Introduction to Bayesian Statistics[1]

STA 260 Spring 2020

[1] This slide show is an open-source document. See last slide for copyright information.

Thomas Bayes (1701-1761)

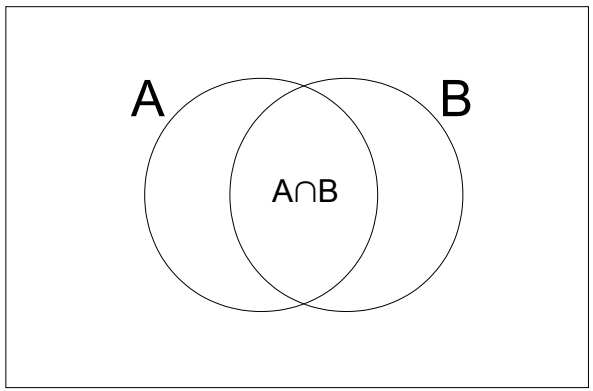
Image from the Wikipedia

Bayes' Theorem

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

Bayes' Theorem

The most elementary version



$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} \\
&= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}
\end{aligned}
$$

There are many versions of Bayes' Theorem

For discrete random variables,

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

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

For continuous random variables

$$
\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} \\
&= \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)\,dt}
\end{aligned}
$$

For X Continuous and Y Discrete

$$f_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} p_{Y|X}(y|t)f_X(t)\,dt}$$

Compare

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\
P(X = x|Y = y) &= \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)} \\
f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)\,dt} \\
f_{X|Y}(x|y) &= \frac{p_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} p_{Y|X}(y|t)f_X(t)\,dt}
\end{aligned}
$$

Philosophy
Bayesian versus Frequentist

- What is probability?
- Probability is a formal axiomatic system (Thank you Mr. Kolmogorov).
- *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.
- Maybe the Bayesian way is more natural.
  - Except for gambling games, is *any* experiment ever carried out repeatedly and independently a large number of times?
  - What's the probability of rain tomorrow?
  - What's the probability I would like fried grasshoppers?
  - What's the probability that these are not my real parents?
  - What is $P(-1.1 < \mu < 5.7)$?

Statistical inference

How it works

- Adopt a probability model for data X.
- Distribution of X depends on a parameter θ.
- Use observed value $X = x$ 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.
- Bayesian 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.
- We base all decisions about the parameter (including estimates) on the posterior distribution.

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.
- For all but the most bone-headed priors.

Bayes' Theorem

Continuous case

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x|t)\pi(t)\,dt}$$

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

- Proportional because $\int_{-\infty}^{\infty} f(x|t)\pi(t)\,dt$ is a constant with respect to θ.
- It's like those problems that say "The random variable X has density $f(x) = k\,e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. What is k?"
- If you can recognize $f(x|\theta)\pi(\theta)$, you don't need to do the integral in the denominator.

Bayes' Theorem

X discrete and θ continuous: Very common.

$$\begin{aligned} \pi(\theta|x) &= \frac{p(x|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} p(x|t)\pi(t)\,dt} \\ &\propto p(x|\theta)\pi(\theta) \end{aligned}$$

Bayes' Theorem

Most General Case

$$E(g(\theta|x)) = \frac{\int g(\theta) f(x|\theta) d\pi(\theta)}{\int f(x|\theta) d\pi(\theta)}$$

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 $i = 1, \ldots, n$,

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}\, I(x_i = 0, 1)$$

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

Prior: The Beta distribution

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\, I(0 \leq x \leq 1)$$

Where $\alpha > 0$ and $\beta > 0$.

