Statistical Models and Inference - Part II

Alberto Garfagnini

Università di Padova

AA 2020/2021 - Stat Lect. 6

Introduction

Arguments treated

- ▷ estimate a posterior probability density over a model parameter given a data set
- ▷ focus on single parameter problems and the use of conjugate priors
- ▷ study how likelihood and prior combine to build posterior probability and how the latter depends on the amount of available data
- ▷ how to assign priors and summarize distributions

Bayesian analysis of coin tossing

Problem

- we have a coin and we toss it n times
- the coin lands heads in r of them

Q is the coin fair ? (i.e. $p = \frac{1}{2}$)

Comment

- no definitive answer exists
- only a probabilistic answer can be provided
- we are looking for

$$P\left(p \mid n, r, M\right)$$

- from Bayes' theorem

$$P\left(p \mid n, r, M\right) = \frac{P\left(r \mid p, n, M\right)\, P\left(p \mid M\right)}{P\left(r \mid n, M\right)}$$

Comment: n is not part of the Prior since it is independent of the number of coin tosses

Coin tossing model and probabilities

Our Measurement Model

- p : probability of getting heads in one toss
- p is constant in all the tosses
- all tosses are independent

The Likelihood

- the appropriate Likelihood is the binomial distribution

$$P\left(r \mid p, n, M\right) = \binom{n}{r} p^r (1-p)^{n-r} \quad \text{with} \quad r \leq n$$

Comment: n is part of the data, but it is on the right side since it is fixed before starting to collect data

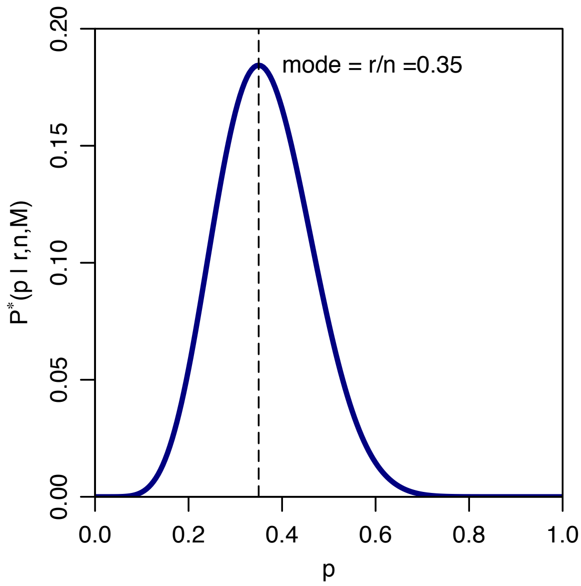
Coin tossing : a uniform Prior

- let's adopt a uniform prior, $P\left(p \mid M\right) \sim \mathcal{U}(0,1)$
- the Posterior pdf is simply proportional to the Likelihood

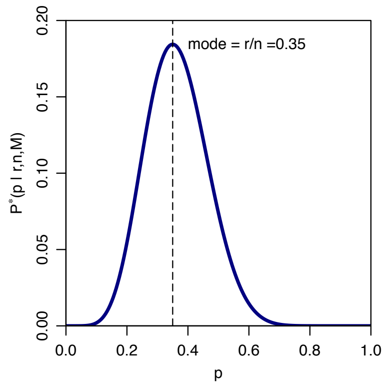
$$P\left(p \mid r,n,M\right) = \frac{1}{Z}\, p^{r}(1-p)^{n-r} = \frac{1}{Z} P^{*}\left(p \mid r,n,M\right)$$

- the normalization factor Z (i.e. the evidence $P\left(r \mid n,M\right)$ does not depend on p
- the mode is at r/n

```
n <- 20
r <- 7
p <- seq(0, 1, length.out = 201)
p.post <- dbinom(x=r, size=n, prob=p)

plot(p, p.post,
     xaxs='i', yaxs='i', col='navy',
     type='l', lty=1, lwd = 3,
     ylim=c(0,0.2),
     xlab="p",
     ylab=expression(paste(P^symbol("*"),
                     "(p | r,n,M)")))
```



