Lecture 27

Part 6 Introduction to Bayesian Statistics

Bayesian Inference

Bayesian Inference

- Now for something completely different.
- Everything we have done up to now is frequentist statistics. Bayesian statistics is very different.
- Bayesians don't do confidence intervals and hypothesis tests. Bayesians don't use sampling distributions of estimators. Modern Bayesians aren't even interested in point estimators.
- So what do they do? Bayesians treat parameters as random variables.

Bayesian Inference

- To a Bayesian probability is the only way to describe uncertainty.
- Things not known for certain — like values of parameters — must be described by a probability distribution.

Bayesian Inference

- Suppose you are uncertain about something. Then your uncertainty is described by a probability distribution called your *prior distribution*.
- Suppose you obtain some data relevant to that thing. The data changes your uncertainty, which is then described by a new probability distribution called your *posterior distribution*.
- The posterior distribution reflects the information both in the prior distribution and the data.
- Most of Bayesian inference is about how to go from prior to posterior.

Bayesian Inference

- The way Bayesians go from prior to posterior is to use the laws of conditional probability, sometimes called in this context *Bayes rule* or *Bayes theorem*.
- Suppose we have a PDF g for the prior distribution of the parameter θ, and suppose we obtain data x whose conditional PDF given θ is f. Then the joint distribution of data and parameters is conditional times marginal

$$f(x|\theta)g(\theta)$$

It is a generalized form of $P(B|A)P(A)$ we have seen.

Bayesian Inference

- The correct posterior distribution, according to the Bayesian paradigm, is the conditional distribution of θ given x, which is joint divided by marginal

