50.034 - Introduction to Probability and Statistics

Week 9 – Cohort Class

January–May Term, 2019

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

CC Week 9 Assigned Readings: (Revise: 7.3, 7.4, 7.5, 7.6)

Outline of Cohort Class

- Recap: Parameter space and parameters as R.V.'s
- Recap: Prior/posterior distributions
- **Mini-quiz 3**

Exercises on the following topics:

- Conjugate priors

Recall: Parameter Space

The parameters of a distribution are numerical attributes whose values determine the distribution completely.

- We have already seen many examples of parameters:
  - Binomial distribution with parameters n and p.
  - Poission distribution with parameter λ.
  - Normal distribution with parameters μ and σ.
  - Bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$.
- Each parameter could be treated as a known constant, an unknown constant, a R.V. whose distribution is known, a R.V. whose distribution is unknown, etc.

Given any parameter θ, the set of all possible values for θ is called the parameter space of θ.

- What is considered "possible" depends on the context.
- If μ is the mean of a normal distribution representing the average height (in cm) of a person, then we could take the parameter space of μ to be the interval $[0, 300]$.

Parameters as random variables

Light Bulb Example: Let $X_1, X_2, \ldots$ be a sequence of iid R.V.'s, each having an exponential distribution with parameter λ. Each X_i represents the lifespan (in hours) of the i-th light bulb.

If we treat λ as a R.V., then the parameter space of λ is the set of all positive real numbers.

- If we assume that the lifespan of every light bulb must be > 1 hour, then since $\frac{1}{\lambda}$ represents the average lifespan, we could restrict the parameter space to just the interval $(0, 1)$.

Question: A lighting expert inspects the light bulb model and advises you that the lifespan of every light bulb must be strictly between 200 hours and 600 hours. Assuming that we believe this lighting expert, what should we restrict the parameter space to?

Answer: The open interval $(\frac{1}{600}, \frac{1}{200})$.

Recall: Prior and posterior distributions

Consider a statistical model with observable R.V.'s $X_1, \dots, X_n$. Let θ be a parameter (possibly one of many parameters) of the joint distribution of $X_1, \dots, X_n$, and treat θ as a random variable.

The prior distribution of θ is the initial distribution specified for θ.

- This is the distribution we specify before observing any data (i.e. before gathering the observed values for $X_1, \dots, X_n$)
- "prior distribution" can simply be called "prior".

After we have some observed values, say $X_1 = x_1, \dots, X_n = x_n$, then the conditional distribution, consisting of all conditional probabilities of the form $\Pr(\theta \in C | X_1 = x_1, \dots, X_n = x_n)$ (over all possible $C \subseteq \mathbb{R}$), is called the posterior distribution of θ.

- "posterior distribution" can simply be called "posterior".

Interpretation: The prior of θ is the initial guess for the distribution of θ, while the posterior of θ is the updated guess, after taking into account the observed values $X_1 = x_1, \dots, X_n = x_n$.

Prior pmf/pdf versus Posterior pmf/pdf

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is treated as a random variable.

- If θ is discrete, then the pmf of θ is called the prior pmf of θ.
- If θ is continuous, then the pdf of θ is called the prior pdf of θ.
- In either case, the pmf/pdf of θ is usually written as $\xi(\theta)$.

Next, suppose we have observed the values $X_1 = x_1, \ldots, X_n = x_n$.

- If θ is discrete, then the posterior pmf of θ is the conditional pmf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- If θ is continuous, then the posterior pdf of θ is the conditional pdf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- In either case, the pmf/pdf is denoted by $\xi(\theta|x_1, \ldots, x_n)$, or more simply, $\xi(\theta|\mathbf{x})$, where $\mathbf{x}$ represents $(x_1, \ldots, x_n)$.

$$\boxed{\begin{array}{c}\text{posterior}\\ \text{distribution}\end{array}} = \boxed{\begin{array}{c}\text{conditional distribution of the}\\ \text{parameter given the data/evidence}\end{array}}$$

Exercise 1 (5 mins)

Pile of parcels: There is a pile of 1000 equally sized parcels in front of you that all look the same to you. Some boxes are heavy ($\geq 1kg$), and some boxes are light ($< 1kg$), and you cannot tell the weight just by looking. You want to determine the proportion of boxes that are heavy, and you decide to test the weights of 10 randomly selected boxes to get more information.