Uniform Prior



Comments

- the curve is not binomial in p, but it is binomial in r
- the posterior is not-normalized: the integral over p is not unity
- we need the normalization factor only if we want to calculate expected values: i.e. mean and variance

- given the un-normalized posterior pdf, $P^{*}\left(p \mid r,n,M\right)$,

$$E[p] = \int_0^1 p \cdot P\left(p \mid r,n,M\right)\, dp = \frac{1}{Z}\int_0^1 p \cdot p^{r}(1-p)^{n-r}\, dp$$

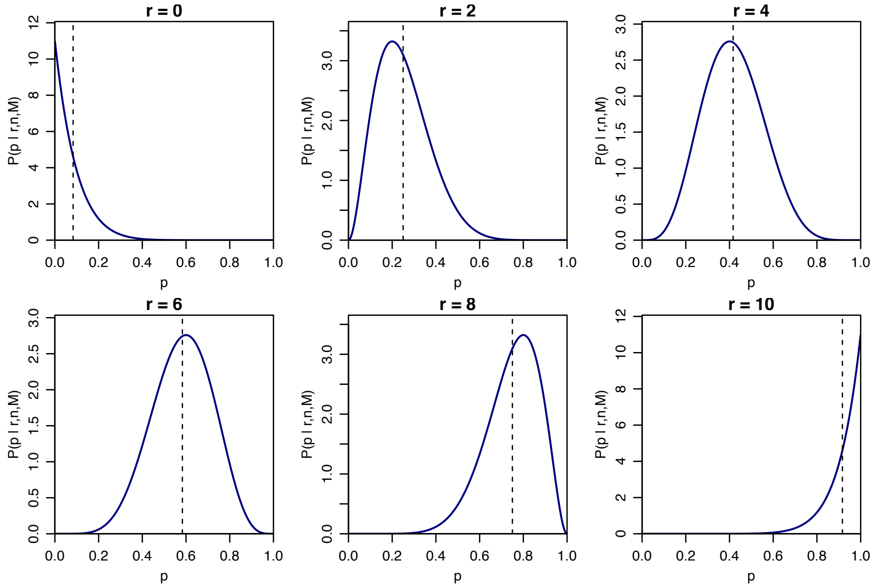
- with

$$Z = \int_0^1 P^{*}\left(p \mid r,n,M\right)\, dp \approx \sum_j P^{*}\left(p_j \mid r,n,M\right)\Delta p_j$$

- estimated using numerical integration

Uniform Prior

```
n <- 10; n.sample <- 2000; delta.p <- 1/n.sample
p <- seq(from=1/(2*n.sample), by=1/n.sample, length.out=n.sample)

for(r in seq(from=0, to=10, by=2)) {
  p.star <- dbinom(x=r, size=n, prob=p)
  p.norm <- p.star/(delta.p*sum(p.star))
  plot(p, p.norm, type="l", lwd=1.5, col='navy',
       xlim=c(0,1), ylim=c(0,1.1*max(p.norm)),
       xaxs="i", yaxs="i", xlab="p", ylab="P(p␣|␣r,n,M)")
  title(main=paste("r␣=",r), line=0.3, cex.main=1.2)
  p.mean <- delta.p*sum(p*p.norm)
  abline(v=p.mean, lty=2)
}
```



- interval $[0, 1]$ is divided into `n.sample` intervals
- un-normalized pdf is evaluated at the center of each point
- a grid of probability is created
- with the normalized posterior, the expected value is computed

