Outline: Introudction to Bayesian inference

1) Classical inference

2) Prior distributions

3) Bayes Theorem

4) Bayesian updating

5 Bayesian confidence intervals

6 Bayesian hypothesis testing

Classical Inference

STATISTICAL INFERENCE:
making conclusions about a "population" from "sample" items drawn from that population

Example:
We wish to estimate the proportion of trees in a large forest which suffer from a particular disease.

A sample of n trees are selected. If θ denote the proportion of trees having the disease in the forest, then each tree in the sample will have the disease, independently of all others in the sample, with probability θ.

Let X be the random variable corresponding to the number of diseased trees in the sample, we will use the observed value of $X = x$, to draw an inference about the population parameter θ.

Classical Inference

This inference can take the form of

a *point estimate* ($\hat{\theta} = 0.1$);

a *confidence interval* (95% confident that θ lies in the range [0.08,1.2]);

a *hypothesis test* (reject the hypothesis that $\theta < 0.07$ at the 5% significance level);

a *prediction* (predict that 15% of trees will be affected by the next year);

a *decision* (decide to identify and remove all infected trees)

Classical Inference

typically, these inferences are made by specifying a probability model,

$$l(x|\theta)$$

which is the likelihood function.

Our model in our example would be

$$X|\theta \sim Bin(n, \theta)$$

maximise the likelihood of $X = x$ with respect to θ – maximum likelihood estimate.

Classical Inference

the parameter θ (though unknown) is being treated as a constant rather than *random* – cornerstone of classical theory

but this leads to problems of interpretation: we'd like a 95% confidence interval of [0.08,1.2] to mean there's a 95% probability that θ lies between 0.08 and 1.2. It *cannot* since θ is fixed, the only random element is the data

so the correct interpretation of the interval is that if we applied our procedure "many times", then in the "long run", the interval we construct will contain θ on 95% of occasions

Bayesian Inference

Fundamental difference: θ is treated as a *random* quantity, inferences are made in terms of *probability* statements.

Inference is based on (*posterior probability distribution*)

$$p(\theta|x)$$

obtained via Baye's rule

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(\theta)p(x|\theta)}{p(x)}$$

Bayesian Inference

denominator

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

is independent of θ for fixed x, and can be considered as a constant

the *unnormalised posterior density* is

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

$p(\theta)$ is a *prior probability distribution*, represents beliefs about the distribution of θ *prior* to having any information about the data. (contentious issue over the pro-cons of Bayesian thinking)

Prior Distribution $p(\theta)$

when trying to estimate θ, we almost always have some knowledge, or belief about the value of θ before we take account of the data

Characteristics of the Bayesian approach

We can identify four fundamental aspects which characterize the Bayesian approach to statistical inference

Prior information: all problems are unique and have their own context. That context derives prior information, and it is the formulation and exploitation of that prior knowledge which sets Bayesian inference apart from classical statistics.

Subjective Probability: Classical statistics hinges on an objective "long-run-frequency" definition of probabilities. Even if this is desirable, which is arguable, it leads to cumbersome inferences. By contrast, Bayesian statistics formalizes explicitly the notion that all probabilities are subjective, depending on an individual's beliefs and knowledge to hand. Thus, a Bayesian analysis is personalistic – unique to the specifications of each individual's prior beliefs. Inference is based on the *posterior* distribution $p(\theta|x)$, whose form will be seen to depend on (through Bayesian theorem) the particulars of the prior specification $p(\theta)$.

Self-consistency: By treating the parameter θ as random, it emerges that the whole development of Bayesian inference stems quite naturally from probability theory only. This has many advantages, and means that all inferential issues can be addressed as probability statements about θ, which then derive directly from the posterior distribution.

no "adhockery": Because classical inference cannot make probability statements about θ, various criteria are developed to judge whether a particular estimator is in some sense "good". This has led to a proliferation of procedures, often in conflict with one another. Bayesian inference sidesteps this tendency to invent *ad hoc* criteria for judging and comparing estimators by relying on the posterior distribution to express in straightforward probabilistic terms the entire inference about the unknown θ.