For your statistical model, you assumed that $X_1, \ldots, X_{10}$ are iid Bernoulli R.V.'s with parameter θ, where $X_i = 1$ if the *i*-th box is heavy, and $X_i = 0$ otherwise. Suppose you assumed for simplicity that θ is a discrete R.V. with two possible values.

- Your initial guess is that θ is either 0.1 with probability 0.2, or 0.9 with probability 0.8.

Questions:

- What is the parameter space of θ?
- What is the prior pmf of θ?

Exercise 1 - Solution

Your initial guess is that θ is either 0.1 with probability 0.2, or 0.9 with probability 0.8.

- This means that you have two guesses for the value of θ.
- You think the "true" value for θ is either 0.1 or 0.9.

Thus, the parameter space of θ is $\{0.1, 0.9\}$.

For your two guesses, how likely do you think each of them is?

- You think that the probability that 0.1 is the "true" value for θ is 0.2.
- You think that the probability that 0.9 is the "true" value for θ is 0.8.

Therefore, the prior pmf of θ is

$$\xi(\theta) = \begin{cases} 0.2, & \text{if } \theta = 0.1; \\ 0.8, & \text{if } \theta = 0.9; \\ 0, & \text{otherwise.} \end{cases}$$

Calculating posterior distributions using Bayes' theorem

Consider a statistical model with observable R.V.'s $X_1, \dots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \dots, X_n$, where θ is a R.V. with parameter space Ω.

Theorem: (**Bayes' theorem**+**Law of total probability** for R.V.'s)

- If $X_1, \dots, X_n$ are **discrete** with joint conditional pmf $p_n(\mathbf{x}|\theta)$ and marginal joint pmf $p(\mathbf{x})$, and if θ is **discrete** with prior pmf $\xi(\theta)$, then the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{p(\mathbf{x})} = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum\limits_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} \quad (\text{for } \theta \in \Omega).$$

- If $X_1, \dots, X_n$ are **discrete** with joint conditional pmf $p_n(\mathbf{x}|\theta)$ and marginal joint pmf $p(\mathbf{x})$, and if θ is **continuous** with prior pdf $\xi(\theta)$, then the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{p(\mathbf{x})} = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\int_{\Omega} p_n(\mathbf{x}|\theta')\xi(\theta')\, d\theta'} \quad (\text{for } \theta \in \Omega).$$

Calculating posterior distributions (continued)

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is a R.V. with parameter space Ω.

Theorem: (**Bayes' theorem**+**Law of total probability** for R.V.'s)

- If $X_1, \ldots, X_n$ are **continuous** with joint conditional pdf $f_n(\mathbf{x}|\theta)$ and marginal joint pdf $f(\mathbf{x})$, and if θ is **discrete** with prior pmf $\xi(\theta)$, then the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{f(\mathbf{x})} = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\displaystyle\sum_{\theta' \in \Omega} f_n(\mathbf{x}|\theta')\xi(\theta')} \quad (\text{for } \theta \in \Omega).$$

- If $X_1, \ldots, X_n$ are **continuous** with joint conditional pdf $f_n(\mathbf{x}|\theta)$ and marginal joint pdf $f(\mathbf{x})$, and if θ is **continuous** with prior pdf $\xi(\theta)$, then the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{f(\mathbf{x})} = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\displaystyle\int_{\Omega} f_n(\mathbf{x}|\theta')\xi(\theta')\,d\theta'} \quad (\text{for } \theta \in \Omega).$$

Exercise 2 (15 mins)

Pile of parcels (continued):
For your statistical model, you assumed that $X_1, \dots, X_{10}$ are iid Bernoulli R.V.'s with parameter θ, where $X_i = 1$ if the *i*-th box is heavy, and $X_i = 0$ otherwise. You assumed that θ is a discrete R.V. with two possible values.

- Your initial guess is that θ is either 0.1 with probability 0.2, or 0.9 with probability 0.8.
- You have already calculated the prior pmf of θ in Exercise 1.

