Section 3

Bayes basics

Outline

These slides cover Chapters 1.3-1.5

- Understand Bayesian learning, i.e., updating the prior using data
- Graphically summarize a univariate posterior distribution
- Summarize a multivariate posterior distribution
- Use Monte Carlo sampling to approximate posteriors
- Compute the posterior predictive distribution

A visit to the doctor's office

- Say you have a scratchy throat and so you go to the doctor to be tested from strep throat
- Denote your true disease status as $\theta = 1$ if you have strep throat and $\theta = 0$ otherwise
- An unknown quantity such as θ that we hope to estimate is called a **parameter**
- Unless the doctor's test is perfect, we will never know θ exactly

A visit to the doctor's office

- Let the **data** Y be the result of the rapid strep test with $Y = 1$ if you test positive and $Y = 0$ otherwise

- The distribution of the data given the parameters is called the **likelihood**

- In this example the likelihood is defined by the false positive rate
$$\text{Prob}(Y = 1|\theta = 0) = p$$
and the false negative rate
$$\text{Prob}(Y = 0|\theta = 1) = q$$

- For now, we assume that we know p and q based on previous analyses

A visit to the doctor's office

- Say you test positive and $Y = 1$, how likely is it that you have strep?

- To formalize this problem statistically, we must first decide whether Y, θ or both are random variables

- Should we treat Y as random?

Y is a random variable e.g.

$$p(Y=1 \mid \theta=0) = p$$

After collecting data we know $Y=1$ & it is fixed

A visit to the doctor's office

- Should we treat θ as random?

θ is fixed
but we are uncertain about θ
so we treat it as a
random variable

Bayesian learning

- Bayesians quantify uncertainty about fixed but unknown parameters by treating them as random variables
- This requires that we set a **prior distribution** $\pi(\theta)$ to summarize uncertainty before observing the data
- The distribution of the observed data given the model parameters is the **likelihood function**, $f(Y|\theta)$
- The likelihood function is the most important piece of a Bayesian analysis because it links the data and parameters

Bayesian learning

- The **posterior distribution** $p(\theta|Y)$ summarizes uncertainty about the parameters given the prior and data
- The reduction in uncertainty from prior to posterior represents **Bayesian learning**
- **Bayes' Theorem** (Bayes' Rule) converts the likelihood and prior to the posterior
- Bayes' Theorem:

$$p(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{m(Y)}$$

where $m(Y) = \int f(Y|\theta)\pi(\theta)d\theta$ is the marginal distribution of the data and can usually be ignored

How to select the prior?

- There is no “true” or “correct” prior
- In some cases expert opinion or similar studies can be used to specify an informative prior
- It would be a waste to discard this information
- If prior information is unavailable, then the prior should be uninformative
- The prior is best viewed as an initial value to a statistical procedure

How to select the prior?

- As we'll see, as Bayesian learning continues and more and more data are collected, the posterior concentrates around the true value for any reasonable prior
- However, in finite sample the prior can have some effect
- We will study several systematic ways to select priors
- However, there is inherent **subjectivity** to selecting the prior
- That is, different analysts may pick different priors and thus have different results

How to select the likelihood?

- The likelihood is the same as in a freqentist analysis
- For example, in a linear regression analysis we might say

$$Y_i|\boldsymbol{\beta}, \sigma^2 \sim \text{Normal}\left(\sum_{j=1}^{p} X_{ij}\beta_j, \sigma^2\right)$$

- Is there “true” or “correct” likelihood?
- Is specification of the likelihood subjective like the prior?

How to select the likelihood?

