Single parameter inference

STA 427/527, Fall 2019, Xin Wang

2.1 Point Estimate

- In Bayesian inference, **point estimator** can be **posterior mean, posterior median and posterior mode**
- Loss function measures the "loss" generated by estimating θ with the estimator $\hat{\theta}$.

1. Linear absoulte loss: $L_1\left(\hat{\theta}, \theta\right) = \left|\hat{\theta} - \theta\right|$
2. Quadratic loss: $L_2\left(\hat{\theta}, \theta\right) = \left(\hat{\theta} - \theta\right)^2$
3. Zero-one loss: $L_3\left(\hat{\theta}, \theta\right) = \begin{cases} 0 & \left|\hat{\theta} - \theta\right| \leq \epsilon \\ 1 & \left|\hat{\theta} - \theta\right| > \epsilon \end{cases}$

- Expected loss:

$$E\left[L\left(\hat{\theta}, \theta\right)|\boldsymbol{y}\right] = \int L\left(\hat{\theta}, \theta\right) p\left(\theta|\boldsymbol{y}\right) d\theta$$

- **Bayesian estimators:**

1. Posterior mean:

$$\min_{\hat{\theta}} E\left[L_2\left(\hat{\theta}, \theta\right) | Y\right] = \min_{\hat{\theta}} \int (\hat{\theta} - \theta)^2 p(\theta|Y)\, d\theta$$

$$\rightarrow \int \left(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2\right) p(\theta|Y)\, d\theta$$

$$= \int \hat{\theta}^2 p(\theta|Y) d\theta - \int 2\hat{\theta}\theta\, p(\theta|Y) d\theta + \int \theta^2 p(\theta|Y) d\theta$$

$$= \hat{\theta}^2 \int p(\theta|Y) d\theta - 2\hat{\theta} \int \theta\, p(\theta|Y) d\theta + C$$

$\int p(\theta|Y) d\theta = 1$ → valid pdf; $\int \theta\, p(\theta|Y) d\theta = E(\theta|Y)$; C = constant w/ respect to $\hat{\theta}$ ⇒ unimportant

$$= \hat{\theta}^2 - 2\hat{\theta} E(\theta|Y)$$

$$\rightarrow \frac{d}{d\hat{\theta}}[\cdots] = 2\hat{\theta} - 2E(\theta|Y)$$

$$\rightarrow 0 = 2\hat{\theta} - 2E(\theta|Y)$$

$$\Rightarrow \boxed{\hat{\theta} = E(\theta|Y) = \int_{\theta} \theta\, p(\theta|Y)\, d\theta}$$ posterior mean

2. Posterior median:

$$\min_{\hat{\theta}} E\left[L_1(\hat{\theta}, \theta)|Y\right] = \min_{\hat{\theta}} \int |\hat{\theta} - \theta|\, p(\theta|Y)\, d\theta$$

$$\Rightarrow \boxed{\hat{\theta}: \int_{-\infty}^{\hat{\theta}} p(\theta|Y)\, d\theta = 0.5}$$ such that

or written as → $p(\theta \leq \hat{\theta}|Y) = 0.5$

3. Posterior mode:

$\max_{\theta} p(\theta|Y)$ → $\hat{\theta}$; two modes: $\hat{\theta}_1$, $\hat{\theta}_2$

→ Always good to check shape of posterior if we have derived it already

- **The coin example**

→ Setup → $Y \sim \text{Binomial}(100, \theta)$ $\theta \sim \text{Uniform}(0,1)$ [$\text{Beta}(1,1)$]

→ Data dist → $p(y|\theta) = \binom{100}{y} \theta^y (1-\theta)^{100-y}$

→ prior dist → $\pi(\theta) = 1$

→ posterior dist → $p(\theta|y) \propto p(y|\theta)\pi(\theta)$

$\propto \theta^y (1-\theta)^{100-y} \cdot 1$ } shown previously

$= \theta^{(y+1)-1}(1-\theta)^{101-y}$

$\theta|y \sim \text{Beta}(y+1, 101-y)$

→ Posterior mean → $E(\theta|y) = \frac{y+1}{(y+1)+(101-y)} = \frac{y+1}{102}$

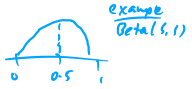
Posterior median → } use R → med = qbeta(0.5, y+1, 101−y)

posterior mode → → mode = optimize(f_x_beta, interval = c(0,1), maximum = TRUE)