Suppose that the observed values are

$$X_1 = 1, X_2 = 1, X_3 = 1, X_4 = \mathbf{0}, X_5 = 1,$$
$$X_6 = 1, X_7 = 1, X_8 = 1, X_9 = 1, X_{10} = 1$$

- Let **x** be the vector of observed values $(1, 1, 1, 0, 1, 1, 1, 1, 1, 1)$.

Question: What is the posterior pmf of θ given the vector of observed values **x**?

Exercise 2 - Solution

From Example 1: The parameter space of θ is $\Omega = \{0.1, 0.9\}$.

Since each X_i is Bernoulli, the conditional pmf of X_i given $\theta = \theta$ is

$$p(x_i|\theta) = \begin{cases} \theta^{x_i}(1-\theta)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

Thus,

$$p(x_i|0.1) = \begin{cases} (0.1)^{x_i}(1-0.1)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

$$p(x_i|0.9) = \begin{cases} (0.9)^{x_i}(1-0.9)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

Exercise 2 - Solution

Using Bayes' theorem (for R.V.'s) and the law of total probability (for R.V.'s), the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum_{\theta' \in \Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} \quad \text{(for } \theta \in \Omega\text{)},$$

where $p_n(\mathbf{x}|\theta)$ is the joint conditional pmf of $X_1, \ldots, X_{10}$.

Since there are two values in Ω, the denominator equals

$$p_n(\mathbf{x}|0.1)\xi(0.1) + p_n(\mathbf{x}|0.9)\xi(0.9),$$

which represents the probability that event $\{(X_1, \ldots, X_n) = \mathbf{x}\}$ occurs, taking into account both possible values in Ω and their corresponding probabilities.

Exercise 2 - Solution (continued)

From 2 slides ago: The conditional pmf of X_i given $\theta = \theta$ is

$$p(x_i|\theta) = \begin{cases} \theta^{x_i}(1-\theta)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

The joint conditional pmf $p_n(\mathbf{x}|\theta)$ is given by:

$$p_n(\mathbf{x}|\theta) = p(x_1|\theta)\cdots p(x_{10}|\theta)$$
$$= \begin{cases} \theta^{(x_1+\cdots+x_{10})}(1-\theta)^{10-(x_1+\cdots+x_{10})}, & \text{if every } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise;} \end{cases}$$

Important Note: $X_1, \ldots, X_{10}$ are iid, so they are **independent**, which means that the joint conditional pmf equals the product of the marginal conditional pmf's.

Exercise 2 - Solution (continued)

We are given that $\mathbf{x} = (1, 1, 1, 0, 1, 1, 1, 1, 1, 1)$, hence

$$p_n(\mathbf{x}|\theta) = p(1, 1, 1, 0, 1, 1, 1, 1, 1, 1|\theta) = \theta^9(1-\theta).$$

Since the possible values for θ are 0.1 and 0.9, we then get that our denominator equals

$$\begin{aligned} & p_n(\mathbf{x}|0.1)\xi(0.1) + p_n(\mathbf{x}|0.9)\xi(0.9) \\ = & (0.1)^9(1-0.1)(0.2) + (0.9)^9(1-0.9)(0.8) \\ = & (0.1)^9(0.9)(0.2) + (0.9)^9(0.1)(0.8). \end{aligned}$$

Therefore, the posterior pmf of θ given $\mathbf{x}$, which is denoted by $\xi(\theta|\mathbf{x})$, is given as follows:

$$\frac{p_n(\mathbf{x}|\theta)\xi(\theta)}{\sum\limits_{\theta'\in\Omega} p_n(\mathbf{x}|\theta')\xi(\theta')} = \begin{cases} \frac{(0.1)^9(0.9)(0.2)}{(0.1)^9(0.9)(0.2)+(0.9)^9(0.1)(0.8)}, & \text{if } \theta = 0.1; \\ \frac{(0.9)^9(0.1)(0.8)}{(0.1)^9(0.9)(0.2)+(0.9)^9(0.1)(0.8)}, & \text{if } \theta = 0.9; \\ 0, & \text{otherwise;} \end{cases}$$

Exercise 2 - Solution (continued)

After simplifying, we get that the posterior pmf of θ is

$$\xi(\theta|\mathbf{x}) \approx \begin{cases} 0.0000000058, & \text{if } \theta = 0.1; \\ 0.9999999942, & \text{if } \theta = 0.9; \\ 0, & \text{otherwise.} \end{cases}$$

Interpretation: Based on experimental evidence, our initial guess is a "very good" guess, and the updated prior distribution reflects our "very good" guess:

- Originally, we guessed that "$\theta = 0.9$" with 80% probability.
- With our experimental evidence, our updated guess becomes "$\theta = 0.9$" with 99.99999942% probability.

