Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Lecture 6 : Bayesian Inference

Ideas of probability

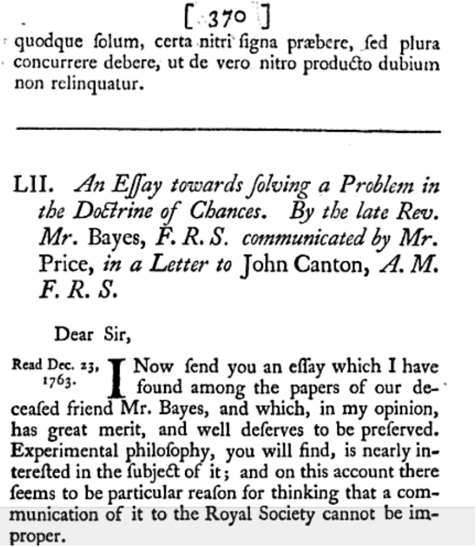
The majority of statistics you have learned so are are called classical or Frequentist. The probability for an event is defined as the proportion of successes in an infinite number of repeatable trials.

By contrast, in Subjective Bayesian inference, probability is a measure of the strength of belief.

We treat parameters as random variables. Before collecting any data we assume that there is uncertainty about the value of a parameter. This uncertainty can be formalised by specifying a pdf (or pmf) for the parameter. We then conduct an experiment to collect some data that will give us information about the parameter. We then use Bayes Theorem to combine our prior beliefs with the data to derive an updated estimate of our uncertainty about the parameter.

The history of Bayesian Statistics

- Bayesian methods originated with Bayes and Laplace (late 1700s to mid 1800s).
- In the early 1920's, Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation i.e. the repeated sampling principle.
- In 1939 Jeffrey's book 'The theory of probability' started a resurgence of interest in Bayesian inference.
- This continued throughout the 1950-60s, especially as problems with the Frequentist approach started to emerge.
- The development of simulation based inference has transformed Bayesian statistics in the last 20-30 years and it now plays a prominent part in modern statistics.



[370]

quodque ſolum, certa nitri ſigna præbere, ſed plura concurrere debere, ut de vero nitro producto dubium non relinquatur.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

Principles of Bayesian inference

Lack of knowledge is understood as uncertainty : modelling via probability distribution

In a statistical model : $X|\theta \sim f(\cdot;\theta)$ with $\theta \in \Theta$

- Ultimately we observe X : the only thing *we know*
- θ is unknown : modelled as a random variable : $\theta \sim \pi$
- $f(\cdot;\theta)$: density of the conditional distribution of X given θ

$$f(X;\theta) = f(X|\theta), \quad \Rightarrow \quad \underbrace{(X,\theta)}_{\text{joint distribution}} \quad \sim f(X|\theta) \times \pi(\theta)$$

- Bayesian Inference based on posterior distribution : conditional distribution of θ given X

Bayesian Inference - Revison

Likelihood $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

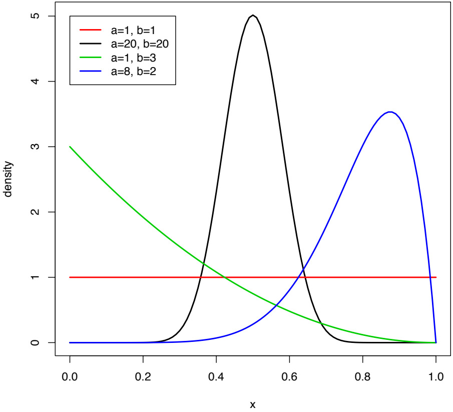
$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta} \Rightarrow \quad \begin{array}{ccc} \pi(\theta \mid x) & \propto & f(x \mid \theta)\pi(\theta) \\ \text{posterior} & \propto & \text{likelihood} \times \text{prior} \end{array}$$

The same form for θ continuous ($\pi(\theta \mid x)$ a pdf) or discrete ($\pi(\theta \mid x)$ a pmf). We call $\int f(x \mid \theta)\pi(\theta)d\theta$ the marginal likelihood.

Example 1

$X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ. Suppose our prior knowledge about ϑ is represented by a Beta distribution on $(0, 1)$, and θ is a trial value for ϑ.

