Single Parameter Models

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Outline

- Bayes Theorem
- Single Parameter Models
- Bayesian Inference Based on Posterior
- Prediction
- Prior Elicitation

Bayes Theorem

- Let A denote an event, and A^c denote its complement. Thus $A \cup A^c = S$ and $A \cap A^c = \emptyset$, where $\mathcal{S}$ is the sample space. We have

$$P(A) + P(A^c) = P(S) \equiv 1$$

- Let A and B are two nonempty events, and $P(A|B)$ denote the probability of A given that B has occurred. From basic probabilities, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

and thus, $P(A \cap B) = P(A|B)P(B)$.
Likewise, $P(A \cap B) = P(B|A)P(A)$ and $P(A^c \cap B) = P(B|A^c)P(A^c)$.

- Observe that

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

  – This is Bayes' Theorem.

Bayes Theorem (cont'd)

- Example: *Suppose 5% of a given population is infected with HIV virus, and that a certain HIV test gives a positive result 98% of the time among patients who have HIV and 4% of the time among patients who do not have HIV. If a given person has tested positive, what is the probability that he/she actually has HIV virus?*

$$\begin{aligned} A_1 &= \text{event person has HIV} \\ B &= \text{event of testing positive} \end{aligned}$$

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_1^c)P(A_1^c)} \\ &= \frac{0.98 \times 0.05}{0.98 \times 0.05 + 0.04 \times 0.95} = 0.563 \end{aligned}$$

- General Bayes's theorem: Let $A_1, \ldots, A_m$ be mutually exclusive and exhaustive events. (Exhaustive means $A_1 \cup \cdots \cup A_m = \mathcal{S}$.) For any event B such that $P(B) > 0$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}, j = 1, \ldots, m.$$

Bayes' Theorem Applied to Statistical Models

- Suppose we have observed data $\mathbf{y}$ which have a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$ that depends upon an unknown vector of parameters $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ that represents the experimenter's opinion about $\boldsymbol{\theta}$.

- *Bayes' theorem applied to statistical model*

$$\begin{aligned} \text{Posterior} \rightarrow p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y},\boldsymbol{\theta})}{m(\mathbf{y})} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\mathbf{y}|\mathbf{t})\pi(\mathbf{t})d\mathbf{t}} \leftarrow \frac{\text{likelihood} \times \text{prior}}{\text{marginal distribution of } \mathbf{y}} \end{aligned}$$

  Θ is the parameter space, i.e. the set of all possible values for $\boldsymbol{\theta}$.

- The marginal distribution of $\mathbf{y}$ is a function of $\mathbf{y}$ alone (nothing with θ), and is often called 'normalizing constant'.

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

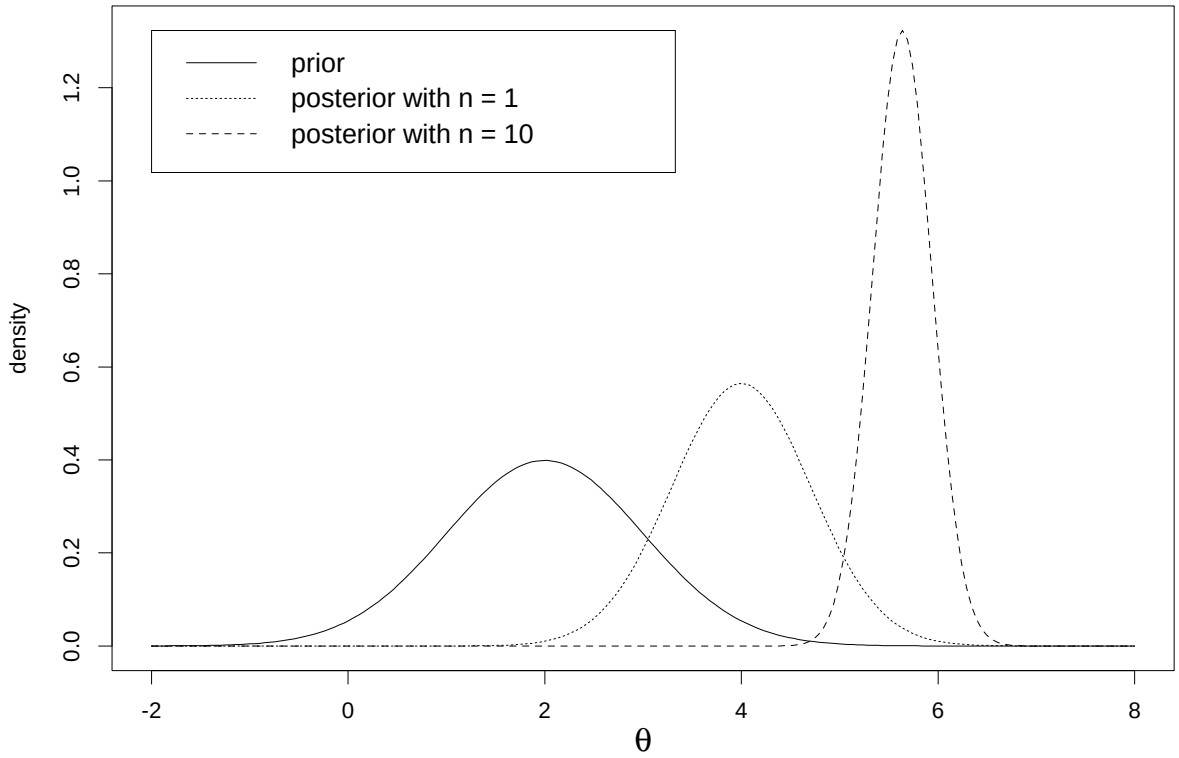
Single Parameter Model: Normal with Known Variance

