# STA 630: Bayesian Inference - Lecture 1

Zeya Wang

University of Kentucky

Spring 2025

# Outline of the Course

- Course overview and introduction
- Bayes rule and example
- The Binomial model
- Bayesian inference for single-parameter models
- Multiple-parameter models
- Posterior simulation and integration
- Markov chain Monte Carlo methods
- Hierarchical models
- Regression Models
- Hierarchical linear models
- Generalized linear models

# Outline of the Course

Main Textbooks:

- Hoff (2009). A first course in Bayesian statistical methods. Springer.
- Gelman et al. (2004). Bayesian Data Analysis, 2nd ed. Chapman & Hall.
- Markov Chain Monte Carlo in Practice (1996). Gilks et al. (eds). Chapman.

# Part 1 - Introduction to Bayesian Statistics, Single- and Multi-Parameter Models

- Review of probability (Chapter 2, P. Hoff's book)
- Bayes rule
- Bayesian inference (prior, likelihood, posterior)
- The Binomial model
- Predictive distributions
- Conjugate models (Poisson and exponential)
- Prior types (objective, subjective, diffuse)
- High density regions and Bayes factors
- Multi-parameter models (Normal and multinomial models)

# Other Parts

- Part 2: MCMC methods
- Part 3: Hierarchical and linear models

# Lecture 1: Introduction to Bayesian Statistics - Outline

- Bayes rule and example
- Bayesian inference
- Prior, likelihood, posterior

# Review of Probability Concepts (Chapters 1 & 2)

- Experiment: phenomenon where outcomes are uncertain - e.g., single throws of a six-sided die.
- Sample space: set of all outcomes of the experiment $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.
- Event: a subset of $\mathcal{S} - A = \{3\}, B = \{3, 4, 5, 6\}$.

# Basic properties of probability

- If $S$ is the sample space, $P(S) = 1$.
- For any event $A, 0 \leq P(A) \leq 1$.
- For any complementary events $A$ and $A^c$,

$$P\left(A^c\right) = 1 - P(A) \quad P(\varnothing) = 1 - P(\mathcal{S}) = 0$$

- For any two events $A$ and $B$, the probability that either $A$ or $B$ will occur is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# Basic properties of probability

- The conditional probability of $A$ given $B$ for any two sets $A$ and $B$ is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } P(B) \neq 0$$

- $A$ and $B$ are independent if

$$P(A \mid B) = P(A) \qquad \text{or equivalently} \qquad P(A \cap B) = P(A)P(B)$$

# Bayes Rule

- Law of total probability

  (i) $P(B) = P(B \mid A)P(A) + P\left(B \mid A^c\right) P\left(A^c\right)$

  (ii) Let $A_1, \ldots, A_n$ be a partition of the sample space, ie, a set of events such that $\bigcup_{i=1}^n A_i = S$ and $A_i \cap A_j = \varnothing$ for $i \neq j$ and $P\left(A_i\right) > 0$ for all $i$. Then, for any event $B$,

$$P(B) = \sum_{i=1}^{n} P\left(B \mid A_i\right) \cdot P\left(A_i\right)$$

# Bayes Rule

- Bayes Theorem

  (i) for two events $P(A \mid B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$

  (ii) for multiple events. Let $B$ and $A_1, \ldots, A_n$ be events where $\bigcup_{i=1}^{n} A_i = S, A_i \cap A_j = \varnothing$ for $i \neq j$, and $P(A_i) > 0, \forall i$. Then

$$P(A_j \mid B) = \frac{P(B \mid A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i) P(A_i)}$$

  [for $(ii)$, $P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{P(B)}$ then use Law of total probability $(ii)$ for $P(B)$ ]

# Example

Diagnostic test (non-controversial, widely accepted use of Bayes rule)

Say you know that HIV has a prevalence in the population of 1/1000. A particular test for HIV has a 95% sensitivity and 98% specificity [1]. What is the probability that someone testing positive actually has HIV?