Mini-quiz 3 (15 mins)

Only writing materials are allowed. No calculators, notes, books, or cheat sheets are allowed. Don't worry, you won't need calculators.

If you are not present in class at the start of the quiz, you will not be given additional time to finish the quiz.

Remarks:

- There are no make-up mini-quizzes! If you arrive in class after the mini-quiz ends, or do not attend that cohort class, you will not have a chance to take the mini-quiz.
- To take into account unforeseen circumstances (e.g. mini-quiz missed due to illness), only the **best 3 of 4** mini-quiz scores will be counted towards your final grade.

Conjugate priors

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

- If $X_1, \ldots, X_n$ are Bernoulli, and if we **start with any beta distribution** as our prior distribution for θ, then our posterior distribution will **remain as a beta distribution**, no matter how many observations we make, and no matter what the observed values are.
- Hence, we say that the family of beta distributions is a conjugate family of prior distributions, or more simply, a family of conjugate priors.

Key Idea: Let Ψ be a family of conjugate priors. If we choose a prior from Ψ, then the posterior will also be in Ψ.

- Remember, your initial guess (i.e. prior distribution) can be ANY distribution. If you choose a complicated prior, then the posterior would be complicated and very tedious to calculate.
- However, if you choose a prior from a family of conjugate priors, then the posterior becomes very easy to calculate!

Conjugate family of prior distributions

Some families of conjugate priors:

Sampling from	**Family of conjugate priors**
Bernoulli distribution	beta distributions
binomial distribution*	beta distributions
geometric distribution	beta distributions
Poisson distribution	gamma distributions
exponential distribution	gamma distributions
normal distribution	normal distributions

*Note: For a fixed known number of trials.

Sampling from various distributions

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (**Sampling from Bernoulli distributions**)
If $X_1, \ldots, X_n$ are Bernoulli, and if θ has the beta prior distribution with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the beta distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n - (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the beta posterior are
  $\alpha' = \alpha + (\text{number of successes}), \quad \beta' = \beta + (\text{number of failures}).$

Theorem: (**Sampling from binomial distributions**)
Let $N \geq 1$ be a known fixed integer. If $X_1, \ldots, X_n$ are binomial with parameters N and θ, and if θ has the beta prior with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the beta distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + Nn - (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the beta posterior are
  $$\alpha' = \alpha + \binom{\text{total number of}}{\text{successes}}, \qquad \beta' = \beta + \binom{\text{total number of}}{\text{failures}}.$$

Sampling from various distributions (continued)

Consider a statistical model where $X_1, \dots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (Sampling from Poisson distributions)
If $X_1, \dots, X_n$ are Poisson, and if θ has the gamma prior distribution with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$ is the gamma distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n$.

- i.e. the new parameters of the gamma posterior are

$$\alpha' = \alpha + \binom{\text{number of new}}{\text{occurrences}}, \qquad \beta' = \beta + \binom{\text{number of time}}{\text{periods}}.$$

Theorem: (Sampling from exponential distributions)
If $X_1, \dots, X_n$ are exponential, and if θ has the gamma prior with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$ is the gamma distribution with parameters $\alpha + n$ and $\beta + (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the gamma posterior are

$$\alpha' = \alpha + \binom{\text{number of}}{\text{experiments}}, \qquad \beta' = \beta + \binom{\text{total time}}{\text{elapsed}}.$$

Sampling from various distributions (continued)

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (**Sampling from normal distributions**)
Let $\sigma > 0$ be a fixed known real number. If $X_1, \ldots, X_n$ are normal with mean θ and variance σ^2, and if θ has the normal prior distribution with mean μ_0 and variance v_0^2, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the normal distribution with mean μ_1 and variance v_1^2 given as follows:

$$\mu_1 = \frac{\sigma^2\mu_0 + v_0^2(x_1 + \cdots + x_n)}{\sigma^2 + nv_0^2},$$

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + nv_0^2}.$$

Exercise 3 (15 mins)

Let X_1, X_2, X_3 be observable R.V.'s that are conditionally iid given the parameter θ, and let $\mathbf{x}$ be the vector of observed values for (X_1, X_2, X_3).

