorname{var}(x|q)\right)$$

$$\operatorname{Var}(x|q) = mq(1-q)$$

So $\mathbb{E}\left(\operatorname{var}(x|q)\right) = \mathbb{E}\left(mq(1-q)\right) = m\left[\mathbb{E}(q - q^2)\right]$

$$= m\left[\mathbb{E}(q) - \mathbb{E}(q^2)\right]$$

$$\operatorname{Var}(q) = \mathbb{E}(q^2) - [\mathbb{E}(q)]^2 \Rightarrow \boxed{\mathbb{E}(q^2) = \operatorname{Var}(q) + [\mathbb{E}(q)]^2}$$

We have

$$\mathbb{E}(\mathrm{var}(X \mid x, \varepsilon)) = m\,\mathbb{E}(\varepsilon) - m\left[\mathrm{var}(\varepsilon) + \mathbb{E}(\varepsilon)^2\right]$$

$$= m\mu - m(v + \mu^2) \qquad (1)$$

where v is the prior variance, μ is the prior mean.

- $\mathrm{Var}(\mathbb{E}(x \mid \cancel{x}, \varepsilon)) = \mathrm{Var}(\mathbb{E}(x \mid \varepsilon)) = \mathrm{Var}(m\varepsilon)$

$$= m^2 \mathrm{Var}(\varepsilon)$$

$$= m^2 v \qquad (2)$$

where $v = \mathrm{Var}(\varepsilon)$

-

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 $> c$.