Statistical Models and Inference - Part I

Alberto Garfagnini

Università di Padova

AA 2023/2024 - Stat Lect. 5

Data Modeling

- we perform experiments and make observations to learn about a phenomenon
- to interpret `data`, we have to model them

Inference

- make general statements about a phenomenon through a model, using noisy and incomplete data
- must describe both the Phenomenon (i.e. Model) and the Measurement Process

▷ Key to Data Modeling: use data together with generative model (theory) and measurement model (experimental practice) to derive consistent probabilistic inferences

Data Modeling

- given some `data`, $\mathtt{D}$, we usually want to perform three actions:

▷ parameter estimation:
for a specific Model $\mathtt{H}$, with parameters θ, infere the values of model parameters, i.e. $P\left(\theta \middle| D\ H\right)$, the parameter posterior pdf

▷ model comparison:
given a set of models $\{H_j\}$, find out which one is best supported by data. This means finding $P\left(H_j \middle| D\right)$, the model posterior probability

▷ prediction:
given a model $\mathtt{H}$, inferred from the data, predict new data at some new location (in the parameter space or time)

Bayesian Model Comparison

- we start by looking at model comparison for the simple case of models with no parameters

▷ using our data $\mathtt{D}$, we look for $P\left(H \middle| D\right)$

- since $\mathrm{H} \cdot \overline{\mathrm{H}} = 0$ and $\mathrm{H} + \overline{\mathrm{H}} = \Omega$, we can write

$$\begin{aligned} P(D) &= P(DH) + P(D\overline{H}) \\ &= P(D \,|\, H)\ P(H) + P(D \,|\, \overline{H})\ P(\overline{H}) \end{aligned}$$

- our quantity of interest, $P\left(H \middle| D\right)$, is related to Bayes' theorem by

$$\begin{aligned} P(H \,|\, D) &= \frac{P(D \,|\, H)\ P(H)}{P(D)} = \frac{P(D \,|\, H)\ P(H)}{P(D \,|\, H)\ P(H) + P(D \,|\, \overline{H})\ P(\overline{H})} \\ &= \frac{1}{1 + \dfrac{P(D \,|\, \overline{H})\ P(\overline{H})}{P(D \,|\, H)\ P(H)}} = \frac{1}{1 + \dfrac{1}{R}} \end{aligned}$$

- with $\quad R = \dfrac{P(D \,|\, H)\ P(H)}{P(D \,|\, \overline{H})\ P(\overline{H})}$ the posterior odd ratio of the models

Bayesian Model Comparison

- it is easy to demonstrate that

$$\frac{P(H \mid D)}{P(\overline{H} \mid D)} = R = \frac{P(D \mid H)\, P(H)}{P(D \mid \overline{H})\, P(\overline{H})}$$

- in order to determine $P(H \mid D)$, we need three quantities:

▷ $P(D \mid H)$: the probability of measuring `D` when `H` is true

▷ $P(D \mid \overline{H})$: the probability of measuring `D` when `H` is not true (i.e. false)

▷ $P(H)$: the probability that `H` is true, independently of the data (and, of course, $P(\overline{H}) = 1 - P(H)) \Rightarrow P(H)$ tells us how probable the model is

- but, shouldn't we have information to tell us that `H` is more likely than $\overline{\texttt{H}}$, we could set

$$P(H) = P(\overline{H})$$

- and R becomes the Bayes factor

$$BF = \frac{P(D \mid H)}{P(D \mid \overline{H})}$$

- i.e. the ratio of the probability of the data under each model

Bayesian Model Comparison

- should we have more models, $\{H_j\}$, with $\sum P(H_j) = 1$, the probability of data becomes

$$P(D) = \sum_j P\left(D \mid H_j\right)\, P\left(H_j\right)$$

- and the posterior probability of model # 1, H_1, becomes

$$P\left(H_1 \mid D\right) = \frac{P\left(D \mid H_1\right)\, P(H_1)}{P(D)}$$

- if we do not have a complete set of models, we cannot compute the posterior probabilities, but we can still compute the odds ratio or Bayes factor between any two models

$$BF = \frac{P(D \mid H_1)}{P(D \mid H_2)} \quad \text{and} \quad R = \frac{P(D \mid H_1)\, P(H_1)}{P(D \mid H_2)\, P(H_2)}$$

Example

Problem

- a test for a disease is 90% reliable
- the probability of testing positive, in absence of the disease, is 0.07
- we know that among people aged 40 to 50 with no symptoms 8 in 1000 have the disease

Q: if a person in his/her 40 tests positive, what is the probability that he/she has the disease ?