Beta prior: $\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I(0 \leq x \leq 1)$

- 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

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)\,\pi(\theta) \\
&= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{(\alpha+\sum_{i=1}^n x_i)-1}(1-\theta)^{(\beta+n-\sum_{i=1}^n x_i)-1}
\end{aligned}$$

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

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

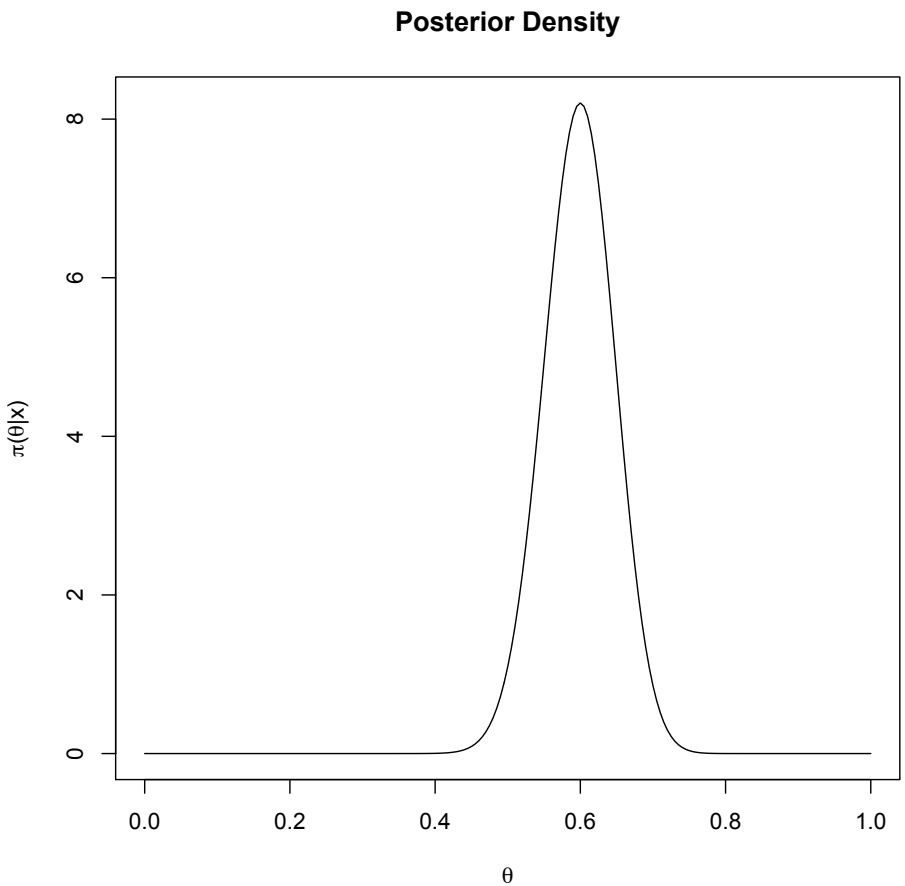
So that's it!

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.

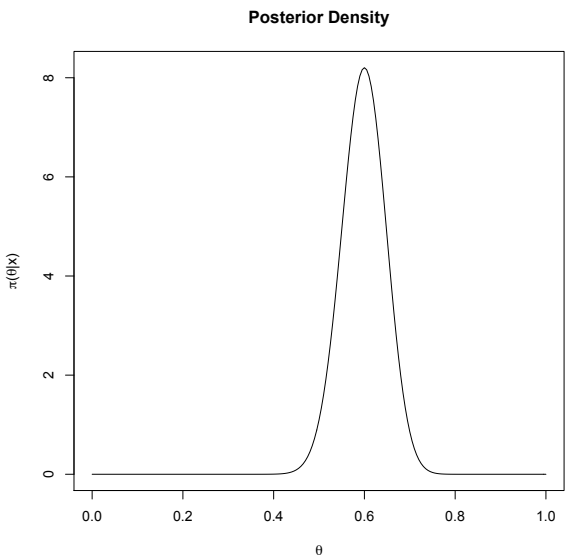
Picture of the posterior

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
$\beta' = \beta + n - \sum_{i=1}^{n} x_i = 41$.



