Aims

- Bayes' Theorem
- Single parameter models

  Binomial Model
- Summarizing Posterior Inference
- Informative Prior Distributions
- Conjugacy
- Exponential Families and Sufficient Statistics
- Example: Bayesian inference under a binomial model

Bayesian Inference

The Rev. Thomas Bayes published in the 1763 a paper entitled *"An essay towards solving a problem in the doctrine of chances"*. This paper introduced the concept of inverse probability.

- $H_1, \dots, H_k$ set of hypotheses
- $P(H_i),\ i = 1, \dots, k$ prior probabilities, $\sum_i P(H_i) = 1$
- $P(A \mid H_i),\ i = 1, \dots, k$ likelihoods of the data A

Bayes' Theorem

$$P(H_i \mid A) = \frac{P(A|H_i)P(H_i)}{\sum_{j=1}^{k} P(A|H_j)P(H_j)}$$

The posterior probability of H_i given A is proportional to the product of the prior probability of H_i and the likelihood of A when H_i is true

Suppose we are interested in two particular hypotheses H_i and H_j. The posterior ratio is given by:

$$\frac{P(H_i|A)}{P(H_j|A)} = \frac{P(A|H_i)}{P(A|H_j)} \times \frac{P(H_i)}{P(H_j)}$$

that is, by the product of the prior odds and the likelihood ratio.

Example: use of the relative frequencies

- There were 12 games with point spreads of 8 points; the outcomes in those games were:

  $$-7, -5, -3, -3, 1, 6, 7, 13, 15, 16, 20, 21$$

  with positive value indicating wins by the favorite and negative values indicating wins by the underdog
- $P(\text{favorite wins} \mid \text{point spread} = 8) = \frac{8}{12}$
- $P(\text{fav.wins by at least 8} \mid \text{p. spread} = 8) = \frac{5}{12}$
- $P(\text{fav.wins by at least 8} \mid \text{p. spread} = 8\&\text{fav. wins}) =$??

Medical diagnosis

- a patient may belong to state H_1 (presence of disease) or H_2 (absence of disease)
- $P(H_1)$ is the prevalence rate of the disease in the population to which the patient is assumed to belong
- information: $D = T$ (presence of disease), or $D = T^c$ absence of disease
- $P(T \mid H_1), P(T^c \mid H_2)$ are the true positive and the true negative rates of the clinical test
- Bayes theorem enables us to understand the manner in which these characteristics of the test combine with the prevalence rate to produce varying degrees of diagnostic discrimination power.

Medical diagnosis (cont)

- goal: assessment of the diagnostic value of scientigraphy, as an indicator of coronary artery disease
- controlled experiment concluded that $P(T \mid H_1) = .9$ and $P(T^c \mid H_2) = .875$ were reasonable order of magnitude for the sensitivity and specificity of the test.
- $P(H_1 \mid D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1)+P(D|H_2)P(H_2)}$

- as a single, overall measure of the discriminatory power of the test, one may consider the difference $P(H_1 \mid T) - P(H_1 \mid T^c)$
- in cases where $P(H_1)$ has very low or very high values (large pop. screening or following an individual patient referred on the basis of suspected coronary disease) then there is limited diagnostic value in the test
- if there is a considerably uncertainty about the presence of coronary disease, $.25 \leq P(H_1) \leq .75$, the test may be expected to provide valuable diagnostic information.

Subjective probability

- to assign probability distributions to parameter like θ, may not be consistent with the usual long-term frequency notion of probability. Let
- $\theta =$*true prob. of success for a new surgical procedure*
- here it is possible to think of θ as the limiting value of the observed success rate as the procedure is independently repeated again and again
- $\theta =$*true proportion of US men who are HIV-positive*
- the long-term frequency notion does not apply, the randomness of θ does not arise from any real world mechanism.

- θ is random only because it is unknown *to us*, though we may have some feelings about it ($\theta = .05$ is more likely than $\theta = .5$

Bayesian analysis is predicated on such a belief in subjective probability, wherein we quantify whatever feelings (however vague) we may have about θ before having looked at the data y in the distribution $p(\theta)$. This distribution is then updated by the data via Bayes' theorem with the resulting posterior distribution — $p(\theta \mid y)$ — reflecting a blend of the information in the data and in the prior

Single parameter models

Binomial Model

- y is the total numbers of successes in a trial
- θ is the probability of success in each trial
- $p(y \mid \theta) \propto \theta^y (1-\theta)^{n-y}$ is the likelihood
- Example: estimating the probability of female birth

  Two hundred years ago it was established that the proportion of female births in European population was less than .5

  Let y the number of girls reported in n recorded births

- By appealing the Binomial model, we are assuming that the n births are conditionally independent given θ, with the probability of a female birth equal to θ for all cases (exchangeability assumption)

How do we perform Bayesian inferences?

