March 14, 2018

CS 361: Probability & Statistics

Inference

The prior

From Bayes' rule, we know that we can express our function of interest as

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

The right hand side contains the likelihood, which we've been working with. Also in the numerator is the so-called **prior probability** of θ

Bayesian inference is useful because it allows us to incorporate prior beliefs we have about the value of θ

Which prior?

$$\underset{\text{Posterior}}{\boxed{P(\theta|\mathcal{D})}} = \frac{\overset{\text{Likelihood}}{P(\mathcal{D}|\theta)}\,\overset{\text{Prior}}{P(\theta)}}{P(\mathcal{D})}$$

A particular kind of good behavior we might insist upon is the following:

1) For a given problem setup, the likelihood function is largely out of our control. E.g. we suppose that our data is from a normal distribution, the likelihood function is going to be normal
2) So the prior is our only degree of freedom
3) We choose a prior that is expressive enough that we can encode arbitrary beliefs about the prior probability of theta — the unknown parameters in our model
4) But choose a prior such that when it is multiplied by the likelihood function, we get a posterior that is of the same random variable type as the prior

A prior satisfying 4) above is called a **conjugate prior** of the likelihood function

Which prior, binomial

The binomial family of distributions is conjugate to the **beta** family of distributions

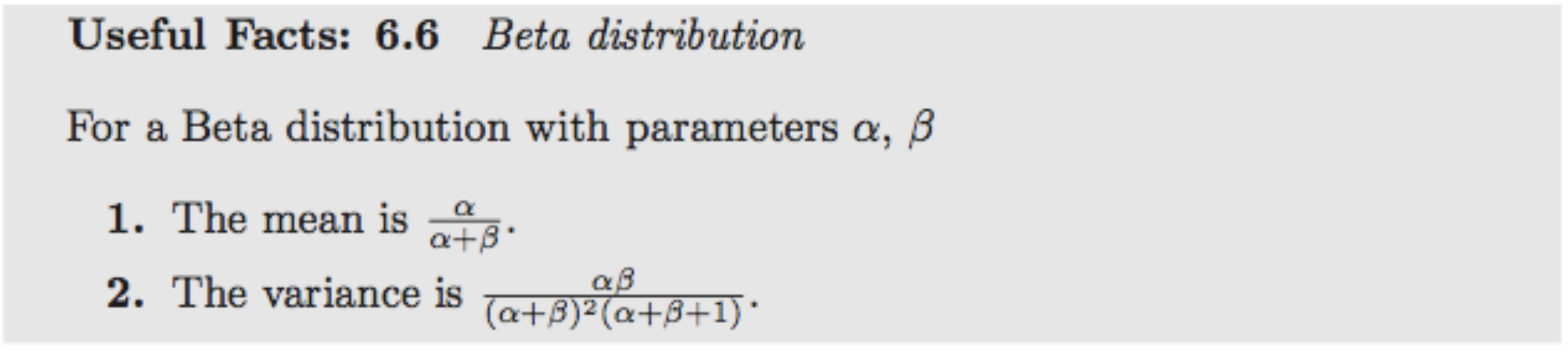
A beta random variable is a continuous random variable defined on $0 \leq x \leq 1$ with parameters alpha > 0 and beta > 0 whose density has the following form

$$p(x; \alpha, \beta) = (\text{constant}) x^{\alpha-1}(1-x)^{\beta-1}$$

The constant is in terms of a special function called the **gamma function** which is a generalization of the factorial function to positive real values rather than just non-negative integers. Details can be found in the first chapter of the book

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Beta distribution

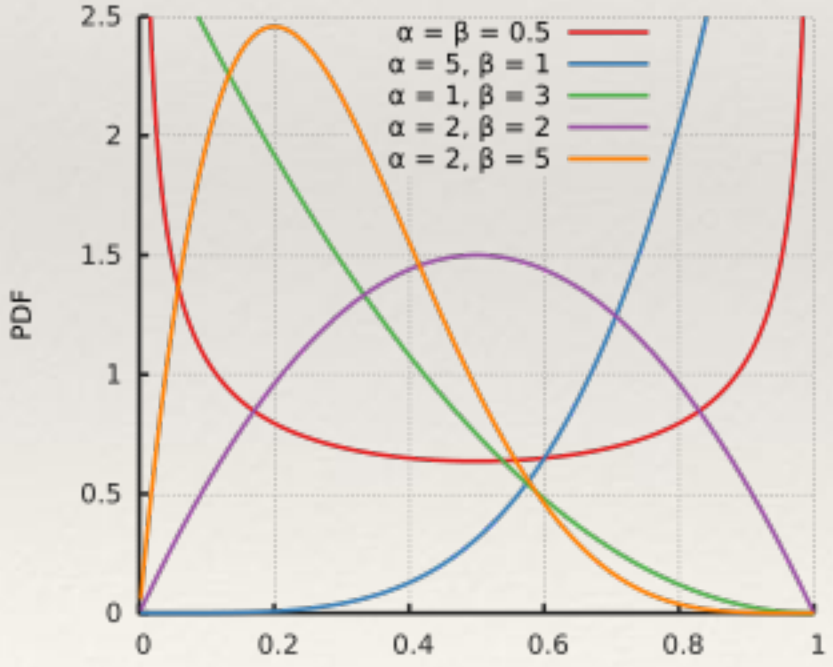


Useful Facts: 6.6 *Beta distribution*

For a Beta distribution with parameters α, β

1. The mean is $\frac{\alpha}{\alpha+\beta}$.
2. The variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}(1-x)^{\beta-1}$$



