Bayesian methods in economics and finance

Motivation: Bayesian statistics and decision making
Bayesian statistics

What is *Bayesian statistics*?
- Principled and logically consistent way to reason under uncertainty
- Especially useful when taking decisions or making predictions

Long-standing debate

Bayesian vs frequentist statistics
What is probability?

Two (competing) philosophies:

- **Frequentist:** Probability of an event is *relative frequency of occurrence* when experiment is repeated infinitely many times.
- **Bayesian:** Probability describes (subjective) degree of belief
What is probability?

Two (competing) philosophies:

- **Frequentist**: Probability of an event is *relative frequency of occurrence* when experiment is repeated infinitely many times.
- **Bayesian**: Probability describes (subjective) degree of belief

Common wisdom

- Frequentist = objective
  Bayesian = subjective
- But:
  - Asymptotics vs finite sample
  - Repeated trials vs decision making
Motivation

Hypothetical situation:

▷ Assume that a patient enters your office
▷ Her test result for a rare disease is positive

Q: Would you suggest an expensive treatment?

What do you need to know for an informed decision?
Classical answer:

- Test between two hypotheses:
  - $H_0$: Patient is healthy (null hypothesis)
  - $H_1$: Patient is infected (alternative)
- Specificity of test, i.e.

$$P(\text{Test} = \text{negative}|H_0)$$

Here: Specificity of 99.9%
Classical answer:

Test between two hypothesis:
- $H_0$: Patient is healthy (null hypothesis)
- $H_1$: Patient is infected (alternative)

Specificity of test, i.e.

$$P(\text{Test} = \text{negative}|H_0)$$

Here: Specificity of 99.9%

Reject null hypothesis at level $\alpha$ if

$$P(\text{Test} = \text{positive}|H_0) = 1 - P(\text{Test} = \text{negative}|H_0) = \frac{1}{1000} < \alpha$$
Bayesian answer

*Basic idea:* Uncertainty about the status of the patient (healthy or infected) can be expressed as a probability distribution. Combine two sources of information:

- **Prior:** How rare is the disease?
  Here: Rare disease can be found in 1 out of 10000 persons, i.e.

  \[ P(\text{Status} = \text{infected}) = \frac{1}{10000} \]

- **Likelihood:** How good is the test?
  - Specificity: \( P(\text{Test} = \text{negative} | \mathcal{H}_0) \)
  - Sensitivity: \( P(\text{Test} = \text{positive} | \mathcal{H}_1) \)
  Here: Sensitivity of 99.9%

Base your decisions on **posterior**, i.e.

\[ P(\text{Status} | \text{Test} = \text{positive}) \]
Bayes rule

Bayes rule is used to calculate posterior probability:

\[
P(S|T) = \frac{P(S)P(T|S)}{P(T)} = \frac{P(S)P(T|S)}{\sum_{S'} P(S')P(T|S')} \propto P(S)P(T|S)
\]

\text{posterior} \propto \text{prior} \times \text{likelihood}

Bayes rule is uncontroversial and follows from the product rule of probability theory:

\[
P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)
\]
Bayesian answer

Want to know posterior probabilities:

\[
P(\text{Status = infected} | \text{Test = positive}) = \frac{P(S = \text{inf}) P(T = \text{pos} | S = \text{inf})}{P(S = \text{inf}) P(T = \text{pos} | S = \text{inf}) + P(S = \text{hea}) P(T = \text{pos} | S = \text{hea})}
\]

\[
= \frac{\frac{1}{10000} \times \frac{999}{1000}}{\frac{1}{10000} \times \frac{999}{1000} + \frac{9999}{10000} \times \frac{1}{1000}}
\]

\[
\approx 0.09
\]

Thus, patient is not very likely to be infected and we would need more evidence before suggesting the expensive treatment!
Bayesian thinking

Bayesian statistics is conceptually simple

\[ \text{posterior} \propto \text{prior} \times \text{likelihood}, \]

but can be computationally demanding

Every type of uncertainty is expressed in terms of probability distributions. This includes

- Statistical models of data, e.g.
  \[ P(\text{Test}|\text{Status} = \text{infected}) \]

- Plausibility of hypothesis, e.g.
  \[ P(\text{Status} = \text{healthy}) \]

- Parameters

In general, probabilities are assigned to logical statements. Conclusions are derived by computing their posterior probabilities.
Decision theory

Bayesian statistics is deeply rooted in decision making. Subjective probabilities can be recovered from betting odds:

- Assume you are willing to accept a bet at 1:19, i.e.
  - You pay 1$ if you loose
  - You get 19$ if you win