Subjective decisions required to specify the likelihood:

1. which predictors? (expert opinion?)
2. which hypothesis to test
3. What is the measure y?
4. how to sample?
5. for x_j
6. transformations of x (x^2)
7. transformations of y ($\log(y)$)
8. Normal errors?
12. p-value threshold
9. n
11. Missing data
10. Outliers

others

prior

My opinions about subjectivity

- Perhaps we should aspire to objectivity, but in most real-life analyses we are forced to accept some subjectivity
- A notable exception is a tightly controlled experiment, although even this is debatable
- However, not all subjective decisions (assumptions) are equal
- If readers disagree with your assumptions they will reject your findings
- It is your job to justify your assumptions theoretically and empirically
- Determining the sensitivity to key assumptions is an important step

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

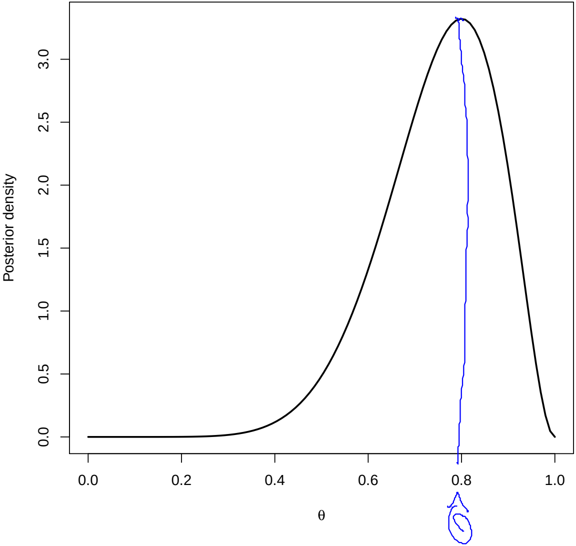
  Likelihood: $Y|\theta \sim \text{Binomial}(N, \theta)$
  Prior: $\theta \sim \text{Uniform}(0, 1)$

- 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.

Summarizing a univariate posterior

- 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

Summarizing a univariate posterior

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

$$p(\theta|Y) = p(Y|\theta)\,p(\theta)/p(Y)$$

- The MAP estimator is

$$\log(p(\theta|Y)) = \log p(Y|\theta) + \log p(\theta) - \log p(Y)$$

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

- If the prior is uniform (i.e., flat) the MAP is the MLE

- The MAP is easier to compute than the posterior mean

Summarizing a univariate posterior

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

$$f(y|\theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y}$$

$$\ell(\theta) = \log f(y|\theta) + \log(p(\theta))$$

$$= \log\binom{n}{y} + \log(\theta^{y}) + \log((1-\theta)^{n-y}) + 0$$

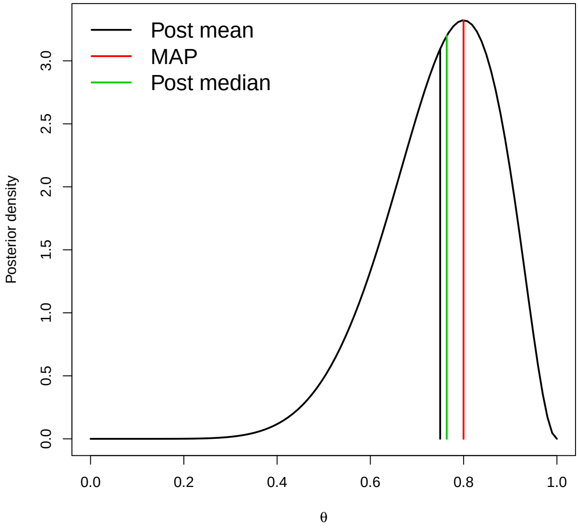
$$= \log\binom{n}{y} + y\log(\theta) + (n-y)\log(1-\theta)$$

$$\ell'(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta} = 0$$

...