$B =$ test is positive, $A =$ have HIV

$P(A) = 1/1000, \quad P(B \mid A) = .95, \quad P\left(B^c \mid A^c\right) = .98$

(or $P\left(B^c \mid A\right) = .05$ false negatives and $P\left(B \mid A^c\right) = .02$ false positives)

$P(A \mid B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)} = \frac{(.95)(.001)}{(.95)(.001)+(.02)(.999)} = 0.045$

Over 95% of the those testing positive will not have HIV: $P\left(A^c \mid B\right)$ (test of limited diagnostic value - low incidence disease)

$P(A) =$ prior belief (disease prevalence); $B =$ data (test result); $P(A \mid B) =$ posterior probability of disease

[1] Sensitivity = prob of testing positive given the disease; specificity = prob of testing negative given that the individual is disease free

## Comments

- Despite the apparent high accuracy of the test, the incidence of the disease is so low that the vast majority of patients who test positive do not have the disease.
- Nonetheless, this is 40 times the proportion before we knew the outcome of the test! The test is not useless, and retesting may improve the reliability of the result.

# More Comments

- Disease prevalence as ”prior belief” that a person has the disease.
- We observe a positive result (i.e., data)
- Bayes rule tells us how the test result should change (update) our belief about the probability of disease in the presence of new evidence.
- We update our belief to a posterior probability of disease.

# More Comments

- More formal notation:
- $\theta =$ disease status ($\theta = 1$ is person has disease, $\theta = 0$ otherwise )
- $X =$ Random variable ($X = 1$ if test positive, $X = 0$ otherwise)
- Probability model for $X : P(X = i \mid \theta = j), \quad i, j = 0, 1$
- Prior belief on $\theta : P(\theta = 1) = 0.001, P(\theta = 0) = .999$
- Likelihood of $X = 1 : P(X = 1 \mid \theta = 0) = 0.02, P(X = 1 \mid \theta = 1) = 0.95$
- Use Bayes rule to update our prior belief to

$$P(\theta = 1 \mid X = 1) = \frac{P(X = 1 \mid \theta = 1)P(\theta = 1)}{P(X = 1 \mid \theta = 1)P(\theta = 1) + P(X = 1 \mid \theta = 0)P(\theta = 0)} = 0.045$$

# Example 2: Paternity dispute

- Suppose you are on a jury considering a paternity suit brought by Suzy Smith's mother against Al Edged.
- Suzy's mother has blood type $\mathrm{O}$ and $\mathrm{Al}$ Edged is type AB .
- You have other information as well. You hear testimony concerning whether Al Edged and Suzy's mother had sexual intercourse during the time that conception could have occurred, about the timing and frequency of such intercourse, about Al Edged fertility, about the possibility that someone else is the father, and so on. You put all this information together in assessing $P(F)$, your probability that Al is Suzy's father.
- The evidence of interest is Suzy's blood type, which turns out to be B (if it were $\mathrm{O}, \mathrm{Al}$ Edged would be excluded from paternity).

# Example 2: Paternity dispute

- According to Mendelian genetics, $P(B \mid F) = \frac{1}{2}$.
- You also accept the blood bank's estimate $P\left(B \mid F^c\right) = 0.09$.
- According to Bayes' rule

$$P(F \mid B) = \frac{P(B \mid F)P(F)}{P(B \mid F)P(F) + P\left(B \mid F^c\right)P\left(F^c\right)}$$

The relationship between our prior probability, $P(F)$, and our posterior probability, $P(F \mid B)$ may be summarized:

| $P(F)$ | 0 | 0.100 | 0.250 | 0.500 | 0.750 | 0.900 | 1 |
|---|---|---|---|---|---|---|---|
| $P(F \mid B)$ | 0 | 0.382 | 0.649 | 0.847 | 0.943 | 0.980 | 1 |

# Discrete Random Variables

- The set of possible values is either finite or countably infinite.
- Associated with a random variable $Y$, with possible values $y_1, y_2, \ldots$, there is a probability mass function, $p(y_i) = P(Y = y_i)$, such that

$$p(y_i) \geq 0 \quad \text{and} \quad \sum_i p(y_i) = 1$$

# Discrete Random Variables

- ... and a cumulative distribution function (cdf), $F(y) = P(Y \leq y), -\infty < y < \infty$. The cdf is non-decreasing and satisfies

$$\lim_{y \to -\infty} F(y) = 0 \text{ and } \lim_{y \to \infty} F(y) = 1.$$

  If $Y$ is a discrete random variable, $F(y)$ is a step function, with jumps occurring at the values of $y$ for which $p(y) > 0$.