- Consider a single observation y from a normal distribution with known variance.
  - *Likelihood*: $y \sim N(y|\theta, \sigma^2)$, $\sigma > 0$ is known.
  - *Prior* on θ: $\theta \sim N(\theta\,|\mu, \tau^2)$, $\mu \in \mathbb{R}$ and $\tau > 0$ are known hyperparameters.
  - *Posterior* distribution of θ:

$$p(\theta|y) = N\left(\theta \mid \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

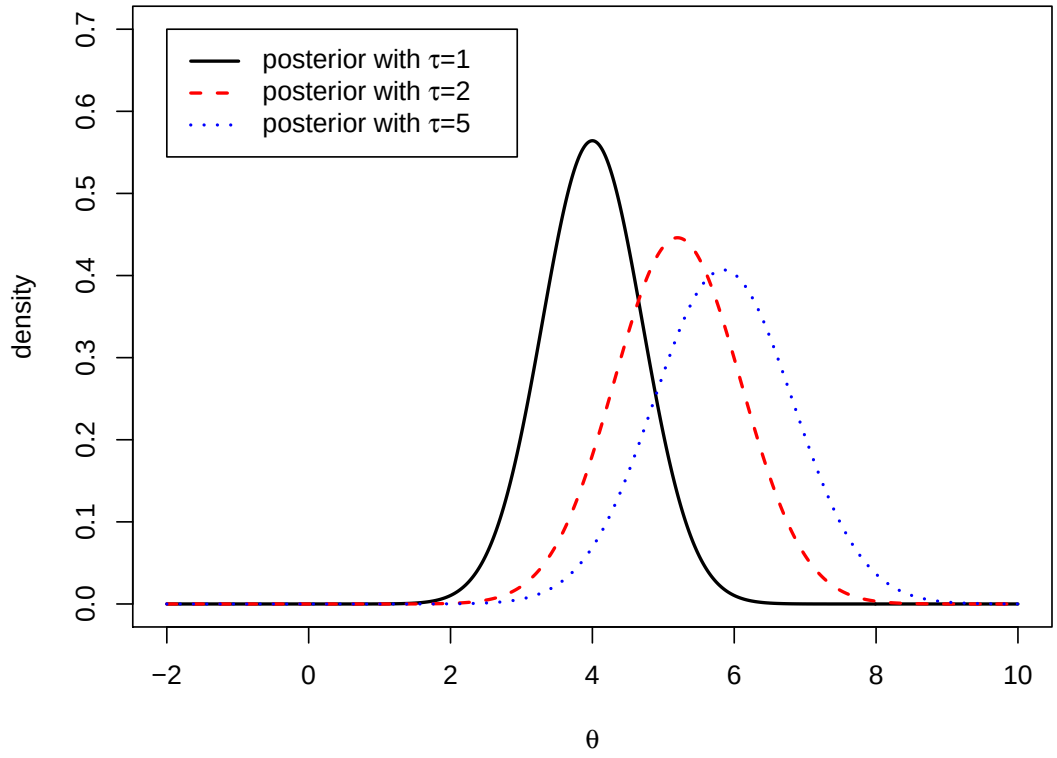
- Write $B = \frac{\sigma^2}{\sigma^2+\tau^2}$, and note that $0 < B < 1$. Then:
  - $E(\theta|y) = B\mu + (1-B)y$, a weighted average of the prior mean and the observed data value, with weights determined sensibly by the variances.
  - $Var(\theta|y) = B\tau^2 \equiv (1-B)\sigma^2$, smaller than τ^2 and σ^2.
  - Precision (which is like "information") is additive:
    $Var^{-1}(\theta|y) = Var^{-1}(\theta) + Var^{-1}(y|\theta)$.

Example: $\mu = 2, \bar{y} = 6, \tau = \sigma = 1$, varying n



- When $n = 1$ the prior and likelihood receive equal weight, so the posterior mean is $4 = \frac{2+6}{2}$.
- When $n = 10$ the data dominate the prior, resulting in a posterior mean much closer to $\bar{y}$.
- The posterior variance also shrinks as n gets larger; the posterior collapses to a point mass on $\bar{y}$ as $n \to \infty$.