Your risk neutral probability of winning is then $\frac{1}{20}$ as it leaves you indifferent between accepting the bet or not, i.e.

$$E[payout] = (1 - \frac{1}{20})(-1$) + \frac{1}{20}19$ = 0$$

*Dutch book argument*:

- A dutch book is a set of bets such that you loose money no matter what happens.
- Coherent betting odds have to fullfil the laws of probability.
In finance Dutch book coherence is known as *no arbitrage*:

- Consider a bet at odds 1:a, i.e. with risk neutral probability $q = \frac{1}{1+a}$
- Equivalent to a lottery ticket which costs $q$ and pays 1 on winning

Now, the total price of tickets winning on disjoint events $A$ and $B$ respectively, must equal the price of ticket winning on $A \cup B$, i.e.

$$q(A) + q(B) = q(A \cup B) \quad \text{for } A \cap B = \emptyset$$

stating that $q$ obeys the sum rule of probability. Similarly, other laws are proved for coherent bets.
Summary: Bayesian statistics is about reasoning under uncertainty. Rationality demands that subjective belief can be modelled as probabilities!

Example: Should you believe this coin is fair?
- Consider a coin that has been tossed 20 times
- Suppose that we observed 15 heads
Classical estimation

- Let $\theta$ denote the (unknown) bias of our coin
- Probability to observe $k$ heads on $n$ tosses is given by the Binomial distribution

$$P(H = k | n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Maximum likelihood estimate $\hat{\theta}_{ML}$ obtained as

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(H = k | n, \theta) = \frac{k}{n}$$

$p$-value that coin is unbiased:

$$P(H \geq k | n, \theta = \frac{1}{2}) = \sum_{i=k}^{n} \binom{n}{k} 2^{-n} \approx 0.02$$
Classical estimation

Classical estimators $\hat{\theta}(D)$ are statistics, i.e. functions of observed data $D$.

Properties of estimators include

- **Bias:**
  \[
  \mathbb{E}_\theta[\hat{\theta}(D)]
  \]
  Estimator is called unbiased if $\mathbb{E}_\theta[\hat{\theta}(D)] = \theta$

- **Variance:**
  \[
  \mathbb{E}_\theta[(\hat{\theta}(D) - \mathbb{E}[\hat{\theta}])^2]
  \]

- **Mean squared error:**
  \[
  \mathbb{E}_\theta[(\hat{\theta}(D) - \theta)^2]
  \]

- **(Asymptotic) consistency:**
  \[
  \lim_{n \to \infty} P(\left| \hat{\theta}(D_n) - \theta \right| < \epsilon) = 1 \quad \forall \epsilon > 0
  \]

Bayesian’s usually do not care about any of these (except for consistency)!
Consider coin bias $\theta$ as a random variable with prior probability distribution $p(\theta)$

Compute the posterior

$$p(\theta|\text{Data}) \propto p(\theta)p(\text{Data}|\theta)$$
Bayesian estimation

- Consider coin bias $\theta$ as a random variable with prior probability distribution $p(\theta)$
- Compute the posterior

\[ p(\theta|Data) \propto p(\theta)p(Data|\theta) \]

Convenient choice for the prior is a Beta distribution

\[ p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \]

- with the Gamma function $\Gamma(x)$ (Note: $\Gamma(x + 1) = x\Gamma(x)$)
- mean $\mathbb{E}[\theta] = \frac{\alpha}{\alpha + \beta}$
- and variance $\mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$
Bayesian estimation

Posterior with beta prior is found to be

\[ p(\theta | k, n) \propto p(\theta | \alpha, \beta)p(H = k | n, \theta) \]
\[ \propto \theta^{\alpha - 1}(1 - \theta)^{\beta - 1} \binom{n}{k} \theta^k(1 - \theta)^{n-k} \]
\[ \propto \theta^{\alpha + k - 1}(1 - \theta)^{\beta + (n-k) - 1} \]

Thus, posterior is again a Beta distribution with parameter \( \alpha_k = \alpha + k, \beta_k = \beta + (n - k) \):

- This is an example of a \textit{conjugate prior}
- The \textit{hyperparameters} \( \alpha, \beta \) can be considered as pseudo-counts, i.e. interpreted as virtual observations
Bayesian learning

Bayesian learning: Updating information from prior to posterior
This can always be done \textit{sequentially}:

1. Assume two independent observations $D_1, D_2$ (or split the one you have), e.g. coin tossed 10 time for 7 heads and then another 10 times for 8 heads