- Specify a prior for θ — we assume $\theta \sim U[0,1]$
- Apply Bayes Rule — $p(\theta \mid y) \propto p(y \mid \theta) p(\theta)$
- Look at $p(\theta \mid y)$ — i.e. mean, variance, regions ..

here we found

$$p(\theta \mid y) \propto \theta^y (1-\theta)^{n-y}$$

this is recognizable like a Beta$(\theta \mid y+1, n-y+1)$.

Prior Prediction: Bernoulli

- $y_i \sim$ Bern(θ) and
- $\theta \sim U[0,1]$
- calculate

$$\begin{aligned} p(y_i = 1) &= \int_0^1 p(y_i = 1 \mid \theta) p(\theta) d\theta \\ &= \int_0^1 \theta p(\theta) d\theta \\ &= E[\theta] = \tfrac{1}{2} \end{aligned}$$

Prior Prediction: Binomial

- $y = \sum_{i=1}^{n} y_i \sim$ Binomial(n, θ) and
- $\theta \sim U[0,1]$
- calculate

$$\begin{aligned} p(y) &= \int_0^1 p(y \mid \theta) p(\theta) d\theta \\ &= \binom{n}{y} \int_0^1 \theta^y (1-\theta)^{n-y} p(\theta) d\theta \\ &= \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \\ &= \frac{1}{n+1} \end{aligned}$$

PS: $\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = 1$

Posterior Prediction

- the posterior predictive distribution of a "future" binary outcome $\tilde{y}$ given the observed n successes $y = \sum_{i=1}^{n} y_i$ is

$$\begin{aligned} P(\tilde{y} = 1 \mid y) &= \int_0^1 P(\tilde{y} = 1 \mid \theta, y) p(\theta \mid y) d\theta \\ &= \int_0^1 \theta p(\theta \mid y) d\theta \\ &= E(\theta \mid y) = \frac{y+1}{n+2} \end{aligned}$$

Frequentist approach

Given $\boxed{\theta}$, what are the probabilities of various possible outcomes of the random variable $\boxed{y}$?

"Weak law of large number" (Bernoulli)

$$y \sim \text{Bin}(n, \theta)$$

$$lim_{n \to \infty} P\left(\left| \frac{y}{n} - \theta \right| > \epsilon \mid \theta \right) \to 0$$

Bayesian approach

Given $\boxed{y}$, what are the probabilities of various possible outcomes of the random variable $\boxed{\theta}$?

Bayes' Rule, y continuous

$$\begin{aligned} p(\theta \mid y) &= \frac{p(\theta)p(y|\theta)}{p(y)} \\ p(y) &= \int p(\theta) p(y \mid \theta) d\theta \end{aligned}$$

- $E(\theta) = E(E(\theta \mid y))$
- $V(\theta) = E(V(\theta \mid y)) + V(E(\theta \mid y))$

Posterior represents a compromise between the prior and the data, and the compromise is controlled by the data as the sample size increases.

Binomial inference

- $E(\theta) = \frac{1}{2}$
- $\hat{\theta} = \frac{y}{n}$
- $E(\theta \mid y) = \frac{y+1}{n+2}$

Summarizing Posterior Inference

- Mean $E(\theta \mid y)$
- Mode $\hat{\theta} := p(\theta \mid y) = \max p(\theta \mid y)$
- Central Interval: *range of values above and below which lies* $(100\ \alpha/2\%)$ *post. prob.*
- Region with Highest Posterior Density: *region of values that contains* $100(1-\alpha)\%$ *of the post. prob. and that the density within the region is never lower than outside*

CI$\neq$HPD when the posterior is bimodal or skewed

CI$=$HPD when the posterior is unimodal and symmetric

Informative Prior Distributions

Population: the prior represents a population of possible parameter values from which the θ of current interest has been drawn

State of Knowledge: we must express our knowledge about θ as if its value could be thought of as a random realization from the prior.

Historical Justification of uniform prior

- Bayes' justification: for the binomial example it leads uniform predictive distribution — $p(y) = 1/n+1$
- Laplace's rationale: *principle of the insufficient reason*, i.e. "if nothing is known about θ then the uniform is appropriate"

Binomial Example

- likelihood: $\theta^y(1-\theta)^{n-y} \propto \text{Bin}(y \mid \theta, n)$
- prior: $\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \text{Beta}(\theta \mid \alpha, \beta)$
- posterior: likelihood $\times$ prior
  $\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \propto \text{Beta}(\theta \mid \alpha+y, \beta+n-y)$

Conjugacy: *the posterior distribution follows the same parametric form in as the prior.* Beta prior is conjugate to the binomial likelihood.

- $E(\theta \mid y) = \frac{\alpha+y}{\alpha+\beta+n}$
- $V(\theta \mid y) = \frac{E(\theta|y)(1-E(\theta|y))}{\alpha+\beta+n+1}$

As n *increases the prior has not influence on the posterior*

Conjugate Prior distributions

- $\mathcal{F}$ is the class of the sampling distributions
- $\mathcal{P}$ is the class of the prior distributions

$\mathcal{P}$ is natural conjugate for $\mathcal{F}$ if $\mathcal{P}$ is the set of all the densities having the same functional form in θ as the likelihood

Conjugate priors are useful because

- it is easy to understand the results (analytic forms of the mean, variance,..)
- simplify calculations
- good starting points
- you can use mixture of conjugate families

We can always use non-conjugate.......

