Chapter 1.4

Summarizing a posterior

Summarizing a univariate posterior

- After selecting the likelihood and prior, all that remains is to summarize the posterior

- Say there is a single parameter, θ

- For example, we say the model is

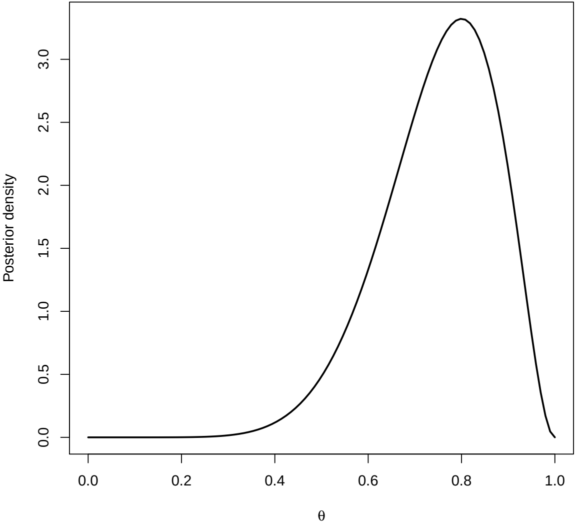
  Likelihood: $Y|\theta \sim \text{Binomial}(N, \theta)$
  Prior: $\theta \sim \text{Uniform}(0, 1)$

- We saw that the posterior is then

$$\theta|Y \sim \text{Beta}(Y + a, N - Y + b)$$

- The posterior is a *distribution* that can be plotted as on the next slide

Summarizing a univariate posterior



In this beta/binomial example $Y = 8$, $N = 10$ and $a = b = 1$

Summarizing a univariate posterior

- A plot of the posterior tells the whole story
- However, to be more concise we typically use a few numerical summaries of the distribution
- This is particularly important when there are many parameters
- The posterior can be summarized like any other distribution, by say the mean, variance, skewness, etc.

Point estimators

- A **point estimator** is a one number summary used to estimate the unknown parameter
- For example, we might use the posterior mean (or median) as the “best guess” of θ
- The posterior mean is

$$\hat{\theta} = \mathrm{E}(\theta|Y) = \int \theta p(\theta|Y) d\theta$$

- For the Beta/Binomial example $\hat{\theta} = \frac{Y+a}{n+a+b}$
- This is an alternative to the sample proportion $\hat{\theta} = Y/n$
- Estimators usually wear hats

MAP estimator

- The posterior mode is the called the maximum a posteriori (MAP) estimator
- The MAP estimator is

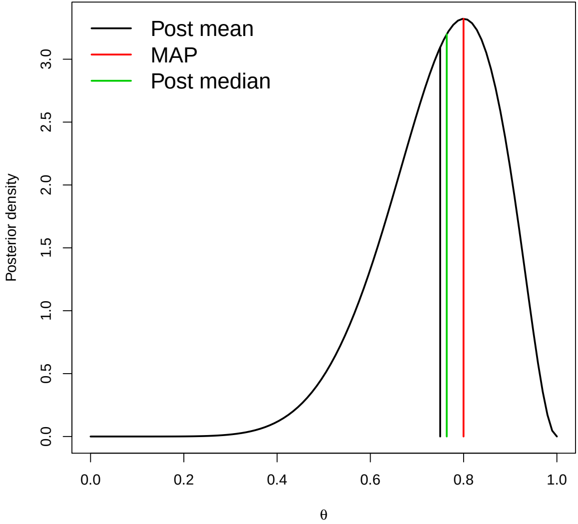
$$\hat{\theta} = \arg\max_{\theta} p(\theta|Y)$$

- Maximizing the log of the posterior is equivalent and easier, so we will usually use this for computation
- The MAP estimator is

$$\hat{\theta} = \arg\max_{\theta} \log[p(\theta|Y)] = \arg\max_{\theta} \log[f(Y|\theta)] + \log[\pi(\theta)]$$

- If the prior is uniform (i.e., flat) the MAP is the MLE
- The MAP is easier to compute than the posterior mean

