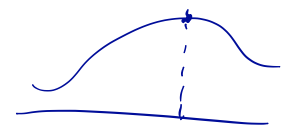
Single parameter inference

STA 427/527, Fall 2019, Xin Wang

$p(\theta|y)$



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

- Expected loss:

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

↑ posterior distribution of θ.

→ It's a function of $\hat{\theta}$.

- **Bayesian estimators:**

  1. Posterior mean:

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

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

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

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

take derivative $2\hat{\theta} - 2E(\theta|y) = 0$

$$\Rightarrow \hat{\theta} = E(\theta|y) = \int \theta\, p(\theta|y)\, d\theta$$

  2. Posterior median:

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

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

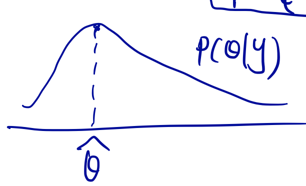
posterior median

$$P(\theta \le \hat{\theta}|y) = 0.5$$

  3. Posterior mode:

$$\max_{\hat{\theta}} p(\hat{\theta}|y)$$

$p(\theta|y)$



$X \sim \text{Beta}(\alpha, \beta) \qquad EX = \dfrac{\alpha}{\alpha+\beta}$

- **The coin example**

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

Data distribution: $p(y|\theta) = \binom{100}{y} \theta^{y} (1-\theta)^{n-y}$

prior: $\pi(\theta) = 1$

posterior: $p(\theta|y) \propto p(y|\theta) \cdot \pi(\theta)$

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

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

posterior mean: $E(\theta|y) = \dfrac{y+1}{y+1+101-y} = \dfrac{y+1}{102}$

posterior median?
posterior mode?] use R $\quad$ qbeta

2.2 Interval Estimation

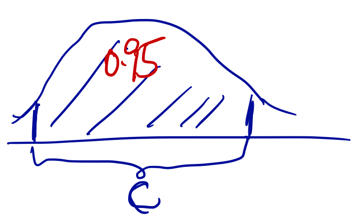
- **Definition:** A $100(1-\alpha)\%$ **credible set (interval)** $\mathcal{C}$ is a subset of the parameter space Θ such that

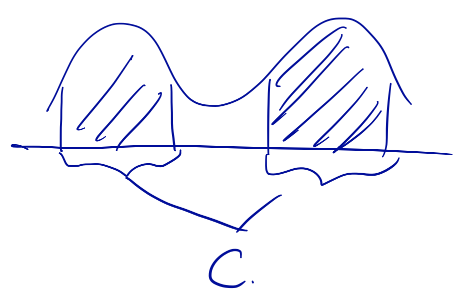
$\alpha = 0.05 \quad 95\%$

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

$\int_{\Theta} p(\theta|y)\, d\theta = 1$

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





For interval.

$\int_{L(y)}^{U(y)} p(\theta|y)\, d\theta = 1-\alpha$

$\alpha = 0.05$

- **Equal tails interval**

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

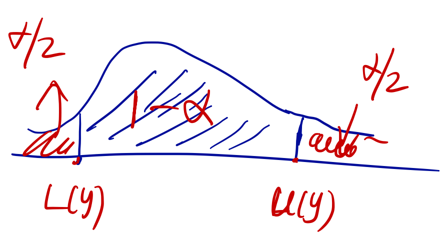
$\int_{-\infty}^{L(y)} p(\theta|y)\, d\theta = \alpha/2$

$L(y) = \text{qbeta}(0.025, y+1, 101-y)$

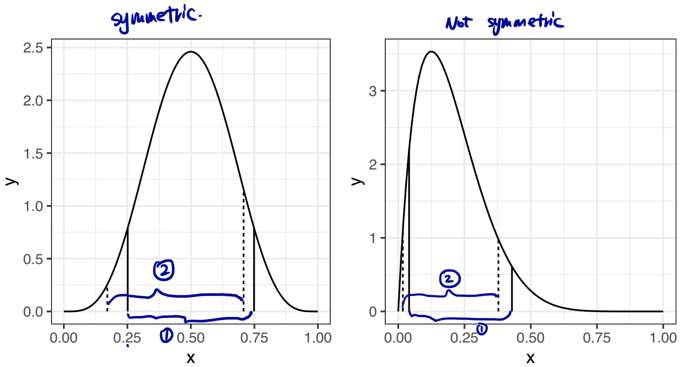
$U(y) = \text{qbeta}(0.975, y+1, 101-y)$

$\int_{U(y)}^{\infty} p(\theta|y)\, d\theta = \alpha/2$

$\Rightarrow \int_{L(y)}^{U(y)} p(\theta|y)\, d\theta = 1-\alpha$



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



① is equal tail

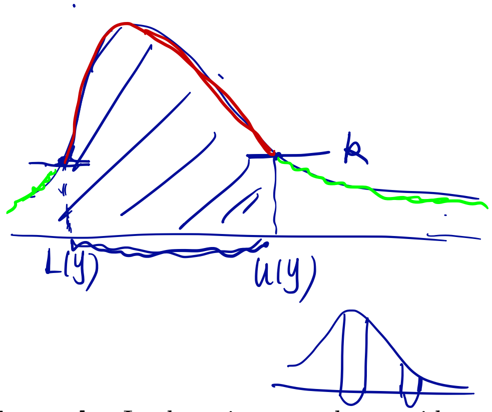
② Not equal tail

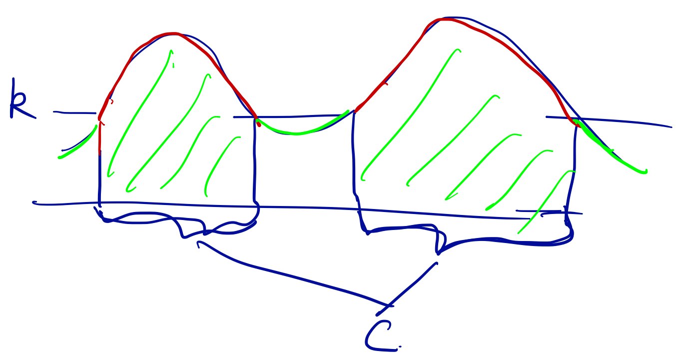
- **Highest posterior density (HPD):** A $100(1-\alpha)\%$ HPD region for θ is a subset $\mathcal{C} \in \Theta$ defined by $\mathcal{C} = \{\theta : p(\theta|\boldsymbol{y}) \geq k\}$, where k is the largest number such that