Background information

- we build the following propositions:
  - D: a person is tested positive
  - H: a person has the disease
- and probabilities
  - $P(D \mid H) = 0.9$
  - $P(D \mid \overline{H}) = 0.07$
  - $P(H) = 0.008$

Example - analytical solution

- we build

$$R = \frac{P(D \mid H)\, P(H)}{P(D \mid \overline{H})\, P(\overline{H})} = \frac{9 \cdot 10^{-1} \times 8 \cdot 10^{-3}}{7 \cdot 10^{-2} \times (1 - 8 \cdot 10^{-3})} = 0.1035$$

- therefore

$$P(H \mid D) = \frac{1}{1 + 1/R} = 0.094$$

- even though a positive test result is quite probable (assuming the person has the disease), it is very unlikely that he/she has the disease
- what is decisive in the computation of $P(H \mid D)$ is the ratio between

$$P(D\, H) = P(D \mid H)\, P(H) = 7.2 \cdot 10^{-3}$$

(positive result, assuming the disease is present)

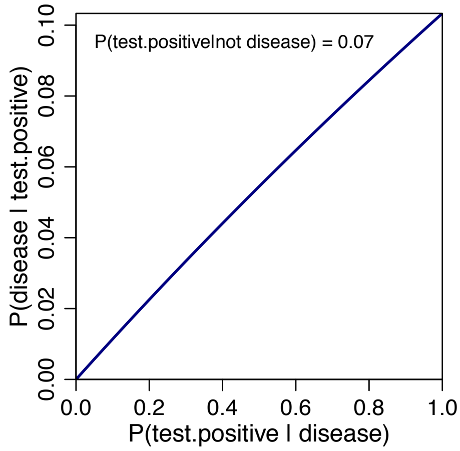
- and

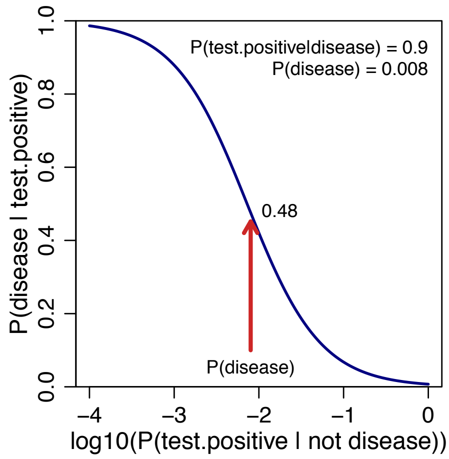
$$P(D\, \overline{H}) = P\left(D \mid \overline{H}\right)\, P\left(\overline{H}\right) = 7 \cdot 10^{-2}$$

(positive result, assuming the disease is absent)

Example - R solution

```
post <- function(p.d.m, p.d.notm, p.m) {
   p.notm <- 1 - p.m
   odds.ratio <- (p.d.m * p.m) /
                 (p.d.notm * p.notm)
   p.m.d <- 1/(1 + 1/odds.ratio)
}

p.d.m <- seq(0, 1, 0.01) # True positive
p.d.notm <- 0.07         # False positive
p.m <- 0.008             # Disease Prior

p.m.d <- post(p.d.m, p.d.notm, p.m)
plot(p.d.m, p.m.d, type='l', lwd=2, col='navy')

p.d.m <- 0.9                   # True positive
p.d.notm <- 10^seq(-4,0, 0.02) # False positive
p.m <- 0.008                   # Disease Prior

p.m.d <- post(p.d.m, p.d.notm, p.m)
plot(log10(p.d.notm), p.m.d, type='l', col='navy')
```