The beta distribution is very expressive

Having alpha=beta=1 would give a uniform prior

Binomial likelihood, beta prior

So if we want to do Bayesian inference against a binomial problem setup. Our likelihood be a binomial distribution. Let's see what happens if we pair that likelihood with a beta prior

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

If our data in this case is that we observed *h* heads in *N* flips, we have

$$p(\theta|\mathcal{D}) \propto \underbrace{\binom{N}{h}\theta^h(1-\theta)^{N-h}}_{\text{Likelihood}}\underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{Prior}}$$

Just focusing on theta, we get

$$p(\theta|\mathcal{D}) \propto \theta^{\alpha+h-1}(1-\theta)^{\beta+N-h-1}$$

Binomial likelihood, beta prior

If we do some clever things to make this a density, we get

$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

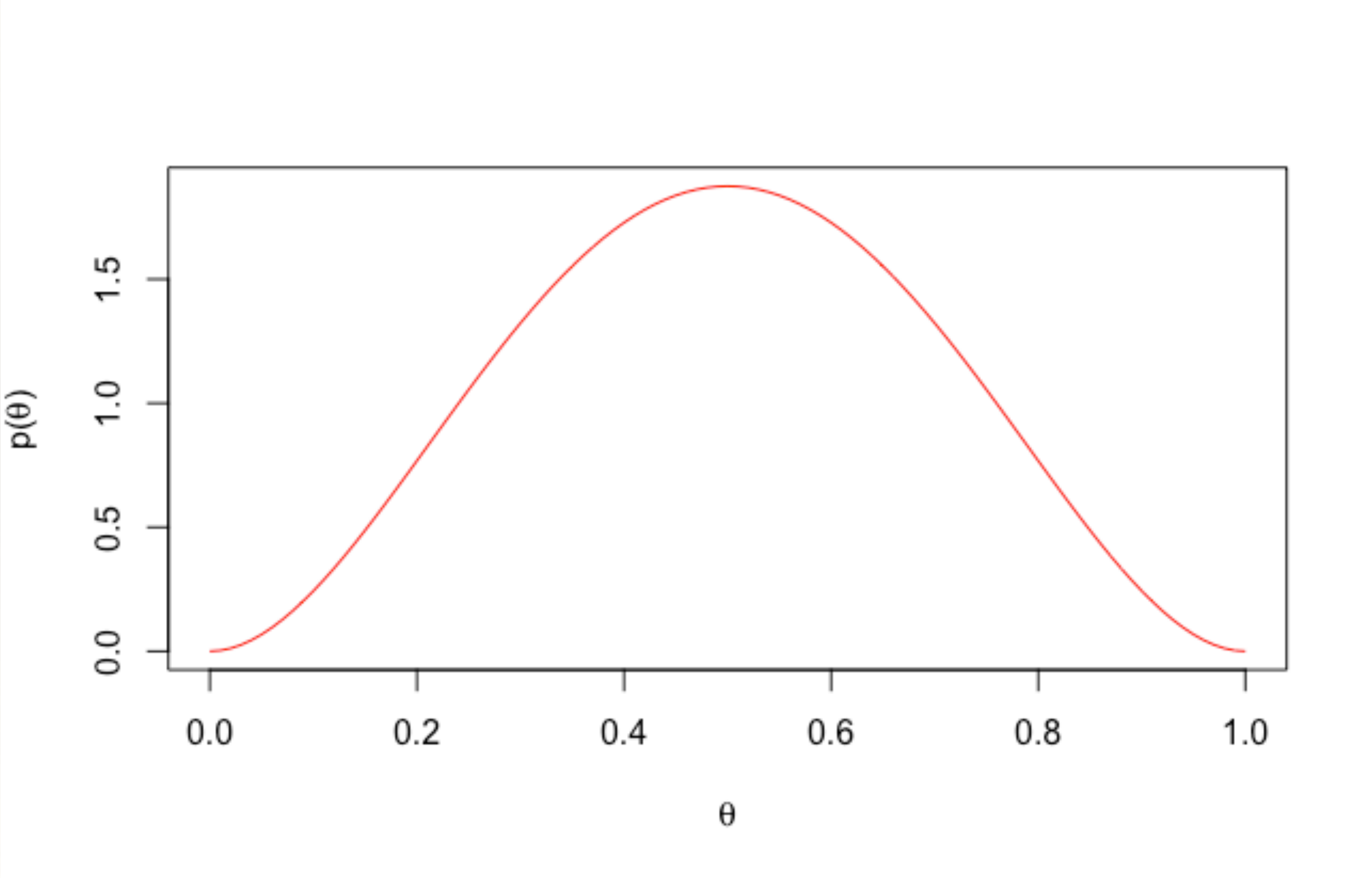
If we pattern match that with our definition of a Beta distribution

$$p(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

which is a beta distribution with parameters $\alpha+h$ and $\beta+(N-h)$

Example: updating

Suppose we have a prior belief that the probability of heads for a coin is governed by a beta distribution with parameters $\alpha = 3, \beta = 3$ this is what the density of theta would look like



