# Introduction to Bayesian Statistics

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

# Introduction

- What we learned in this semester conceptually is called *frequentist* approach.
- In the frequentist approach, parameters are treated as unknown non-random constants.
- Probability statements are about observable random variables.
- For example, if the 95% CI is denoted as

$$P(L(X) < \theta < U(X)) = 0.95,$$

  the probability measure $P$ is about $X$, not $\theta$.
- Give it a try: How do we interpret the confidence interval?

# Introduction (cont'd)

- In Bayesian approach, parameters such as $\theta$ are conceptualized as random variables.
- Their distribution is called *prior distribution*.
- The prior distribution can be interpreted as our belief or knowledge about $\theta$ before observing $X$.
- It can also be interpreted as a *plausibility* function.
- The interpretation of the prior are key differences between different Bayesian schools.

# Notations

- Prior, pdf or pmf: $\pi(\theta)$, completely known and specified in advance.
- Likelihood: $f(x|\theta)$, conditional distribution of $X$ given $\theta$.
- Posterior: $\pi(\theta|x)$, conditional distribution of $\theta$ given $X$. It can be expressed as

$$\pi(\theta|x) = \pi(x)f(x|\theta)/m(x),$$

where $m(x)$ is the marginal distribution of $X$,

$$m(x) = \int_{\Theta} \pi(\theta)f(x|\theta)d\theta.$$

- The integral is replaced by a summation if $\theta$ is discrete.

# Binomial Bayes Estimation

- Let $X_1, \ldots, X_n$ be iid Bernoulli($p$). Then $Y = \sum_{i=1}^{n} X_i$ is binomial($n, p$).
- We assume that the prior distribution on $p$ is beta($\alpha, \beta$).
- The joint distribution of $Y$ and $p$ is

$$f(y, p) = f(y|p)\pi(p)$$

- The marginal distribution of $Y$ is

$$m(y) = \int_0^1 f(y, p)dp$$

- The posterior distribution is

$$\pi(p|y) = \frac{f(y, p)}{m(y)}$$

# Binomial Bayes Estimation (cont'd)

- The posterior is beta($y + \alpha, n - y + \beta$).
- How to estimate $p$?
- The mean of the posterior is

$$\hat{p}_B = \frac{y + \alpha}{\alpha + \beta + n},$$

which can be written as

$$\hat{p}_B = \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right).$$

- This is a weighted mean between sample mean and prior mean.
- Notice that, one can have different options for the prior.

# MSE of Binomial Bayes Estimator

- The MSE of $\hat{p}$, the MLE, as an estimator of $p$, is

$$\mathrm{E}(\hat{p}-p)^2 = \mathrm{Var}\bar{X} = \frac{p(1-p)}{n}.$$

- The MSE of the Bayes estimator of $p$ is

$$\mathrm{E}(\hat{p}_B-p)^2 = \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i+\alpha}{\alpha+\beta+n}\right) + \left\{\mathrm{E}\left(\frac{\sum_{i=1}^n X_i+\alpha}{\alpha+\beta+n}\right)-p\right\}^2$$

- Choose $\alpha=\beta=\sqrt{n/4}$ yields

$$\mathrm{E}(\hat{p}_B-p)^2 = \frac{n}{4(n+\sqrt{n})^2}.$$

- For small $n$, $\hat{p}_B$ is the better choice; for large $n$, $\hat{p}$ is the better choice.

# Conjugate Priors

- In the binomial example above, both the prior and the posterior distributions were in the beta family.
- A family of priors that leads to posteriors in the same family is called a *conjugate family*.
- Such priors are called *conjugate priors*.
- **Example (Normal Bayes estimators)** Let $X \sim n(\theta, \sigma^2)$, and suppose that the prior distribution of $\theta$ is $n(\mu, \tau^2)$.
- The posterior distribution of $\theta$ is also normal, with mean and variance given by

$$\mathrm{E}(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu$$

and

$$\mathrm{Var}(\theta|x) = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}.$$

# Conjugate Priors (cont'd)

- If the random sample is extended to $X_1, \ldots, X_n$, the posterior mean and variance become

$$\mathrm{E}(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu$$

and

$$\mathrm{Var}(\theta|x) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n},$$

respectively.

- What do we learn from the binomial and normal Bayes estimation?