2. Then,

\begin{align*}
p(\theta|D_1, D_2) \propto p(\theta)p(D_1, D_2|\theta) \\
= p(\theta)p(D_2|\theta)p(D_1|\theta)
\end{align*}

since $D_1$ and $D_2$ are conditionally independent given $\theta$

\begin{align*}
\propto p(\theta|D_1)p(D_2|\theta)
\end{align*}

Thus, posterior $p(\theta|D_1)$ serves as prior when learning from $D_2$
Illustration of coin toss example:

- In R the density of the Beta distribution is available as
  \[
  \text{dbeta}(x, \alpha, \beta) \]

- Above example can be plotted as follows:

```r
n <- 20; k <- 15
alpha <- 1; beta <- 1
x <- seq(0, 1, 0.01)
# Posterior
plot(x, dbeta(x, alpha + k, beta + (n-k)), type='l', col='red')
# Prior
lines(x, dbeta(x, alpha, beta), type='l', col='blue')
```
Point estimates

Posterior $p(\theta|D)$ summarizes knowledge about $\theta$ after data $D$ was observed
How can we construct a point estimate, i.e. $\hat{\theta}_{Bayes}$?
  ▶ Could simply take posterior mean, median or mode . . .
Point estimates

Posterior $p(\theta|D)$ summarizes knowledge about $\theta$ after data $D$ was observed.

How can we construct a point estimate, i.e. $\hat{\theta}_{Bayes}$?

- Could simply take posterior mean, median or mode . . .
- More principled approach considers estimation as a decision problem:
  - Define loss function $L : \Theta \times \Theta \rightarrow \mathbb{R}$ that specifies cost of estimating $\hat{\theta}$ when true parameter was $\theta$.
    Loss function satisfies $L(\hat{\theta}, \theta) \geq 0$ with equality if and only if $\hat{\theta} = \theta$.
  - Point estimate is decision rule which minimizes expected loss, also called *Bayes risk*:
    \[
    \hat{\theta}_{Bayes} = \arg\min_{\hat{\theta}(D)} \mathbb{E}[L(\hat{\theta}(D), \theta)]
    \]
    Note: Expectation is taken over joint distribution $p(\theta, D) = p(\theta)p(D|\theta)$.
Point estimates

Common loss functions include

- **0-1-loss** (for categorical values):

\[
L(\hat{\theta}, \theta) = \begin{cases} 
0 & \text{if } \hat{\theta} = \theta \\
1 & \text{if } \hat{\theta} \neq \theta
\end{cases}
\]

\[\Rightarrow \hat{\theta}_{\text{Bayes}} \text{ is posterior mode}\]

- **Squared error**:

\[
L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2
\]

\[\Rightarrow \hat{\theta}_{\text{Bayes}} \text{ is posterior mean}\]

- **Absolute error**:

\[
L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|
\]

\[\Rightarrow \hat{\theta}_{\text{Bayes}} \text{ is posterior median}\]
Point estimates

Posterior mean minimizes mean squared error, i.e.

$$\arg\min_{\hat{\theta}(D)} \mathbb{E}_{p(\theta,D)}[(\hat{\theta}(D) - \theta)^2] = \mathbb{E}_{p(\theta|D)}[\theta]$$

Proof:

$$\mathbb{E}_{p(\theta,D)}[(\hat{\theta}(D) - \theta)^2] = \int p(\theta,D)(\hat{\theta}(D) - \theta)^2 d\theta dD$$

$$= \int p(D) \int p(\theta|D)(\hat{\theta}(D) - \theta)^2 d\theta dD$$

Now, for each $D$:

$$\frac{\partial}{\partial \hat{\theta}} \int p(\theta|D)(\hat{\theta} - \theta)^2 d\theta = 2 \int p(\theta|D)(\hat{\theta} - \theta)d\theta$$

$$= 2 \left( \int p(\theta|D)\hat{\theta}d\theta - \int p(\theta|D)\theta d\theta \right)$$

Setting derivative to zero, it follows that

$$\hat{\theta}(D) = \int p(\theta|D)\theta d\theta = \mathbb{E}_{p(\theta|D)}[\theta]$$
Point estimates

Back to coin example:

- Posterior is Beta distribution with parameters $\alpha + k$ and $\beta + (n - k)$
- Posterior mean is

$$\frac{\alpha + k}{\alpha + k + \beta + (n - k)} = \frac{\alpha + k}{\alpha + \beta + n}$$

- Defining $m = \alpha + \beta$ posterior mean can be written as

$$\frac{m}{n + m} \frac{\alpha}{\alpha + \beta} + \frac{n}{n + m} \frac{k}{n}$$

- Convex combination between prior mean $\theta_0 = \frac{\alpha}{\alpha + \beta}$ and ML estimate $\hat{\theta}_{ML} = \frac{k}{n}$
- Weights correspond to relative number of (pseudo-)observations
Sampling properties

Classical, i.e. sampling properties of Bayesian estimators:

- Generally biased: Pulls estimate towards prior mean
  Note: Bias reduces variance!