Example: updating

Now, if we observed 10 coin flips and saw 7 heads, our posterior would be given by this ugly formula

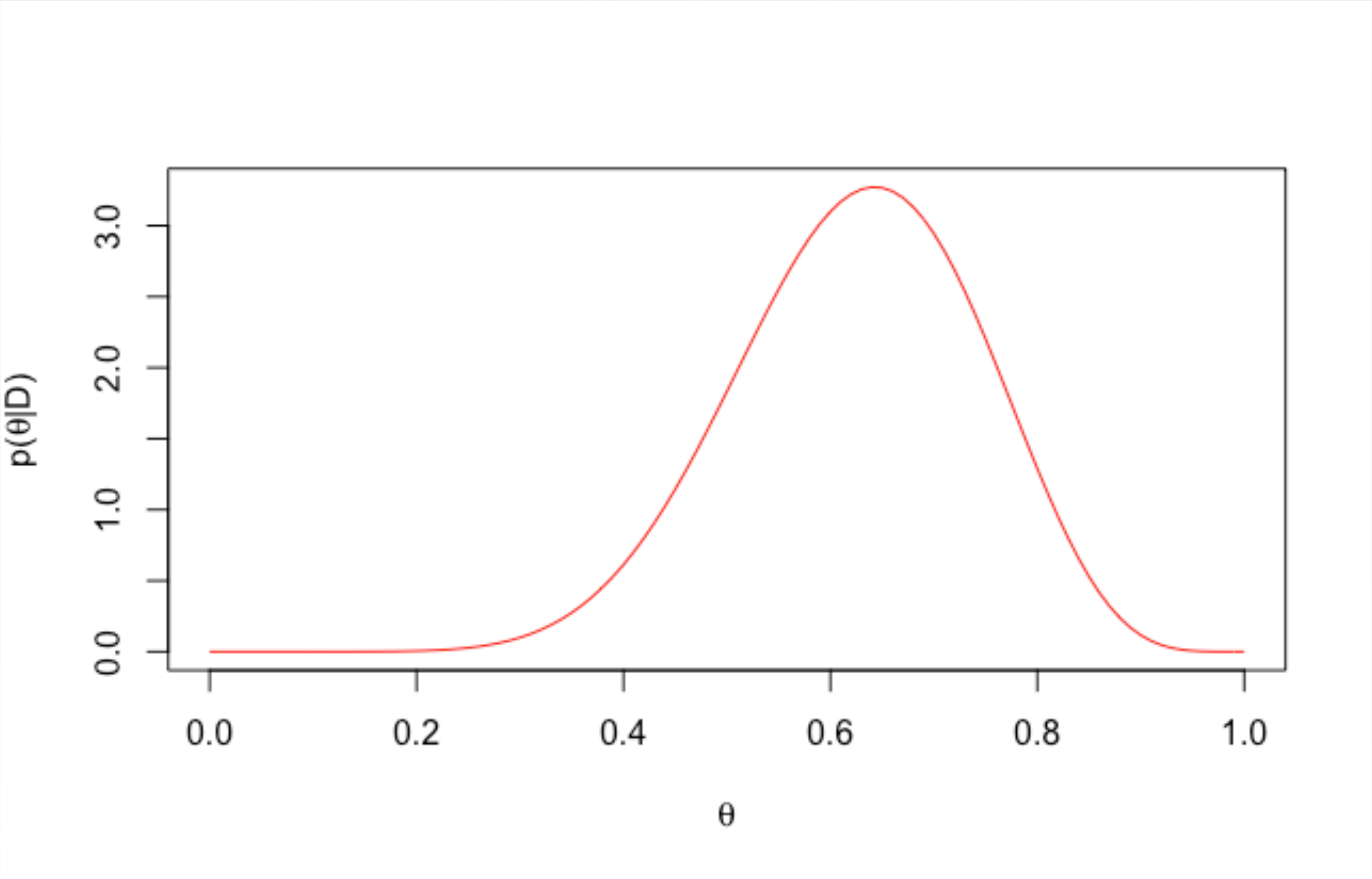
$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

But we can ignore the ugliness and just realize that starting with $\alpha = 3, \beta = 3$ and then observing 7 heads and 3 tails, we will have a beta distribution with $\alpha = 10, \beta = 6$

We always just add the number of heads we saw to alpha and the number of tails we saw to beta. We don't have to actually think about multiplying the prior and likelihood every time since the beta distribution is a conjugate prior for the binomial!

