STA 321
Spring 2016

Lecture 7

Thursday, Feb 4th

Bayes Theorem

- Bayesian statistics named after Rev. Thomas Bayes (1702-1761)
- Bayes Theorem for probability events A and B

$$p(A \mid B) = \frac{p(B \mid A)\,p(A)}{p(B)}$$

- Or for a set of mutually exclusive and exhaustive events (i.e. $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i \mid B) = \frac{p(B \mid A_i)\,p(A_i)}{\sum_j p(B \mid A_j) P(A_j)}$$

Bayesian Inference

In Bayesian inference there is a fundamental distinction between

- Observable quantities *x,* i.e. the data
- Unknown quantities θ

θ can be statistical parameters, missing data, latent variables…

- Parameters are treated as random variables

In the Bayesian framework we make probability statements about model parameters

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

Prior distributions

As with all statistical analyses we start by positing a model which specifies $p(x|\theta)$

This is the **likelihood** which relates all variables into a '**full probability model**'

However from a Bayesian point of view :

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data
- Therefore we specify a **prior distribution** $p(\theta)$

Note this is like the prevalence in the example

Posterior Distributions

Also *x* is known so should be conditioned on and here we use Bayes theorem to obtain the conditional distribution for unobserved quantities given the data which is known as the **posterior distribution**.

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

The prior distribution expresses our uncertainty about θ **before** seeing the data.

The posterior distribution expresses our uncertainty about θ **after** seeing the data.

Conjugate posterior and prior

- When the posterior is in the same family as the prior we have *conjugacy*. Examples include:

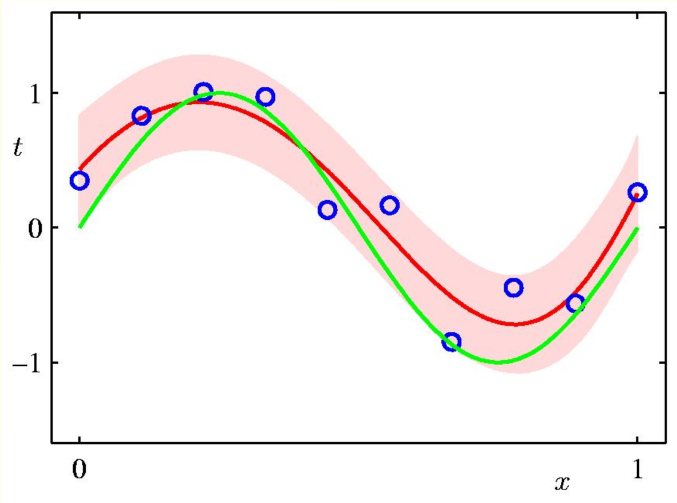
Likelihood	Parameter	Prior	Posterior
Normal	Mean	Normal	Normal
Normal	Precision	Gamma	Gamma
Binomial	Probability	Beta	Beta
Poisson	Mean	Gamma	Gamma

Parametric Distributions

- Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$
- Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Representation: $\boldsymbol{\theta}^\star$ or $p(\boldsymbol{\theta})$?

- Recall Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}$$



Binary Variables (1)

- Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

- Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

Binary Variables (2)

- N coin flips:

$$p(m \text{ heads}|N, \mu)$$

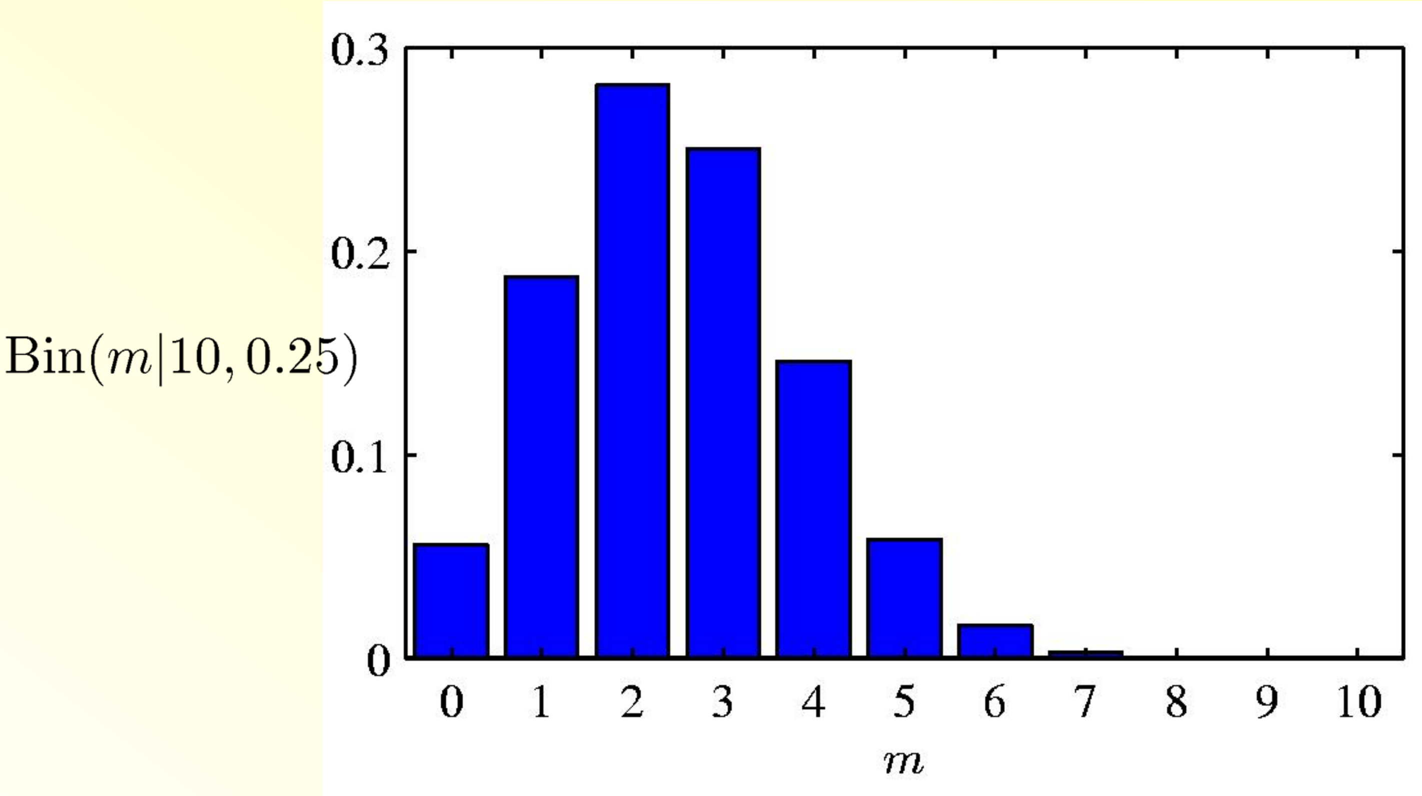
- Binomial Distribution

$$\mathrm{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N, \mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \,\mathrm{Bin}(m|N, \mu) = N\mu(1-\mu)$$

Binomial Distribution



Parameter Estimation (1)

- ML for Bernoulli
- Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)
- $$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$

Parameter Estimation (2)

- Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$
- Prediction: *all* future tosses will land heads up

- Overfitting to D

Beta Distribution

- Distribution over $\mu \in [0, 1]$.