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)\,d\theta}$$

- Often we do not need to do the integral. If we recognize that

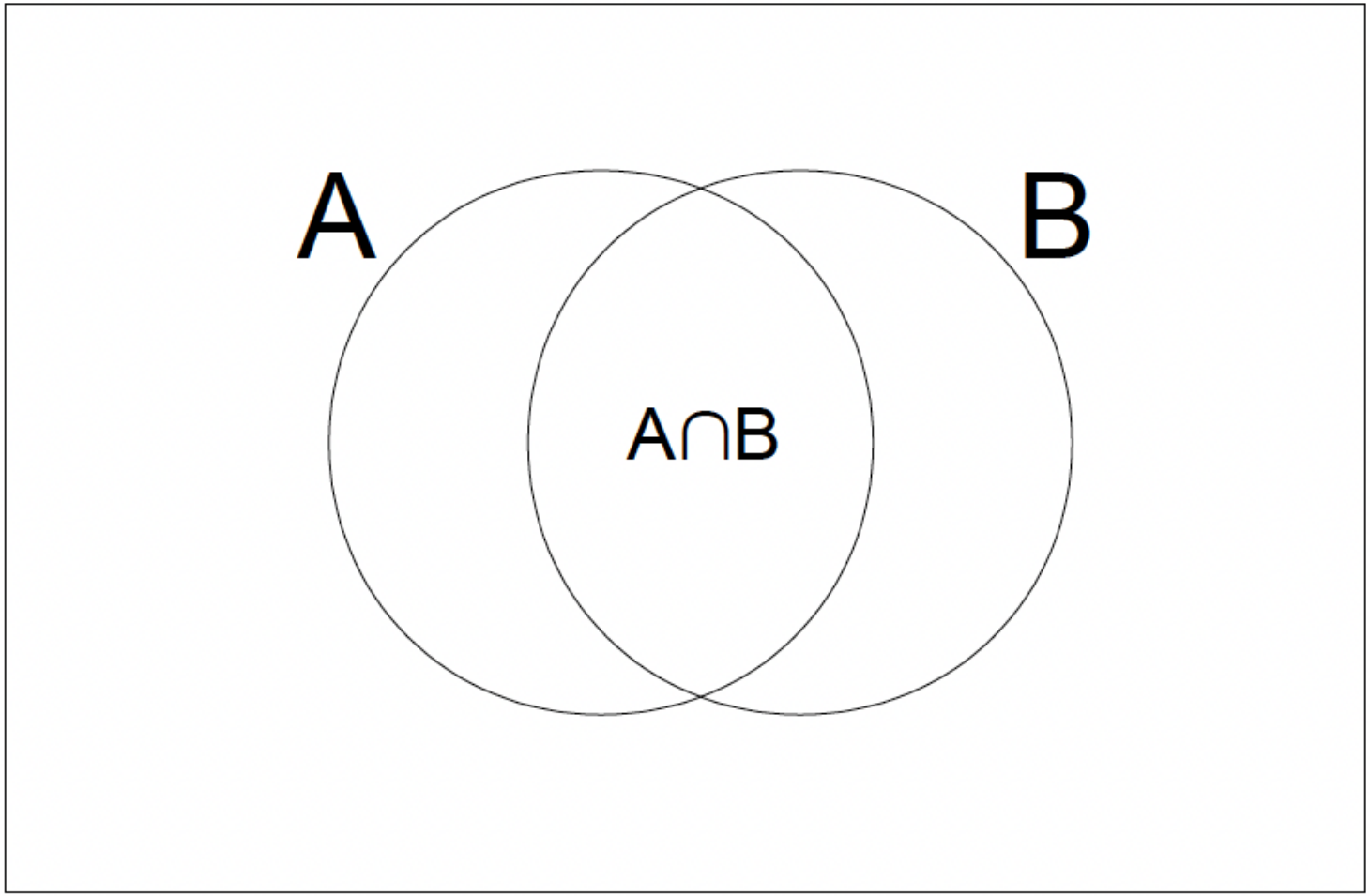
$$\theta \mapsto f(x|\theta)g(\theta)$$

is, except for constants, the PDF of a brand name distribution, then that distribution must be the posterior.

Bayes' Theorem

- Bayes' Theorem is about conditional probability.
- It has statistical applications.

Conditional Probability



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

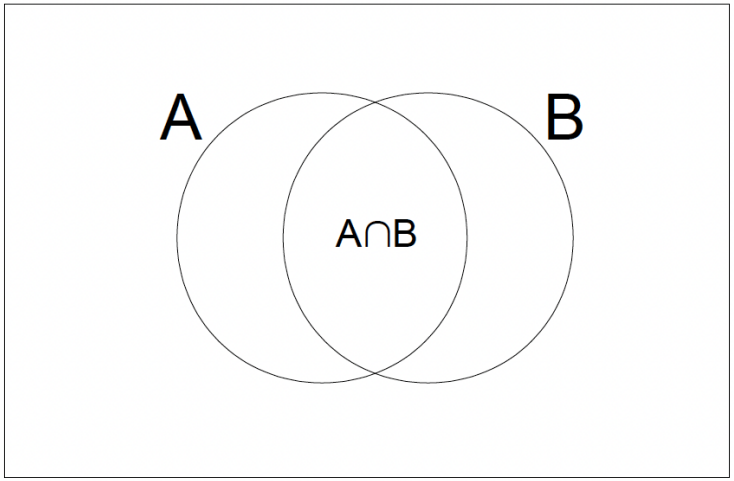
Multiplication Rule

From $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, get $P(A \cap B) = P(A|B)P(B)$.

From $P(B|A) = \dfrac{P(B \cap A)}{P(A)}$, get $P(A \cap B) = P(B|A)P(A)$

Bayes' Theorem

The most elementary version



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

$$= \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)}$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Define "events" in Terms of Random Variables

Instead of A, B, etc.

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Continuous Random Variables

We have conditional densities:

$$f_{y|x}(y|x) = \frac{f_{xy}(x,y)}{f_x(x)}$$

Different Versions of Bayes' Theorem

For discrete random variables,

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$= \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)}$$

Continuous Random Variables

$$f_{x|y}(x|y) = \frac{f_{xy}(x,y)}{f_y(y)}$$

$$= \frac{f_{y|x}(y|x)f_x(x)}{\int f_{y|x}(y|t)f_x(t)\,dt}$$

Compare

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)}$$

$$f_{x|y}(x|y) = \frac{f_{y|x}(y|x)f_x(x)}{\int f_{y|x}(y|t)f_x(t)\,dt}$$

Philosophy

Bayesian versus Frequentist

- What is probability?
- Probability is a formal axiomatic system.
- *Of what* is probability a model?