- The mean or expected value of a discrete random variable $Y$ with probability mass function $p(y)$ is given by

$$\mu = E(Y) = \sum_i y_i p\left(y_i\right)$$

# Continuous random variable

- For continuous random variables, the set of possible values is uncountable.
- The probability density function (pdf), $f(y)$, of a continuous random variable, $Y$, with support $\mathcal{S}$ is an integrable function such that:

  (a)

  $$f(y) > 0, \text{ if } y \in \mathcal{S} \quad f(y) = 0, \text{ if } y \notin \mathcal{S}$$

  (b)

  $$\int_S f(y) dy = 1$$

  (c)

  $$P(a \leq Y \leq b) = \int_a^b f(y) dy$$

# Continuous random variable

- For a continuous random variable, the cumulative distribution function (cdf) is given by

  $$P(Y \leq a) = F(a) = \int_{-\infty}^{a} f(y)dy$$

  The cdf of a continuous random variable, $F(y)$, is continuous and monotonically non-decreasing.

- If $a < b$, we obtain

  $$P(a \leq Y \leq b) = \int_{a}^{b} f(y)dy = F(b) - F(a)$$

- The mean or expected value of a continuous random variable $Y$ with pdf $f(y)$

  $$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy$$

# Joint distributions

- $X, Y$ discrete r. v. with values $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$, respectively. Their joint probability mass function $p_{X,Y}(x, y)$ is $p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$
- The marginal probability mass function of one random variable is obtained from the joint frequency distribution by

$$p_X(x) = \sum_j p_{X,Y}(x, y_j) \quad p_Y(y) = \sum_i p_{X,Y}(x_i, y).$$

- The conditional probability that $X = x_i$, given that $Y = y_j$ is,

$$p_{X|Y}(x \mid y) = P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}.$$

This can be re-expressed as $p_{X,Y}(x, y) = p_{X|Y}(x \mid y) p_Y(y)$.

Summing both sides over all values of $y$, we get a very useful application of the law of total probability:

# Continuous random variables:

- Suppose that $X$ and $Y$ are two continuous random variables. Their joint probability density function, $f(x, y)$, is the surface such that for any region $A$ in the $xy$-plane,

$$P((X, Y) \in A) = \iint_A f(x, y) dxdy$$

- The marginal probability density function of one random variable is obtained from the joint pdf

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

# Continuous random variables:

- The conditional density functions of $Y$ given $X$ is defined to be

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}, \qquad \text{if } 0 < f_X(x) < \infty$$

The joint density can be expressed in terms of the marginal and conditional densities as:

$$f_{XY}(x,y) = f_{Y|X}(y \mid x) f_X(x)$$

Integrating both sides over $y$ allows the marginal density of $X$ to be expressed as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy$$

which is the law of total probability for the continuous case.

# Bayes Rule for continuous random variables

- Rule of total probability:

$$f(x) = \int f(x \mid y) f(y) dy$$

- Bayes rule:

$$f(y \mid x) = \frac{f(x \mid y) f(y)}{f(x)} = \frac{f(x \mid y) f(y)}{\int f(x \mid y) f(y) dy}$$

Note on notation: From now on, $p(x)$ for generic $X$ (discrete or continuous).

# Exchangeability and de Finetti's theorem

- Let $p\left(x_1, \ldots, x_n\right)$ be the joint distribution of $X_1, \ldots, X_n$ and let $\pi_1, \ldots, \pi_n$ be a permutation of the indices $1, \ldots, n$.
- If $p\left(x_1, \ldots, x_n\right) = p\left(x_{\pi_1}, \ldots, x_{\pi_n}\right)$ for all permutations, then $X_1, \ldots, X_n$ are exchangeable.
- de Finetti's Theorem:

  Let $X_1, X_2, \ldots$ be a sequence of random variables. If for any $n$, $X_1, \ldots, X_n$ are exchangeable, then there exists a prior distribution $p(\theta)$ and sampling model $p(x \mid \theta)$ such that

$$p\left(x_1, \ldots, x_n\right) = \int_{\Theta} \left\{ \prod_1^n p\left(x_i \mid \theta\right) \right\} p(\theta) d\theta$$

# Exchangeability and de Finetti's theorem

$$\left.\begin{array}{rl} X_1, \ldots, X_n \mid \theta & \overset{\text{iid}}{\sim} p(x \mid \theta) \\ \theta & \sim p(\theta) \end{array}\right\} \Leftrightarrow X_1, \ldots, X_n \text{ are exchangeable.}$$

This is applicable if $X_1, \ldots, X_n$ are

- outcomes of a repeatable experiment
- sampled from an infinite population without replacement
- sampled from a finite population of size $N \gg n$ without replacement

# What would not be an exchangeable sequence?

- Consider the case of ”streaks” in sports, where a team that has just won its previous game is more likely to win the next, and conversely, a team that has just lost a game is more likely to loose the next.
- In this case,

  $$p(1,1,1,0,0,0)$$

  would not be believed to equal

  $$p(1,0,1,0,1,0)$$

  and thus the joint probability is not preserved under permutation.
- Thus, the sequence would not be regarded as exchangeable.