$\mu = 2, \bar{y} = 6, n = 1, \sigma = 1$, **varying** τ



- When $\tau = 1$ the prior is as informative as likelihood, so the posterior mean is $4 = \frac{2+6}{2}$.
- When $\tau = 5$ the prior is almost flat over the likelihood region, and thus is dominated by the likelihood.
- As τ increases, the prior becomes "flat" relative to the likelihood function. Such prior distributions are called "noninformative" priors.

Deriving the Posterior

- We can find the posterior distribution of the normal mean θ via Bayes Theorem

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

- Note that $m(y)$ does NOT depend on θ, and thus is just a constant. That is,

$$p(\theta|y) \propto f(y|\theta)\pi(\theta).$$

  The final posterior is $A \cdot f(y|\theta)\pi(\theta)$, such that

$$\int A \cdot f(y|\theta)\pi(\theta)d\theta = 1$$

- Question: Consider n independent observations $\mathbf{y} = (y_1, \ldots, y_n)$ from the normal distribution $f(y_i|\theta) = N(y_i|\theta, \sigma^2)$, and prior $\pi(\theta) = N(\theta|\mu, \tau^2)$. What is the posterior of θ now?

Bayes and Sufficiency

- Recall that $T(\mathbf{y})$ is sufficient for θ if the likelihood can be factored as

$$f(\mathbf{y}|\theta) = h(\mathbf{y})g(T(\mathbf{y})|\theta).$$

- Implication in Bayes:

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta) \propto g(T(\mathbf{y})|\theta)\pi(\theta)$$

  Then $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|T(\mathbf{y})) \Rightarrow$ we may work with $T(\mathbf{y})$ instead of the entire dataset $\mathbf{y}$.

- Again, consider n ind. observations $\mathbf{y} = (y_1, \ldots, y_n)$ from the normal distribution $f(y_i|\theta) = N(y_i|\theta, \sigma^2)$, and prior $\pi(\theta) = N(\theta|\mu, \tau^2)$.

  Since $T(\mathbf{y}) = \bar{y}$ is sufficient for θ, we have that $p(\theta|\mathbf{y}) = p(\theta|\bar{y})$.

- We know that $f(\bar{y}|\theta) = N(\theta, \frac{\sigma^2}{n})$, this implies that

$$p(\theta|\bar{y}) = N\left(\theta \,\middle|\, \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}\mu + \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{y}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

Single Parameter Model: Binomial Data

- Example: Estimating the probability of a female birth. The currently accepted value of the proportion of female births in large European-race populations is **0.485**. Recent interest has focused on factors that may influence the sex ratio.

- We consider a potential factor, the maternal condition *placenta previa*, an unusual condition of pregnancy in which the placenta is implanted low in the uterus obstructing the fetus from a normal vaginal delivery.

- Observation: An early study concerning the sex of placenta previa births in Germany found that of a total of 980 births, 437 were female.

- Question: How much evidence does this provide for the claim that the proportion of female births in the population of placenta previa births is less than the proportion of female births in the general population?

Example: Probability of a female birth given placenta previa

- Likelihood: Let

$$\begin{aligned} \theta &= \text{prob. of a female birth given placenta previa} \\ Y_i &= \begin{cases} 1 & \text{if a female birth} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

- Let $X = \sum_{i=1}^{980} Y_i$. Assuming independent births and constant θ, we have $X|\theta \sim Binomial(980, \theta)$,