The implementation of Bayes' Theorem in practice can be computationally difficult, mainly as a result of the normalising integral in the denominator.

For some choices of prior-likelihood combination, this integral can be avoided, but in general, specialised techniques are required to simplify this calculation.

Binomial Bayes estimation

Example: (Binomial sample) Suppose our likelihood model is $x \sim Bin(n, \theta)$, and we wish to make inferences about θ.

So

$$p(x|\theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}; x = 0, \ldots, n$$

the choice of prior for θ will vary from problem to problem. We proceed by considering a possible family of prior distributions which, give rise to simple computations.

Binomial Bayes estimation

So, suppose we can represent our prior beliefs about θ by a Beta distribution:

$$\theta \sim Be(p, q)$$

so that

$$\begin{aligned} p(\theta) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\theta^{p-1}(1-\theta)^{q-1} \quad (0 \leq \theta \leq 1) \\ &\propto \theta^{p-1}(1-\theta)^{q-1} \end{aligned}$$

then by Bayes Theorem

$$\begin{aligned} p(\theta|x) &\propto p(\theta)p(x|\theta) \\ &\propto \theta^{x}(1-\theta)^{n-x} \times \theta^{p-1}(1-\theta)^{q-1} \\ &= \theta^{p+x-1}(1-\theta)^{q+n-x-1} \end{aligned}$$

Binomial Bayes estimation

since we know that $p(\theta|x)$ is a proper density function, it must be the case that

$$\theta|x \sim Be(p+x, q+n-x)$$

the effect of the data is to modify the parameters of the beta distribution from their prior values of (p, q) to the posterior values of $(p+x, q+n-x)$, we have avoided the need to calculate any integrals.

An estimate for θ is the mean of the posterior distribution, which would give us the Bayes estimator for θ

$$\hat{\theta} = \frac{x+p}{p+q+n}$$

Binomial Bayes estimation

Notice that the prior has mean

$$p/(p+q)$$

which is the best estimate without having seen the data.

Ignoring prior information, we would probably use $\theta = x/n$ as our estimate of θ.

The Bayes estimate of θ combines all of this information, since we can write $\hat{\theta}$ as

$$\hat{\theta} = \left(\frac{n}{p+q+n}\right)\left(\frac{x}{n}\right) + \left(\frac{p+q}{p+q+n}\right)\left(\frac{p}{p+q}\right)$$

Binomial Bayes estimation

So, $\hat{\theta}$ is a linear combination of the prior mean and the sample mean, with the weights determined by p, q, n.

When estimating the Binomial parameter, it is not necessary to choose a prior distribution from the Beta family. However, there was a certain advantage to choosing the beta family, since we have a closed-form expression for the estimator. In general, for any sampling distribution, there is a natural family of prior distributions, called the **Conjugate family**.

Conjugate priors

Conjugate Prior: if $\mathcal{F}$ is a class of sampling distributions $p(x|\theta)$ and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta|x) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

Conjugate priors

Example:(Gamma sample) Suppose $X_1, \ldots, X_n$ are independent variables having the $Ga(k, \theta)$ distribution, where k is known.

Then

$$l(\theta; x) \propto \theta^{nk} \exp\{-\theta \sum x_i\}.$$

this form (as a function of θ) suggests we could take a prior of the form

$$p(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

i.e., $\theta \sim Ga(p, q)$, then by Bayes' theorem

$$f(\theta|x) \propto \theta^{p+nk-1} \exp\{-(q + \sum x_i)\theta\},$$

so $\theta|x \sim Ga(p+nk, q+\sum x_i)$, is in the same family as the prior distribution.