MAP estimator



In this beta/binomial example $Y = 8$, $N = 10$ and $a = b = 1$

MAP estimator

Assuming $Y|\theta \sim Binomial(n, \theta)$ and $\pi(\theta) = 1$, find the MAP estimator of θ

- The likelihood is $f(Y|\theta) = \binom{n}{Y}\theta^{Y}(1-\theta)^{n-Y}$
- The log likelihood is

$$\log[f(Y|\theta)] = \log\left[\binom{n}{Y}\right] + Y\log(\theta) + (n - Y)\log(1-\theta)$$

- The prior is $\pi(\theta) = 1$
- The log prior is $\log[\pi(\theta)] = 0$
- Therefore, the MAP estimator is

$$\hat{\theta} = \arg\max_{\theta} \log[p(\theta|Y)] = \log\left[\binom{n}{Y}\right] + Y\log(\theta) + (n-Y)\log(1-\theta)$$

MAP estimator

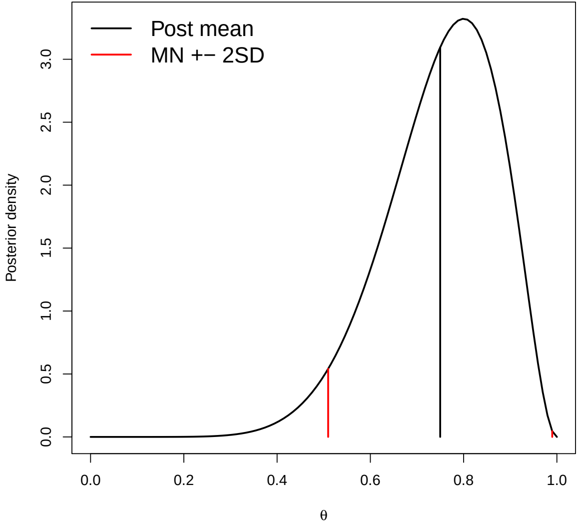
Assuming $Y|\theta \sim Binomial(n, \theta)$ *and* $\pi(\theta) = 1$*, find the MAP estimator of* θ

- To find the MAP estimator we take a derivative, set it so zero and solve
- $\frac{d}{d\theta} \log[p(\theta|Y)] = \frac{Y}{\theta} - \frac{n-Y}{1-\theta} = 0$
- Solving for θ gives the MAP estimator is $\hat{\theta} = Y/n$
- This is the sample proportion, which is also the MLE

Uncertainty measures

- Sometimes a point estimate is sufficient, but more often we need to quantify uncertainty
- The **posterior standard deviation** is one measure of uncertainty
- If the posterior is approximately normal, then the mean plus/minus two standard deviation units captures 95% of the posterior probability
- The posterior standard deviation is analogous to but fundamentally different than the frequentist **standard error**
- The standard error is the standard deviation of $\hat{\theta}$'s sampling distribution

Uncertainty measures



In this beta/binomial example $Y = 8$, $N = 10$ and $a = b = 1$

Credible intervals

- In addition to standard error, uncertainty can be quantified using a **credible interval**
- The interval (l, u) is a $100(1-\alpha)\%$ posterior credible interval/set if

$$\text{Prob}(l < \theta < u | Y) = 1 - \alpha$$

- Interpretation of a 95% credible interval: "given the data and prior, I am 95% certain that θ is between l and u"
- This is analogous but different than a **confidence interval**

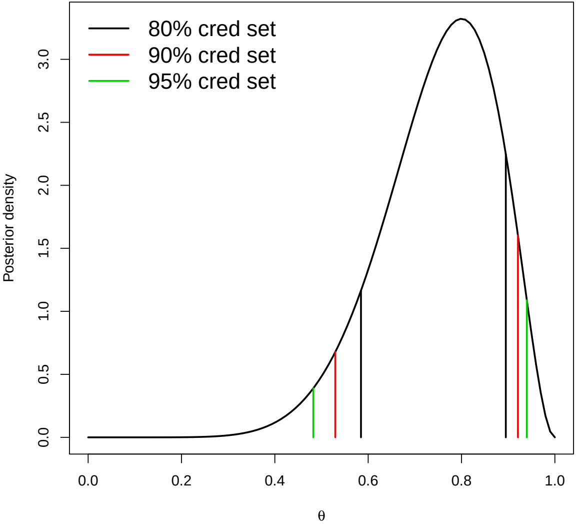
Credible intervals

- Credible sets are not unique
- Let q_τ be the τ quantile of the posterior so that