1. If each X_i is Bernoulli, $\mathbf{x} = (1, 1, 0)$, and the prior of θ is beta with parameters 3 and 8, then what is the posterior distribution of θ given $\mathbf{x}$?
2. If each X_i is Poisson, $\mathbf{x} = (3, 0, 2)$, and the prior of θ is gamma with parameters 2 and 2, then what is the posterior distribution of θ given $\mathbf{x}$?
3. If each X_i is exponential, $\mathbf{x} = (1.1, 1.2, 1.3)$, and the prior of θ is gamma with parameters 3 and 3, then what is the posterior distribution of θ given $\mathbf{x}$?
4. If each X_i is normal with mean θ and variance 25, the prior of θ is normal with mean 4 and variance 8, and $\mathbf{x} = (2.4, -1, 4.6)$, then what is the posterior distribution of θ given $\mathbf{x}$?

Exercise 3 - Solution

1. We are given the following information:
   - Prior distribution is beta with parameters 3 and 8.
   - $X_1 = 1$, $X_2 = 1$, $X_3 = 0$ (i.e. 2 successes, 1 failures)

   Thus, the posterior distribution is the **beta distribution** with parameters $3 + 2 = \mathbf{5}$ and $8 + 1 = \mathbf{9}$.
2. We are given the following information:
   - Prior distribution is gamma with parameters 2 and 2.
   - $X_1 = 3$, $X_2 = 0$, $X_3 = 2$ (i.e. 5 new occurrences, 3 time periods)

   Thus, the posterior distribution is the **gamma distribution** with parameters $2 + 5 = \mathbf{7}$ and $2 + 3 = \mathbf{5}$.
3. We are given the following information:
   - Prior distribution is gamma with parameters 3 and 3.
   - $X_1 = 1.1$, $X_2 = 1.2$, $X_3 = 1.3$ (i.e. 3 pairs, total time elapsed $= 3.6$)

   Thus, the posterior distribution is the **gamma distribution** with parameters $3 + 3 = \mathbf{6}$ and $3 + 3.6 = \mathbf{6.6}$.

Exercise 3 - Solution (continued)

4. We are given the following information:
   - Each X_i has mean θ and known variance $\sigma^2 = 25$.
   - Prior distribution is normal with mean $\mu_0 = 4$ and variance $v_0^2 = 8$.
   - $X_1 = 2.4$, $X_2 = -1$, $X_3 = 4.6$ (the sum of observed values is 6)

Thus, the posterior distribution is the **normal distribution** with mean μ_1 and variance v_1^2 given by:

$$\mu_1 = \frac{\sigma^2 \mu_0 + v_0^2 (x_1 + \cdots + x_n)}{\sigma^2 + n v_0^2} = \frac{(25)(4) + 8(6)}{25 + 3(8)} = \frac{148}{49} \approx 3.020,$$

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2} = \frac{(25)(8)}{25 + 3(8)} = \frac{200}{49} \approx 4.082.$$

Mid-term Exam Statistics

- Mean = 58.78
- Standard deviation = 16.93
- Median = 60
- Max = 90
- Min = 17

Breakdown by questions:

	Qn 1	**Qn 2**	**Qn 3**	**Qn 4**	**Qn 5**
Mean	15.12	15.70	12.50	12.73	2.78
Median	17	16	13	15	2
s.d.	4.62	3.61	5.62	5.28	3.00
Max	20	20	20	20	20
Min	4	6	1	0	0

Summary

- Recap: Parameter space and parameters as R.V.'s
- Recap: Prior/posterior distributions
- **Mini-quiz 3**

Exercises on the following topics:

- Conjugate priors

Note: Prof Gemma will be teaching from next week onwards.