Coin tossing : a Beta Prior

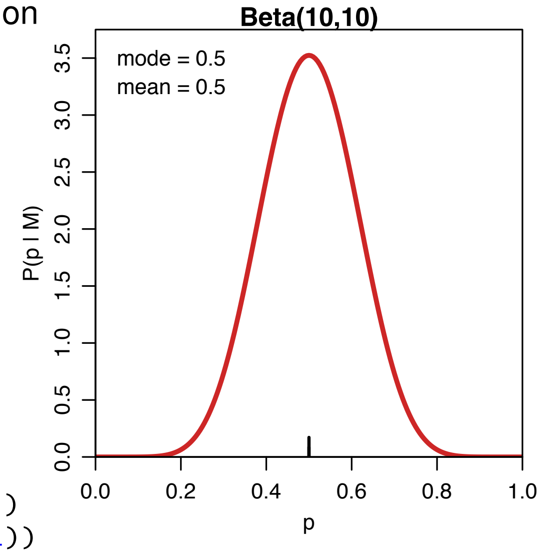
- given a random coin, we may believe the coin is fair, or close to fair
- an appropriate probability density function is the Beta distribution

$$P\left(p \,\middle|\, r, n, M\right) = \frac{1}{\mathrm{B}(\alpha, \beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \quad \text{with } \alpha > 0\,, \beta > 0$$

Note: for $\alpha = \beta = 1$ we get a uniform distribution

- if $\alpha = \beta$ the function is symmetric, and the mean and mode are 0.5
- the larger α (when $\alpha \geq 1$), the narrower the distribution

```
alpha <- 10; beta  <- 10
p <- seq(0, 1, length.out = 201)
p.prior <- dbeta(p, alpha, beta)
plot(p, p.prior, xaxs='i', yaxs='i',
     col='navy', type='l', lty=1, lwd = 3,
     ylim=c(0,3.75),
     xlab="p", ylab=paste("P(p␣|␣M)"),
     main=paste("Beta(",alpha,",",beta,")"))
mode <- (alpha - 1)/(alpha + beta - 2)
lines(c(mode, mode), c(0, 0.2), lty=5, lwd=2)
mean <- alpha/(alpha + beta)
lines(c(mean, mean), c(0, 0.2), lty=2, lwd=2)
text(0.05, 3.5, adj=0, paste("mode␣=␣", mode))
text(0.05, 3.25, adj=0, paste("mean␣=␣", mean))
```



Beta Prior

- multiplying the Prior by the likelihood, and absorbing the terms not depending on p in the constant term Z, we get

$$\begin{aligned} P\left(p \mid r, n, M\right) &= \frac{1}{Z}\, p^{r}(1-p)^{n-r} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= \frac{1}{Z}\, p^{r+\alpha-1}(1-p)^{n-r+\beta-1} \end{aligned}$$

- multiplying the Posterior with this Likelihood, we get the same form for the Posterior (another Beta distribution)
- the normalization constant is

$$Z = \mathrm{B}(r+\alpha, n-r+\beta)$$

- we say the Prior and Posterior are conjugate distributions

▷ the Prior is the *conjugate Prior* for this Likelihood function

Beta Prior

- if we start with a Beta Prior with parameters α_p and β_p, and then measure r heads in n tosses, the Posterior is a Beta functions with parameters

$$\alpha = \alpha_p + r \quad \text{and} \quad \beta = \beta_p + n - r$$

- mean and mode for the Posterior are

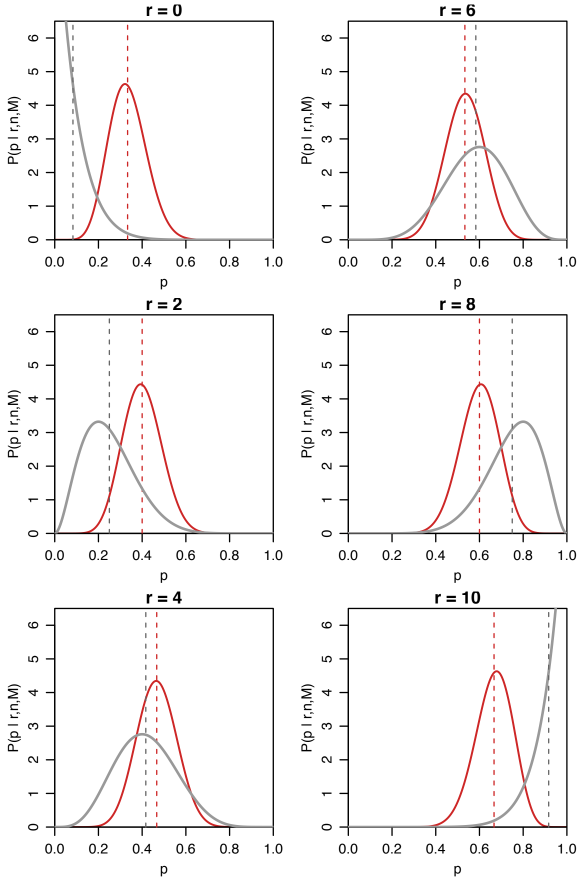
$$\text{mean} = \frac{\alpha_p + r}{\alpha_p + \beta_p + n} \quad \text{and} \quad \text{mode} = \frac{\alpha_p + r - 1}{\alpha_p + \beta_p + n - 2}$$