$$\text{Prob}(\theta < q_\tau | Y) = \tau$$

- Then $(q_{0.00}, q_{0.95})$, $(q_{0.01}, q_{0.96})$, etc. are all valid 95% credible sets
- The **equal-tailed** interval is $(q_{\alpha/2}, q_{1-\alpha/2})$
- The **highest posterior density** interval searches for the smallest interval that contains the proper probability

Credible intervals



In this beta/binomial example $Y = 8$, $N = 10$ and $a = b = 1$

Hypothesis testing

- **Hypothesis tests** are conducted by simply computing the posterior probability of each hypothesis
- Say the null hypothesis is $\mathcal{H}_0 : \theta \leq 0.5$ and the alternative is $\mathcal{H}_1 : \theta > 0.5$
- The posterior probability of the null hypothesis is

$$\text{Prob}(\theta < 0.5|Y) = \int_0^{0.5} p(\theta|Y)d\theta$$

- We reject the null if its probability is small
- In a Bayesian analysis we can say "Given the data and prior the probability that the null hypothesis is true is 0.02"
- This is analogous to but different than the **p-value**

Hypothesis testing

```
> # Data
> Y <- 8; n <- 10
> # The posterior is theta|Y~Beta(A,B)
> A <- Y+1; B <- n-Y+1
> # Posterior mean
> A/(A+B)
[1] 0.75
> # Posterior standard deviation
> sqrt(A*B/((A+B)*(A+B)*(A+B+1)))
[1] 0.1200961
> # Posterior 95% credible interval
> qbeta(c(0.025,0.975),A,B)
[1] 0.4822441 0.9397823
> # Posterior probability that theta<0.5
> pbeta(0.5,A,B)
[1] 0.03271484
```

Monte Carlo approximations

- **Monte Carlo (MC) sampling** is a useful tool for summarizing a posterior
- For univariate cases is it not particularly useful, but in harder problems is the best approach available
- In MC sampling we draw S samples from the posterior,

  $$\theta^{(1)}, ..., \theta^{(S)} \sim p(\theta|Y)$$

  and use these samples to approximate the posterior
- For example, the posterior mean and variance are approximated by the sample mean and variance of the $\theta^{(s)}$

Monte Carlo approximations

- MC sampling facilitates studying **transformations** of parameters

- For example, the odds corresponding to θ are $\gamma = \theta/(1 - \theta)$

- How to approximate the posterior mean and variance of γ?

- We simply transform each draw to the odds

$$\gamma^{(1)} = \frac{\theta^{(1)}}{1 - \theta^{(1)}}, \ldots, \gamma^{(S)} = \frac{\theta^{(S)}}{1 - \theta^{(S)}}$$

  and use these draws to approximate γ's posterior

Monte Carlo approximations

```
> # Data
> Y <- 8; n <- 10
> # The posterior is theta|Y~Beta(A,B)
> A <- Y+1; B <- n-Y+1
> # MC sampling
> theta <- rbeta(100000,A,B)
> # Approximate the posterior mean and SD
> mean(theta);sd(theta)
[1] 0.749792
[1] 0.1201799
> # Transform to odds
> gamma <- theta/(1-theta)
> # Approximate the posterior mean and SD
> mean(gamma);sd(gamma)
[1] 4.483378
[1] 4.720541
```

Summarizing multivariate posteriors

- A univariate posterior is captured by simple plot
- When there are many parameters this is impossible
- Say $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$
- Ideally we reduce to the univariate marginal posteriors

$$p(\theta_1|Y) = \int ... \int p(\theta_1, ..., \theta_p|Y)d\theta_2, ..., d\theta_p$$