Of what is Probability a Model?

Two answers

- Frequentist: Probability is long-run relative frequency.
- Bayesian: Probability is degree of subjective belief.

Statistical Inference

How it works

- Adopt a probability model for data set Y.
- Distribution of Y depends on a parameter θ.
- Use observed value $Y = y$ to decide about θ.
- Translate the decision into a statement about the process that generated the data.
- Bayesians and Frequentists agree so far, mostly.
- The description above is limited to what frequentists can do.
- Bayes methods can generate more specific recommendations.

What is a Parameter?

- To the frequentist, it is an unknown constant.
- To the Bayesian since we don't know the value of the parameter, it's a random variable.

Unknown Parameters are Random Variables

To the Bayesian

- That's because probability is subjective belief.
- We model our uncertainty with a probability distribution, $\pi(\theta)$.
- $\pi(\theta)$ is called the *prior* distribution.
- Prior because it represents the statistician's belief about θ *before* observing the data.
- The distribution of θ after seeing the data is called the *posterior* distribution.
- The posterior is the conditional distribution of the parameter given the data.

Bayesian Inference

- Model is $p(x|\theta)$ or $f(x|\theta)$.
- Prior distribution $\pi(\theta)$ is based on the best available information.
- But yours might be different from mine. It's subjective.
- Use Bayes' Theorem to obtain the posterior distribution $\pi(\theta|x)$.
- As the notation indicates, $\pi(\theta|x)$ might be the prior for the next experiment.

Subjectivity

- Subjectivity is the most frequent objection to Bayesian methods.
- The prior distribution influences the conclusions.
- Two scientists may arrive at different conclusions from the same data, *based on the same statistical analysis*.
- The influence of the prior goes to zero as the sample size increases

Bayes' Theorem

Continuous case

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|t)\pi(t)\,dt}$$

$$\propto f(x|\theta)\pi(\theta)$$

Once You Have the Posterior Distribution, You Can . . .

- Give a point estimate of θ. Why not $E(\theta|X = x)$?
- Test hypotheses, like $H_0 : \theta \in H$.
- Reject H_0 if $P(\theta \in H|X = x) < P(\theta \notin H|X = x)$. Why not?
- We should be able to do better than "Why not?"

Decision Theory

- Any time you make a decision, you can lose something.
- Risk is defined as expected loss.
- Goal: Make decisions so as to minimize risk.
- Or if you are an optimist, you can maximize expected utility.

Decisions

$$d = d(x) \in \mathcal{D}$$

- d is a decision.
- It is based on the data.
- It is an element of a *decision space*.

Decision Space $\mathcal{D}$

- It is the set of possible decisions that might be made based on the data.
- For estimation, $\mathcal{D}$ is the parameter space.
- For accepting or rejecting a null hypothesis, $\mathcal{D}$ consists of 2 points.
- Other kinds of kinds of decision are possible, not covered by frequentist inference.
- What kind of chicken feed should the farmer buy?

Loss Function

$$L = L(d(x), \theta) \geq 0$$

When X and θ are random, L is a real-valued random variable.

Law of Total Expectation $\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}(X)$

Recall that we have $\mathrm{E}(X) = \int x \Pr[X = x]\, dx$ and $\mathrm{E}(X|Y = y) = \int x \Pr[X = x|Y = y]\, dx$

$$
\begin{aligned}
\mathrm{E}(\mathrm{E}(X|Y)) &= \int \left(\int x \Pr[X = x|Y = y]\, dx \right) \Pr[Y = y]\, dy \\
&= \int \int x \Pr[X = x, Y = y]\, dx\, dy \\
&= \int x \left(\int \Pr[X = x, Y = y]\, dy \right) dx \\
&= \int x \Pr[X = x]\, dx \\
&= \mathrm{E}(X)
\end{aligned}
$$

Risk is Expected Loss

$L = L(d(x), \theta)$

$$E(L) = E(E[L|X])$$

$$= \int \left(\int L(d(x), \theta)\, d\pi(\theta|x) \right) dP(x)$$

- Any decision $d(x)$ that minimizes posterior expected loss for all x also minimizes overall expected loss (risk).
- Such a decision is called a *Bayes decision*.
- **This is the theoretical basis for using the posterior distribution.**
- We need an example.