# Bayesian Inference formal notation

- Motivation is to combine inference from data with prior information
- The probability model determines the likelihood of the data as a function of $\theta$, i.e., $x_1, \ldots, x_n \sim p(\cdot \mid \theta)$, then $L(\theta) \propto \prod_i p\left(x_i \mid \theta\right)$ (exchangeability).
- In the Bayesian point of view $\theta$ has a probability distribution, $\theta \sim \pi(\theta)$, that reflects our uncertainty about it.
- Inference on the unknown, $\theta$, is made conditional on all relevant known information (e.g., the data $x = (x_1, \ldots, x_n)$ ). Bayes theorem allows us to condition upon the data to calculate a posterior distribution

$$p(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int p(x \mid \theta)\pi(\theta)d\theta} \propto L(\theta)\pi(\theta)$$

# Bayesian Inference formal notation

- $p(x) = \int p(x \mid \theta)\pi(\theta)d\theta$ is the normalizing constant that makes $p(\theta \mid x)$ integrate to 1 . It is also the marginal distribution of the data.
- All inference about $\theta$ must be based on the posterior distribution, often summarized through point estimates (mean, median, mode) or interval estimates (lower and upper $\alpha/2$ percentiles)

# Differences between frequentist and Bayesian statistics

- Bayesian statistics: uncertainty is quantified by determining how prior opinion about parameter values changes in light of the observed data.
- Classical view: uncertainty about, for example, parameter estimates is quantified by investigating how such estimates vary in repeated sampling from the same population.
- Data sets which might have observed, but were not, are irrelevant to a Bayesian in making inference. The only relevant data set is the one observed. Bayesians, on the other hand, need to specify their priors.
- The Bayesian approach has deep historical roots but required the algorithmic developments of the late 1980s before it became useful.
- The old sterile Bayesian-frequentist debates are a thing of the past. Most data analysts take a pragmatic point of view and use whatever is most useful.

# Difficulties with the Bayesian Approach

- It requires the specification of a prior distribution for all unknowns.
- A Bayesian analysis is subjective: two people with different priors observe same data and yet reach different conclusions on $\theta$.

  **Counters**:

  - When there is concrete prior knowledge it should be used!
  - Use objective priors, for example "noninformative" or vague priors that express ignorance or little knowledge.
  - When a large amount of data is available the prior has little influence on the posterior, unless it is very "peaked".
  - "Reality": Scientists often disagree due to different knowledge they have. Bayes methods provides a way of formally incorporating this information in the decision making process.
- Bayesian methods involve high-dimensional integrals. No longer a serious concern, after the advent of MCMC methods (time consuming but often worth the effort, as they allow fitting complex models without resorting to large sample approximations).

# Reverend Bayes

The term BAYESIAN derives from Thomas Bayes, a British mathematician and a Presbyterian minister (ca. 1702-1761) who lived in Tunbridge Wells (Kent).

# SUMMARY: Bayes theorem applied to statistical models

- Bayesian inference is based on the following premise/axiom: "Uncertainties about all unknown quantities are expressed by a joint probability distribution (prior/posterior distributions). Statistical inference about an unknown, $\theta$, is made conditional on all relevant known information (e.g., data)".
- Motivation: Combine inference from data with prior information.
- Probability model:

$$x \mid \theta \sim p(x \mid \theta)$$

- $\theta$ unknown model parameters, missing data, events we did not observe directly or exactly (latent variable)
- In the Bayesian point of view $\theta$ has a probability distribution

$$\theta \sim \pi(\theta)$$

that reflects our uncertainty about it.

# What we will learn

**Posterior $\propto$ Likelihood $\times$ Prior**

- How do I quantify my prior information?
  - conjugate choices
  - diffuse choices which assign probability more or less evenly over large regions of the parameter space
- How do I assess the effect of my prior beliefs?
  - sensitivity analyses across alternative specifications can reveal stability (or not) to prior models.
- How do I do integrals? For most problems $p(x)$ does not have a closed form.
  - conjugate choices
  - Markov chain Monte Carlo methods

# Lecture 2: The Binomial Model Outline

- A single-parameter model: The Binomial model
- Credible intervals and Highest posterior density regions
- Predictive distribution

# The Binomial model

Example: A drug is tested on 21 patients and it is found to have a positive effect in 18 patients. What is the probability that the drug is effective?