Poisson distribution

Example: (Poisson sample) Suppose $X_1, \ldots, X_n$ are independent Poisson(θ) variables. The

$$\begin{aligned} l(\theta; x) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \\ &\propto e^{-n\theta}\theta^{\sum x_i} \end{aligned}$$

Suppose the prior beliefs can be represented by a Gamma distribution, $\theta \sim Ga(p, q)$, so

$$p(\theta) = \frac{q^p}{\Gamma(p)}\theta^{p-1}\exp\{-q\theta\} \quad (\theta > 0)$$

Then by Bayes Theorem,

$$\begin{aligned} p(\theta|x) &\propto \frac{q^p}{\Gamma(p)}\theta^{p-1}\exp\{-q\theta\} \times \exp\{-n\theta\}\theta^{\sum x_i} \\ &\propto \theta^{(p+\sum x_i-1)}\exp\{-(q+n)\theta\} \end{aligned}$$

and so

$$\theta|x \sim Ga(p + \sum_{i=1}^{n} x_i, q+n)$$

another Gamma distribution whose parameters are modified by the data through $\sum_{i=1}^{n} x_i$ and n.

Normal mean

Example: (Normal mean): Let $X_1, \ldots, X_n$ be a set of independent variables from $N(\theta, \sigma^2)$ where σ^2 is known.

Then

$$p(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{-\frac{(x_i - \theta)^2}{2\sigma^2}\}$$

giving a likelihood

$$l(\theta; x) \propto \exp\{-\frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2\sigma^2}\}$$

Take as prior, $\theta \sim N(b, d^2)$, then our posterior

$$\begin{aligned}
p(\theta|x) &\propto \exp\{-\frac{(\theta-b)^2}{2d^2}\}\exp\{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\} \\
&= \exp\{-\frac{(\theta-b)^2}{2d^2} - \frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\} \\
&= \exp\{-\frac{\theta^2-2b\theta+b^2}{2d^2} - \frac{\sum_{i=1}^n x_i^2-2n\bar{x}\theta+n\theta^2}{2\sigma^2}\} \\
&\propto \exp\{-\frac{1}{2}[\theta^2(\frac{1}{d^2}+\frac{n}{\sigma^2}) - 2\theta(\frac{b}{d^2}+\frac{n\bar{x}}{\sigma^2})]\} \\
&\propto \exp\{-\frac{1}{2}(\frac{1}{d^2}+\frac{n}{\sigma^2})[\theta - \frac{\frac{b}{d^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}]^2\}
\end{aligned}$$

thus,

$$\theta|x \sim N(\frac{\frac{b}{d^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{d^2}+\frac{n}{\sigma^2}})$$

Credibility Intervals

the idea of a credibility interval is to give an analogue of a confidence interval in classical statistics.

the reasoning is that point estimates give no measure of accuracy, so it is preferable to give an interval within which it is "likely" that the parameter lies.

this causes problems in classical statistics since parameters are not regarded as random, so it is not possible to give an interval with the interpretation that there is a certain probability that the parameter lies in the interval

Credibility Intervals

there is no such difficulty in the Bayesian approach because parameters are treated as random, thus a region $C_\alpha(x)$ is a $100(1-\alpha)\%$ *credible region* for θ if

$$\int_{C_\alpha(x)} p(\theta|x)d\theta = 1-\alpha$$

that is, there is a probability of $1-\alpha$, based on the posterior distribution, that θ lies in $C_\alpha(x)$.

one difficulty with credible intervals (in common with confidence intervals) is that they are not uniquely defined. Any region with probability $1-\alpha$ will do.

Credibility Intervals

since we want the interval to contain only the "most plausible" values of the parameter, it is usual to impose additional constraint which is that the width of the interval should be as small as possible.

This amounts to an interval (or region) of the form

$$C_\alpha(x) = \{\theta : p(\theta|x) \geq \gamma\}$$

where γ is chosen to ensure that

$$\int_{C_\alpha(x)} p(\theta|x)d\theta = 1 - \alpha$$

Credibility Intervals

such regions are called "highest posterior density regions".