Coffee Taste Test

A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of $n = 100$ coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked "A" and "B." Half the time the new blend will be in cup A, and half the time it will be in cup B. Management wants to know if there is a difference in preference for the two blends.

Model: The Conditional Distribution of X Given θ

Letting θ denote the probability that a consumer will choose the new blend, treat the data $X_1, \ldots, X_n$ as a random sample from a Bernoulli distribution. That is, independently for $i = 1, \ldots, n$.
θ is a random variable.

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

for $x_i = 0$ or $x_i = 1$.

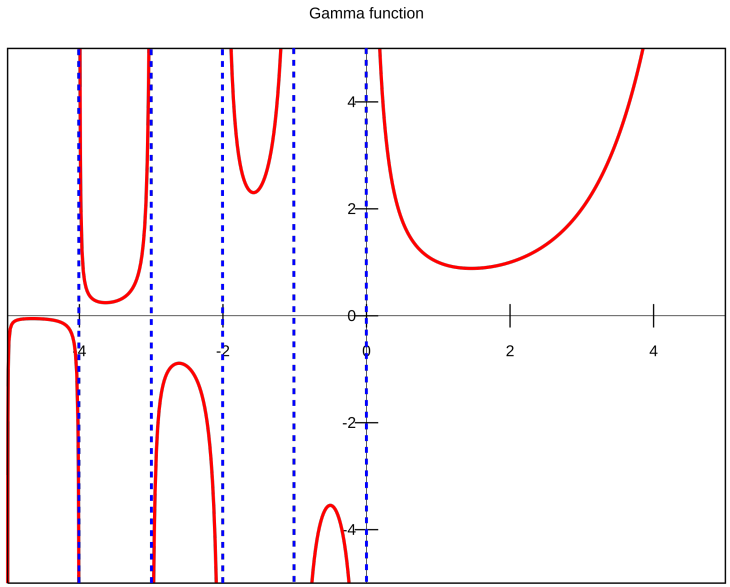
$$\begin{aligned} p(x|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

Prior: The Beta distribution

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

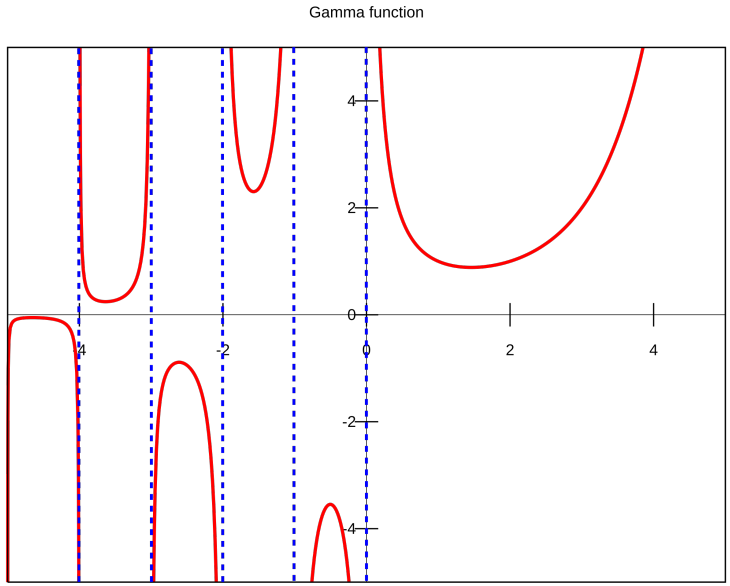
For $0 < \theta < 1$, and zero otherwise.
Note $\alpha > 0$ and $\beta > 0$

