Introduction to Bayesian Statistics[1]

STA 442/2101 Fall 2018

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

Thomas Bayes (1701-1761)

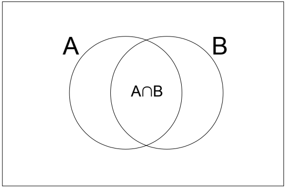
Image from the Wikipedia

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,

$$
\begin{aligned}
P(X = x|Y = y) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\
&= \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)}
\end{aligned}
$$

For continuous random variables

$$
\begin{aligned}
f_{x|y}(x|y) &= \frac{f_{xy}(x,y)}{f_y(y)} \\
&= \frac{f_{y|x}(y|x)\,f_x(x)}{\int f_{y|x}(y|t)\,f_x(t)\,dt}
\end{aligned}
$$

Compare

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

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.

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.
- Bayes methods can generate more specific recommendations.

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

Bayes' Theorem
Continuous case

$$
\begin{aligned}
\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|t)\pi(t)\,dt} \\
&\propto f(x|\theta)\pi(\theta)
\end{aligned}
$$

Bayes' Theorem
General case

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

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

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

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

Risk is Expected Loss

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

$$
\begin{aligned}
E(L) &= E(E[L|X]) \\
&= \int \left(\int L\,(d(x), \theta)\, d\pi(\theta|x) \right) dP(x)
\end{aligned}
$$

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 $i = 1, \ldots, n$,

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

for $x_i = 0$ or $x_i = 1$, and zero otherwise.

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

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 $\beta > 0$

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

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

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

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

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

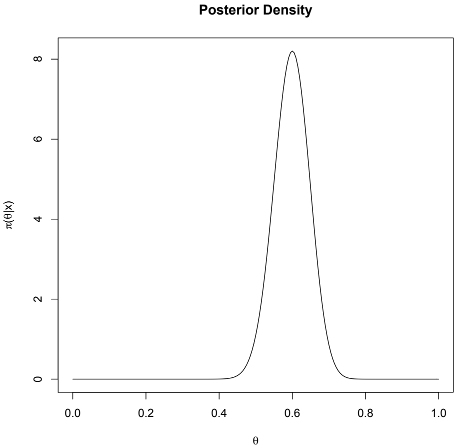
So that's it!

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



Estimation

- Question: How should I estimate θ?
- Answer to the question is another question: What is your loss function?
- First, what is the decision space?
- $\mathcal{D} = (0, 1)$, same as the parameter space.
- $d \in \mathcal{D}$ is a guess about the value of θ.
- The loss function is up to you, but surely the more you are wrong, the more you lose.
- How about squared error loss?
- $L(d, \theta) = k(d - \theta)^2$
- We can omit the proportionality constant k.

Minimize expected loss

$L(d, \theta) = (d - \theta)^2$

Denote $E(\theta|X = x)$ by μ. Then

$$
\begin{aligned}
E\left(L(d,\theta)|X=x\right) &= E\left((d-\theta)^2|X=x\right) \\
&= E\left((d-\mu+\mu-\theta)^2|X=x\right) \\
&= \ldots \\
&= E\left((d-\mu)^2|X=x\right) + E\left((\theta-\mu)^2|X=x\right) \\
&= (d-\mu)^2 + Var(\theta|X=x)
\end{aligned}
$$

- Minimal when $d = \mu = E(\theta|X = x)$, the posterior mean.
- This was general.
- The Bayes estimate under squared error loss is the posterior mean.

Back to the example

Give the Bayes estimate of θ under squared error loss.

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

```
> 61/(61+41)

[1] 0.5980392
```

Hypothesis Testing

$\theta > \frac{1}{2}$ means consumers tend to prefer the new blend of coffee.

Test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

- What is the loss function?
- When you are wrong, you lose.
- Try zero-one loss.