Example: updating

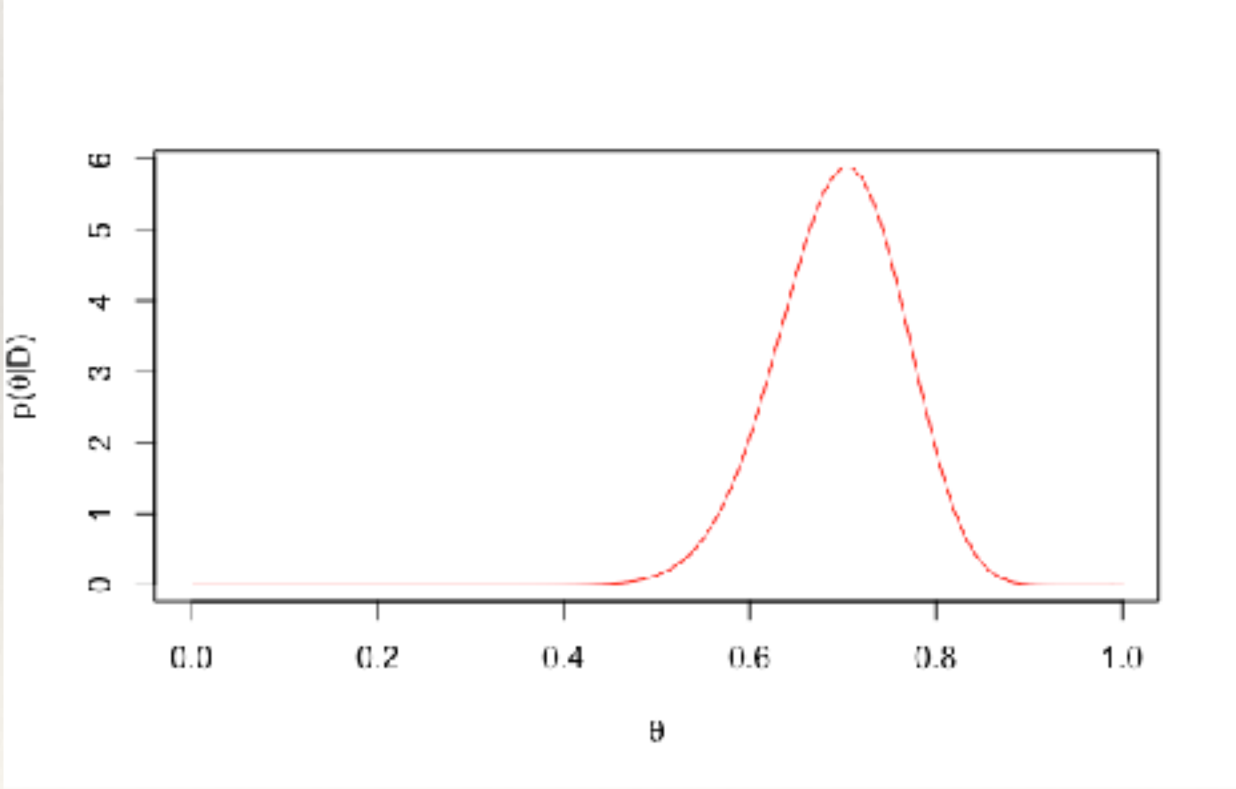
This is what our posterior distribution now looks like



Example: updating

Now here is the real power of conjugate priors. What if we had the same coin and after observing the last 10 flips we got to observe 30 more and this time we saw 22 heads?

We can use the beta posterior from the last slide as the beta prior in this slide. Doing so would give a new posterior that's a beta with $\alpha = 32, \beta = 14$



Beta prior

Since multiplying a likelihood by a conjugate prior gives a posterior that's in the same family as the prior, this makes it easier to have a probabilistic model that we just update with new data as it comes in instead of re-calculating the posterior from scratch every time we have new data

If we have new data we use our old posterior as the prior for our new estimate

Furthermore our example showed that the parameters in a beta distribution just wind up counting heads and tails.

If we started off with a prior with values alpha and beta, we observe a bunch of data in perhaps separate updates, the final posterior we will have will just be a beta distribution with parameters alpha + number of heads and beta + number of tails we have seen so far

Conjugate priors

The beta is also a conjugate prior for geometric and Bernoulli random variables

The exponential and Poisson distributions have another distribution called the **gamma distribution** as their conjugate prior which we will consider next

The normal distribution is conjugate to itself which we will not show in depth

Poisson's prior

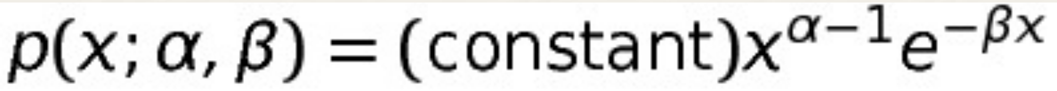
We liked the Beta distribution because it assigned probabilities in the range [0,1] and when multiplied by the binomial likelihood gave another Beta distribution.