- if we compare the result with that obtained with a Uniform Prior ($\mathcal{U}(0,1) \sim \mathrm{Beta}(\alpha=1, \beta=1)$), we get

$$\text{mean} = \frac{1+r}{2+n} \quad \text{and} \quad \text{mode} = \frac{r}{n}$$

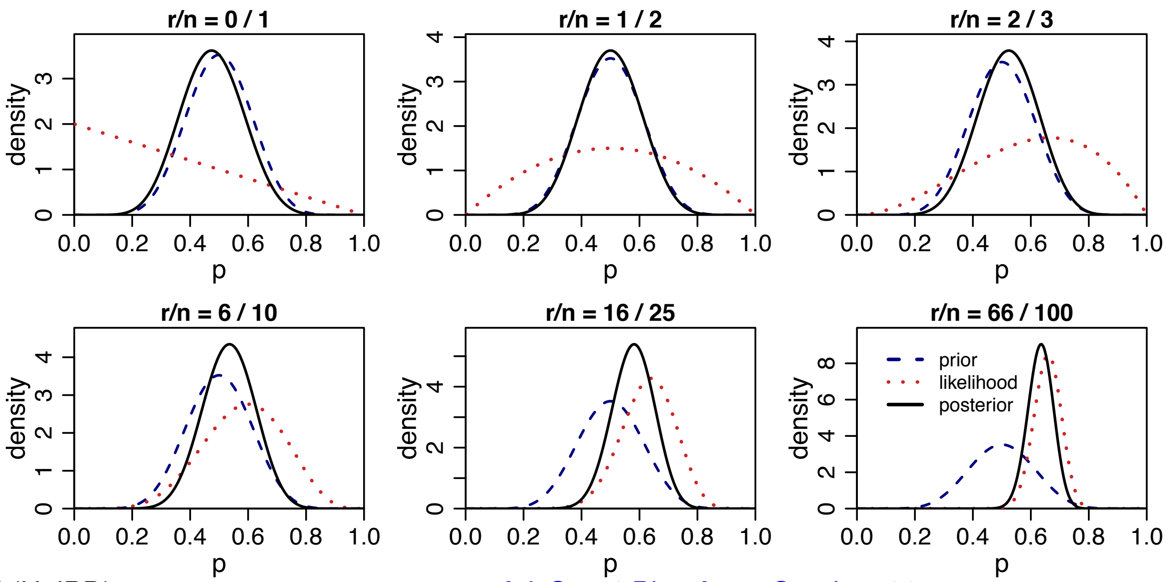
Beta Prior vs Uniform Prior

```
n <- 10;
alpha.prior <- 10;  beta.prior  <- 10
n.sample <- 2000;  delta.p <- 1/n.sample

p <- seq(from=1/(2*n.sample),
         by=1/n.sample, length.out=n.sample)

par(mfrow=c(3,3))

for(r in seq(from=0, to=10, by=2)) {
  post.beta <- dbeta(x=p,
                     alpha.prior+r,
                     beta.prior+n-r)
  plot(p, post.beta, type="l", lwd=1.5,
       col='firebrick3', ...)
  p.mean.b <- delta.p*sum(p*post.beta)
  abline(v=p.mean.b,
         col='firebrick3',lty=2)

  # overplot posterior with Unif Prior
  post.unif <- dbinom(x=r, size=n, prob=p)
  lines(p,
        post.unif/(delta.p*sum(post.unif)))
  p.norm.u <- post.unif/
              (delta.p*sum(post.unif))
  p.mean.u <- delta.p*sum(p*p.norm.u)
  abline(v=p.mean.u, col="grey60", lty=2)
}
```