	Loss $L(d_j, \theta)$	
Decision	When $\theta \leq \theta_0$	When $\theta > \theta_0$
$d_0 : \theta \leq \theta_0$	0	1
$d_1 : \theta > \theta_0$	1	0

Compare expected loss for d_0 and d_1

	Loss $L(d_j, \theta)$	
Decision	When $\theta \leq \theta_0$	When $\theta > \theta_0$
$d_0 : \theta \leq \theta_0$	0	1
$d_1 : \theta > \theta_0$	1	0

Note $L(d_0, \theta) = I(\theta > \theta_0)$ and $L(d_1, \theta) = I(\theta \leq \theta_0)$.

$$\begin{aligned} E\left(I(\theta > \theta_0)|X = x\right) &= P\left(\theta > \theta_0|X = x\right) \\ E\left(I(\theta \leq \theta_0)|X = x\right) &= P\left(\theta \leq \theta_0|X = x\right) \end{aligned}$$

- Choose the smaller posterior probability of being wrong.
- Equivalently, reject H_0 if $P(H_0|X = x) < \frac{1}{2}$.

Back to the example

Decide between $H_0 : \theta \leq 1/2$ and $H_1 : \theta > 1/2$ under zero-one loss.

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

Want $P(\theta > \frac{1}{2}|X = x)$

```
> 1 - pbeta(1/2,61,41) # P(theta > theta0|X=x)

[1] 0.976978
```

How much worse is a Type I error?

	Loss $L(d_j, \theta)$	
Decision	When $\theta \leq \theta_0$	When $\theta > \theta_0$
$d_0 : \theta \leq \theta_0$	0	1
$d_1 : \theta > \theta_0$	k	0

To conclude H_1, posterior probability must be at least k times as big as posterior probability of H_0.
$k = 19$ is attractive.

A realistic loss function for the taste test would be more complicated.

Computation

- Inference will be based on the posterior.
- Must be able to calculate $E(g(\theta)|X = x)$
- For example, $E(L(d, \theta)|X = x)$
- Or at least
$$\int L(d, \theta) f(x|\theta)\pi(\theta)\, d\theta.$$
- If θ is of low dimension, numerical integration usually works.
- For high dimension, it can be tough.

Monte Carlo Integration to get $E(g(\theta)|X = x)$

Based on simulation from the posterior

Sample $\theta_1, \ldots, \theta_m$ independently from the posterior distribution

and calculate

$$\frac{1}{m}\sum_{j=1}^{m} g(\theta_j) \stackrel{a.s.}{\to} E(g(\theta)|X = x)$$

By the Law of Large Numbers.

Large-sample confidence interval is helpful.

Sometimes it's Hard

- If the posterior is a familiar distribution (and you know what it is), simulating values from the posterior should be routine.
- If the posterior is unknown or very unfamiliar, it's a challenge.

The Gibbs Sampler

Geman and Geman (1984)

- $\theta = (\theta_1, \ldots, \theta_k)$ is a random vector with a (posterior) joint distribution.
- It is relatively easy to sample from the conditional distribution of each component given the others.
- Algorithm, say for $\theta = (\theta_1, \theta_2, \theta_3)$:
  First choose starting values of θ_2 and θ_3 somehow. Then,
  - Sample from the conditional distribution of θ_1 given θ_2 and θ_3. Set θ_1 to the resulting number.
  - Sample from the conditional distribution of θ_2 given θ_1 and θ_3. Set θ_2 to the resulting number.
  - Sample from the conditional distribution of θ_3 given θ_1 and θ_2. Set θ_3 to the resulting number.

  Repeat.

Output

- The Gibbs sampler produces a sequence of random $(\theta_1, \theta_2, \theta_3)$ vectors.
- Each one depends on the past only through the most recent one.
- It's a Markov process.
- Under technical conditions (Ergodicity), it has a stationary distribution that is the desired (posterior) distribution.
- Stationarity is a $\to \infty$ concept.
- In practice, a "burn in" period is used.
- The random vectors are sequentially dependent.
- Time series diagnostics may be helpful.
- Retain one parameter vector every "n" iterations, and discard the rest.

Copyright Information