- **Bernstein-von Mises theorem:**
  
  *Bayesian estimate is asymptotically consistent, i.e. posterior mean converges to true parameter.*
  
  *Posterior approximately normal in large n limit*

Under some technical conditions (valid for finite dimensional parameter spaces) and assuming that prior support includes true parameter
Admissibility

Define the expected loss of an estimator $\hat{\theta}$ with respect to $L$ and true parameter $\theta$ as

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta [L(\hat{\theta}(D), \theta)]$$

Note: Expectation over sampling distribution $p(D|\theta)$

- Say that $\hat{\theta}'$ dominates $\hat{\theta}$ if

$$R(\hat{\theta}', \theta) \leq R(\hat{\theta}, \theta) \quad \forall \theta$$

where the inequality is strict for some $\theta$

- Estimator $\hat{\theta}$ is called **admissible** if there is no estimator which dominates it.

Could argue that any reasonable estimator should be admissible (for some loss function)
Admissibility

Theorem:

Every Bayes estimator is admissible

Proof: Let \( \hat{\theta} \) be a Bayes estimator minimizing the Bayes risk \( \mathbb{E}[L(\hat{\theta}, \theta)] \) with prior \( p(\theta) \).
Assume that \( \hat{\theta}' \) dominates \( \hat{\theta} \). Then,

\[
\mathbb{E}[L(\hat{\theta}', \theta)] = \int p(D, \theta) L(\hat{\theta}', \theta) dD d\theta \\
= \int p(\theta) R(\hat{\theta}', \theta) \\
< \int p(\theta) R(\hat{\theta}, \theta) \\
= \mathbb{E}[L(\hat{\theta}, \theta)]
\]

contradicting the assumption that \( \hat{\theta} \) is a Bayes estimator.
Admissibility

- Converse is also true: If $\hat{\theta}$ is admissible than it must be a (generalized) Bayes estimator for some prior $p(\theta)$.
  Every reasonable estimator is based on implicit prior assumptions!

- Many widely used estimators are not admissible, e.g. James-Stein theorem:

  \[ \text{The empirical mean } \hat{\mu} = \frac{1}{N} \sum_{i} x_i \text{ is inadmissible in dimension } > 2 \]

  Dominated by estimator shrinking towards zero!
Likelihood principle

Summary of Bayesian estimation:

▶ Posterior combines information from prior and likelihood of data:

\[ \text{posterior} \propto \text{prior} \times \text{likelihood} \]

▶ Prior represents (subjective) belief/information before data are obtained

Sequential learning: Prior can arise from data learned about previously

Likelihood principle:

Data enters via likelihood \( p(D|\theta) \) only, i.e. inference from \( D_1 \) and \( D_2 \) is the same when \( p(D_1|\theta) = p(D_2|\theta) \)
**Likelihood principle**

Summary of Bayesian estimation:

- Posterior combines information from prior and likelihood of data:
  
  $posterior \propto prior \times likelihood$

- Prior represents (subjective) belief/information before data are obtained
  
  Sequential learning: Prior can arise from data learned about previously

**Likelihood principle:**

Data enters via likelihood $p(D|\theta)$ only, i.e. inference from $D_1$ and $D_2$ is the same when $p(D_1|\theta) = p(D_2|\theta)$

**Q:** Would your estimate change if I told you that

- coin was tossed 20 times
- or coin was tossed until 15 heads were obtained?
## Summary

<table>
<thead>
<tr>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong> $\theta$</td>
<td>Unknown, but fixed</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Likelihood $p(D</td>
</tr>
<tr>
<td><strong>Inference</strong></td>
<td>Point estimate $\hat{\theta}(D)$</td>
</tr>
<tr>
<td></td>
<td>▶ Repeated sampling</td>
</tr>
<tr>
<td></td>
<td>▶ Null hypothesis testing</td>
</tr>
<tr>
<td><strong>Logic</strong></td>
<td>Reason based on</td>
</tr>
<tr>
<td></td>
<td>▶ prob. of potential data</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Course outline

- Motivation:
  - Bayesian statistics and decision making
  - Common distributions and conjugate priors

- Linear regression:
  - Data modeling with *Stan*
  - Bayesian model selection and sparsity priors

- Data modeling:
  - Vector Autoregression (VAR)
  - Volatility models
  - Hierarchical models
  - Bayesian nonparametrics

- Algorithms:
  - Sampling methods
  - Variational approximations

Project work:

- Pick topic or paper
- Writeup of 5-10 pages