↳ = function(x){dbeta(x, α, β)}

output ⇒ $ maximum = mode

example Beta(5,1)



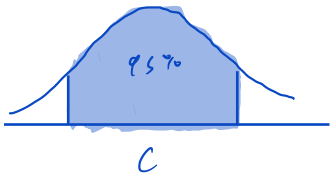
2.2 Interval Estimation

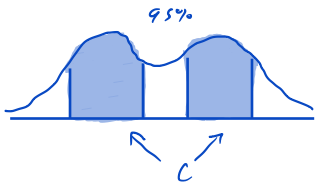
- **Definition:** A $100(1-\alpha)\%$ **credible set (interval)** $\mathcal{C}$ is a subset of the parameter space Θ such that

$\alpha = 0.05 \Rightarrow 95\%$

$$\int_{\mathcal{C}} p(\theta|\boldsymbol{y})\, d\theta = 1-\alpha \qquad \int_{\Theta} p(\theta|\underline{y})\, d\theta = 1$$

If the parameter space is discrete, we replace the integral with sum.





(doesn't have to be a continuous range)

→ Now the 95% represents an actual probability (which is different than C% confident in frequentist)

⇒ e.g. given the data we have, there is a 95% chance that our parameter is in our set

→ said another way → $P(L(\underline{y}) \le \theta \le U(\underline{y}))$

→ In Bayesian, we no longer use confidence intervals b/c the interpretation of the confidence level $100(1-\alpha)\%$ doesn't match what we want

(i.e. capture rate of parameter w/ repeated sampling)

⇒ So instead we use a "credible set"

Types of intervals

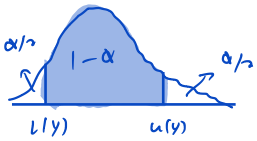
- **Equal tails interval**

→ Bounds L(y) & U(y) are just quantiles, so can easily find them if have posterior dist

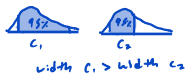
→ just use R or solve integrals

$$\left.\begin{aligned}\int_{-\infty}^{L(y)} p(\theta|\underline{y})\, d\theta &= \alpha/2 \\ \int_{U(y)}^{\infty} p(\theta|\underline{y})\, d\theta &= \alpha/2\end{aligned}\right\} \Rightarrow \int_{L(y)}^{U(y)} p(\theta|\underline{y})\, d\theta = 1-\alpha$$

→ If we simulate our data ⇒ then the interval is an approximation (monte carlo)



Bounds can shift, just need middle $(1-\alpha)/100$



width c_1 > width c_2

⇒ But there is an optimal credible set

  - Based on quantiles of the posterior distribution.
  - Not always the optimal, it could be wider if the posterior distribution is extremely skewed.

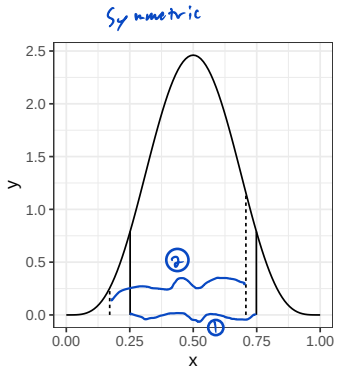
↳ Drawbacks of equal tails interval → than non-equal tails intervals

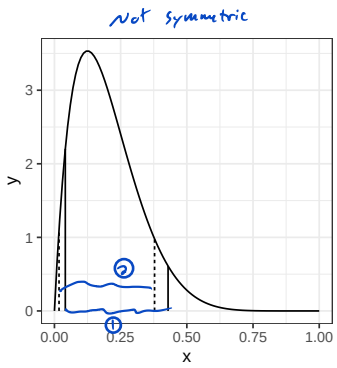
→ obviously want the narrowest interval

→ Advantages

→ If posterior dist is symmetric & unimodal, then equal tails interval is optimal (i.e. better than any other type of interval)

→ Easy to find/calculate





① is equal tail

② not equal tail

equivalent to equal tails when posterior is unimodal & symmetric, but if don't have $q^*()$ function in R, hard to compute ⇒ Equal tails preferred