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$



Example 1

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x},\ x = 0, \dots, n$$

Posterior probability density

$$\begin{aligned}\pi(\theta \mid x) &\propto \text{likelihood} \times \text{prior} \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{x}(1-\theta)^{n-x} \\ &= \theta^{a+x-1}(1-\theta)^{n-x+b-1}\end{aligned}$$

Here posterior has the same form as the prior (conjugacy) with updated parameters a, b replaced by $a+x, b+n-x$, so

$$\pi(\theta \mid x) = \frac{\theta^{a+x-1}(1-\theta)^{n-x+b-1}}{B(a+x, b+n-x)}$$

A bit on the controversy between Bayesian and frequentist inference

(Some) interesting aspects of Bayesian inference

- Unified framework : all statements are coherent, natural notions of uncertainty
- Fondamentally based on modelling : both sampling model $(X|\theta)$ and prior model $\pi(\theta)$
  It can also be viewed as a downfall : more demanding
- You condition on what you observe (i.e. what you know) and integrate out what you do not know
- However statistics is not a religion : take whatever works

Why we should think conditionally on X : a simple example

Frequentist statements of uncertainty are often unconditional
A confidence interval is a set-valued function $C(X) \subseteq \Theta$ of the data X which covers the parameter $\theta \in C(X)$ a fraction $1-\alpha$ of repeated draws of X taken under the null H_0.

This is not the same as the statement that, given data $X = x$, the interval $C(x)$ covers θ with probability $1-\alpha$. But this is the type of statement we might wish to make. (observe that this statement makes sense iff θ is a r.v.)

Example 1 Suppose $X_1, X_2 \sim U(\theta - 1/2, \theta + 1/2)$ so that $X_{(1)}$ and $X_{(2)}$ are the order statistics. Then $C(X) = [X_{(1)}, X_{(2)}]$ is a $\alpha = 1/2$ level CI for θ. Suppose in your data $X = x$, $x_{(2)} - x_{(1)} > 1/2$ (this happens in an eighth of data sets). Then $\theta \in [x_{(1)}, x_{(2)}]$ with probability one.

The likelihood principle

The likelihood principle Suppose that two experiments relating to θ, E_1, E_2, give rise to data y_1, y_2, such that the corresponding likelihoods are proportional, that is, for all θ

$$L(\theta; y_1, E_1) = cL(\theta; y_2, E_2).$$

then the two experiments lead to identical conclusions about θ.

Key point MLE's respect the likelihood principle i.e. the MLEs for θ are identical in both experiments. But moment estimators do not, nor do p-values Bayesian inference does respect the likelihood principle:
If two likelihood functions are proportional, then any constant cancels top and bottom in Bayes rule, and the two posterior distributions are the same.

Example

A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$\begin{aligned} L(p; y_1, E_1) &= \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1-y_1} \quad \textit{binomial} \\ L(p; n_2, E_2) &= \binom{n_2-1}{y_2-1} p^{y_2}(1-p)^{n_2-y_2} \quad \textit{negative binomial} \end{aligned}$$

If $n_1 = n_2 = n$, $y_1 = y_2 = y$ then $L(p; y_1, E_1) \propto L(p; n_2, E_2)$.
So MLEs for p will be the same under E_1 and E_2.

Example

But significance tests contradict : eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$.

The p-value based on E_1 is

$$P\left(Y \leq y|\theta = \frac{1}{2}\right) = \sum_{k=0}^{y} \binom{n}{k} (1/2)^k (1-1/2)^{n-k} (= 0.073)$$

while the p-value based on E_2 is

$$P\left(N \geq n|\theta = \frac{1}{2}\right) = \sum_{k=n}^{\infty} \binom{k-1}{y-1} (1/2)^k (1-1/2)^{n-k} (= 0.033)$$

so different conclusions at significance level 0.05.

Note The p-values disagree because they sum over portions of two different sample spaces.

Bayesian inference & decision theory

Bayesian inference is based on the posterior distribution

But how do we construct estimators ? measures of uncertainty ? tests ?

Using decision theory

A first introduction to decision theory : loss, risks & Bayes estimates

- sampling Model $X|\theta \sim f(X;\theta)$, Decision set : $\mathcal{D}$

Definition (Loss & risks)

- A loss function is any function

$$L : \Theta \times \mathcal{D} \to \mathbb{R}^+$$

- Frequentist risk of $\delta(.) : \mathcal{X} \to \mathcal{D}$

$$R(\theta, \delta) = E_X[L(\theta, \delta(X)); \theta] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x;\theta) dx$$

- Posterior risk

$$\rho(\pi, \delta|x) = \mathbb{E}^{\pi}[L(\theta, \delta(x))|x] = \int_{\Theta} L(\theta, \delta(x)) \pi(\theta|x) d\theta$$

Integrated risk and Bayes estimator

Definition (and Theorem)

- Integrated risk

$$r(\pi, \delta) = \int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta = \int_{\mathcal{X}} \rho(\pi, \delta|x)m(x)dx$$

where $m(x)$ is the marginal distribution of X.