- Parameter of interest: $\theta =$ probability of drug being effective, $0 < \theta < 1$
- Probability model: $X =$ number of successes in $n$ (binary) trials, $X \sim \text{Binomial}(n, \theta)$ with $n = 21$

$$p(X = x \mid \theta) = \left( \begin{array}{c} n \\ x \end{array} \right) \theta^x (1 - \theta)^{n-x}$$

- Likelihood (observe $X = 18$ ): $L(\theta) \propto \theta^{18}(1 - \theta)^3$

$$\left[ x_1, \ldots, x_{21} \sim \text{Bernoulli}(\theta) \text{ then } L(\theta) \propto \prod_i p\left(x_i \mid \theta\right) \right]$$

# The Binomial model

- Use uniform prior, $\pi(\theta) = 1$ for $0 < \theta < 1$
- Posterior: $p(\theta \mid X = x) \propto L(\theta)\pi(\theta) \propto \theta^{18}(1-\theta)^3 \rightarrow \text{Beta}(19, 4)$

$$\theta \mid X = x \sim \text{Beta}(x+1, n+1-x)$$

- Now consider $\theta \sim \text{Beta}(\alpha, \beta)$, that is $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ (we call $\alpha$ and $\beta$ hyperparameters - Beta $(1,1) \equiv U[0,1]$ )
- Posterior $\propto$ likelihood $\times$ prior

$$p(\theta \mid x) \propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

(recognize this as the kernel of a Beta dist)

- Therefore $\theta \mid x \sim \text{Beta}(\alpha + x, \beta + n - x)$

$$p(\theta \mid x) = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \times \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

# The Binomial model

- First example of a conjugate prior - results in a posterior of the same parametric form (typically same functional form of the likelihood). Note that we avoided the computation of $p(x)$ !
- Posterior mean as point estimate of $\theta$. Also, $100(1-\alpha)\%$ credible interval as $(a, b) : P(a \leq \theta \leq b \mid x) = 1 - \alpha$
- Case $\theta \sim \text{Beta}(1, 1)$ : Posterior mean, $19/(19+4) = .83$; 95% credible interval: $(0.65, 0.95)$ as 2.5 th and 97.5 th percentiles of $\text{Beta}(19, 4)$ - read as $\theta \in (0.65, 0.95)$ with .95 probability

# Practice 1: Ex from Gelman et al. book, page 39, sec. 2.5

Question: Is the proportion of female births in the population of the placenta previa births - $\theta$ - less than .485 , the proportion of female births in the general population?
A study in Germany found that of a 980 placenta previa births, 437 were female
Likelihood: $L(\theta) \propto \theta^{437}(1-\theta)^{980-437}$; Prior: $\theta \sim \text{Beta}(1,1)$
Posterior: $p(\theta \mid x) \propto \theta^{437}(1-\theta)^{980-437}$, that is $\theta \mid x \sim \text{Beta}(438, 544)$.
Post mean $= .446; 95\%$ c.i. is $(0.415, 0.477).P(\theta \leq .485 \mid x) = .993$

```
y<-437 ; n<-980;
theta.mean <-(y+1) /(n+2)
[1] 0.4460285 # posterior mean
qbeta(c(.025,.25,.50,.75,.975),438,544) # percentiles
[1] 0.41506550 .4353081  0.4459919  0.4567090  0.4771998
postdraws <- rbeta (1000,438,544) # via simulation
quantile(postdraws, probs =c(0.025, .25, .50, .75, .975) )
0.4134468  0.4349219  0.4467121  0.4575036  0.4760245
```

# Practice 1: Ex from Gelman et al. book, page 39, sec. 2.5

- Posterior mean: weighted average of prior mean and sample proportion with weights proportional to $\alpha + \beta$ and $n$, respectively (" $\alpha + \beta$ " determines prior precision; often called 'prior sample size')

$$E[\theta \mid x] = \frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \frac{\alpha}{\alpha + \beta} + \left(\frac{n}{\alpha + \beta + n}\right) \frac{x}{n}$$

- Posterior variance:

$$\text{var}[\theta \mid x] = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E[\theta \mid x](1 - E[\theta \mid x])}{\alpha + \beta + n + 1}$$