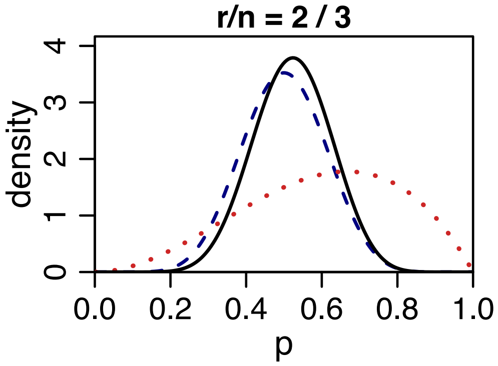


Posterior evolution with data size

- the outcome of only few coin flips tells us little about the fairness of a coin. Our state of knowledge after the analysis of the data is strongly dependent on what we knew or assumed a priori
- as the evidence grows, we are eventually led to the same conclusions irrespective of our initial beliefs
- the posterior pdf is then dominated by the likelihood function
- the choice of the prior becomes largely irrelevant



Posterior Evolution, R code




```
alpha.prior <- 10;  beta.prior  <- 10
Nsamp <- 200

delta.p <- 1/Nsamp
p <- seq(from=1/(2*Nsamp),
         by=1/Nsamp,
         length.out=Nsamp)
p.prior <- dbeta(x=p,
                alpha.prior,
                beta.prior)

n.str <- readline("Enter␣n␣extractions:␣")
n.seq <- as.numeric(unlist(strsplit(n.str, ",")))

# Loop over the vector
for (n in n.seq) {
  r <- as.integer((2/3) * n)

  p.like  <- dbinom(x=r, size=n, prob=p)
  p.like  <- p.like/(delta.p*sum(p.like))
  p.post  <- dbeta(x=p, shape1=alpha.prior+r, shape2=beta.prior+n-r)

  plot(p, p.prior, type="l", xlim=c(0,1), ...)

  lines(p, p.like, col='firebrick3',lwd=2, lty=3)
  lines(p, p.post, lwd=1.5)
  title(main=paste("r/n␣=",r,"/",n), line=0.3, cex.main=1.2)
...
}
```


Parameters best estimates and reliability

- once the posterior is determined, we wish to summarize our inference on a parameter with two numbers:
  - the best estimates
  - and a measure of its reliability
- probability distribution associated with the parameter ⇒ a measure of how much we believe the result lies in the neighborhood of that point
- Best estimate ➜ maximum of the posterior pdf

$$\theta_\circ = \mathrm{MAX}\left\{P\left(\theta \mid D, M\right)\right\}$$

- which means

$$\left.\frac{dP}{d\theta}\right|_{\theta_\circ} = 0 \quad \text{and} \quad \left.\frac{d^2P}{d\theta^2}\right|_{\theta_\circ} < 0$$

- to get a measurement of the reliability of our 'best estimate', we need to look at the spread of the posterior pdf around $\theta_\circ$

Parameters best estimates and reliability

- let's consider a Taylor expansion of the posterior pdf around $\theta_\circ$
- rather than working with the pdf, the calculations will be done with the natural logarithm

$$\begin{aligned} L &= \ln P\left(\theta \,\middle|\, D, M\right) \\ &= L(\theta_\circ) + \frac{1}{2} \left.\frac{d^2 P}{d\theta^2}\right|_{\theta_\circ} (\theta - \theta_\circ)^2 + \ldots \end{aligned}$$