The free parameter in a Poisson distribution is λ and it corresponds to the rate or intensity of the number of events we expect to observe per interval of time or space

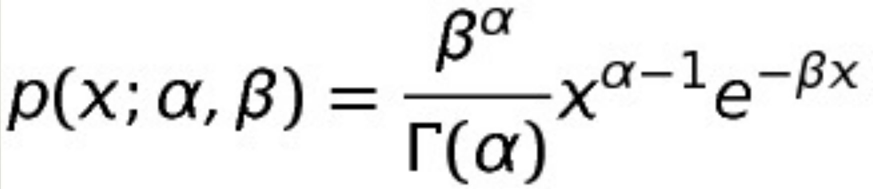
Our prior for a Poisson distribution, then, will need to be able to assign a probability to any valid value of λ . In other words it will need to be defined for $x \geq 0$

Gamma distribution

It turns out that the family of distributions with parameters alpha > 0 and beta > 0 called the gamma distribution is the conjugate prior to the Poisson. Its density is

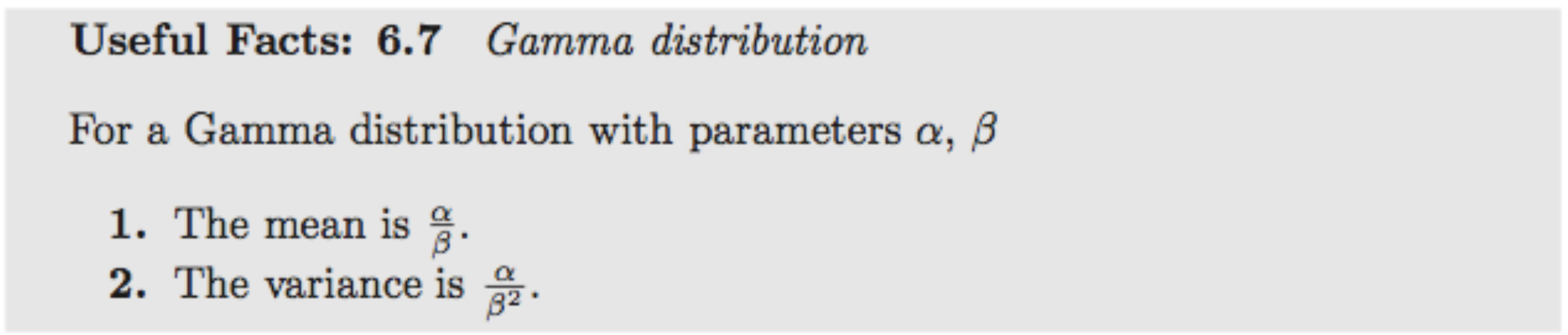


$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}e^{-\beta x}$$



$$p(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$$

And we have



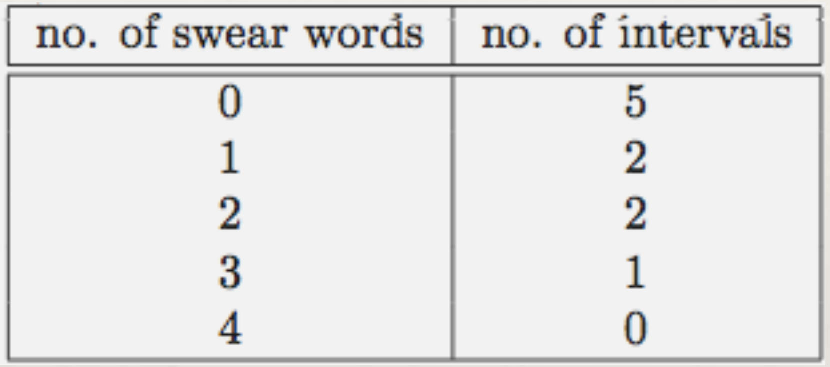
Useful Facts: 6.7 *Gamma distribution*

For a Gamma distribution with parameters α, β

1. The mean is $\frac{\alpha}{\beta}$.
2. The variance is $\frac{\alpha}{\beta^2}$.

Example

Suppose we are watching a speech by a politician who swears a lot. We model the politician’s swearing with a Poisson distribution. Here are how many swear words we hear in the first 10 intervals in the politician’s speech



no. of swear words	no. of intervals
0	5
1	2
2	2
3	1
4	0

For our likelihood we get

$$P(\mathcal{D}|\theta) = \left(\frac{\theta^0 e^{-\theta}}{0!}\right)^5 \left(\frac{\theta^1 e^{-\theta}}{1!}\right)^2 \left(\frac{\theta^2 e^{-\theta}}{2!}\right)^2 \left(\frac{\theta^3 e^{-\theta}}{3!}\right)^1$$

Or

total number of swears observed

number of intervals observed

$$P(\mathcal{D}|\theta) \propto \theta^{\boxed{9}} e^{-\boxed{10}\theta}$$

Example

$$P(\mathcal{D}|\theta) \propto \theta^9 e^{-10\theta}$$

If we multiplied this likelihood by a gamma prior with parameters alpha and beta, we would get

$$P(\mathcal{D}|\theta)P(\theta) \propto \theta^9 e^{-10\theta}\theta^{\alpha-1}e^{-\beta\theta}$$ or $$P(\theta|\mathcal{D}) \propto \theta^{(\alpha+9)-1}e^{-(\beta+10)\theta}$$

Any distribution that's proportional to $\theta^{\alpha-1}e^{-\beta\theta}$ is a gamma distribution

