50.034 - Introduction to Probability and Statistics

Week 9 – Lecture 15

January–May Term, 2019

Lecture 15 Assigned Readings: 5.8, 7.3, 7.4

Outline of Lecture

- Gamma function and beta function
- Beta distribution
- Conjugate priors, conjugate family of prior distributions
- Gamma distribution
- Estimator and estimate
- Loss function
- Bayes estimator/estimate

Gamma function and beta function

Motivation: There are some integrals that keep on appearing in probability and statistics, so we give them names.

The gamma function is the function $\Gamma(z)$ defined by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt \quad (\text{for } z > 0).$$

- "Γ" is the capitalization of the greek alphabet "γ" ("gamma").
- **Fact:** If $z > 1$, then $\Gamma(z) = (z-1)\Gamma(z-1)$.
- **Special Case:** $\Gamma(n) = (n-1)!$ for all positive integers n.
  - i.e. $\Gamma(z)$ is a generalization of the factorial function!

The beta function is the bivariate function $B(x, y)$ defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt \quad (\text{for } x, y > 0).$$

- "B" is the capitalization of the greek alphabet "β" ("beta").
- **Fact:** $B(x, y)$ is finite for all $x, y > 0$. (The proof is not too easy..)

Formula: $B(x, y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ for all $x, y > 0$.

Beta distribution

A continuous R.V. X is called beta if its pdf is given by:

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 \leq x \leq 1; \\ 0, & \text{otherwise;} \end{cases}$$

for some real numbers $\alpha, \beta > 0$.

- We say that X is the beta R.V. with parameters α and β.
- Its distribution is called beta distribution.
- $\mathbf{E}[X] = \frac{\alpha}{\alpha+\beta}$ and $\operatorname{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- **Special Case:** The beta distribution with parameters $\alpha = 1$ and $\beta = 1$ is precisely the uniform distribution on $[0, 1]$.

Common Use: To model the uncertainty about the probability of success of a Bernoulli process (usually with many Bernoulli trials).

- e.g. the likelihood that a candidate will win an election.
- e.g. the likelihood that a movie achieves or exceeds a certain approval rating.

Example 1

A new movie titled "Predator vs Prey" has been released. Based on the data of sneak preview approval ratings collected, the overall worldwide approval rating of the movie could be modeled as a beta R.V. X with parameters $\alpha = 42$ and $\beta = 60$.

- **Note:** We will later interpret and understand the meaning of the parameters α and β of a beta distribution (see Slide 10).

Question: Find the probability that the overall worldwide approval rating of this movie is at least 50%.

Example 1 (continued)

Solution: We are given that X is a beta R.V. with parameters $\alpha = 42$ and $\beta = 60$.

Using the formula $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ for the beta function, we get

$$B(42, 60) = \frac{\Gamma(42)\Gamma(60)}{\Gamma(102)} = \frac{(41!)(59!)}{101!},$$

thus the pdf of X is

$$f(x) = \begin{cases} \frac{1}{B(42,60)} x^{42-1}(1-x)^{60-1}, & \text{if } 0 \leq x \leq 1; \\ 0, & \text{otherwise;} \end{cases}$$

$$= \begin{cases} \frac{101!}{(41!)(59!)} x^{41}(1-x)^{59}, & \text{if } 0 \leq x \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

Example 1 (continued)

Therefore,

$$\begin{aligned}\Pr(X \geq 0.5) &= 1 - \Pr(X < 0.5) = 1 - \int_{-\infty}^{0.5} f(x)\,dx \\ &= 1 - \int_{0}^{0.5} \frac{101!}{(41!)(59!)} x^{41}(1-x)^{59}\,dx \\ &\approx 0.03638.\end{aligned}$$

In other words, the probability that the eventual worldwide approval rating of this movie is at least 50% is approximately 0.03638.

Remarks on calculation: In general, calculating values for the cdf of a beta R.V. requires numerical approximation, e.g. via the use of a graphing calculator or some computing software.

- This is because in general, there is no "nice" formula for the cdf of a beta R.V.