$$\theta = \frac{y}{n}$$

Summarizing a univariate posterior



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

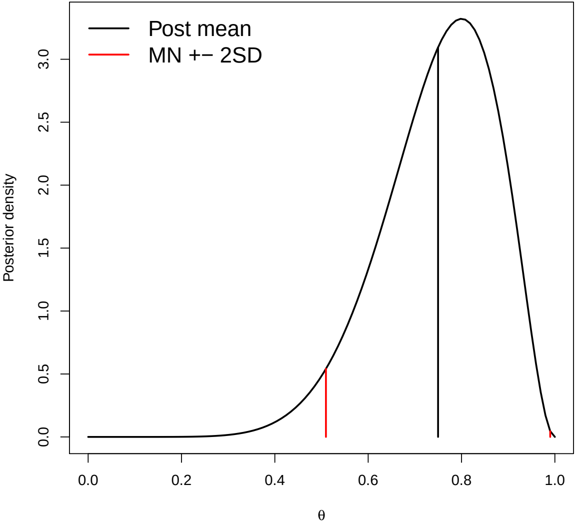
Summarizing a univariate posterior

- Sometimes a point estimate is sufficient, but more often we need to quantify uncertainty
- The **posterior standard deviation** is one measure of uncertainty

  quantifies uncertainty in θ give the observed data.
- If the posterior is approximately normal, then the mean plus/minus two standard deviation units captures 95% of the posterior probability
- The posterior SD ~~mean~~ is analogous to but fundamentally different than the frequentist **standard error**
- The standard error is the standard deviation of $\hat{\theta}$'s sampling distribution

  quantifies randomness in Y over repeated experiments

Summarizing a univariate posterior



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

Summarizing a univariate posterior

- 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 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**

Summarizing a univariate posterior

- Credible sets are not unique

- Let q_τ be the τ quantile of the posterior so that

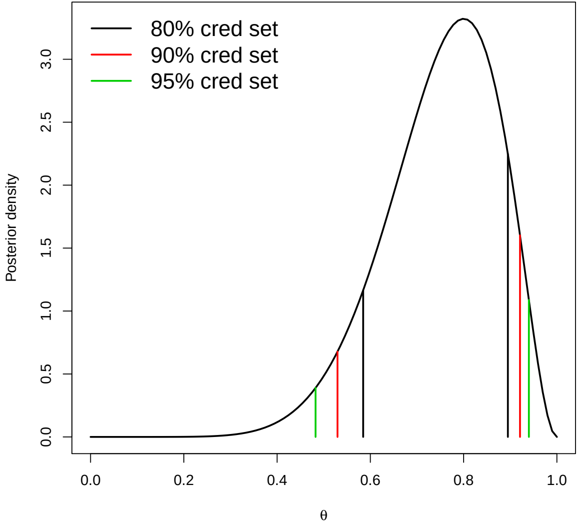
$$\text{Prob}(\theta < q_\tau | Y) = \tau$$

- 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

Summarizing a univariate posterior



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

Summarizing a univariate posterior

- **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**

Summarizing a univariate posterior

```
> # 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
```

Summarizing a univariate posterior

- **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 $\theta^{(s)}$

Summarizing a univariate posterior

- 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

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

  and use these draws to approximate γ's posterior

Summarizing a univariate posterior

```
> # 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
```

Homework

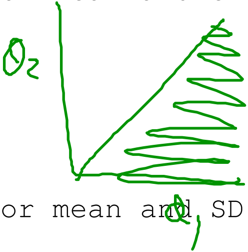
$Y_1|\theta_1 \sim Bin(n,\theta_1)$ $\theta_1 \sim Beta$

$Y_2|\theta_2 \sim Bin(n,\theta_2)$ $\theta_2 \sim Beta$

$P(\theta_1 > \theta_2 | Y_1, Y_2)$

① Draw $\theta_1^{(s)}$ + $\theta_2^{(s)}$

② Compute proportion of draws w $\theta_1^{(s)} > \theta_2^{(s)}$



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: $Y_i|\mu, \sigma \sim \mathrm{N}(\mu, \sigma^2)$ independent over $i = 1, ..., n$
- Priors: $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ independent of $\sigma^2 \sim \mathrm{InvGamma}(a, b)$

  $p(Y_1, ... Y_n | \mu, \sigma^2)$
- 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