# Hypothesis Testing

- Suppose we want to test $H_0 : p \in A$ versus $H_1 : p \in A^c$.
- We can use the posterior $\pi(p|y)$ to compute the probability

$$a_0 = P(p \in A|y)$$

- Reject $H_0$ when $a_0 > 1/2$.
- **(Normal Bayesian Test)** Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We will reject $H_0$ if and only if

$$P(\theta \leq \theta_0|\boldsymbol{x}) > 1/2.$$

- Since $\pi(\theta|\boldsymbol{x})$ is symmetric, $H_0$ will be rejected if $E(\theta|\boldsymbol{x}) > \theta_0$, i.e.,

$$\bar{X} > \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}.$$

# Hypothesis Testing (cont'd)

- If the type I error is considered more serious than the type II error we may change our cutoff "1/2" to a smaller number.
- **(Bayes Factor)** A Bayesian measure of evidence against the null hypothesis ($H_0$), and in favor of an alternative hypothesis ($H_1$), is called *Bayes Factor*, which is defined by

$$\text{BF} = \frac{P(H_1|\boldsymbol{x})/P(H_0|\boldsymbol{x})}{P(H_1)/P(H_0)}.$$

- This factor can be interpreted as the ratio of posterior odds of $H_1$ against the prior odds of $H_1$.

# Hypothesis Testing (cont'd)

- According Kass and Raftery (1995), $1 < \text{BF} \leq 3$ provides "weak" evidence, $3 < \text{BF} \leq 20$ provides "positive" evidence, $20 < \text{BF} \leq 150$ provides "strong" evidence, and $\text{BF} > 150$ provides "very strong" evidence in favor of $H_1$.

# Bayes Factor: An Example

- Assume the survival time for advanced-stage colorectal cancer follows
$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0.$$
- For unknown $\lambda$, one is willing to assume the prior of $\lambda$ follows
$$\pi(\beta) = \beta e^{-\beta \pi}, \quad x > 0, \quad \beta > 0.$$
- If $\beta = 1$ and $x = 3$, what is the Bayes factor for testing $H_0 : \lambda \geq 1$ versus $H_1 : \lambda < 1$?
- What is the strength of evidence in favor of $H_1$ according to the scale proposed by Kass and Raftery (1995)?

# Bayes Factor: An Example (cont'd)

- It may be easier to get the marginal CDF of $X$, which equals

$$E\{F(x|\lambda)\} = \int_0^\infty F(x|\lambda)\pi(\lambda)d\lambda = 1 - \left(1 + \frac{x}{\beta}\right)^{-1}.$$

- The posterior distribution hence is

$$\pi(\lambda|x) = \frac{f(x|\lambda)\pi(\lambda)}{m(x)} = \frac{\lambda\beta e^{-(x+\beta)\lambda}}{\beta^{-1}(1 + x/\beta)^{-2}} = \lambda(x+\beta)^2 e^{-(x+\beta)\lambda},$$

  which is Gamma$(2, (x+\beta)^{-1})$ ($\pi(\lambda)$ conjugate prior?).
- Since $P(\lambda < \lambda^*|x) = 1 - \{\lambda^*(x+\beta) + 1\}e^{-(x+\beta)\lambda^*}$, one can have

$$\text{BF} = \frac{P(H_1|\boldsymbol{x})P(H_0)}{P(H_0|\boldsymbol{x})P(H_1)} = \frac{(1 - 5e^{-4})e^{-1}}{5e^{-4}(1 - e^{-1})} = 5.77.$$

# Interval Estimation

- Quantile of $\pi(p|x)$ can be used to compute interval estimators.
- The resulting intervals are called *Bayesian credible intervals* or *credible set* (C&B).
- For a $1-\alpha$ credible interval, we choose $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ with $\alpha = \alpha_1 + \alpha_2$.
- Define $L(x)$ and $U(x)$ to be $\alpha_1$ and $1-\alpha_2$ quantiles of $\pi(p|x)$, respectively.
- The intervals can be one-sided by taking either $\alpha_1$ or $\alpha_2$ to be zero.

# Normal Credible Set

- In the previous normal example,

$$\pi(\theta|\bar{x}) = n(\delta^B(\bar{x}), \sigma^2(\theta|\bar{x})).$$

- The $1-\alpha$ credible set for $\theta$ is given by

$$1-\alpha = P\left(\delta^B(\bar{x}) - z_{\alpha/2}\sigma(\theta|\bar{x}) \leq \theta \leq \delta^B(\bar{x}) + z_{\alpha/2}\sigma(\theta|\bar{x})\right).$$

- How about the coverage probability of this region in frequentist sense? One can have

$$\begin{aligned} &P(|\theta - \delta^B(\bar{x})| \leq z_{\alpha/2}\sigma(\theta|\bar{x})) \\ &\quad = P\left(-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}}\right), \end{aligned}$$

where $\gamma = \sigma^2/(n\tau^2)$ and $Z = \sqrt{n}(\bar{X}-\theta)/\sigma$.