Recall: Prior and posterior distributions (Lecture 13)

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Let θ be a parameter (possibly one of many parameters) of the joint distribution of $X_1, \ldots, X_n$, 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 $X_1, \ldots, X_n$)
- "prior distribution" can simply be called "prior".

After we have some observed values, say $X_1 = x_1, \ldots, X_n = x_n$, then the conditional distribution, consisting of all conditional probabilities of the form $\Pr(\theta \in C | X_1 = x_1, \ldots, X_n = x_n)$ (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 $X_1 = x_1, \ldots, X_n = x_n$.

Technicality: Conditionally iid R.V.'s vs iid R.V.'s

Consider a statistical model with observable **iid** R.V.'s $X_1, \ldots, X_n$, each with parameter θ.

- If we treat θ as a R.V., then each X_i is conditional on the given value of θ.
- Hence, when we say that $X_1, \ldots, X_n$ are iid, we technically mean that $X_1, \ldots, X_n$ are iid, given the value of θ.
- So to be more accurate, we should actually say that $X_1, \ldots, X_n$ are conditionally iid given θ.

However, "conditional on the given value of θ" would already be implicitly assumed if we treat θ as a R.V.

- Thus, as long as it is understood that θ is treated as a R.V. (e.g. by the mention of a prior or posterior distribution of θ), we could simply say "$X_1, \ldots, X_n$ are iid R.V.'s" to mean exactly the same as "$X_1, \ldots, X_n$ are conditionally iid given θ".
- Conversely, when we say that "$X_1, \ldots, X_n$ are conditionally iid R.V.'s, given the parameter θ", it should be understood that θ is treated as a R.V.

Beta distribution as a prior distribution

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

Theorem: If the prior distribution of θ is the beta 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 $\beta + n - (x_1 + \cdots + x_n)$.

Interpretation:

- θ is a parameter of our Bernoulli process, treated as a R.V.
- We started with the initial guess that θ follows the beta distribution with parameters α and β.
- Given observed values $X_1 = x_1, \ldots, X_n = x_n$, there are a total of $(x_1 + \cdots + x_n)$ successes and $n - (x_1 + \cdots + x_n)$ failures.
- We update our guess for the distribution of θ to a new beta distribution, with parameters

  $\alpha' = \alpha + (\text{number of successes})$, $\beta' = \beta + (\text{number of failures})$.

Starting from scratch

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

- To complete our specification of the statistical model, we have to decide on a prior distribution for θ.
- **Question:** What if we have absolutely no information that could help us choose a prior distribution for θ?

Note: Since θ is the parameter of a Bernoulli R.V., we know that $0 \leq \theta \leq 1$ by definition.

- Without additional information, we have no reason to choose any specific value over any other specific value in $[0, 1]$.
- Hence, we could "start from scratch" and set the prior distribution to be the uniform distribution on $[0, 1]$.
- Such a distribution is precisely the beta distribution with parameters $\alpha = 1$ and $\beta = 1$.

The data-driven approach

To update our distribution of θ, we could (and should) gather experimental evidence.
Suppose we are given the observed values $X_1 = x_1, \ldots, X_n = x_n$.
Then by our previous theorem, we can update our posterior distribution to the beta distribution with parameters

$$\alpha' = 1 + (x_1 + \cdots + x_n), \quad \beta' = 1 + n - (x_1 + \cdots + x_n).$$

- Remember: We started with $\alpha = 1$, $\beta = 1$ (uniform distribution).

Example: To get a good initial idea of the approval rating of "Predator vs Prey", we could hold a sneak preview and get data.

- Perhaps the approval rating was 41% among 100 people (i.e. 41 liked the movie and 59 dislike the movie).
- Then we could update our posterior distribution to the beta distribution with parameters 42 and 60.

Possible interpretation for Example 1: The parameters $\alpha = 42$ and $\beta = 60$ were chosen, perhaps because initial data from 100 people consists of 41 likes and 59 dislikes.

Sequential observations

In many experiments, observations are obtained sequentially.