Exponential Families and Sufficient Statistics

Probability distributions that belong to an exponential family have natural conjugate prior distribution

The class $\mathcal{F}$ is an exponential family if all its members have the form:

$$p(y \mid \theta) \propto g^n(\theta) \exp(\phi(\theta) t(y))$$

the natural conjugate prior is:

$$p(\theta) \propto g^\eta(\theta) \exp(\phi(\theta)\nu)$$

posterior ... very easy:

$$p(\theta \mid y) \propto g^{(n+\eta)}(\theta) \exp\left(\phi(\theta)(t(y) + \nu)\right)$$

- Binomial is an exponential family with $\phi(\theta) = \text{logit}(\theta)$
- How about Normal, Cauchy, lognormal, Poisson?

- The exponential families are the only classess of distributions that have natural conjugate prior distribution because they have a fixed number of sufficient statistics.

Example

- Question: How much evidence supports the hypothesis that the proportion of female births in the population of the placenta previa births – θ – is less than .485, the proportion of female births in the general populations?
- Study in Germany found that of a 980 placenta previa births, 437 were female
- likelihood: $\theta^{437}(1-\theta)^{980-437}$
- prior: $U[0,1] = \text{Beta}(\theta \mid 1, 1)$
- posterior: $\theta^{437}(1-\theta)^{980-437} = \text{Beta}(\theta \mid 438, 544)$
- $P[\theta \leq .485 \mid y] = .9928$


```
#posterior inference under a binomial model
#example pag 39 sec 2.5
binomial.beta_function(w=1,theta=seq(0.3,.6,.001),alpha=1,beta=1){
  if (w == 1) postscript("/home/biostats/fdominic/teaching/BM/binomialbeta.p
  plot(theta,dbeta(theta,436,543),type="l",xlab="theta",ylab="",
       xaxs="i",yaxs="i",yaxt="n",bty="n",cex=2)  #likelihood
  lines(theta,dbeta(theta,alpha,beta),lty=2)      #prior
  lines(theta,dbeta(theta,437+alpha,543+beta),lty=2) #posterior
  abline(v=.485)
  CI_quantile(rbeta(1000,437+alpha,543+beta),probs=c(.025,.975))
  abline(v=CI[1],lty=2)
  abline(v=CI[2],lty=2)
  par(oma=c(0,0,0,0))
  par(mfrow=c(1,1))
  if (w == 1) dev.off()}
}
```

Bayesian inference under a Binomial model

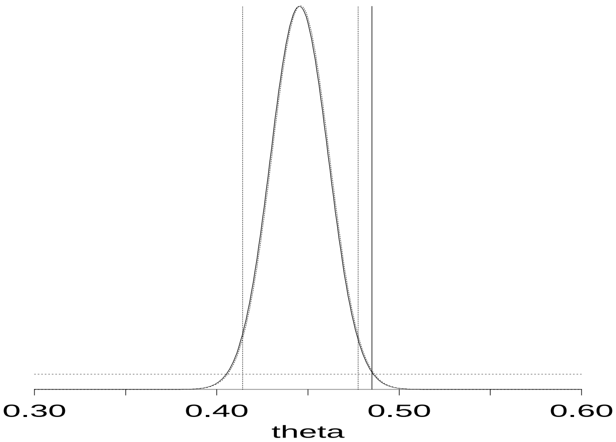

Figure 1: Likelihood, prior, posterior and 95% posterior interval for θ.

Sensitivity analysis

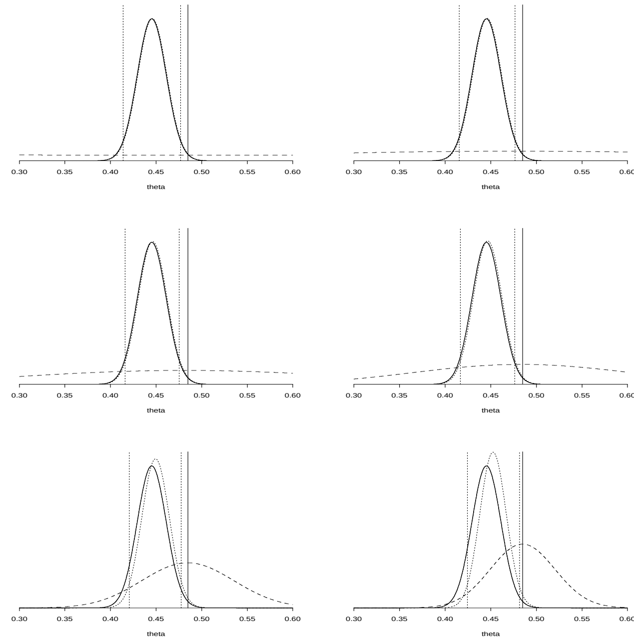

Figure 2: Likelihood, prior, posterior and 95% posterior interval for θ for different prior specifications.