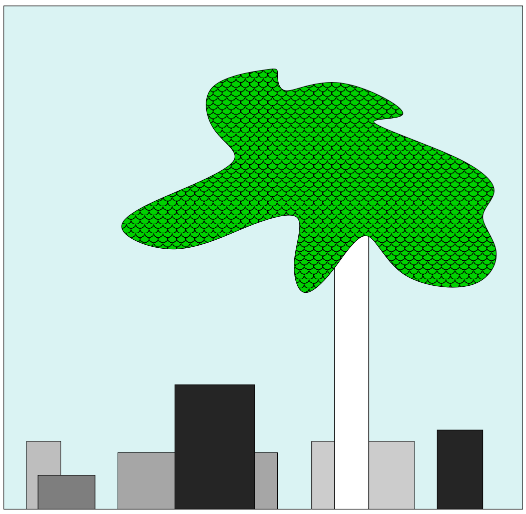
- only once the false positive rate drops below the base rate ($P(H)$) does the test starts to be useful

Occam's Razor

Look at the picture

Q_1 How many boxes are in the picture ?

Q_2 How many boxes are in the vicinity of the tree ?



What is Occam's Razor

- If several explanations are compatible with a set of observation, Occam's razor advises us to buy the simplest.

➜ Accept the simplest explanation that fits the data

But why ?

- a theory with mathematical beauty is more likely to be correct than an ugly one that fit some experimental data (P.A.M. Dirac)
- choerent inference, embodied by bayesian probability, automatically embodies Occam's razor, quantitatively

Model comparison and Occam's Razor

- we evaluate the plausibility of two alternative theories: H_1 and H_2
- we derived the probability ratio between the two theories as

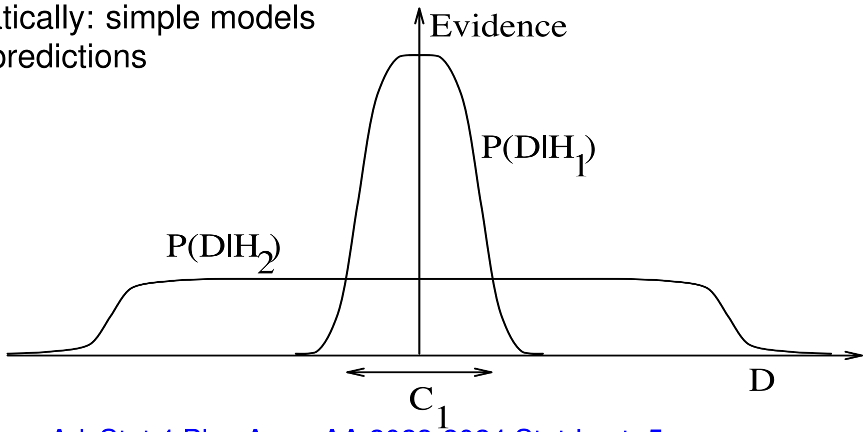
$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)\, P(H_1)}{P(D \mid H_2)\, P(H_2)}$$

- how is this related to Occam's razor when H_1 is a simpler model than H_2 ?

  the ratio

$$\frac{P(D \mid H_1)}{P(D \mid H_2)}$$

  which depends on the data embodies Occam's razor automatically: simple models tend to make precise predictions



Example

(1)

- we have a sequence of numbers:

  `-1, 3, 7, 11`
- we want to predict the next two numbers in the sequence

  Answer$_1$: `15` and `19`

  i.e. add `4` to the previous number

➜ the sequence is an arithmetic progression

  Answer$_2$: `-19.9` and `1043.8`

  i.e. we apply $-x^3/11 + 9x^2/11 + 23/11$ to the previous number

➜ the sequence is generated by a cubic function

Example

- let's compare the two models: H_g (geometric) versus H_c (cubic)
- in general, an arithmetic progression is more frequent than a cubic, but since this would out a bias in the ratio $P(H_g)/P(H_c)$, we assume they have equal probabilities
- let's compute $P(D \mid H_g)$:

$$P(D \mid H_g) = \frac{1}{101}\frac{1}{101} = 10^{-4}$$

(we have assumed that these number could be anywere in the interval $[-50, 50]$)

- to compute $P(D \mid H_c)$, we use the same probability distribution for evaluating the coefficients

$$P(D \mid H_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{5}\right)\left(\frac{2}{101}\frac{1}{50}\right) = 2.5 \cdot 10^{-12}$$