Which is to say our posterior would be a gamma with parameters $\alpha + 9$ and $\beta + 10$

Poisson & gamma

To generalize, then, for a Poisson likelihood, if we begin with a prior that is a gamma with parameters alpha and beta

And then we observe *N* intervals with a total of *k* events

The posterior probability of the Poisson parameter is a gamma random variable with parameters alpha + k and beta + N

MAP inference

We've done all of this so that if we encounter a distribution, we know a good prior to choose so that we can write down $P(\theta|\mathcal{D})$ after observing some data

Recall that the point estimate task involves choosing a value for theta. The so-called MAP (maximum a-posteriori) estimate is

$$\hat{\theta} = \arg\max_{\theta} P(\theta|\mathcal{D})$$

Binomial MAP

If we have observed some binomial data, say we have seen *h* heads in *N* coin flips and we have a beta prior with parameters alpha and beta, then we have

$$P(\theta|\mathcal{D}) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h)-1}(1-\theta)^{(\beta+N-h)-1}$$

If we differentiate with respect to theta and set equal to 0, we will get a MAP estimate of

$$\hat{\theta} = \frac{\alpha-1+h}{\alpha+\beta-2+N}$$

Note that if we had started off with a prior of alpha=beta=1, we would have a uniform prior and our MAP estimate would be the same as the MLE estimate we derived before

Poisson MAP

If we have observed some Poisson data, say we have seen *k* total events in *N* total intervals and we have a gamma prior with parameters alpha and beta, then we have

$$P(\theta|\mathcal{D}) = \frac{(\beta+N)^{(\alpha+k)}}{\Gamma(\alpha+k)}\theta^{(\alpha+k)-1}e^{-(\beta+N)\theta}$$

Differentiating with respect to theta, setting equal to zero and solving for theta gives a MAP estimate of

$$\hat{\theta} = \frac{\alpha-1+k}{\beta+N}$$

MAP caveats

As we see more and more data, the prior tends to matter less and less. Which is to say our MAP estimate is very close to the MLE estimate for a large number of data items. Hence it might not be worth the trouble thinking about priors unless we have a small amount of data

Justifying the choice of prior can be hard. We chose ones that were mathematically convenient—that gave us a nice posterior—but does nature use conjugate priors? Maybe not for whichever problem you're considering

Simulation for MLE confidence intervals

MLE

When we were doing MLE estimation, we had a dataset and a probability distribution with some unknown parameter(s)

The data that we observe determined our max likelihood estimate for those unknown parameters. Had we seen different data, we would have gotten a different estimate.

Our likelihood function was not a probability function of theta, though. So how can we give an interval estimate of theta?

When we don't have an easy to analyze probability model of our estimator, we should be thinking of using something like the bootstrap method from before

Parametric bootstrap

Probability distributions and densities can be used to calculate probabilities, but they can also be used with software to generate data. Generating data from a known distribution is called **sampling from the distribution**

Instead of collecting multiple datasets we can observe the dataset we have, construct the distribution with our max likelihood estimate, and then use that distribution to generate many datasets with the same size as our original dataset

For each “synthetic” dataset we generate like this, we can see what the MLE would have been had that been the data we observed. We can use this set of MLEs to construct an interval estimate

Generating multiple MLEs this way is a process called the **parametric bootstrap**