- **Highest posterior density (HPD):** A $100(1-\alpha)\%$ HPD region for θ is a subset $\mathcal{C} \in \Theta$ defined by $\mathcal{C} = \{\theta : p(\theta|\boldsymbol{y}) \geq k\}$, where k is the largest number such that

$$\int_{\theta: p(\theta|\boldsymbol{y}) \geq k} p(\theta|\boldsymbol{y})\, d\theta = 1 - \alpha$$

Disadvantages
- Hard to compute if we don't have the inverse CDF for the posterior distribution → would have to write code to search → i.e. find θ: $p(\theta|\underline{y}) \geq k$
- Not guaranteed to be an interval (could be a set)

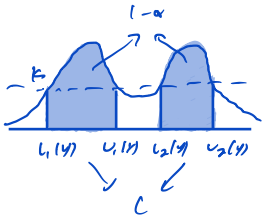
Advantages → good to use for skewed distributions

→ probability statement → $p(l(y) \leq \theta \leq u(y)|\underline{y}) = 1-\alpha \iff \theta \in [l(y), u(y)] \iff p(\theta|\underline{y}) \geq k$

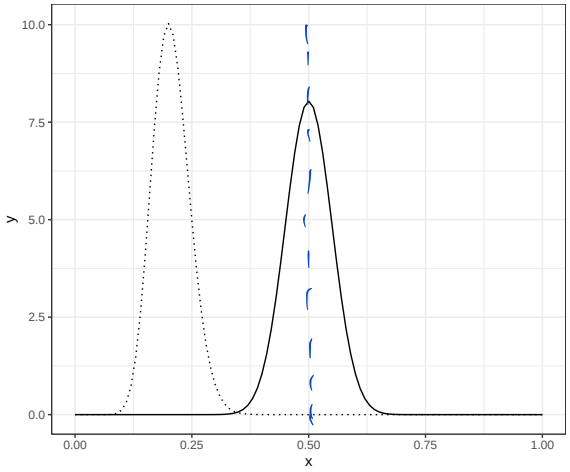
all density values outside of $[l(y), u(y)]$ are smaller than k

simplest case (unimodal)

also can be for more complex sets



Example: In the coin example, consider $y = 50$ and $y = 30$



→ $\theta|\underline{y} \sim Beta(y+1, 101-y)$ $\theta \in (0,1)$

→ if $y = 50$ → $\theta|\underline{y} \sim Beta(51, 51)$ ⇒ symmetric ⇒ equal tails interval is optimal

→ if $y = 30$ → $\theta|\underline{y} \sim Beta(31, 71)$ ⇒ not symmetric ⇒ need HPD interval to be optimal

→ It is good practice to (roughly) graph the posterior density so that we know which type of interval will be narrowest

Comparison of point estimates

-> point estimates for frequentist & Bayesian will be very close if have a uniform prior

-> Posterior mean -> always is a compromise (i.e. weighted average) between the prior distribution ($E(\theta)$) & the likelihood function ($MLE\ \hat{\theta}$)
↳ for mean $\Rightarrow \hat{\theta}_{MLE} = \bar{y}$ sample mean

↳ e.g. algebraically can rewrite

$$E(\theta|y) = \underbrace{k}_{\text{weight data}}\underbrace{\hat{\theta}_{MLE}}_{} + \underbrace{(1-k)}_{\text{weight prior}}\underbrace{E(\theta)}_{}$$

-> k usually some function of sample size & hyperparameters

-> Bin example -> $y|\theta \sim Binomial(n,\theta)$, $\theta \sim Beta(\alpha,\beta)$ -> Post mean $E(\theta|y) = \frac{y+\alpha}{n+\alpha+\beta}$ (if $\alpha=\beta=1$ $= \frac{y+1}{102}$ shown in notes)

$= \underbrace{\left(\frac{n}{n+\alpha+\beta}\right)}_{\text{weight}}\underbrace{\frac{y}{n}}_{\text{data}} + \underbrace{\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)}_{\text{weight}}\underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior}}$

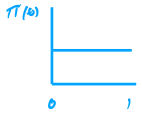
-> Mixing weight on data increases as n increases
$\Rightarrow$ more data $\Rightarrow$ lesser influence of prior

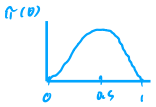
Selection of prior

-> $\pi(\theta)$ is a distribution that should match the parameter space
$\Rightarrow$ if the parameter space is discrete, then the prior should be a discrete dist