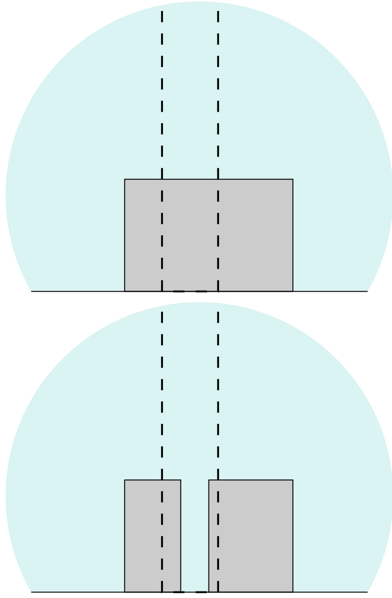
- the ratio favours the geometric hypothesis H_g

$$P(D \mid H_g)/P(D \mid H_c) = 40 \cdot 10^6/1$$

Occam's Razor and the *box/boxes behind the tree*

- let's go back to our example:

Q_1 Are there one or two boxes behind the tree ?



- H_1 says there is one box with four free parameters (three coordinates and colour)
- H_2 says there are two boxes (eight free parameters)
- putting some constraint on the colour (any of 16 values) and of the possible height (20 values in pixels), we get:

$$\frac{P(D \mid H_1)P(H_1)}{P(D \mid H_2)P(H_2)} = \frac{\frac{1}{20}\frac{1}{20}\frac{1}{20}\frac{1}{16}}{\frac{1}{20}\frac{1}{20}\frac{10}{20}\frac{1}{16}\frac{1}{20}\frac{1}{20}\frac{10}{20}\frac{1}{16}} = \frac{1000}{1}$$

the more complex model has four extra parameters for size and colours: it has to pay two big Occam factors (1/20 and 1/6) for the suspicious coincidence that the two boxes have exactly the same colour and exact height

D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003

Data Modeling with Parametric Models

- generative model : theory predicting observable data from model parameters
  - the model just studied did not have any parameter: it was either `true` or `false`
- the simplest generative model is a straight line

$$f(x; a, b) = a + b \cdot x$$

- but our measurements will differ from the model due to noise

$$y = f(x; a, b) + \epsilon$$

- and the noise model - we call it the measurement model - has also parameters
  - given our set of data $\mathrm{D} = \{y_j\}$ at specified values $\{x_j\}$, we want to infer the values of the parameters for the generative model
  - in some cases we want to find the best set of parameters that predicts the data
  - but data are noisy $\rightarrow$ there is no unique solution
- we look for the probability distributions of the parameters, $P\left(\theta \,\middle|\, D\ H\right)$, also called parameter posterior pdf. Thanks to Bayes' theorem

$$P\left(\theta \,\middle|\, D\ H\right) = \frac{P\left(D \,\middle|\, \theta\ H\right)\ P\left(\theta \,\middle|\, H\right)}{P\left(D \,\middle|\, H\right)}$$

The Likelihood

- $P\left(D \,\middle|\, \theta\ H\right)$ is the Likelihood probability
  - it is a key function since it describes both the phenomenon and the data
  - it tells us the probability of getting the data we measured, given some value of the parameters
- `H` specifies:
  - a generative model ← the equation for the straight line $f(x; a, b)$
  - a measurement model ← how the measurement of y at a given x differs from $f(x; a, b)$ due to noise
- the measurement model describes ϵ in $y = f(x; a, b) + \epsilon$
  - example: Gaussian distribution with variance σ^2. The Likelihood for any measurement is

$$P\left(y \,\middle|\, \theta\ H\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - f(x; a, b))^2}{2\sigma^2}\right)$$

- telling us that the measurement has a Gaussian distribution about the true value
- $\theta = \theta(a, b; \sigma)$ is the union of the generative and measurements models

The Prior

- $P\left(\theta \mid H\right)$ is the Prior probability
- it encapsulates all the information we have, independent of the data
- it is called Prior because is the background information we have before obtaining the Data
- different people may have different information, or different opinion on what prior information is important
- this is not a weakness of inference
- it just reflects reality: we do not only use our immediate measurements to reach scientific conclusions