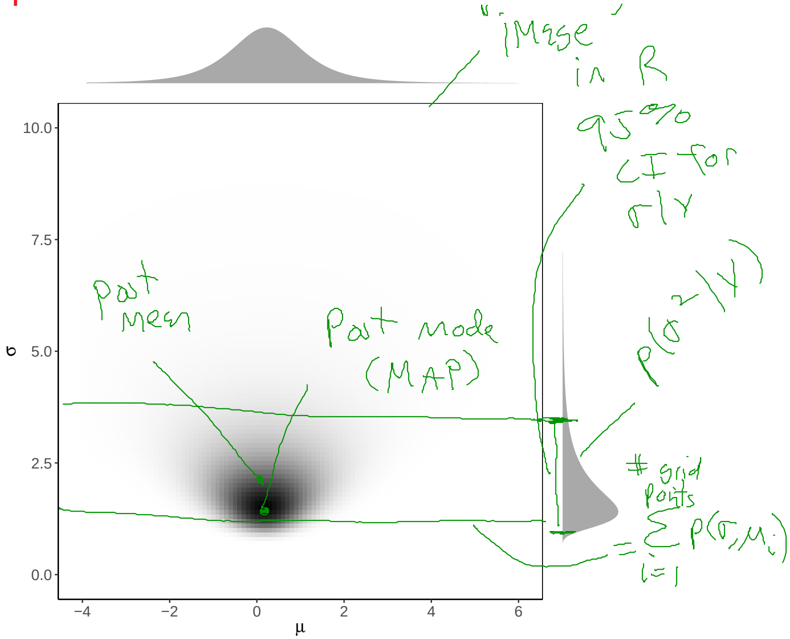
- The number of grid points is m^p where p is the number of parameters in the model

- 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

$p(\mu|Y) = \int p(\mu, \sigma^2) d\sigma^2$

	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)

$p(\sigma|y)$

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

Sample SD

- 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

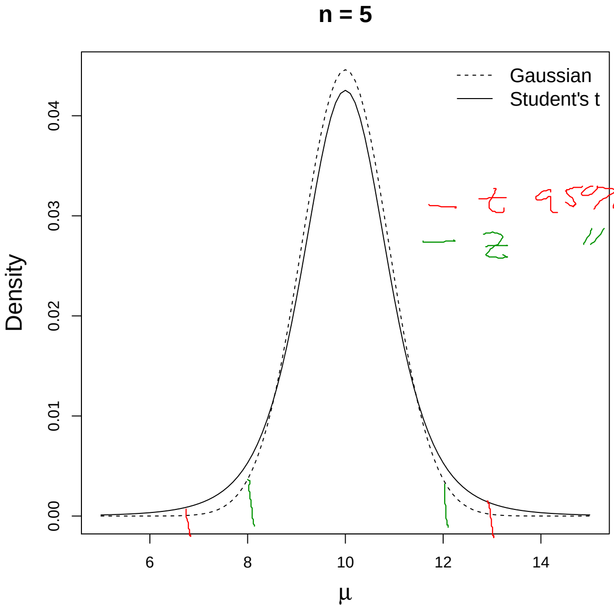
Bayesian analysis of a normal mean

- The Bayesian estimate of μ is its marginal posterior mean
- The interval estimate is the 95% posterior 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