-> Also the ranges of Θ & $\pi(\theta)$ should match

-> Example priors -> uninformative uniform(0,1) vs Beta($\alpha=2$, $\beta=2$)
$\Rightarrow$ we suspect a higher prob of ≈ 0.5 based on previous knowledge



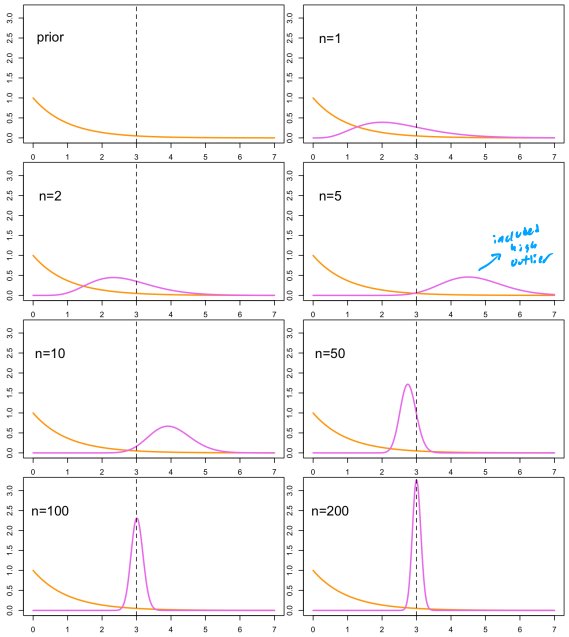


-> Effect of choosing the wrong prior

-> choosing a really bad prior can impact the final result a lot

-> with not a lot of data, a strong prior will dominate (i.e heavy weight) the result (in terms of the shape & location of the posterior)

-> But if have lots of data, the data will dominate the resulting posterior instead of the prior



$\Rightarrow$ If have several candidates for choice of prior, it is better to just use a uninformative prior rather than guessing/hoping that our choice of a good prior is good

In frequentist regression → $Y_i = X_i^T \beta + \varepsilon_i$ ⇒ estimate $\hat{\beta}$

→ observe $\tilde{x}_i^T$ & predict $\tilde{y}_i^T$ → $\hat{E}(\tilde{Y}_i) = \tilde{X}_i^T \hat{\beta}$

$\tilde{y}_i = \hat{E}(\tilde{Y}_i) + \tilde{\varepsilon}_i$

2.3 Prediction

- Suppose $\tilde{y}$ is an estimate of the future observation.

$\underset{\sim}{y} = (y_1, \ldots, y_n)$

Data distribution: $p(\boldsymbol{y}|\theta)$ } → $p(\theta|\underset{\sim}{y})$ → posterior
Prior: $\pi(\theta)$ ↳ Bayes is conditional on the data we have
Future observation: $p(\tilde{y}|\boldsymbol{y})$ ⇒ so for prediction, we are wanting to predict a new obs given the data we currently have

↓ future value ↳ vector

- **Definition:** The **posterior predictive distribution** of the future observaition is

$$p(\tilde{y}|\boldsymbol{y}) = \int_{\Theta} p(\tilde{y}, \theta|\boldsymbol{y})\, d\theta = \int_{\theta} p(\tilde{y}|\theta)\, p(\theta|\underset{\sim}{y})\, d\theta$$

→ integrate out θ to get a marginal of $\tilde{y}$ (the future obs), but still based on the data we have

future obs | given dataset | data dist | posterior

→ in Bayes, everything is written given $\underset{\sim}{y}$

→ this is the joint dist of future obs + our knowledge of θ

need to know both of these for prediction

rewrite this as conditional to use the info we have

just sample some values if want a vector of future obs

→ prior predictive distribution

→ $p(\tilde{y}) = \int_{\theta} p(\tilde{y}|\theta)\, \pi(\theta)\, d\theta$

start

→ this is predicting a future $\tilde{y}$ w/o data, so only using knowledge from the prior

update

→ predictive credible interval → $\int_{L(y)}^{U(y)} p(\tilde{y}|\underset{\sim}{y})\, d\tilde{y} = 1-\alpha$ (uses current data) (distribution)

- **Example:** The coin example: Suppose a coin was tossed 100 times and 50 heads were obtained. What is the chance if another head is obtained for another toss?