$$\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

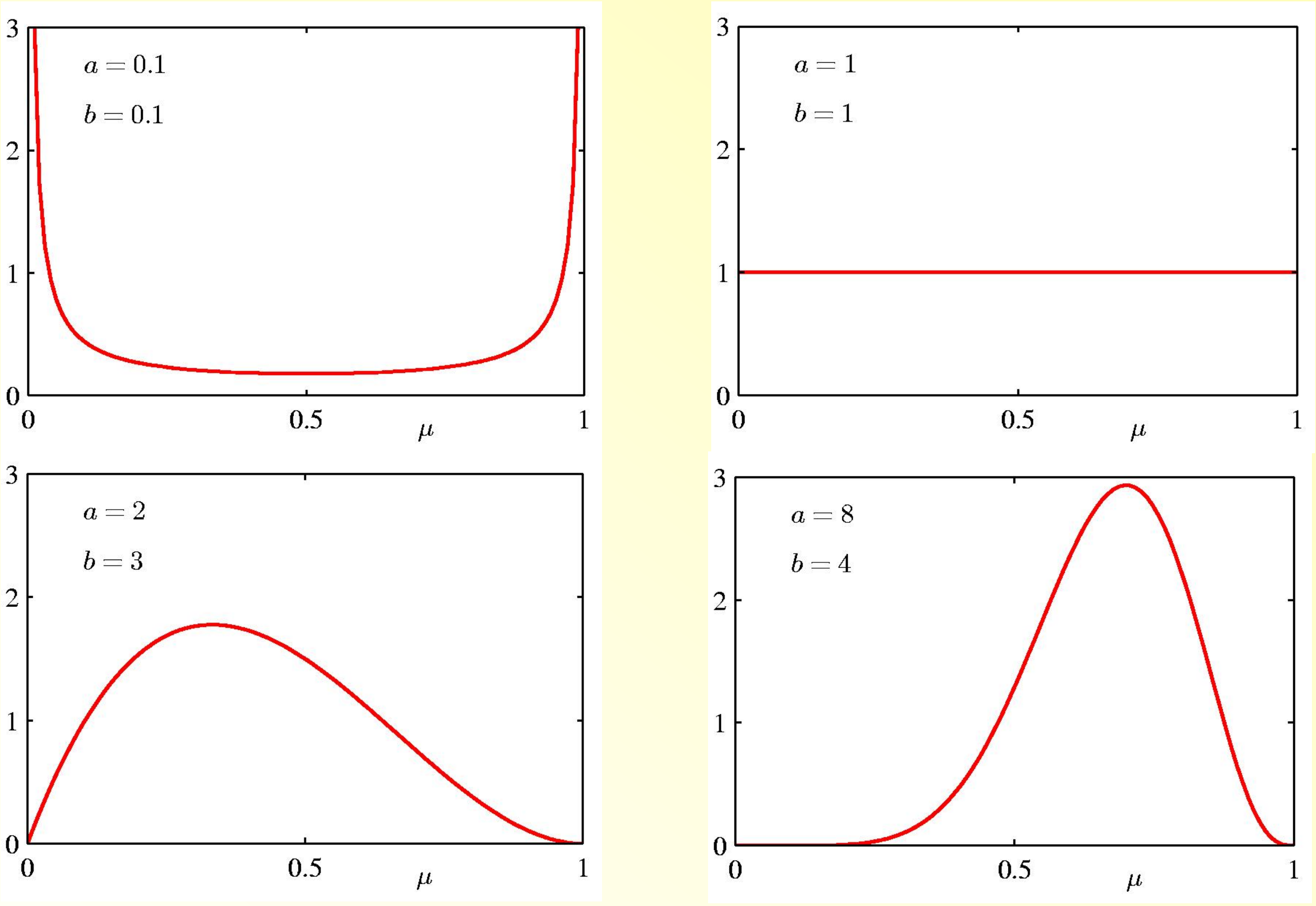
Bayesian Bernoulli

$$
\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1} (1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}
$$

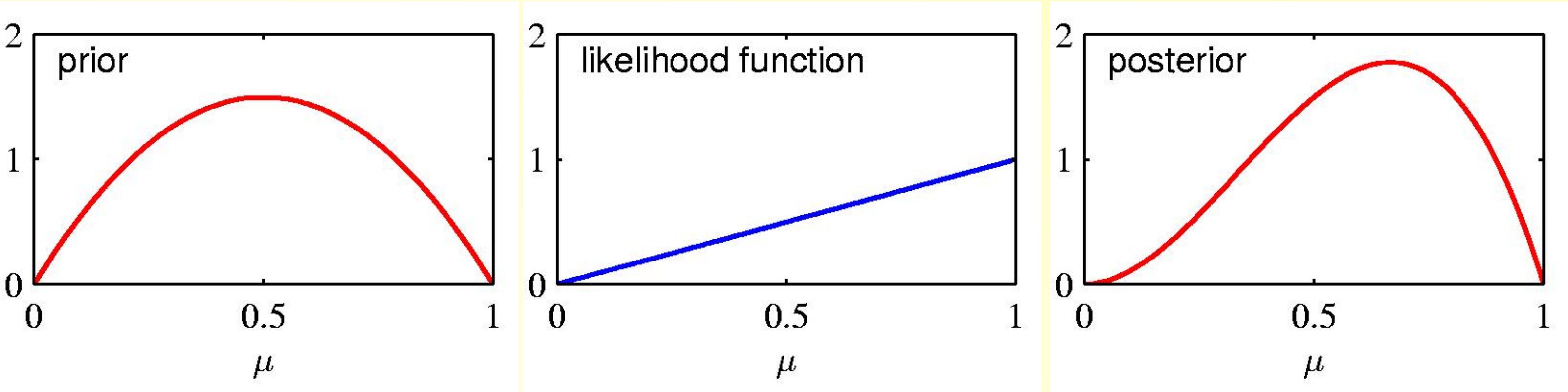
$$
a_N = a_0 + m \qquad b_N = b_0 + (N - m)
$$

The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

Beta Distribution



Prior · Likelihood = Posterior



Properties of the Posterior

As the size of the data set, N , increase

$$
\begin{aligned}
a_N &\rightarrow m \\
b_N &\rightarrow N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0
\end{aligned}
$$

Example

- Experiment: Toss a coin with P(H) = q n times.

$$X_i = \begin{cases} 1 & \text{if } H \\ 0 & \text{else.} \end{cases}$$

- Want to estimate q using the posterior distribution.
- Note that the number of heads X has Binomial distribution:

$$p(x) = \binom{n}{x} q^x (1-q)^{n-x}$$

Example cont.

- Note that for some constant a and b:

$$p(q) \propto q^a (1-q)^b$$

- With the normalize constant we have the prior:

$$p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}$$

$B(\alpha,\beta)$ is Beta function of α and β

Example cont.