- The same ideas we used for univariate models then apply
- However, computing these integrals is often challenging

Bayesian one-sample t-test

- In this section we will study the one-sample t-test in depth
- Likelihood: $Y_i|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ independent over $i = 1, ..., n$
- Priors: $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ independent of $\sigma^2 \sim \text{InvGamma}(a, b)$
- The joint (bivariate PDF) of (μ,σ^2) is proportional to

$$\left\{\sigma^n \exp\left[-\frac{\sum_{i=1}^n (Y_i-\mu)^2}{2\sigma^2}\right]\right\}\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right](\sigma^2)^{a-1}\exp(-\frac{b}{\sigma^2})$$

- How to summarize this complicated function?

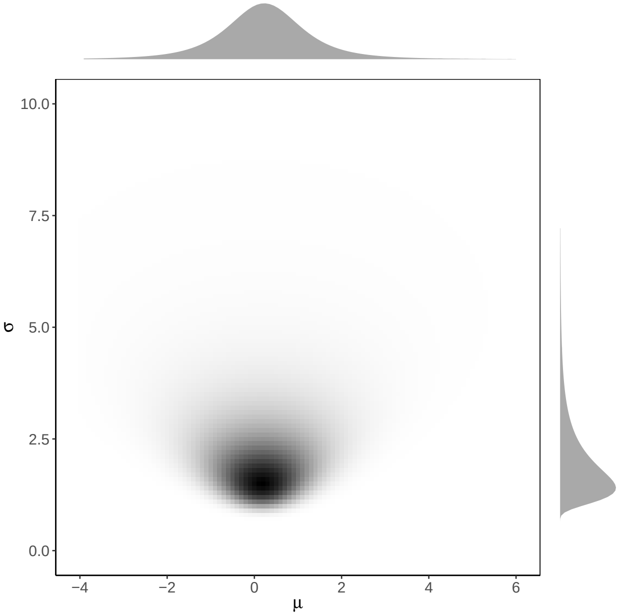
Plotting the posterior on a grid

- For models with only a few parameters we could simply plot the posterior on a grid
- That is, we compute $p(\mu, \sigma^2 | Y_1, ..., Y_n)$ for all combinations of m values of μ and m values of σ^2
- The number of grid points is m^p where p is the number of parameters in the model
- The posterior is plotted on the next slide for

  $Y_1 = 2.68$, $Y_2 = 1.18$, $Y_3 = -0.97$, $Y_4 = -0.98$, $Y_5 = -1.03$

  and uniform priors over the plotting window

Bivariate posterior



Summarizing the results in a table

- Typically we are interested in the marginal posterior

$$f(\mu|\mathbf{Y}) = \int_0^\infty p(\mu, \sigma^2|\mathbf{Y})d\sigma^2$$

  where $\mathbf{Y} = (Y_1, ..., Y_n)$

- This accounts for our uncertainty about σ^2
- We could also report the marginal posterior of σ^2
- Results are usually given in a table with marginal mean, SD, and 95% interval for all parameters of interest
- The marginal posteriors can be computed using numerical integration

Summarizing the results in a table

	Posterior mean	Posterior SD	95% credible set
μ	0.17	1.31	(-2.49, 2.83)
σ	2.57	1.37	(1.10, 6.54)

Frequentist analysis of a normal mean

- In frequentist statistics the estimate of the mean is $\bar{Y}$
- If σ is known the 95% interval is