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

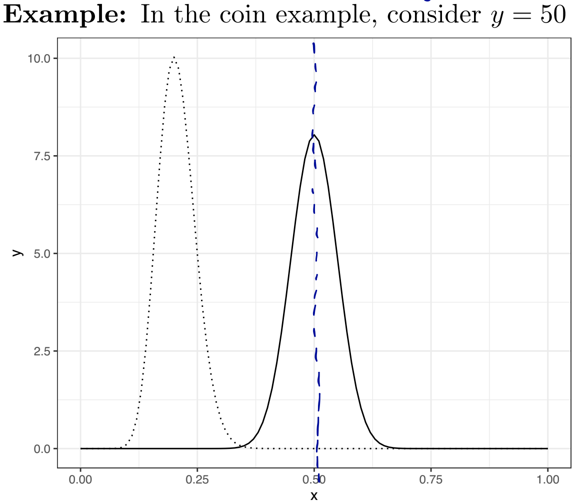
  - Hard to compute if we don't have the inverse CDF for the posterior distribution
  - Not guaranteed to be an interval

$P(L(y) < \theta < U(y) | y) = 1-\alpha$ $\quad \theta \in [L(y), U(y)]$ $\quad p(\theta|y) \geq k$





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



$\theta|y \sim Beta(y+1, 101-y)$ $\quad \theta \in (0,1)$

$y = 50$ $\quad \theta|Y=50 \sim Beta(51, 51)$

symmetric.

$y = 30$

$\theta|Y=30 \sim Beta(31, 71)$

In the frequentist. $Y_i = x_i^T\beta + \varepsilon_i$ $\hat{\beta}$

2.3 Prediction

$\tilde{x}_i \to \tilde{y}_i$ $\hat{E}(\tilde{Y}_i) = \tilde{x}_i^T\hat{\beta}$

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

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

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

Data distribution: $p(\boldsymbol{y}|\theta)$ $\to$ $p(\theta|y) \Leftarrow$ posterior.
Prior: $\pi(\theta)$
Future observation: $p(\tilde{y}|\boldsymbol{y})$

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

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

predictive credible interval: $\int_{L(y)}^{U(y)} \underbrace{p(\tilde{y}|y)}_{\text{distribution.}}\, d\tilde{y} = 1-\alpha$

prior predictive distribution: $p(\tilde{y}) = \int_\Theta p(\tilde{y}|\theta)\, \pi(\theta)\, d\theta$

- **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: $\theta|y \sim \text{Beta}(y+1, 101-y)$ $y = 50$

$\sim \text{Beta}(51, 51)$ $\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right.$

posterior predictive distribution: $p(\tilde{y}|y) = ?$

data distribution from your model.

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

$$= \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} \cdot \frac{\Gamma(102)}{\Gamma(51)\Gamma(51)} \theta^{50}(1-\theta)^{50}\, d\theta$$

$$= \frac{\Gamma(102)}{\Gamma(51)\Gamma(51)} \int_0^1 \theta^{\boxed{\tilde{y}+51}-1} (1-\theta)^{\boxed{52-\tilde{y}}-1}\, d\theta$$

$$= \frac{\Gamma(102)}{\Gamma(51)\Gamma(51)} \cdot \frac{\Gamma(\tilde{y}+51)\,\Gamma(52-\tilde{y})}{\Gamma(103)}$$

α integer. $\Gamma(\alpha) = (\alpha-1)!$

$\tilde{y} = 0, 1$ If $\tilde{y} = 0$

$$p(0|y) = \frac{\Gamma(102)}{\cancel{\Gamma(51)}\Gamma(51)} \frac{\cancel{\Gamma(51)}\,\Gamma(52)}{\Gamma(103)} = \frac{51}{102} = 0.5$$

$$p(1|y) = 1 - p(0|y) = 0.5$$

Use simulation to obtain posterior predictive distribution.

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

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

2 steps: for $m = 1, 2, \cdots, M$.

Step 1: Simulate $\theta^{(m)} \sim p(\theta|y)$.

Step 2: Simulate $\tilde{y}^{(m)} \sim p(\tilde{y}|\theta^{(m)})$

$\tilde{y}^{(1)}, \tilde{y}^{(2)} \ldots \tilde{y}^{(M)}$

---

coin example: $M = 10{,}000$

Step 1: $\theta^{(m)} \sim \text{Beta}(51, 51)$

Step 2: $\tilde{y}^{(m)} \sim \text{Bern}(\theta^{(m)})$

$\Rightarrow \tilde{y}^{(1)} \ldots \tilde{y}^{(M)}$

---

what's the prob of having 2 heads if tossing the coin for 5 times?

$\pi(\theta)$ distribution

$\downarrow$

knowledge belief. on the distribution of θ.

coin $\theta \in (0,1)$

Unif $(0,1)$ $\pi(\theta)=1$ $0<\theta<1$