In general they are found numerically

Example: (Normal means)

Let $X_1, \ldots, X_n$ independent variables from $N(\theta, \sigma^2)$ (σ^2 known) with a prior for θ of the form $\theta \sim N(b, d^2)$.

with this construction we obtain the posterior

$$\theta|x \sim N(\frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}})$$

now, since the Normal distribution is uni-modal and symmetric it follows that the $100(1-\alpha)\%$-highest posterior density region for θ is:

$$(\frac{\frac{b}{d^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}) \pm z_{\alpha/2}(\frac{1}{\frac{1}{d^2}+\frac{n}{\sigma^2}})^{\frac{1}{2}},$$

where $z_{\alpha/2}$ is the appropriate percentage point of the standard normal $N(0,1)$ distribution. notice that as $n \to \infty$ this interval becomes

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n},$$

which is precisely the $100(1-\alpha)\%$ confidence interval for θ obtained in classical inference.

Credibility Intervals

Example:

Suppose $x \sim Bin(n, \theta)$ with the prior

$$\theta \sim Beta(p, q)$$

this gives the posterior distribution

$$\theta | x \sim Beta(p + x, q + n - x)$$

thus the $100(1 - \alpha)\%$ highest posterior density interval $[a, b]$ satisfies:

$$\frac{1}{Be(p + x, q + n - x)} \int_a^b \theta^{p+x-1} (1 - \theta)^{q+n-x-1} d\theta = 1 - \alpha$$

and

$$\frac{1}{Be(p+x,q+n-x)}a^{p+x-1}(1-\alpha)^{q+n-x-1}$$
$$=\frac{1}{Be(p+x,q+n-x)}b^{p+x-1}(1-b)^{q+n-x-1}=\gamma$$

generally this has to be solved numerically. (except in the special case $x=0$, where an analytical solution can be obtained)

Bayesian Hypothesis testing

In a hypothesis testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true.

Remember that $p(\theta|x)$ is a probability distribution for a random variable. Hence the posterior probabilities

$$p(\theta \in \Theta_0|x) = P(H_0 \text{ is true}|x)$$

and

$$p(\theta \in \Theta_0^c|x) = P(H_1 \text{ is true } |x)$$

can be computed.

One way a Bayesian hypothesis tester may choose to use the posterior distribution to decide to accept H_0 as true if

$$p(\theta \in \Theta_0|x) > p(\theta \in \Theta_0^c|x)$$

and to reject H_0 otherwise.

That is, the rejection region is

$$\{x : p(\theta \in \Theta_0^c | x) > 1/2\}.$$

Alternatively, if the Bayesian hypothesis tester wishes to guard against falsely rejecting H_0, he may decide to reject H_0 only if $p(\theta \in \Theta_0^c | x)$ is greater than some larger number, e.g., 0.99.

Example

Let $X_1, \ldots, X_n$ be iid $N(\theta, \sigma^2)$ and let the prior distribution on θ be $N(\mu, \tau^2)$, where σ^2, μ, τ^2 are known. Consider testing

$$H_0 : \theta \leq \theta_0$$

against

$$H_1 : \theta > \theta_0.$$

The posterior distribution $\pi(\theta|\bar{x})$ is $N(\frac{n\tau^2\bar{x}+\sigma^2\mu}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2})$.

If we decide to accept H_0 if and only if $p(\theta \in \Theta_0|x) > p(\theta \in \Theta_0^c|x)$, then we will accept H_0 if and only if

$$1/2 \leq p(\theta \in \Theta_0|x) = P(\theta < \theta_0|x).$$

Since the posterior distribution is symmetric, this is true if and only if the mean of the posterior is less than or equal to θ_0. Therefore H_0 will be accepted as true if

$$\bar{X} \leq \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}$$

and H_1 will be accepted as true otherwise.