Gamma Function



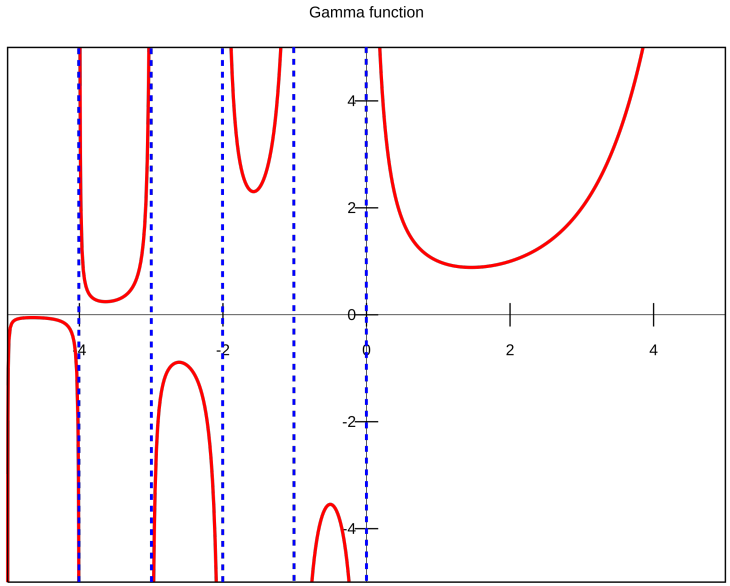
- The Gamma function extends factorials to real numbers
- Definition: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$
- Key property: $\Gamma(n) = (n-1)!$ for positive integers
- Fundamental recursive relation: $\Gamma(z+1) = z\Gamma(z)$

Visualizing the Gamma Function



- Smooth curve through factorial points
- Has a minimum at $x \approx 1.46163$
- Poles at non-positive integers
- For real numbers: $\Gamma(x+1) = x\Gamma(x)$
- Related to many special functions in mathematics

Important Properties



- Reflection formula: $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$
- Special values:
  - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
  - $\Gamma(1) = 1$
  - $\Gamma(2) = 1$
  - $\Gamma(n+1) = n!$ for natural numbers

Applications of the Gamma Function

- Statistics and Probability:
  - Gamma distribution
  - Chi-square distribution
  - Student's t-distribution
- Physics:
  - Quantum mechanics
  - Statistical mechanics

Application: Reliability Engineering

- We want to know lifetime of electronic components
- We can use Gamma distribution
  - Shape parameter (k): wear-out characteristics
  - Scale parameter (θ): time scale
  - PDF: $f(x) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)}$
  - Helps predict failure rates and maintenance schedules

Prior: The Beta distribution

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

For $0 < \theta < 1$, and zero otherwise.
Note $\alpha > 0$ and $\beta > 0$

Beta Distribution: A Group of Shape-Shifting Distribution

- Beta distribution is a continuous probability distribution on interval [0,1]
- Two parameters: α (alpha) and β (beta)
- Highly flexible shape: can be U-shaped, bell-shaped, or skewed
- Perfect for modeling probabilities and proportions

Effects of Beta Parameters

- $\alpha = \beta = 1$: Uniform distribution
- $\alpha = \beta > 1$: Bell-shaped, symmetric
- $\alpha > \beta$: Right-skewed
- $\alpha < \beta$: Left-skewed
- Larger parameters: More concentrated distribution
- Mean of distribution: $\frac{\alpha}{\alpha+\beta}$

Click-Through Rate Prediction for ad/email/online content

In Online Advertising:

- Problem: Predicting ad click-through rates (CTR)
- $\alpha =$ clicks $+ 1$
- $\beta =$ (views without clicks) $+ 1$
- Benefits:
  - Naturally handles uncertainty
  - Gets more accurate with more data
  - Perfect for online learning
  - Provides confidence intervals