Example

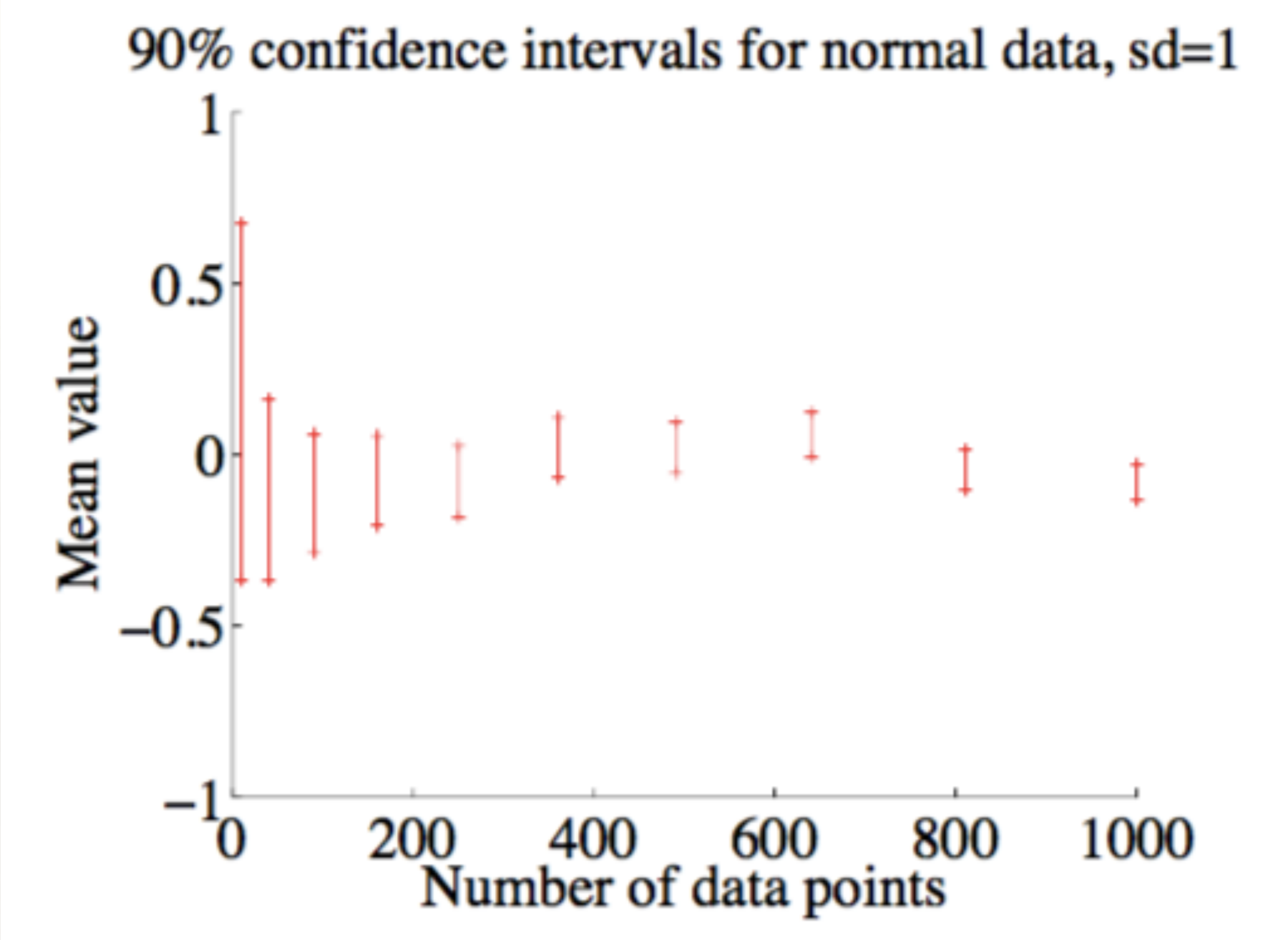
Suppose we have a dataset that comes from a binomial distribution. How will we give a 90% confidence interval for an estimate of the parameter of this distribution?

First we compute the max likelihood estimate of theta. Then we generate a dataset using software to get random samples from a binomial distribution with parameter theta

For this new dataset we calculate what the max likelihood estimate had this been the dataset we had seen and record our estimate

We do this a large number of times, getting a list of MLE estimates for theta and we report back the interval that contains 90% of these estimates, where 5% of the estimates are larger and 5% are smaller than the bounds of the interval