- Bayes estimator : For all $x \in \mathcal{X}$

$$\delta^{\pi}(x) = \text{argmin}_{\delta \in \mathcal{X}} \rho(\pi, \delta|x)$$

When it exists,

$$\delta^{\pi} = \text{argmin}\left\{r(\pi, \delta), \delta : \mathcal{X} \to \mathcal{D}\right\}$$

Proof : Fubini

$$\begin{aligned} r(\pi,\delta) &= \int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \int_\mathcal{X} L(\theta,\delta(x)) f(x;\theta)\pi(\theta)d\theta dx \\ &= \int_\mathcal{X}\int_\Theta L(\theta,\delta(x))\frac{f(x;\theta)\pi(\theta)}{m(x)}d\theta m(x)dx \\ &= \int_\mathcal{X}\int_\Theta L(\theta,\delta(x))\pi(\theta|x)d\theta m(x)dx \\ &= \int_\mathcal{X} \rho(\pi,\delta|x)m(x)dx \end{aligned}$$

Let δ^π such that for all x and all $\delta(x)$

$$\begin{aligned} &\rho(\pi,\delta(x)|x) \geq \rho(\pi,\delta^\pi(x)|x) \\ &\quad\Rightarrow \quad \int_\mathcal{X} \rho(\pi,\delta|x)m(x)dx \geq \int_\mathcal{X}\rho(\pi,\delta^\pi|x)m(x)dx \end{aligned}$$

Examples

- Quadratic loss function $L(\theta, \delta) = \|\theta - \delta\|_2^2$

$$\delta^\pi(x) = E^\pi(\theta|x) = \int_\Theta \theta\pi(\theta|x)d\theta$$

- L_1 loss: $L(\theta, \delta) = \sum_j |\theta_j - \delta_j| = \|\theta - \delta\|_1$

$\delta^\pi(x)$ is the posterior median (componentwise)

- $0 - 1$ loss function $\Theta = \{0, \cdots, k\}$ $k \geq 1$ and $L(\theta, \delta) = \mathbb{1}_{\theta \neq \delta}$

$$\delta^\pi(x) = \text{argmax}\{\pi(\theta|x), \quad \theta \in \{0, \cdots, k\}\}$$

Example 1 —- continued $X \sim \mathcal{B}(n, p)$

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \; \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior mean and variance are

$$\delta^\pi(x) = \frac{a+X}{a+b+n}, \; V(\theta|x) = \frac{(a+X)(b+n-X)}{(a+b+n)^2(a+b+n+1)}$$

Suppose $X = n$ and we set $a = b = 1$ for our prior. Then posterior mean is

$$\frac{n+1}{n+2}$$

i.e. when we observe events of just one type then our point estimate is not 0 or 1 (which is sensible especially in small sample sizes).

Example 1

For large n, the posterior mean and variance are approximately

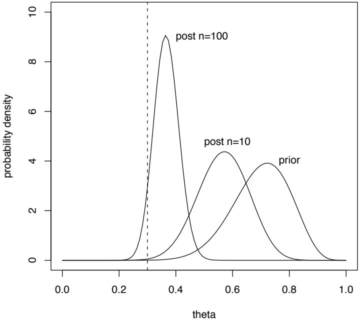
$$\frac{X}{n},\ \frac{X(n-X)}{n^3}$$

In classical statistics

$$\widehat{\theta} = \frac{X}{n},\ \frac{\widehat{\theta}(1-\widehat{\theta})}{n} = \frac{X(n-X)}{n^3}$$

Example 1

Still supposing $X \sim \text{Bin}(n, \vartheta)$, with Beta prior such that mean is 0.7 with std 0.1. Suppose we observe $X = 3$ when $n = 10$ and then $X = 30$ when $n = 100$.



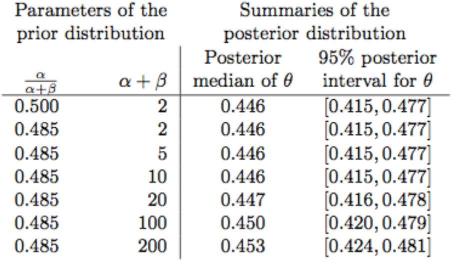
As n increases, the likelihood overwhelms information in prior.

Example: estimating the probability of female birth given placenta previa

Result of german study: 980 birth, 437 females. In general population the proportion is 0.485.
Using a uniform (Beta(1,1)) prior, posterior is Beta(438,544).

$$\text{post. mean } = 0.446 \quad \text{post. std dev } = 0.016$$
$$\text{central 95\% post. interval } = [0.415, 0.477]$$

Sensibility to proposed prior. $\alpha + \beta - 2 =$ "prior sample size".



Parameters of the prior distribution		Summaries of the posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	Posterior median of θ	95% posterior interval for θ
0.500	2	0.446	$[0.415, 0.477]$
0.485	2	0.446	$[0.415, 0.477]$
0.485	5	0.446	$[0.415, 0.477]$
0.485	10	0.446	$[0.415, 0.477]$
0.485	20	0.447	$[0.416, 0.478]$
0.485	100	0.450	$[0.420, 0.479]$
0.485	200	0.453	$[0.424, 0.481]$