$$\bar{Y} \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$$

where z is the quantile of a normal distribution

- If σ is unknown the 95% interval is

$$\bar{Y} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$$

where t is the quantile of a t-distribution

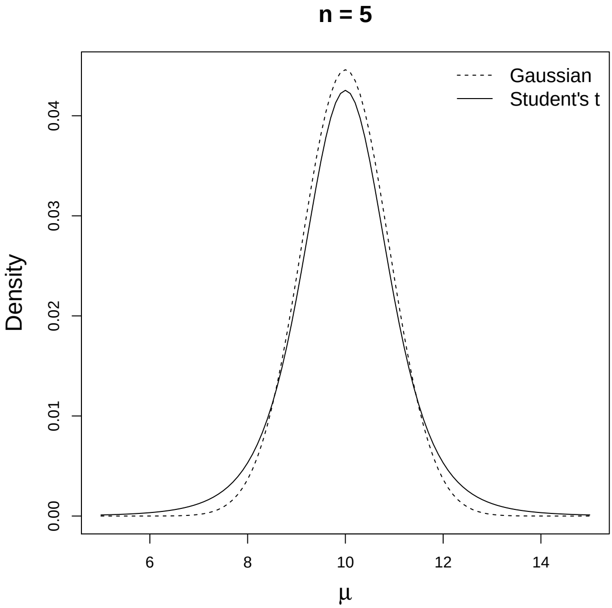
Bayesian analysis of a normal mean

- The Bayesian estimate of μ is its marginal posterior mean
- The interval estimate is the 95% posterior credible interval
- If σ is known the posterior of $\mu|\mathbf{Y}$ is Gaussian and the 95% interval is
$$\mathrm{E}(\mu|\mathbf{Y}) \pm z_{0.975}\mathrm{SD}(\mu|\mathbf{Y})$$
- If σ is unknown the marginal (over σ^2) posterior of μ is t with $\nu = n + 2a$ degrees of freedom.
- Therefore the 95% interval is
$$\mathrm{E}(\mu|\mathbf{Y}) \pm t_{0.975,\nu}\mathrm{SD}(\mu|\mathbf{Y})$$
- See "Marginal posterior of μ" the online derivations

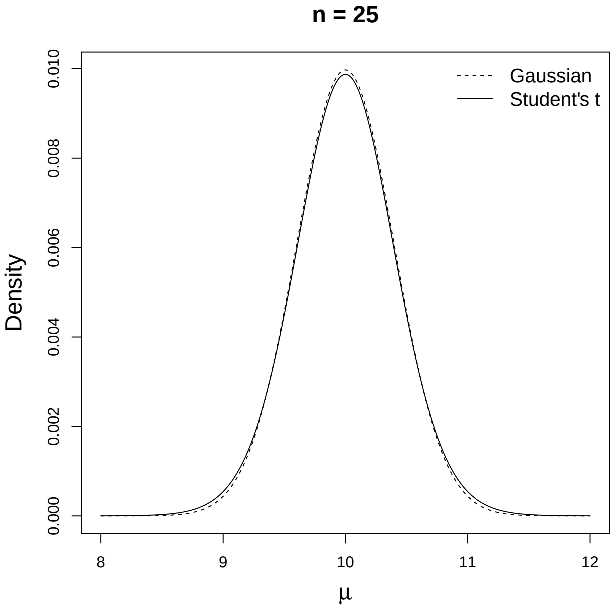
Bayesian analysis of a normal mean

- The following two slides give the posterior of μ for a data set with sample mean 10 and sample variance 4
- The Gaussian analysis assumes $\sigma^2 = 4$ is known
- The t analysis integrates over uncertainty in σ^2
- As expected, the latter interval is a bit wider

Bayesian analysis of a normal mean



Bayesian analysis of a normal mean



Example with two parameters

- Assume $N = 10$ patients are given treatment and $N = 10$ are given control
- Let θ_1 and θ_2 be the survival probabilities for the two treatment groups
- We pick priors $\theta_1, \theta_2 \sim \text{Uniform}(0, 1)$
- We observe $Y_1 = 5$ survivals and $Y_2 = 8$ survivals for treatment, respectively
- Our goal is to determine if treatment improves survival, i.e., $\theta_2 > \theta_1$
- This requires summarizing the bivariate posterior $p(\theta_1, \theta_2 | Y_1, Y_2)$

Methods for dealing with multiple parameters

- We want to compute Prob$(\theta_2 > \theta_1 | Y_1, Y_2)$
- We could to an integral or grid approximation
- Both often fail when there are many parameters
- We need new tools!
- Monte Carlo sampling will be a key tool
- We'll spend a month on this in the computing section