Two possible approaches:

- Update the posterior distribution sequentially.
  - $\xi(\theta|x_1)$ becomes prior pmf/pdf before X_2 is observed.
  - $\xi(\theta|x_1, x_2)$ becomes prior pmf/pdf before X_3 is observed.
  - $\xi(\theta|x_1, x_2, x_3)$ becomes prior pmf/pdf before X_4 is observed.
  - etc.
- Update the posterior distribution just once collectively.
  - Start with prior pmf/pdf $\xi(\theta)$.
  - Collect all observed values $X_1 = x_1, \ldots, X_n = x_n$.
  - Then update the posterior pmf/pdf with $\xi(\theta|x_1, x_2, \ldots, x_n)$.

Equivalence of the two approaches:

If $X_1, \ldots, X_n$ are conditionally iid given θ, then both approaches give the same posterior distribution after n observed values.

- **Consequence:** If $X_1, \ldots, X_n$ are Bernoulli, and θ is a beta R.V., then we can sequentially update the posterior distribution:
  - Given a new observed value x_k, replace (α, β) by $(\alpha + 1, \beta)$ if $x_k = 1$, and replace (α, β) by $(\alpha, \beta + 1)$ if $x_k = 0$.

Conjugate priors

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

- **Important Note:** If we start with any beta distribution as our prior distribution, 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.

General statistical models:

Suppose $X_1, \ldots, X_n$ are **arbitrary** observable R.V.'s that are conditionally iid given the parameter θ. Let Ω be the parameter space of θ, and let Ψ be a family of distributions on Ω.

- A prior distribution selected from Ψ is called a conjugate prior if its corresponding posterior distribution always remain in the same family Ψ, given **any** observed values.
- If every prior distribution from Ψ is a conjugate prior, then we say that Ψ is a conjugate family of prior distributions, or that Ψ is a family of conjugate priors.
  - e.g. the beta distributions form a family of conjugate priors.

Hyperparameters

If a family of conjugate priors is parametrized further by some parameters, then these parameters are also called hyperparameters.

- In other words, parameters of parameters are hyperparameters
- The term "hyperparameters" is used especially to avoid confusion with the parameters of the observable R.V.'s.
- Hyperparameters associated to prior distributions are called prior hyperparameters, while hyperparameters associated to posterior distributions are called posterior hyperparameters.

Theorem: (Equivalent reformulation of theorem on slide 9)
Let $X_1, \ldots, X_n$ be observable **Bernoulli** R.V.'s that are conditionally iid, given that the parameter θ has a beta distribution with prior hyperparameters α and β. Then, given the observed values $X_1 = x_1, \ldots, X_n = x_n$, the posterior hyperparameters of θ are $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n - (x_1 + \cdots + x_n)$.

Conjugate family of prior distributions

Suppose $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

- **What we know:** If $X_1, \ldots, X_n$ are **Bernoulli**, then the family Ψ of beta distributions is a family of conjugate priors.
- We then say that Ψ is closed under sampling from the Bernoulli distribution.
  - i.e. sampling from the Bernoulli distribution for $X_1, \ldots, X_n$ does not change what family the posterior distribution is in.

There are other 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

Gamma distribution

A continuous R.V. X is called gamma if its pdf is given by:

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0; \end{cases}$$

for some real numbers $\alpha, \beta > 0$.

- We say that X is the gamma R.V. with parameters α and β.
- Its distribution is called gamma distribution.
- $\mathbf{E}[X] = \frac{\alpha}{\beta}$ and $\mathrm{var}(X) = \frac{\alpha}{\beta^2}$.

Special Case: The gamma distribution with parameters $\alpha = 1$ and β is precisely the exponential distribution with parameter β.

Common Use: To model the uncertainty about the parameter or mean of a Poisson process.

- e.g. the likelihood that the average number of people visiting the Apple store in any given one-hour period exceeds a certain target number.

Sampling from Poisson distribution

Theorem: (Equivalent reformulation of theorem on slide 9) Let $X_1, \ldots, X_n$ be observable **Poisson** R.V.'s that are conditionally iid given the parameter θ. Suppose that θ is a gamma R.V. with prior hyperparameters α and β. Then, given the observed values $X_1 = x_1, \ldots, X_n = x_n$, the posterior hyperparameters of θ are $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n$.