Comments

- $L(\theta_\circ)$ is a constant and tells us nothing about the slope of the posterior pdf
- the linear term in $(\theta - \theta_\circ)$ is missing since we are expanding about a maximum
- the quadratic term is the dominant factor and it determines the width of the pdf

- ignoring higher order contributions and taking the exponential of the Taylor expansion

$$P\left(\theta \,\middle|\, D, M\right) \sim A \; \exp\left[\frac{1}{2} \left.\frac{d^2 P}{d\theta^2}\right|_{\theta_\circ} (\theta - \theta_\circ)^2\right]$$

with A, a normalization constant

Parameters best estimates and reliability

- we have approximated our posterior pdf by a Gaussian distribution

$$P\left(\theta \,\middle|\, \theta_\circ, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(\theta - \theta_\circ)^2}{\sigma^2}\right]$$

- comparing the two functions, we get

$$\left.\frac{d^2 L}{d\theta^2}\right|_{\theta_\circ} = -\frac{1}{\sigma^2} \quad \Rightarrow \quad \sigma = \left(-\left.\frac{d^2 L}{d\theta^2}\right|_{\theta_\circ}\right)^{-1/2}$$

- our inference about the quantity of interest is

$$\theta = \theta_\circ \pm \sigma$$

- with:
  - $\theta_\circ$ our best estimate for θ
  - σ a measurement of its reliability
- for a Gaussian distribution

$$P\left(\left|\theta - \theta_\circ\right| \leq \sigma \,\middle|\, DM\right) \sim 0.67$$

$$P\left(\left|\theta - \theta_\circ\right| \leq 2\sigma \,\middle|\, DM\right) \sim 0.95$$

Parameters estimates, coin example, Uniform Prior

- the Posterior is

$$P\left(p \mid r, n, M\right) \propto p^r (1-p)^{n-r}$$

- taking the natural logarithm

$$L = \text{const} + r \ln p + (n-r) \ln(1-p)$$

$$\frac{dL}{dp} = \frac{r}{p} - \frac{n-r}{1-p} \quad \text{and} \quad \frac{d^2L}{dp^2} = -\frac{r}{p^2} - \frac{n-r}{(1-p)^2}$$

- from the request of a maximum

$$\frac{dL}{dp} = 0 \quad \Rightarrow \quad p_\circ = \frac{r}{n}$$

- the reliability is given by the second derivative

$$\left.\frac{d^2L}{dp^2}\right|_{p_\circ} = -\frac{r}{p_\circ^2} - \frac{n-r}{(1-p_\circ)^2} = -\frac{n}{p_\circ(1-p_\circ)}$$

- therefore

$$\sigma = \left(-\left.\frac{d^2L}{d\theta^2}\right|_{\theta_\circ}\right)^{-1/2} = \sqrt{\frac{p_\circ(1-p_\circ)}{n}} = \frac{1}{n}\sqrt{\frac{r(n-r)}{n}}$$

Parameters estimates, coin example, Beta Prior

- the Posterior is

$$P\left(p \mid r, n, M\right) \propto p^{r+\alpha-1} (1-p)^{n-r+\beta-1}$$

- taking the natural logarithm

$$L = \text{const} + (r+\alpha-1) \ln p + (n-r+\beta-1) \ln(1-p)$$

$$\frac{dL}{dp} = \frac{r+\alpha-1}{p} - \frac{n-r+\beta-1}{1-p} \quad \text{and} \quad \frac{d^2L}{dp^2} = -\frac{r+\alpha-1}{p^2} - \frac{n-r+\beta-1}{(1-p)^2}$$

- from the request of a maximum

$$\frac{dL}{dp} = 0 \quad \Rightarrow \quad p_\circ = \frac{r+\alpha-1}{n+\alpha+\beta-2}$$

- the reliability is given by the second derivative

$$\left.\frac{d^2L}{dp^2}\right|_{p_\circ} = -\frac{r+\alpha-1}{p_\circ^2} - \frac{n-r+\beta-1}{(1-p_\circ)^2} = -(\alpha+\beta+n-2)\,\frac{\alpha+r}{\alpha+r-1}$$