Comparing proportions

- The model is
  - $Y_1|\theta_1 \sim \text{Binomial}(N, \theta_1)$
  - $Y_2|\theta_2 \sim \text{Binomial}(N, \theta_2)$
  - $\theta_1, \theta_2 \sim \text{Beta}(1, 1)$

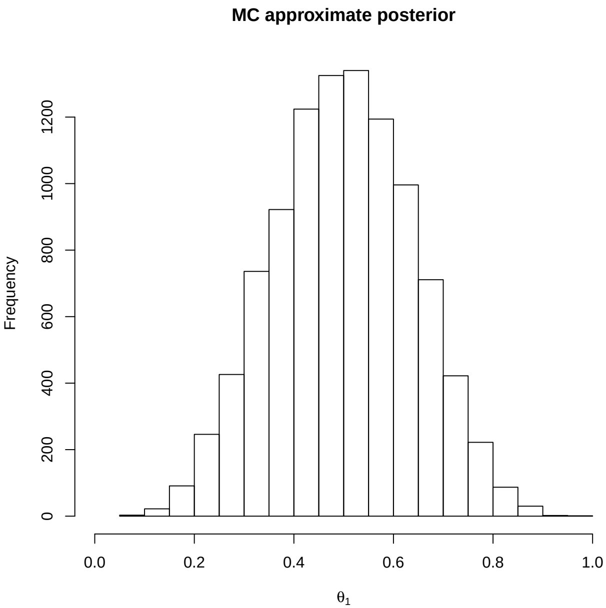
- The posterior is $\theta_1 | Y_1, Y_2 \sim Beta(Y_1 + 1, N - Y_1 + 1)$ independent of $\theta_2 | Y_1, Y_2 \sim Beta(Y_2 + 1, N - Y_2 + 1)$

- The next slide approximates posterior probability that $\theta_2 < \theta_1$ using MC sampling

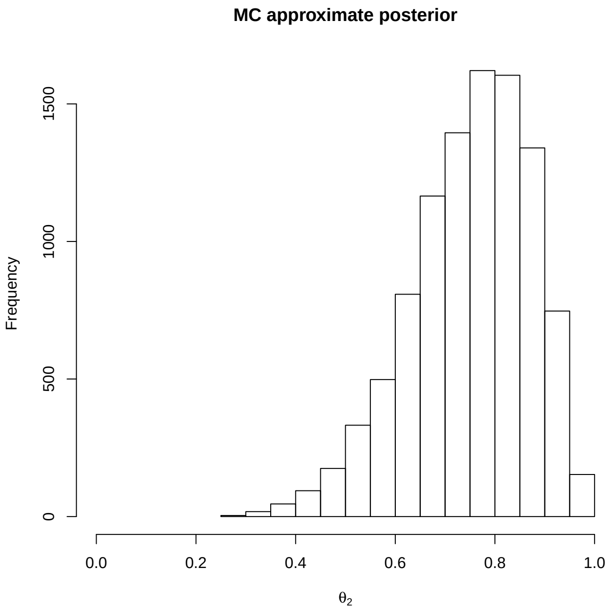
Summarizing a posterior using MC sampling

```
> N <- 10; Y1 <- 5; Y2 <- 8 # Data
> S      <- 10000   # Number of MC samples
> theta1 <- rbeta(S,Y1+1,N-Y1+1)
> theta2 <- rbeta(S,Y2+1,N-Y2+1)
>
> hist(theta1,xlab=expression(theta[1]),
+      main="MC approximate posterior")
> (Y1+1)/(N+2)# True post mean
[1] 0.5
> mean(theta1) # MC estimate
[1] 0.499631
>
> plot(theta1,theta2,xlim=c(0,1),ylim=c(0,1),
+      xlab=expression(theta[1]),
+      ylab=expression(theta[2]))
> mean(theta2>theta1)
[1] 0.9058
```

Marginal posterior, $p(\theta_1 | Y_1, Y_2)$



Marginal posterior, $p(\theta_2 | Y_1, Y_2)$



Joint posterior, $p(\theta_1, \theta_2 | Y_1, Y_2)$