- Now we will compute the posterior dist. h is the number of heads and t is the number of tails (h + t = n).

$$P(h, t|q) = \binom{h+t}{h} q^h (1-q)^t$$

$$P(q|h,t) = \frac{P(h,t|q)P(q)}{\int P(h,t|q)P(q)dq}$$

$$=$$

Example cont.

- Now we will compute the posterior dist. h is the number of heads and t is the number of tails (h + t = n).

$$P(h,t|q) = \binom{h+t}{h} q^h (1-q)^t$$

$$P(q|h,t) = \frac{P(h,t|q)P(q)}{\int P(h,t|q)P(q)dq}$$

$$= \frac{q^{h+\alpha-1}(1-q)^{t+\beta-1}}{B(h+\alpha,t+\beta)}$$

Another example

- A store owner models the number of customers arriving at the store by Poisson distribution with unknown rate θ
- The owner assigns the distribution of θ a gamma prior distribution with parameter 3 and 2.
- Let X be the number of customers during one hour. If X=3 is observed, what is the distribution of θ ?

Poisson distribution

- We have a likelihood function:

$$P(x|\theta) = \frac{\theta^x \exp(-\theta)}{x!}$$

- For sample X1, … Xn we have

$$P(x_1, \ldots x_n|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} \exp(-\theta)}{x_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i} \exp(-n\theta)}{\prod_{i=1}^{n} x_i!}$$

so we have

$$P(x_1, \ldots x_n|\theta) \propto \theta^{\sum_{i=1}^{n} x_i} \exp(-n\theta)$$

Example cont.

- Note that $P(\theta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$
- Thus we have

$$P(\theta|x_1, \ldots x_n) \propto \theta^{\sum_{i=1}^n x_i + \alpha - 1} \exp(-(\beta + n)\theta)$$

- Using the normalize constant we have

$$P(\theta|x_1, \ldots x_n) = \frac{\theta^{\sum_{i=1}^n x_i + \alpha - 1} \exp(-(\beta + n)\theta)(n + \beta)^{\sum_{i=1}^n x_i + \alpha}}{\Gamma(\sum_{i=1}^n x_i + \alpha)}$$

= Gamma($\sum_{i=1}^{n} x_i + \alpha$, $n + \beta$).

Exponential distribution

- We have a likelihood function:

$$P(x|\theta) = \theta \exp(-\theta)$$

- For sample X1, … Xn we have

$$P(x_1, \ldots x_n|\theta) = \theta^n \exp(-\theta(\sum_{i=1}^{n} x_i))$$

Also we have

$$P(\theta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$$

Exponential dist. cont.

Thus we have

$$P(\theta|x_1,\ldots x_n) \propto \theta^{\alpha+n-1}\exp(-(\beta+\sum_{i=1}^{n} x_i)\theta)$$

- Using the normalize constant we have

$P(\theta|x_1,\ldots x_n)$

= Gamma($\alpha+n$, $\beta+\sum_{i=1}^{n} x_i$).

Examples of Bayesian Inference using the Normal distribution

Known variance, unknown mean

It is easier to consider first a model with 1 unknown parameter. Suppose we have a sample of Normal data: $x_i \sim N(\mu, \sigma^2), i = 1, \ldots, n.$

Let us assume we know the variance, σ^2 and we assume a prior distribution for the mean, μ based on our prior beliefs:

$\mu \sim N(\mu_0, \sigma_0^2)$ Now we wish to construct the posterior distribution $p(\mu|x)$.

Posterior for Normal distribution mean

So we have

$$p(\mu) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu - \mu_0)^2 / \sigma_0^2)$$

$$p(x_i \mid \mu) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i - \mu)^2 / \sigma^2)$$

and hence

$$p(\mu \mid x) = p(\mu) p(x \mid \mu)$$

$$= (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu - \mu_0)^2 / \sigma_0^2) \times$$

$$\prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i - \mu)^2 / \sigma^2)$$

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

Posterior for Normal distribution mean (continued)

For a Normal distribution with response *y* with mean θ and variance ϕ we have

$$f(y) = (2\pi\phi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(y-\theta)^2/\phi\}$$

$$\propto \exp\{-\tfrac{1}{2}y^2\phi^{-1} + y\theta/\phi + cons\}$$

We can equate this to our posterior as follows:

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

$$\rightarrow \phi = (1/\sigma_0^2 + n/\sigma^2)^{-1} \text{ and } \theta = \phi(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2)$$

Conjugate posterior and prior

- When the posterior is in the same family as the prior we have *conjugacy*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	Mean	Normal	Normal
Normal	Precision	Gamma	Gamma
Binomial	Probability	Beta	Beta
Poisson	Mean	Gamma	Gamma