Interpretation:

- We started with the initial guess that θ follows the gamma distribution with parameters α and β.
- Given observed values $X_1 = x_1, \ldots, X_n = x_n$, there are a total of $(x_1 + \cdots + x_n)$ occurrences of a certain event, over the n different time periods, each of fixed length.
- We update our guess for the distribution of θ to a new gamma distribution with the following parameters:

  $\alpha' = \alpha + (\text{number of new occurrences}) = \alpha + (x_1 + \cdots + x_n)$,

  $\beta' = \beta + (\text{number of time periods}) = \beta + n$

Example 2

Suppose the arrival of customers in a shop is modeled by a Poisson distribution with a rate of θ customers per minute. (θ is unknown.)

To determine a value for θ, consider a statistical model where $X_1, \ldots, X_6$ are observable Poisson R.V.'s that are conditionally iid given the parameter θ (i.e. θ is a R.V.).

- Each X_i represents the number of customers that arrive in the shop during the i-th randomly selected one-minute period.

Suppose that θ is a gamma R.V. with prior hyperparameters $\alpha = 3$ and $\beta = 2$.

Questions:

- Given that $X_1 = 3, X_2 = 2, X_3 = 0, X_4 = 5, X_5 = 1, X_6 = 2$, what is the posterior pdf of θ?

Example 2 - Solution

Method 1: (Tedious method!) We are given that the prior pdf is

$$\xi(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, & \text{if } \theta \geq 0; \\ 0, & \text{if } \theta < 0. \end{cases}$$

where $\alpha = 3$ and $\beta = 2$. Since $\Gamma(3) = 2! = 2$, this simplifies to:

$$\xi(\theta) = \begin{cases} 4\theta^2 e^{-2\theta}, & \text{if } \theta \geq 0; \\ 0, & \text{if } \theta < 0. \end{cases}$$

Let $\mathbf{x} = (3, 2, 0, 5, 1, 2)$ be the vector of observed values. Since $X_1, \ldots, X_n$ are discrete, it follows from Bayes' theorem that the posterior pdf of θ is

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

where $p(\mathbf{x}|\theta)$ is the **likelihood function** of θ given $\mathbf{x}$.

Example 2 - Solution

Since each X_i is Poisson, the marginal conditional pmf of X_i given θ is

$$p(x_i|\theta) = \begin{cases} \dfrac{\theta^{x_i} e^{-\theta}}{x_i!}, & \text{if } x_i = 0, 1, 2, \ldots; \\ 0, & \text{otherwise.} \end{cases}$$

Since $X_1, \ldots, X_6$ are conditionally iid given θ, we can use the observed values $\mathbf{x} = (3, 2, 0, 5, 1, 2)$ to compute the likelihood function of θ as follows:

$$\begin{aligned} p(\mathbf{x}|\theta) &= p(3|\theta)p(2|\theta)p(0|\theta)p(5|\theta)p(1|\theta)p(2|\theta) \\ &= \left(\frac{\theta^3 e^{-\theta}}{3!}\right)\left(\frac{\theta^2 e^{-\theta}}{2!}\right)\left(\frac{\theta^0 e^{-\theta}}{0!}\right)\left(\frac{\theta^5 e^{-\theta}}{5!}\right)\left(\frac{\theta^1 e^{-\theta}}{1!}\right)\left(\frac{\theta^2 e^{-\theta}}{2!}\right) \\ &= \frac{\theta^{13} e^{-6\theta}}{2880} \end{aligned}$$

Example 2 - Solution

Therefore, the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\xi(\theta)}{\int_0^\infty p(\mathbf{x}|\theta')\xi(\theta')\,d\theta'} = \frac{(\frac{1}{2880}\theta^{13}e^{-6\theta})(4\theta^2e^{-2\theta})}{\int_0^\infty (\frac{1}{2880}t^{13}e^{-6t})(4t^2e^{-2t})\,dt}$$

$$= \frac{\frac{1}{720}\theta^{15}e^{-8\theta}}{\frac{1}{720}\int_0^\infty t^{15}e^{-8t}\,dt} \approx (215.25)\theta^{15}e^{-8\theta}$$

if $\theta \geq 0$, and $\xi(\theta|\mathbf{x}) = 0$ otherwise.