$$f(x|\theta) = \binom{980}{x} \theta^x (1-\theta)^{980-x} .$$

- Consider a beta prior distribution for θ

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} .$$

Example: Probability of a female birth given placenta previa

- The posterior distribution can be obtained via

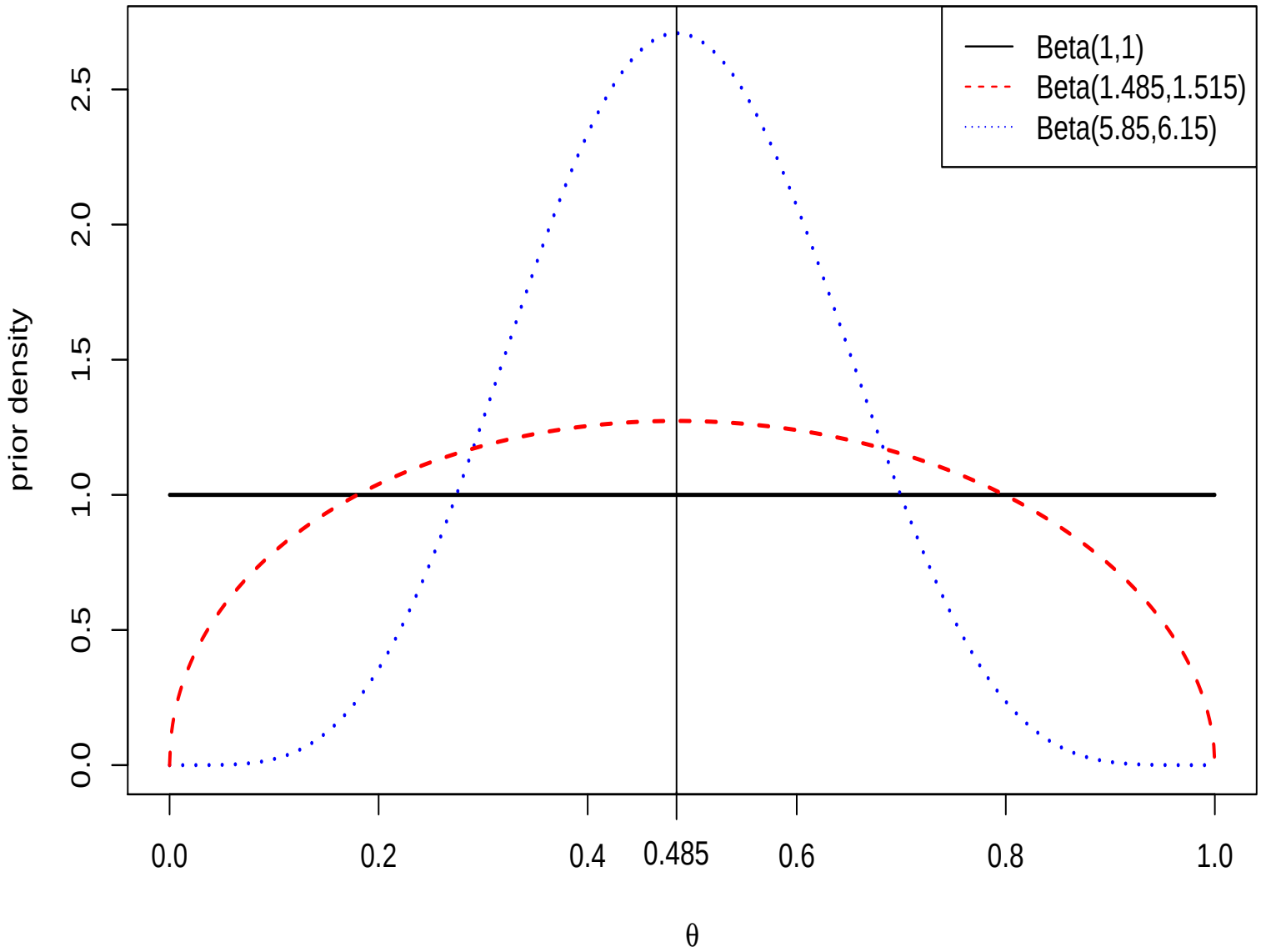
$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\,\pi(\theta) \\ &= \binom{980}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{x+\alpha-1}(1-\theta)^{980-x+\beta-1} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{980-x+\beta-1}\,. \end{aligned}$$

- The only distribution function that is proportional to the above is $Beta(x+\alpha, 980-x+\beta)$!

$$\theta|X \sim Beta(x+\alpha,\ 980-x+\beta)$$

- Beta distributions are conjugate priors for Binomial likelihood models!

Three different beta priors



Bayesian Inference

- Now that we know what the posterior is, we can use it to make inference about $\boldsymbol{\theta}$.
- The "big three" class of classical, or frequentist, inference are
  1. Point estimation
  2. Confidence interval (CI)
  3. Hypothesis testing
- Each of them has its analog in the Bayesian world.

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.
- Mode is often easiest to compute (no integration), but is often least representative of "middle", especially for one-tailed distributions.
- Mean has the opposite property, tending to "chase" heavy tails (just like the sample mean $\bar{X}$)
- Median is probably the best compromise overall, though can be awkward to compute, since it is the solution θ^{median} to

$$\int_{-\infty}^{\theta^{median}} p(\theta|x)\, d\theta = \frac{1}{2}\,.$$

Posterior estimates

Prior distribution	Posterior Mode	Mean	Median
$Beta(1,1)$	0.44592	0.44603	0.44599
$Beta(1.485, 1.515)$	0.44596	0.44607	0.44603
$Beta(5.85, 61.5)$	0.44631	0.44642	0.44639

The classical point estimate is $\hat{\theta}_{MLE} = \frac{437}{980} = 0.44592$.

Remark:

1. A Bayes point estimate is a weighted average of a common frequentist estimate and a parameter estimate obtained only from the prior distribution.
2. The Bayes point estimate "shrinks" the frequentist estimate toward the prior estimate.
3. The weight on the frequentist estimate tends to 1 as n tends to infinity.

Bayesian Inference: Interval Estimation

- The Bayesian analogue of a frequentist CI is referred to as a credible set: a $100 \times (1-\alpha)\%$ credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$P(C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \geq 1-\alpha.$$

- Unlike the classical confidence interval, it has a proper probability interpretation: "The probability that $\boldsymbol{\theta}$ lies in C is $(1-\alpha)$"

- Two principles used in constructing credible set C:
  - The volume of C should be as small as possible.
  - The posterior density should be greater for every $\boldsymbol{\theta} \in C$ than it is for any $\boldsymbol{\theta} \notin C$.

  The two criteria turn out to be equivalent.