Beta distribution

$Beta(\alpha, \beta)$ has density

$$f(\theta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Observation:
The coefficient is a normalizing factor, so if we have a pdf

$$f(\theta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

then

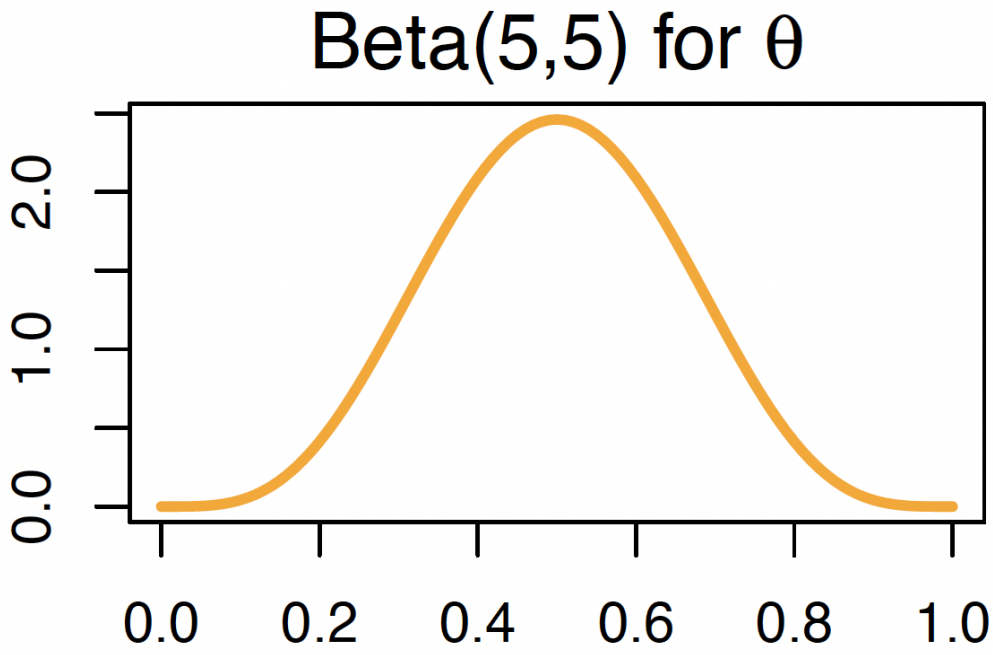
$$\theta \sim \text{beta}(\alpha, \beta)$$

and

$$c = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$$

Board question preamble: beta priors

Suppose you are testing a new medical treatment with unknown probability of success θ. You don't know that θ, but your prior belief is that it's probably not too far from 0.5. You capture this intuition with a beta(5,5) prior on θ.



Board question: beta priors

To sharpen this distribution you take data and update the prior.

- $Beta(\alpha, \beta) : f(\theta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Treatment has prior $f(\theta) \sim$ beta$(5, 5)$

Board question: beta priors

- Suppose you test it on 10 patients and have 6 successes. Find the posterior distribution on θ. Identify the type of the posterior distribution.
- Suppose you recorded the order of the results and got SSSFFSSSSFF. Find the posterior based on this data.
- Using your answer to (2) give an integral for the posterior predictive probability of success with the next patient.
- Use what you know about pdf's to evaluate the integral without computing it directly.

Solution

- 1. Prior pdf is $f(\theta) = \frac{9!}{4!4!}\theta^4(1-\theta)^4 = c_1\theta^4(1-\theta)^4$.

hypoth.	prior	likelihood	Bayes numer.	posterior
θ	$c_1\theta^4(1-\theta)^4 d\theta$	$\binom{10}{6}\theta^6(1-\theta)^4$	$c_3\theta^{10}(1-\theta)^8 d\theta$	beta(11,9)

We know the normalized posterior is a beta distribution because it has the form of a beta distribution ($c\theta^{a-1}(1-\theta)^{b-1}$ on [0,1]) so by our earlier observation it must be a beta distribution.