More comments about the posterior distribution

Beta(α', β') with $\alpha' = \alpha + \sum_{i=1}^n x_i$ and $\beta' = \beta + n - \sum_{i=1}^n x_i$

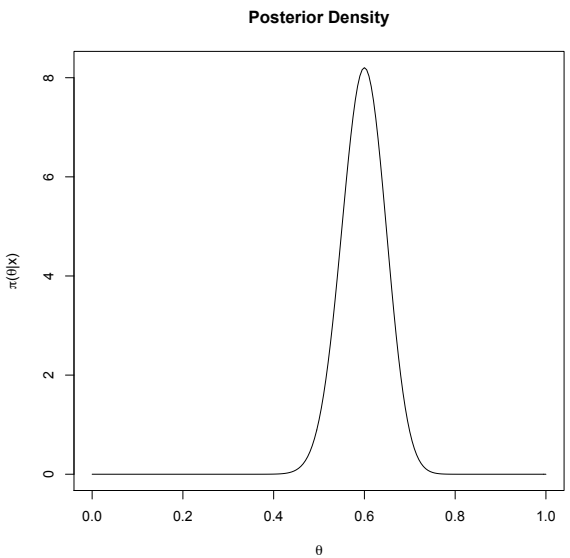


- The prior was uniform, but the posterior is much more concentrated.
- Peak is close to $\overline{x} = 0.6$.
- Note

$$
\begin{aligned}
E(\Theta|X) &= \frac{\alpha'}{\alpha' + \beta'} \\
&= \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n} \\
&= \frac{\alpha + n\overline{X}_n}{\alpha + \beta + n} \\
&= \frac{\alpha/n + \overline{X}_n}{\alpha/n + \beta/n + 1} \\
&\xrightarrow{p} \theta
\end{aligned}
$$

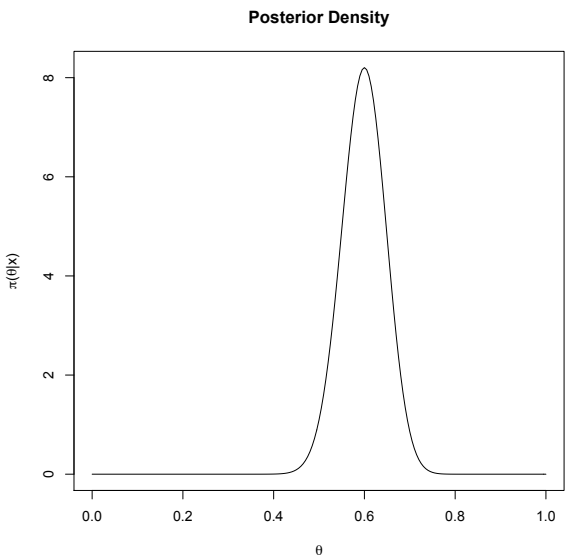
How should we estimate θ?

If we really insist on a point estimate



- To some pure Bayesians, the estimate is the posterior, period.
- The posterior expected value is a natural choice.
- The posterior mode is also popular.
- Because $\pi(\theta|x) \propto p(x|\theta)\ \pi(\theta)$ $= L(\theta, x)\ \pi(\theta)$, the posterior mode is a lot like the MLE.
- Actually, because $E(\Theta|x) \to \theta$ and $Var(\Theta|x) \to 0$, a randomly chosen point from the posterior is also a consistent estimator.
- More sophisticated answers are available from Bayesian decision theory. What's the cost of being wrong?

Test $H_0 : \theta = \frac{1}{2}$



- Because the prior probability of $\Theta = \frac{1}{2} = 0$, so is the posterior probability.
- The frequentist null hypothesis $H_0 : \theta = \frac{1}{2}$ is unbelievable.
- How about comparing $P(\Theta < \frac{1}{2}|x)$ to $P(\Theta > \frac{1}{2}|x)$?
- For this example, $P(\Theta < \frac{1}{2}|x) = 0.023$
- Again, more sophisticated answers are available from Bayesian decision theory.
- If being wrong either way is equally costly and it doesn't matter how much you're wrong, comparing the posterior probabilities is the optimal decision rule.

Copyright Information