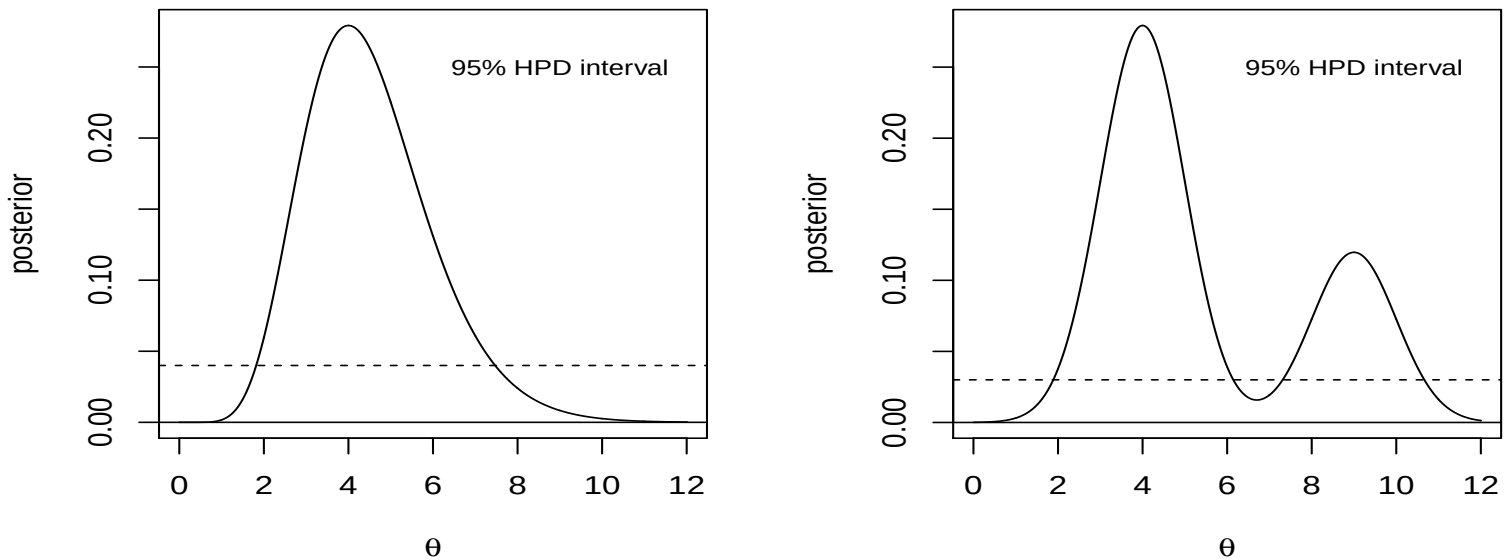
HPD Credible Interval

- Definition: The $100(1-\alpha)\%$ highest posterior density (HPD) credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \ p(\boldsymbol{\theta}|\mathbf{y}) \geq k(\alpha)\} \ ,$$

where $k(\alpha)$ is the largest constant for which

$$P(C|\mathbf{y}) \geq 1-\alpha \ .$$



- An HPD credible set has the smallest volume of all sets of the same α level.

Equal-tail Credible Interval

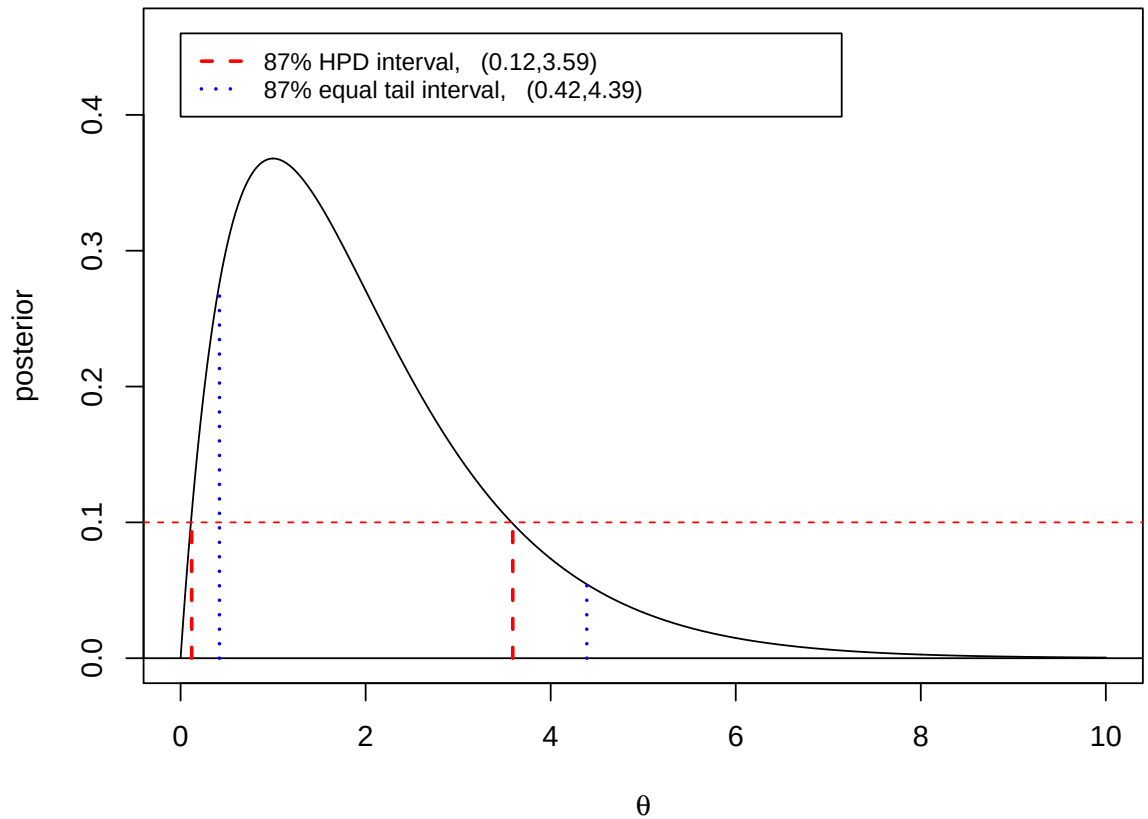
- Simpler alternative: the equal-tail set, which takes the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$.
- Specifically, consider q_L and q_U, the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$:

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2 \text{ and } \int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = \alpha/2\,.$$

  Then clearly $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$; our confidence that θ lies in (q_L, q_U) is $100 \times (1-\alpha)\%$. Thus this interval is a $100 \times (1-\alpha)\%$ credible set for θ.
- This interval is usually wider than HPD interval, but easier to compute (just two quantiles), and also transformation invariant.

Interval Estimation: Example

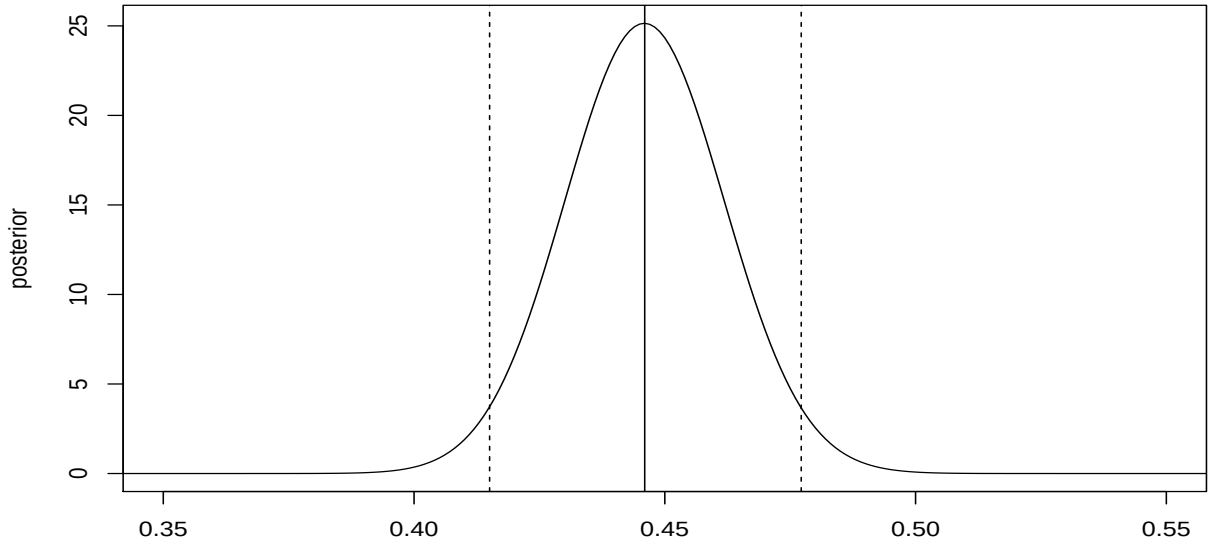
Using a $Gamma(2, 1)$ posterior distribution and $k(\alpha) = 0.1$:



Equal tail intervals do not work well for multi-mode posteriors.

Example: probability of a female birth

$f(X|\theta) = Bin(980, \theta)$, $\pi(\theta) = Beta(1, 1)$, $x_{obs} = 437$



Plot the posterior $Beta(x_{obs} + 1, n - x_{obs} + 1) = Beta(438, 544)$ in R:

```
theta <- seq(from=0, to=1, by=0.01)
xobs <- 437;  n <- 980;
plot(theta,dbeta(theta,xobs+1,n-xobs+1),type="l")
```

Add 95% equal-tail Bayesian CI (dotted vertical lines):

```
abline(v=qbeta(.5, xobs+1, n-xobs+1))
abline(v=qbeta(c(.025,.975),xobs+1,n-xobs+1),lty=2)
```

Bayesian Hypothesis Testing

To test hypothesis of H_0 versus H_1:

- Classical approach bases accept/reject decision on

$$\text{p-value} = P\{T(\mathbf{Y}) \text{ more "extreme" than } T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\} ,$$

  where "extremeness" is in the direction of H_A

- Several *troubles* with this approach:
  - hypotheses must be nested
  - p-value can only offer evidence against the null
  - p-value is not the "probability that H_0 is true" (but is often erroneously interpreted this way)
  - As a result of the dependence on "more extreme" $T(\mathbf{Y})$ values, two experiments with identical likelihoods could result in different p-values, violating the Likelihood Principle!