- 2. The answer is the same. The only change is that the likelihood has a coefficient of 1 instead of a binomial coefficient.
- 3. The posterior on θ is beta(11,9) which has density

$$f(\theta|\text{data}) = \frac{19!}{10!8!}\theta^{10}(1-\theta)^8.$$

Solution continued

The law of total probability says that the posterior predictive probability of success is

$$P(\text{success}|\text{data}) = \int_0^1 f(\text{success}|\theta) \cdot f(\theta|\text{data}) d\theta$$

$$= \int_0^1 \theta \cdot \frac{19!}{10!8!} \theta^{10} (1-\theta)^8 d\theta = \int_0^1 \frac{19!}{10!8!} \theta^{11} (1-\theta)^8 d\theta$$

Solution continued

4. We compute the integral in (3) by relating it to the pdf of beta(12,9): $\frac{20!}{11!8!}\theta^{11}(1-\theta)^7$. Since the pdf of beta(12,9) integrates to 1 we have

$$\int_0^1 \frac{20!}{11!8!}\theta^{11}(1-\theta)^7 = 1 \quad \Rightarrow \quad \int_0^1 \theta^{11}(1-\theta)^7 = \frac{11!8!}{20!}.$$

Thus

$$\int_0^1 \frac{19!}{10!8!}\theta^{11}(1-\theta)^8 d\theta = \frac{19!}{10!8!} \cdot \frac{11!8!}{20!} = \boxed{\frac{11}{20}}.$$

Conjugate priors

We had

- Prior $f(\theta)d\theta$: beta distribution
- Likelihood $p(x|\theta)$: binomial distribution
- Posterior $f(\theta|x)d\theta$: beta distribution

The beta distribution is called a conjugate prior for the binomial likelihood.
That is, the beta prior becomes a beta posterior and repeated updating is easy!

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

- Supported on $[0, 1]$.
- $E(\theta) = \frac{\alpha}{\alpha+\beta}$
- $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- Can assume a variety of shapes depending on α and β.
- When $\alpha = \beta = 1$, it's uniform.
- Bayes used a Bernoulli model and a uniform prior in his posthumous paper.

Posterior Distribution

$$\pi(\theta|x) \propto p(x|\theta)\,\pi(\theta)$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{(\alpha+\sum_{i=1}^{n} x_i)-1}(1-\theta)^{(\beta+n-\sum_{i=1}^{n} x_i)-1}$$

Proportional to the density of a Beta(α', β'), with

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i$$

$$\beta' = \beta + n - \sum_{i=1}^{n} x_i$$

Conjugate Priors

- Prior was Beta(α, β).
- Posterior is Beta(α', β').
- Prior and posterior are in the same family of distributions.
- The Beta is a *conjugate prior* for the Bernoulli model.
- Posterior was obtained by inspection.
- Conjugate priors are very convenient.
- There are conjugate priors for many models.
- There are also important models for which conjugate priors do not exist.

Concept Question

Suppose your prior $f(\theta)$ in the bent coin example is Beta(6,8). You flip the coin 7 times, getting 2 heads and 5 tails. What is the posterior pdf $f(\theta|x)$?

- Beta(2,5)
- Beta(3,6)
- Beta(6,8)
- Beta(8,13)

We saw in the previous board question that 2 heads and 5 tails will update a beta(α, β) prior to a beta$(\alpha + 2, \beta + 5)$ posterior.

answer: (4) beta(8,13).

Continuous priors, continuous data

Bayesian update tables:

hypoth.	prior	likelihood	Bayes numerator	posterior $f(\theta\|x)d\theta$
θ	$f(\theta)d\theta$	$f(x\|\theta)$	$f(x\|\theta)f(\theta)d\theta$	$\frac{f(x\|\theta)f(\theta)d\theta}{f(x)}$
total	1		$f(x)$	1

$$f(x) = \int f(x|\theta)f(\theta)d\theta$$

Picture of the posterior

Suppose 60 out of 100 consumers picked the new blend of coffee beans.

Posterior is Beta, with $\alpha' = \alpha + \sum_{i=1}^{n} x_i = 61$ and $\beta' = \beta + n - \sum_{i=1}^{n} x_i = 41$.