# Bayesian Optimality

- One can obtain the smallest credible interval with a specific coverage probability.
- We would like to find the set $C(x)$ that satisfies
  (a) $\int_{C(x)} \pi(\theta|x)dx = 1-\alpha$,
  (b) $\text{size}(C(x)) \leq \text{size}(C'(x))$,
  for any set $C'(x)$ satisfying $\int_{C'(x)} \pi(\theta|x)dx \geq 1-\alpha$.
- Using Theorem 9.3.2 in C&B, we can conclude if the posterior density $\pi(\theta|x)$ is unimodal, then for a given value of $\alpha$, the shortest credible interval for $\theta$ is given by

$$\{\theta : \pi(\theta|x) \geq k\}, \quad \text{where} \quad \int_{\{\theta:\pi(\theta|x)\geq k\}} \pi(\theta|x)d\theta = 1-\alpha.$$

- We call this *highest posterior density* (HPD) region.

# Decision Theory

- Estimating $\theta$ can be viewed as a decision or an action.
- A loss function $L(\theta, \hat{\theta})$ quantifies the penalty for choosing $\hat{\theta}$ when the true value is $\theta$.
- Two types of loss functions: *squared-error loss*

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2,$$

and *absolute-error loss*

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|.$$

- One may also consider *weighted* loss function

$$L(\theta, \hat{\theta}) = \omega(\theta)|\hat{\theta} - \theta|^r,$$

for some $\omega(\theta) \geq 0$ and $r > 0$.

# Risk Function

- Formally, let $X_1, \ldots, X_n$ be a random sample from distribution $f(x|\theta)$, $\theta \in \Theta \subseteq \Re$, and let $\delta(x)$ be an estimator of $\theta$.
- The loss function $L(\theta, \delta(x)) \geq 0$ is defined over $\Theta \times D \to \Re^+$.
- The *risk function* represents the expected loss over the sample space $D$, which is defined by

$$R(\theta, \delta(x)) = E_X\{L(\theta, \delta(x))\} = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx.$$

- Given two decision rules $\delta^*(x)$ and $\delta(x)$, if $R(\theta, \delta^*) \leq R(\theta, \delta(x))$ $\forall \theta \in \Theta$, and $R(\theta, \delta^*) < R(\theta, \delta(x))$ at at least one $\theta \in \Theta$, we will call $\delta^*(x)$ is better than $\delta(x)$.

# Minimax Decision Rule

- We will call a decision rule $\delta^*(x)$ a *minimax decision rule* if

$$\sup_{\theta} R(\theta, \delta^*(x)) = \inf_{\delta} \sup_{\theta} R(\theta, \delta(x))$$

- **Example** On a rainy day a teacher has three choices: ($a_1$) to take an umbrella and face the possible prospect of carrying it around the sunshine; ($a_2$) to leave the umbrella at home and perhaps get drenched; ($a_3$) to just give up the lecture and stay at home.
- Let $\Theta = \{\theta_1, \theta_2\}$ and $\theta_1$ corresponds to rain, and $\theta_2$ to no rain.
- The following table give the losses for the decision problem:
- The weather report that depends on $\theta$ as follows:
- Find the minimax rule to help the teacher make a decision.

# Minimax Decision Rule (cont'd)

- There are 9 decisions when you saw the weather outside.
- The risk function:

$$R(\theta_j, \delta_i) = E\{L(\theta_j, \delta_i)\} = \sum_{k=1}^{2} L(\theta_j, \delta_{ik}) P(W_k|\theta_j),$$

  where $\delta_{ik}$ is the action $\{a_1, a_2, a_3\}$ you take when you saw $W_1$ (rain) or $W_2$ (shine).
- The conclusion: Bring the umbrella no matter rain or shine.

# Bayes Decision Rule

- Add a probability measure on the parameter, $\pi(\theta)$.
- Bayes risk with respect to $\pi(\theta)$:

$$r^B(d) = E\{R(\theta, \delta(x))\} = \int_\Theta R(\theta, \delta(x))\pi(\theta)d\theta.$$

- If there exists a decision function $\delta^*(x)$, satisfying

$$r^B(\delta^*) = \inf_\delta r^B(\delta),$$

  we will call $\delta^*(x)$ is Bayes decision function with respect to $\pi(\theta)$
- Show that $\inf_\delta r^B(\delta)$ has the same solution as

$$\inf_\delta \int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta.$$