Method 2: Let $\mathbf{x} = (3, 2, 0, 5, 1, 2)$ be the vector of observed values. Since the prior hyperparameters are $\alpha = 3$ and $\beta = 2$, the posterior hyperparameters are $\alpha' = 3 + (3 + 2 + 0 + 5 + 1 + 2) = 16$ and $\beta' = 2 + 6 = 8$. Thus the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \frac{8^{16}}{\Gamma(16)}\theta^{15}e^{-8\theta} = \frac{8^{16}}{15!}\theta^{15}e^{-8\theta} \approx (215.25)\theta^{15}e^{-8\theta}$$

if $\theta \geq 0$, and $\xi(\theta|\mathbf{x}) = 0$ otherwise.

Estimation

Given a prior distribution and some observed values, we know how to compute the posterior distribution.

- What is "the best possible" **estimate** for the value of θ?
  - Does this question even make sense? What is "best"?

Example: If θ represents the likelihood that a candidate will win an election, we would be more interested in finding a single value $\theta = \theta_0$ that "best" represents this likelihood, rather than specifying the whole distribution of θ.

- **Intuition:** This estimate θ_0 should depend on experimental evidence that we have collected.
  - Different observed values for the experiment should give different estimates.

Estimation is a kind of statistical inference that involves finding an approximate value (i.e. estimate) for a hypothetically observable parameter, or more generally, for a latent R.V.

- **Recall:** A latent R.V. is a function of observable R.V.'s.

Estimators and Estimates

Definition: Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ.

- An estimator of θ is a real-valued function $\delta(X_1, \ldots, X_n)$.
- Given δ and specific observed values $X_1 = a_1, \ldots, X_n = a_n$, the real number $\delta(a_1, \ldots, a_n)$ is called an estimate of θ.

Note: The definition of "estimator" is deliberately broad and allows for all kinds of estimators, whether "good" or "bad". We shall see several kinds of estimators in this course.

- An estimator is a function, while an estimate is a real number!
- An estimator is a function of n R.V.'s, so it is a R.V. itself.
- Since $X_1, \ldots, X_n$ are observable, an estimator is a **statistic**.
  - Recall: A statistic is a function of observable R.V.'s.

Note: Given a single parameter θ (treated as a R.V.), there could be many possible estimates θ_0 of θ that could be considered a good "representation" of θ, even for the same observed values.

A suitable choice of an estimator would typically depend on the context of the real-world problem that your statistical model is modeling.

Loss function

To determine whether an estimate is "good", we have to agree on some criterion. A commonly used criterion is given in terms of a loss function, defined as any real-valued bivariate function $L(x, y)$.

Interpretation: Given some parameter θ, a statistician would want to find an estimate of θ that is as close to the "true" value of θ as possible, otherwise the statistician incurs a "loss".

- If the "true" value of the R.V. θ equals the value θ, and if a is the estimate of θ found, then $L(\theta, a)$ is a measure of this **loss**.
- Typically, the larger the approximation error $|\theta - a|$, the larger the value of $L(\theta, a)$.

Some other common names equivalent to loss function:

- Cost function, objective function, fitness function.

Some Examples: (Note: There are many kinds of loss functions!)

- The squared error loss function is $L(x, y) = (x - y)^2$.
- The absolute error loss function is $L(x, y) = |x - y|$.

Bayes estimator

Suppose $X_1, \ldots, X_n$ are observable R.V.'s whose joint distribution is parametrized by a parameter θ with parameter space Ω. Let $L(x, y)$ be a loss function, and let $\xi(\theta|\mathbf{x})$ be the posterior pmf/pdf of θ given some vector $\mathbf{x}$ of observed values for $(X_1, \ldots, X_n)$.