- therefore

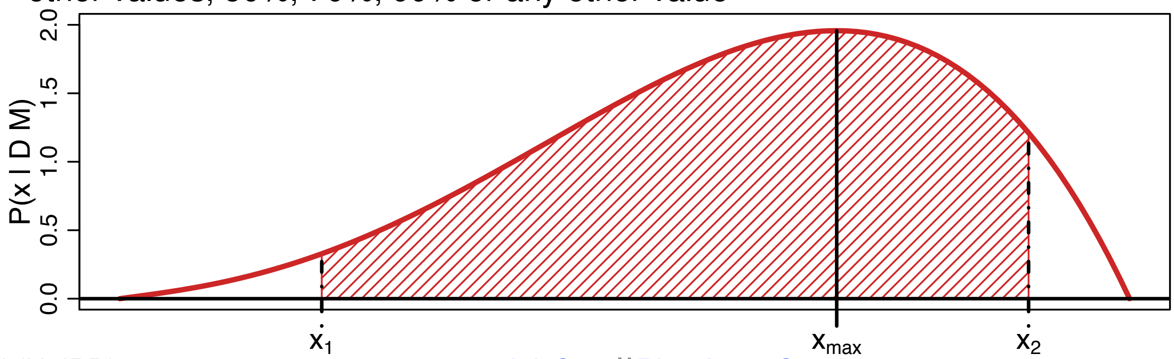
$$\sigma = \left(-\left.\frac{d^2L}{d\theta^2}\right|_{\theta_\circ}\right)^{-1/2} = \frac{1}{\alpha+\beta+n-2}\sqrt{\frac{\alpha+r-1}{\alpha+r}}$$

Asymmetric Posterior pdfs

- our derivation of the reliability of the parameter estimate (i.e. the error) relies on the validity of the quadratic expansion
- this is usually a reasonable approximation
- however there are times when the posterior pdf is markedly asymmetric
- while the maximum of the posterior can still be regarded as giving the best estimate, the concept of symmetric error bars does not seem appropriate
- a good way to express the reliability is through a confidence interval

$$P\left(x_1 \leq x < x_2 \,\middle|\, D, M\right) = \int_{x_1}^{x_2} P\left(x \,\middle|\, D, M\right)\, dx \sim 0.95$$

- Why 95% confidence level ?
- it is traditionally seen as a reasonable value, but nothing stops us from quoting other values, 50%, 70%, 99% or any other value



Assigning Priors

- probabilistic inference provides answers to well-posed problems

  but
- it does not define our models
- it does not define the priors
- or tell us which data to collect and how

- with the coin example we learned how the posterior pdf depends on both the prior and the likelihood

  ➜ when data are poor, the prior plays a more dominant role

How do we assign a Prior ?

1) a prior should incorporate any relevant information we have about the problem

   (➜ we implicitly use priors all the time in every day life)
2) some principles can help us to adopt an appropriate prior

Principle of insufficient reason

- also called the principle of indifference
- if we have a set of mutually exclusive outcomes, and we do not expect any one of them more likely, we should assign equal probabilities

Assigning Priors

Maximum Entropy

- it is based on the idea of finding the least informative (most entropic) distribution, given certain information
- example:

  if only mean and variance are known, it shows that the Gaussian is the least informative distribution

Empirical Bayes

- priors are estimated from some general properties of the data
- we can take the posterior from one analysis to be the prior of the next analysis, if they involve independent data
- the final posterior will be identical to having combined the two data sets together with the original prior
- let D_1 and D_2 be two independent data sets

$$
\begin{aligned}
P\left(\theta \,\middle|\, D_1 D_2\right) &\propto P\left(D_1 D_2 \,\middle|\, \theta\right) P(\theta) \\
&\propto \underbrace{P\left(D_2 \,\middle|\, \theta\right)}_{\text{likelihood for } D_2} \underbrace{P\left(D_1 \,\middle|\, \theta\right) \times P(\theta)}_{\text{posterior from } D_1}
\end{aligned}
$$