# Practice 1: Ex from Gelman et al. book, page 39, sec. 2.5

- Posterior dist: comprise between prior and data depending on relative weight of $\alpha + \beta$ with respect to $n$
- With $\alpha$ and $\beta$ fixed, as $n$ increases prior has less influence on posterior $E[\theta \mid x]$ dominated by $\frac{x}{n}$ and $\text{var}[\theta \mid x]$ by $\frac{1}{n}\frac{x}{n}\left(1 - \frac{x}{n}\right)$
- For $n$ fixed, larger $\alpha + \beta$ values increase prior precision - more and more 'informative' prior with respect to the data
- Let's look at posterior summaries for different $\alpha$ and $\beta$ - this is called sensitivity analysis.

# Highest Posterior Density Regions

- The $(1-\alpha)100\%$ credible interval is defined as the interval $(a, b)$ of the upper and lower $1-\alpha/2$ percentiles. It is the range of values above and below which lies exactly $(1-\alpha)100\%$ of the posterior density.
- The $(1-\alpha)100\%$ highest posterior density (HPD) region is defined as the region that contains $(1-\alpha)100\%$ of the highest area of the posterior density. It is the region of values that contains $(1-\alpha)100\%$ of the posterior density and such that the density inside the region in never lower than that outside. HPD regions have smallest possible volume in the parameter space. They give the highest probabilities of containing the parameter for a given volume. They may not have symmetric tails (skewed dist) and may not be an interval (bimodal dist).
- HPD region $\equiv$ credible interval for symmetric, unimodal and concave distributions
- For skewed or bimodal distributions it is hard to construct HPD regions.

# Example

- $X \sim \text{Binomial}(10, \theta)$. Observe $X = 2$ successes. For $\theta \sim \text{Beta}(1, 1)$ we have $\theta \mid x \sim \text{Beta}(1 + 2, 1 + 8)$.
- 95% credible interval: $(0.06, 0.52)$, i.e. after seeing the data we believe that $\theta \in (0.06, 0.52)$ with .95 probability.
- 95% HPD region: $(0.04, 0.048)$ (narrower) - see plot (adjust horizontal line ”by hand” to get .95)

# Example

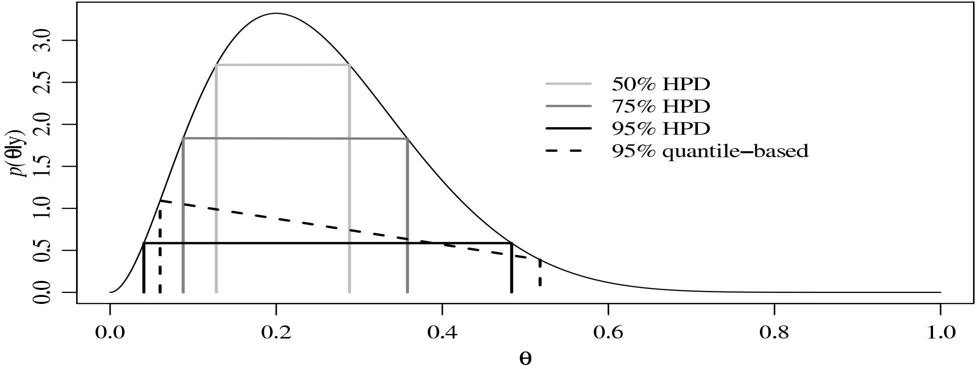

Figure : 95% credible interval (dashed line) and HPD regions

# Predictive distribution

Suppose we want to calculate the predictive dist of a future outcome $z$, given the data, that is $p(z \mid x)$. Sample observations are typically assumed conditionally independent, i.e. $z$ independent of $x$ given $\theta$

$$p(z \mid x) = \int p(z, \theta \mid x) d\theta = \int p(z \mid \theta, x) p(\theta \mid x) d\theta = \int p(z \mid \theta) p(\theta \mid x) d\theta$$

(this is an expected value over the post dist, $E_{\theta|x}[p(z \mid \theta)]$ )

Back to Binomial example of Practice 1 , with $\theta \sim U[0,1]$. The probability of a future success is

$$p(z = 1 \mid x) = \int p(z = 1 \mid \theta) p(\theta \mid x) d\theta = \int_0^1 \theta p(\theta \mid x) d\theta = E[\theta \mid x] = \frac{x+1}{n+2}$$

For $n = 21$ and $x = 18$ we have $p(z = 1 \mid x) = \frac{19}{23}$