- For every estimate a of θ obtained from $\mathbf{x}$, the Bayes risk of θ given a and $\mathbf{x}$ is the "expected loss" given by the expectation

$$\mathbf{E}[L(\theta, a)|\mathbf{x}] = \sum_{\theta \in \Omega} L(\theta, a)\xi(\theta|\mathbf{x}) \quad \text{or} \quad \int_{\Omega} L(\theta, a)\xi(\theta|\mathbf{x})\, d\theta.$$

- Note that different estimates a give different expected losses.
- Different observed $\mathbf{x}$ also give different expected losses.
- If $\delta^*(X_1, \ldots, X_n)$ is a real-valued function of $X_1, \ldots, X_n$ such that for every possible observed vector $\mathbf{x}$, the real number $\delta^*(\mathbf{x})$ is an estimate that minimizes the Bayes risk, then we say that $\delta^*(X_1, \ldots, X_n)$ is a Bayes estimator of θ.

Bayes estimator and Bayes estimate

Equivalent definition: A Bayes estimator of θ is a real-valued function $\delta^*(X_1, \ldots, X_n)$ such that for every possible observed vector $\mathbf{x}$, we set $\delta^*(\mathbf{x})$ to be a value such that the Bayes risk (or expected loss) is minimized over all possible estimates, i.e.

$$\mathbf{E}[L(\theta, \delta^*(\mathbf{x}))|\mathbf{x}] = \min_{a \in \mathbb{R}} \mathbf{E}[L(\theta, a)|\mathbf{x}].$$

Once the vector $\mathbf{x}$ of observed values is actually observed, we say that the real number $\delta^*(\mathbf{x})$ is a Bayes estimate of θ.

Interpretation: For a Bayes estimator to make sense, we first have to fix a loss function $L(x, y)$.

- We interpret $L(\theta, a)$ as the incurred loss of an estimate a, provided that θ is the "true" value of θ.
- We do not know the "true" value of θ, so we can consider all possible θ in Ω, and compute the expected incurred loss.
- A Bayes estimator is thus any function that assigns an estimate that minimizes this expected incurred loss, no matter what observed values we obtain.

Bayes estimators and squared error loss function

Suppose $X_1, \ldots, X_n$ are observable R.V.'s whose joint distribution is parametrized by a parameter θ. Let $L(x, y)$ be a loss function, and let $\xi(\theta|x_1, \ldots, x_n)$ be the posterior pmf/pdf of θ given the observed values $X_1 = x_1, \ldots, X_n = x_n$.

Theorem: If $L(x, y) = (x - y)^2$ is the **squared error loss function**, then the **Bayes estimator** of θ is the function

$$\delta^*(x_1, \ldots, x_n) = \mathbf{E}[\theta|x_1, \ldots, x_n]$$

defined on all possible $\mathbf{x} = (x_1, \ldots, x_n)$.

- In other words, $\delta^*(x_1, \ldots, x_n)$ is the **mean of the posterior distribution** of θ, also known as the posterior mean of θ.
- More simply, we could write that the Bayes estimator is the function $\mathbf{E}[\theta|X_1, \ldots, X_n]$.

Example 3

(**Statistical model from Example 2**):
Let $X_1, \ldots, X_6$ be observable Poisson R.V.'s that are conditionally iid given the parameter θ, where θ has the gamma distribution with prior hyperparameters 3 and 2.

Given that $X_1 = 3, X_2 = 2, X_3 = 0, X_4 = 5, X_5 = 1, X_6 = 2$, find the Bayes estimate of θ with respect to the squared error loss function $L(x, y) = (x - y)^2$.

Solution: Let $\mathbf{x} = (3, 2, 0, 5, 1, 2)$ be the vector of observed values. In Example 2, we computed that the posterior hyperparameters of θ are $\alpha = 16$ and $\beta = 8$. Thus, the Bayes estimate is

$$\mathbf{E}[\theta|\mathbf{x}] = \mathbf{E}[\text{posterior distribution}] = \frac{\alpha}{\beta} = \frac{16}{8} = 2.$$

In other words, the arrival of customers in Example 2 can be estimated as having a rate of 2 customers per minute.

Summary

- Gamma function and beta function
- Beta distribution
- Conjugate priors, conjugate family of prior distributions
- Gamma distribution
- Estimator and estimate
- Loss function
- Bayes estimator/estimate

Reminder:

- There is **Mini-quiz 3** (15mins) this week during cohort class.
  - Tested on materials from Lectures 11–13 only.
  - As mentioned during Lecture 13, the focus for Mini-quiz 3 will be on materials covered in Lecture 13.