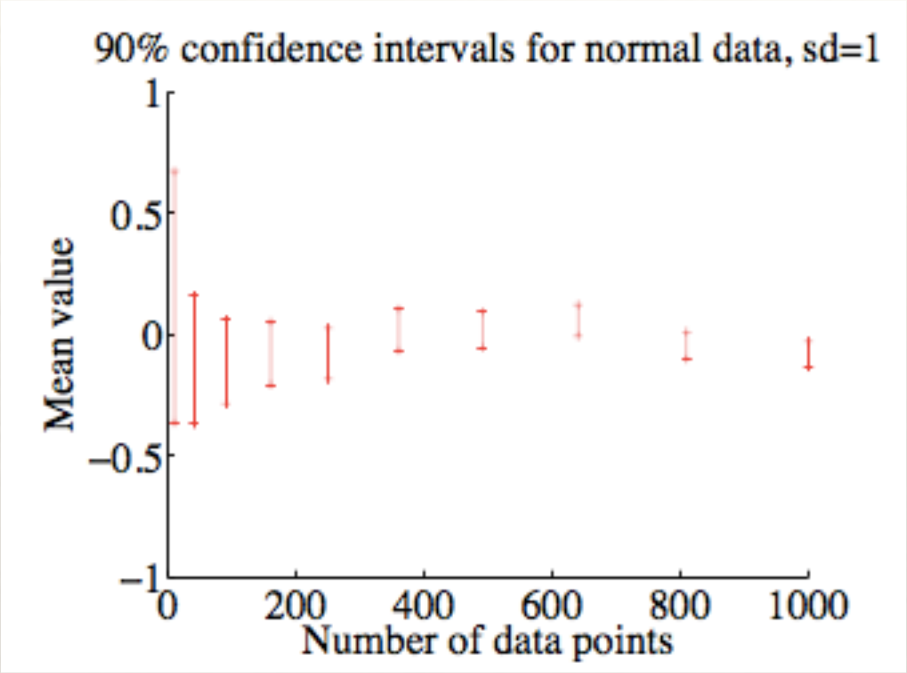
Example

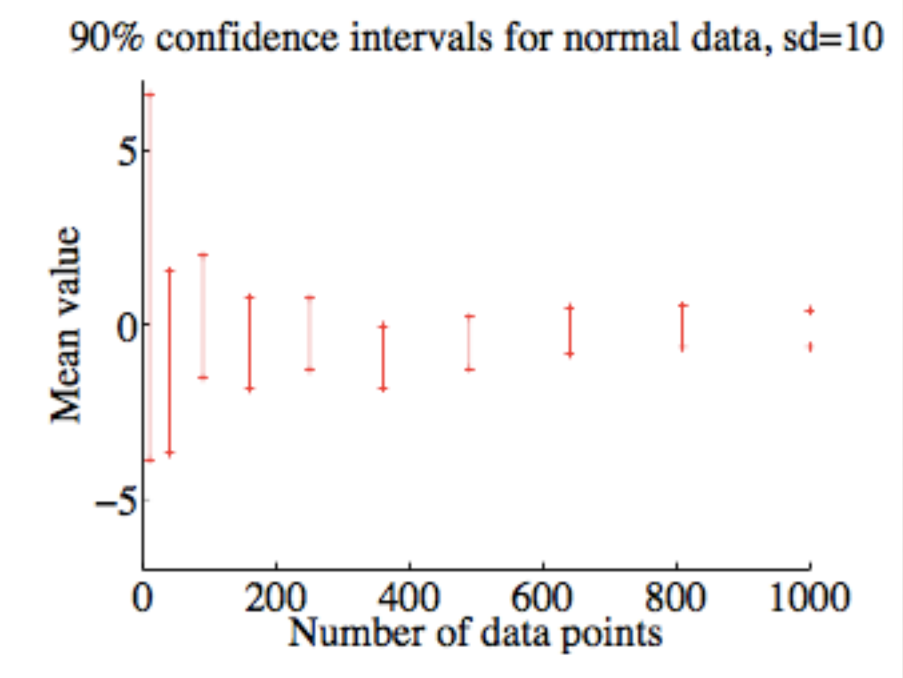


Parametric bootstrap applied to estimating the mean of a normal distribution with known standard deviation

# of datapoints in the original dataset=# of datapoints in the simulated dataset

Example





In general the size of the interval behaves like $\frac{\sigma}{\sqrt{N}}$

Bootstrap

In general the size of the interval behaves like $\frac{\sigma}{\sqrt{N}}$

This might sound like cheating, but realize that we aren't getting anything for free here. Constructing these synthetic datasets does not improve the accuracy of our point estimate in any way. And the N above is the size of our original dataset

All we are doing is answering the question "hey, what if we had gotten a slightly different dataset, how much would our estimate have changed?"

Confidence intervals for Bayesian parameter estimates

Bayesian estimation

When we did Bayesian parameter estimation, we had a dataset and a probability distribution with some unknown parameter(s)

We chose a prior distribution and combined it with the data we observed to get a function with which we could assign probabilities to values of the unknown parameter

We had a distribution for the data—with its unknown parameter—and a distribution for the parameter itself

Bayesian estimation

We then chose the value for the unknown parameter that maximized this posterior probability distribution

We could have asked other questions of our posterior, though. For example we could have said “given the data, what is the probability that theta was in the interval [a,b]”

We would have gotten our answer by computing

$$\int_a^b p(\theta|\mathcal{D})\,d\theta$$

Example

We have a coin with unknown probability of coming up heads. Starting with a uniform Beta prior, suppose we flip the coin 10 times and observe 7 heads. What is the probability that theta is between 0.5 and 0.8?

Our posterior is given by

$$p(\theta|\mathcal{D}) = \frac{\Gamma(12)}{\Gamma(8)\Gamma(4)}\theta^7(1-\theta)^3$$

And we have

$$\int_{0.5}^{0.8} p(\theta|\mathcal{D})\, d\theta \approx 0.73$$

Example

We have a coin with unknown probability of coming up heads. Starting with a uniform Beta prior, suppose we flip the coin 10 times and observe 7 heads. Construct an interval [a,b] such that $P(\theta \leq a|\mathcal{D}) \approx 0.05$ and $P(\theta \geq b|\mathcal{D}) \approx 0.05$. This is the 90% confidence interval for theta

We have to get some software to solve this for us, but we want to solve for a in

$\int_{-\infty}^{a} P(\theta|\mathcal{D})\, d\theta \approx 0.05$ And b in $\int_{b}^{\infty} P(\theta|\mathcal{D})\, d\theta \approx 0.05$

Doing so yields the interval [0.435, 0.865]

Bayesian confidence intervals

If the specified level of confidence we are looking for is **1-2u** for $0 \leq u \leq 0.5$ we must find an a and b such that $P(\{\theta \in [a,b]\}|\mathcal{D}) = 1 - 2u$

For any value on the RHS the a and b aren't unique. The reason we are writing it as 1-2u instead of just p is that we will usually want to find the interval with

$$P(\{\theta \leq a\}) = \int_{-\infty}^{a} P(\theta|\mathcal{D})d\theta = u$$

and

$$P(\{\theta \geq b\}\,|\mathcal{D}) = \int_{b}^{\infty} P(\theta|\mathcal{D})d\theta = u.$$

Which is valid since

$$P(\{\theta \in [a,b]\}\,|\mathcal{D}) = 1 - P(\{\theta \leq a\}\,|\mathcal{D}) - P(\{\theta \geq b\}\,|\mathcal{D})$$