Bayes Factor

- The quantity commonly used to test hypotheses in Bayesian framework is the Bayes factor:

$$
\begin{aligned}
BF &= \frac{P(H_1|\mathbf{y})/P(H_0|\mathbf{y})}{P(H_1)/P(H_0)} \Leftarrow \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} \\
&= \frac{P(H_1,\mathbf{y})/m(\mathbf{y})/P(H_1)}{P(H_0,\mathbf{y})/m(\mathbf{y})/P(H_0)} \\
&= \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)} \\
&= \frac{\int_{\Theta_{H_0}} p(\mathbf{y}|\boldsymbol{\theta},H_0)\pi(\boldsymbol{\theta}|H_0)d\boldsymbol{\theta}}{\int_{\Theta_{H_1}} p(\mathbf{y}|\boldsymbol{\theta},H_1)\pi(\boldsymbol{\theta}|H_1)d\boldsymbol{\theta}}
\end{aligned}
$$

Bayes Factor vs Likelihood Ratio Test

- Bayes factors can also be written in a similar form to the likelihood ratio test:

$$BF = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)}$$

- We integrate over the parameter space instead of maximizing over it.
- The Bayes factor reduces to a likelihood ratio test in case of a simple vs. simple hypothesis test, i.e. $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta = \theta_1$
- Other advantages of Bayes factor:
  - The BF does NOT require nested models.
  - The BF has a nice interpretation: large values of BF favors H_0.

Interpretation of Bayes Factor

- Possible interpretations

BF	Strength of evidence
1 to 3	barely worth mentioning
3 to 20	positive
20 to 150	strong
> 150	very strong

- These are subjective interpretations and not uniformly agreed upon.

Example: Probability of a female birth

- Data: $x = 437$ out of $n = 980$ placenta previa births were female. We test the hypothesis that $H_0 : \theta \geq 0.485$ vs. $H_1 : \theta < 0.485$.
- Choose the uniform prior $\pi(\theta) = Beta(1, 1)$, and the prior probability of H_1 is
$$P(\theta < 0.485) = 0.485.$$
- The posterior is $p(\theta|x) = Beta(438, 544)$, and the posterior probability of H_1 is
$$P(\theta < 0.485|x = 437) = 0.993$$
- The Bayes factor is
$$BF = \frac{0.993/(1 - 0.993)}{0.485/(1 - 0.485)} = 150.6,$$
strong evidence in favor of H_1, a substantial lower proportion of female births in population of placenta previa births than in the general population.

BP Limitations and Alternatives

- Limitations:
  - NOT well-defined when the prior $\pi(\boldsymbol{\theta}|H)$ is improper
  - may be sensitive to the choice of prior.
- Alternatives:
  - Modified *BF*: partial Bayes factor, fractional Bayes factor
  - Conditional predictive distribution

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \int f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \ ,$$

    which will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is.

  - Penalized likelihood criteria: the Akaike information criterion (AIC), Bayesian information criterion (BIC), or Deviance information criterion (DIC).

Bayesian Prediction

- We are often interested in predicting a future observation, y_{n+1}, given the observed data $\mathbf{y} = (y_1, \ldots, y_n)$. A necessary assumption is exchangability.

- **Exchangability:** Given a parametric model $f(Y|\theta)$, observations $y_1, \ldots, y_n, y_{n+1}$ are conditionally independent, i.e. the joint distribution density $f(y_1, \ldots, y_{n+1})$ is invariate to permutation of the indexes.

- Under the assumption, we can predict a future observation, y_{n+1}, conditional on the observed data

$$p(y_{n+1}|\mathbf{y}) = \int f(y_{n+1}|\theta)p(\theta|\mathbf{y})d\theta \ ,$$

  $p(y_{n+1}|\mathbf{y})$ is known as posterior predictive distribution.

- The frequentist would use $f(y_{n+1}|\widehat{\theta})$ here, which is asymptotically equivalent to $p(y_{n+1}|\mathbf{y})$ above (i.e., when $p(\theta|\mathbf{y})$ is a point mass at $\widehat{\theta}$).