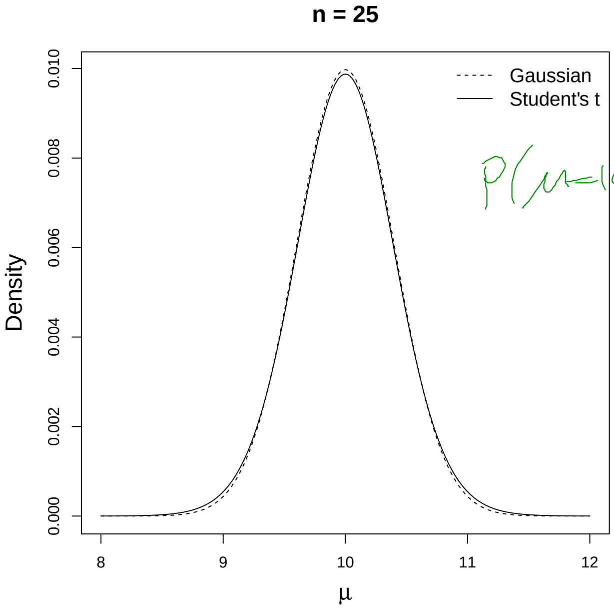
Bayesian analysis of a normal mean



— t 95% cred set

— z ″

Bayesian analysis of a normal mean



$P(\mu = 10 \mid Y) = 0$

Bayesian one sample t-test

- The one-sided test of $H_1 : \mu \leq 0$ versus $H_2 : \mu > 0$ is conducted by computing the posterior probability of each hypothesis
- This is done with the `pt` function in `R`

point null

- The two-sided test of $H_1 : \mu = 0$ versus $H_2 : \mu \neq 0$ is conducted by either
  - Determining if 0 is in the 95% posterior interval
  - Bayes factor (later)

Methods for dealing with multiple parameters

- In this case, we were able to compute the marginal posterior in closed form (a t distribution)
- We were also able to compute the posterior on a grid
- For most analyses the marginal posteriors will not be a nice distributions, and a grid is impossible if there are many parameters
- We need new tools!
- We'll spend a month on this in the computing section

Bayesian prediction

- Often the objective is to predict a future event
- Example: Last spring we planted $n = 10$ seedlings and $Y = 2$ survived the winter, if we plant n again this year what is the probability at least one will survive the winter?
- Let Y^* be the predicted value and θ be the true survival probability
- If the parameters were known then we would predict

  $$Y^*|\theta \sim \text{Binomial}(10, \theta)$$

  and thus $\text{Prob}(Y > 0) = 1 - (1 - \theta)^{10}$
- Of course, if we knew the parameters we would be doing probability and not statistics

Bayesian prediction

- One approach for accounting for parametric uncertainty is a "plug-in" approach
- That is, if $\hat{\theta}$ is an estimate, then $Y^* \sim f(Y|\hat{\theta})$
- Example: $\hat{\theta} = 2/10$ and Prob$(Y > 0) = 1 - (1 - 0.2)^{10}$
- If $\hat{\theta}$ has small uncertainty this is fine
- Otherwise, this underestimates uncertainty in Y^*

Bayesian prediction

- For the sake of prediction, the parameters are not of interest
- They are vehicles by which the data inform about the predictive model
- The **Posterior Predictive Distribution** (PPD) averages over their posterior uncertainty

$f(Y^*, \theta | Y)$

$$f(Y^*|Y) = \int f(Y^*|\theta) f(\theta|Y) d\theta$$

- This properly accounts for parametric uncertainty
- The input is data, the output is a prediction distribution

Bayesian prediction

- Monte Carlo sampling approximates the PPD
- Say $\theta^{(1)}, ..., \theta^{(S)}$ are samples from the posterior
- If we make a sample for Y^* for each $\theta^{(s)}$,

$$Y^{*(s)} \sim f(Y|\theta^{(s)})$$

  then the $Y^{*(s)}$ are samples from the PPD

- The posterior predictive mean is approximated by the sample mean of the $Y^{*(s)}$
- The probability that $Y^* > 0$ is approximated by the sample proportion of the $Y^{*(s)}$ that are non-zero

Bayesian prediction

```
> # Data
> Y <- 2; n <- 10

> # The posterior is theta|Y~Beta(A,B)
> A <- Y+1; B <- n-Y+1
>
> # Plug in estimate of P(Ystar>0)
> 1-dbinom(0,10,.2)
[1] 0.8926258
>
> # Approximate the PPD using MC sampling
> theta <- rbeta(100000,A,B)
> Ystar <- rbinom(100000,10,theta)
> mean(Ystar>0)
[1] 0.87454
```

$\hat{\theta} = 0.2$

Bayesian prediction