→ posterior → $\theta|y \sim \text{Beta}(y+1, 101-y)$

→ if $y=50$ → $\sim \text{Beta}(51, 51)$ → remember $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0<x<1$ (Beta)

→ posterior predictive distribution

→ $p(\tilde{y}|\underset{\sim}{y}) = ???$

$= \int_0^1 p(\tilde{y}|\theta)\, p(\theta|\underset{\sim}{y})\, d\theta$

↳ data dist from model ($n=1$ for this example)

$= \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} \frac{\Gamma(103)}{\Gamma(51)\Gamma(51)} \theta^{50}(1-\theta)^{50}\, d\theta$

$= \frac{\Gamma(103)}{\Gamma(51)\Gamma(51)} \int_0^1 \theta^{(51+\tilde{y})-1} (1-\theta)^{(52-\tilde{y})-1}\, d\theta$

$= \left[\frac{\Gamma(\tilde{y}+51)\Gamma(52-\tilde{y})}{\Gamma(103)} \right]$

→ $\tilde{y}|\theta \sim$ Bernoulli (single trial) → if $y=0$ → $p(\tilde{y}=0|\underset{\sim}{y}) = \frac{\Gamma(102)}{\Gamma(51)\Gamma(51)} \frac{\Gamma(51)\Gamma(52)}{\Gamma(103)} = \frac{51}{102} = 0.5$

⇒ if $y=1$ ⇒ $p(\tilde{y}=1|\underset{\sim}{y}) = 1 - p(\tilde{y}=0|\underset{\sim}{y})$

↓ $= 0.5$

⇒ if $y=50$ ⇒ our posterior for θ suggests equal prob for heads & tails

→ so it makes sense that prob of tails for $\tilde{y} = 0.5$

Posterior predictive distribution

→ In the previous example, the posterior dist had a nice, closed form solution

⇒ could easily do prediction

→ If it wasn't nice ⇒ can use simulation to obtain the posterior predictive distribution

using simulation

→ Suppose $p(\theta|\underset{\sim}{y})$ is the posterior & we have/can get samples from $p(\theta|\underset{\sim}{y})$

→ $p(\tilde{y}|\underset{\sim}{y}) = \int_\theta p(\tilde{y}|\theta)\, p(\theta|\underset{\sim}{y})\, d\theta$

→ 2 steps to simulate sample of $\tilde{y}$

for $m = 1, 2, \dots, M$ → the mth sample point generated

1) simulate $\theta^{(m)} \sim p(\theta|\underset{\sim}{y})$
2) simulate $\tilde{y}^{(m)} \sim p(\tilde{y}|\theta^{(m)})$ → data dist, just getting a new y value, $\tilde{y}$

→ now y conditioned on previous draw of θ

< repeat >

→ Then we have $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \dots, \tilde{y}^{(M)}$

→ Coin example → $M = 10{,}000$ ($y = 50$)

→ step 1) $\theta^{(m)} \sim \text{Beta}(51, 51)$

→ step 2) $\tilde{y}^{(m)} \sim \text{Bernoulli}(\theta^{(m)})$

⇒ $\tilde{y}^{(1)}, \dots, \tilde{y}^{(10000)}$ → this is called a "Monte Carlo" sample ⇒ simulating independent samples from the posterior (regular posterior or posterior predictive)

→ Monte Carlo just means applying statistical formulas to approximate continuous integrals

e.g.) $E[\theta|\underset{\sim}{y}] = \frac{1}{M}\sum \theta|\underset{\sim}{y}$

$p(\theta \le \theta_0|\underset{\sim}{y}) = \frac{1}{M}\sum I[\theta \le \theta_0|\underset{\sim}{y})$

→ Then we can approximate probabilities of interest using our simulated/generated sample

→ e.g) what's the probability of having 3 heads if toss the coin 5 times??

→ simulate $\tilde{y}^{(m)} \sim \text{Bernoulli}(\theta^{(m)})$ 5 times

↳ this is just getting a sample rather than an individual $\tilde{y}$ at a time

A Markov chain (will learn later) is when have dependent samples

e.g) $\theta^{(m)} \sim p(\theta^{(m-1)}|\underset{\sim}{y})$

(dependent on previous draw)

⇒ sequentially updating the posterior

↑

$(\theta^{(3)} \perp\!\!\!\perp \theta^{(1)})$

form of posterior predictive dist

→ A conjugate prior has nothing to do w/ the form of the posterior predictive distribution

→ usually the form of the posterior predictive is not nice

⇒ which is why we often use simulation to obtain future samples