Example: Predicting the sex of a future birth

- Given a $Beta(1,1)$ prior, the posterior of θ is $Beta(438,544)$. The posterior predictive distribution for the sex of a future birth is thus

$$p(y^*|\mathbf{y}) = \int_0^1 \theta^{y^*}(1-\theta)^{1-y^*} \cdot \frac{\Gamma(982)}{\Gamma(438)\Gamma(544)} \theta^{437}(1-\theta)^{543} d\theta$$

- This is known as the beta-binomial distribution. Mean and variance of the posterior predictive distribution can be obtained by

$$E(y^*|\mathbf{y}) = E(E(y^*|\theta,\mathbf{y})|\mathbf{y}) = E(\theta|\mathbf{y}) = 0.446$$

$$\begin{aligned} var(y^*|\mathbf{y}) &= E(var(y^*|\theta,\mathbf{y})|\mathbf{y}) + var(E(y^*|\theta,\mathbf{y})|\mathbf{y}) \\ &= E(\theta(1-\theta)|\mathbf{y}) + var(\theta|\mathbf{y}) \end{aligned}$$

Prior Elicitation

- A Bayesian analysis can be subjective in that two different people may observe the same data $\mathbf{y}$ and yet arrive at different conclusions about $\boldsymbol{\theta}$ when they have different prior opinions on $\boldsymbol{\theta}$.
  - Main criticism from frequentists.
- How should one specify a prior (countering to subjectivity)?
  - Objective and informative: e.g. Historical data, data from pilot experiments.
    "Today's posterior is tomorrow's prior"
  - Noninformative: priors meant to express ignorance about the unknown parameters.
  - Conjugate: posterior and prior belong to the same distribution family.

Noninformative Prior

- Meant to express ignorance about the unknown parameter or have minimal impact on the posterior distribution of $\boldsymbol{\theta}$.
- Also referred as **vague prior** or **flat prior**.
- Example: $\theta =$ true probability of success for a new surgical procedure, $0 \leq \theta \leq 1$. A noninformative prior is $\pi(\theta) = \mathit{Unif}(0, 1)$.
- Example 2.3: $y_1, \ldots y_n \sim N(y_i|\theta, \sigma^2)$, σ is known, $\theta \in \mathtt{R}$. A noninformative prior is $\pi(\theta) = 1, -\infty \leq \theta \leq \infty$.
  - This is an improper prior: $\int_{-\infty}^{\infty} \pi(\theta) d(\theta) = \infty$.
  - An improper prior may or may not lead to a proper posterior.
  - The posterior of θ in Example 2.3 is $p(\theta|\mathbf{y}) = N\left(\bar{y}, \frac{\sigma^2}{n}\right)$, which is proper and is equivalent to the likelihood.

Jeffreys Prior

- Another noninformative prior, given in the univariate case by

$$p(\theta) = [I(\theta)]^{1/2},$$

where $I(\theta)$ is the expected Fisher information in the model, namely

$$I(\theta) = -E_{\mathbf{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right].$$

- Jeffreys prior is improper for many models. It may be proper, however, for certain models.
- Unlike the uniform, the Jeffreys prior is invariant to 1-1 transformations. That is, computing the Jeffreys prior for some 1-1 transformation $\gamma = g(\theta)$ directly produces the same answer as computing the Jeffreys prior for θ and subsequently performing the usual Jacobian transformation to the γ scale.

Conjugate Priors

- Defined as one that leads to a posterior distribution belonging to the same distributional family as the prior:
  - normal prior is conjugate for normal mean
  - beta prior is conjugate for binomial proportion
- Conjugate priors are convenient, computationally, but are rarely possible in complex settings
  - In higher dimensions, priors that are conditionally conjugate are often available (and helpful).
- We can often guess the conjugate prior by looking at the likelihood as a function of θ.

Another Example of Conjugate Prior

- Suppose that X is distributed Poisson(θ), so that

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

- A reasonably flexible prior for θ is the *Gamma*(α, β) distribution,

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha},\ \theta > 0, \alpha > 0,\ \beta > 0,$$

- The posterior is then

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \\ &\propto \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}. \end{aligned}$$

There is one and only one density proportional to the very last function, *Gamma*$(x + \alpha, (1 + 1/\beta)^{-1})$ density. Gamma is the conjugate family for the Poisson likelihood.

Common Conjugate Families

Likelihood	Conjugate Prior
$Binomial(N, \theta)$	$\theta \sim beta(\alpha, \lambda)$
$Poisson(\theta)$	$\theta \sim gamma(\delta_0, \gamma_0)$
$N(\theta, \sigma^2)$, σ^2 is known	$\theta \sim N(\mu, \tau^2)$
$N(\theta, \sigma^2)$, θ is known	$\tau^2 = 1/\sigma^2 \sim gamma(\delta_0, \gamma_0)$
$Exp(\lambda)$	$\lambda \sim gamma(\delta_0, \gamma_0)$
$MVN(\boldsymbol{\theta}, \Sigma), \Sigma$ is known	$\boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}, V)$
$MVN(\boldsymbol{\theta}, \Sigma), \boldsymbol{\theta}$ is known	$\Sigma \sim Inv - Wishart(\nu, V)$

Prior Distribution: Summary

- The prior distribution plays an important role in a Bayesian analysis.
- People generally prefer non or minimally informative priors, in practice, but this is not as simple as it sounds.
- Auxillary data are often available but can overwhelm the data, in some cases
  - Will see in the lab practice ...