The Posterior

- $P\left(\theta \mid D\ H\right)$ is the Posterior probability
- it is the pdf over the model parameters, given `data` and background information
- from Bayes' theorem

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- the proportionality is through $P\left(D \mid H\right)$, a normalization factor which is independent of θ. Therefore:

$$P\left(\theta \mid D\ H\right) = \frac{1}{Z} P\left(D \mid \theta\ H\right) P\left(\theta \mid H\right)$$

- with $Z = P\left(D \mid H\right)$
- from a conceptual point of view, inference is really that strightforward
- Bayesian inference is the process of improving our knowledge of the model paramaters by using the data

▷ we update the Prior using the Likelihood to obtain the Posterior

The Evidence

- $P\left(D \mid H\right)$ is the Evidence

- is the denominator of Bayes's equation and it gives the probability of observing the Data `D`, assuming the model `H` to be true, for any values of θ

$$P\left(D \mid H\right) = \int P\left(D \mid \theta H\right) P\left(\theta \mid H\right) d\theta$$

- evidence plays a key role in model comparison
- as a normalization constant, it is very important if we want to compute certain quantities from the posterior
- sometimes the integral can be calculated analytically, but for many real-world problems, we have to resort to numerical integration → Markov Chain Monte Carlo

Bayesian Inference of repeated Bernoulli trials

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. $\pi = \frac{1}{2}$)

Comment

- no definitive answer exists
- only a probabilistic answer can be provided

- we are looking for

$$P\left(\pi \,\middle|\, n, r, H\right)$$

- from Bayes' theorem

$$P\left(\pi \,\middle|\, n, r, H\right) = \frac{P\left(r \,\middle|\, \pi, n, H\right)\; P\left(\pi \,\middle|\, H\right)}{P\left(r \,\middle|\, n, H\right)}$$

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

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

The Likelihood

- the appropriate Likelihood is the binomial distribution

$$P\left(r \,\middle|\, \pi, n, H\right) = \binom{n}{r} \pi^r (1-\pi)^{n-r} \quad \text{with} \quad r \leq n$$

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

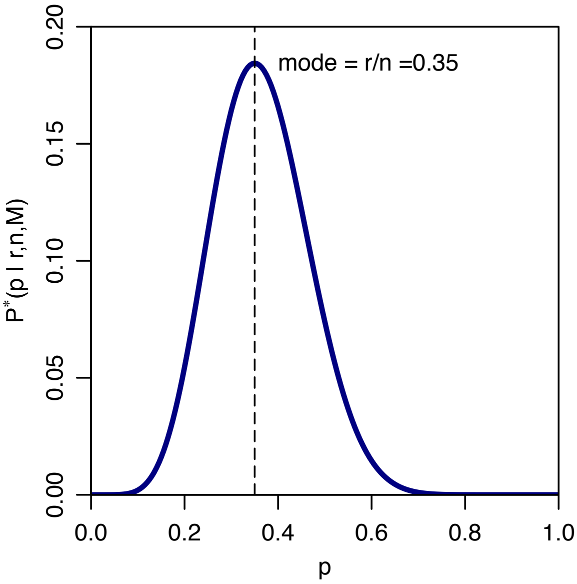
Coin tossing : a uniform Prior

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

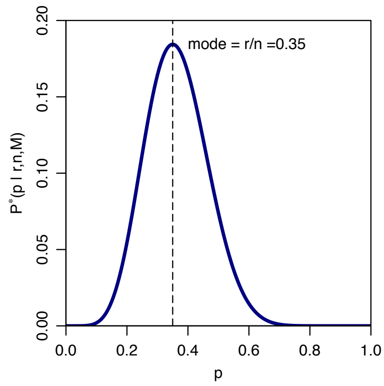
$$P\left(\pi \mid r,n,H\right) = \frac{1}{Z}\,\pi^r(1-\pi)^{n-r} = \frac{1}{Z}P^*\left(\pi \mid r,n,H\right)$$

- the normalization factor Z (i.e. the evidence $P\left(r \mid n,H\right)$ does not depend on π
- 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,H)")))
```



Uniform Prior



Comments

- the curve is not binomial in π, but it is binomial in r
- the posterior is not-normalized: the integral over π 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, $P^*(\pi \mid r,n,H)$,

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

- with

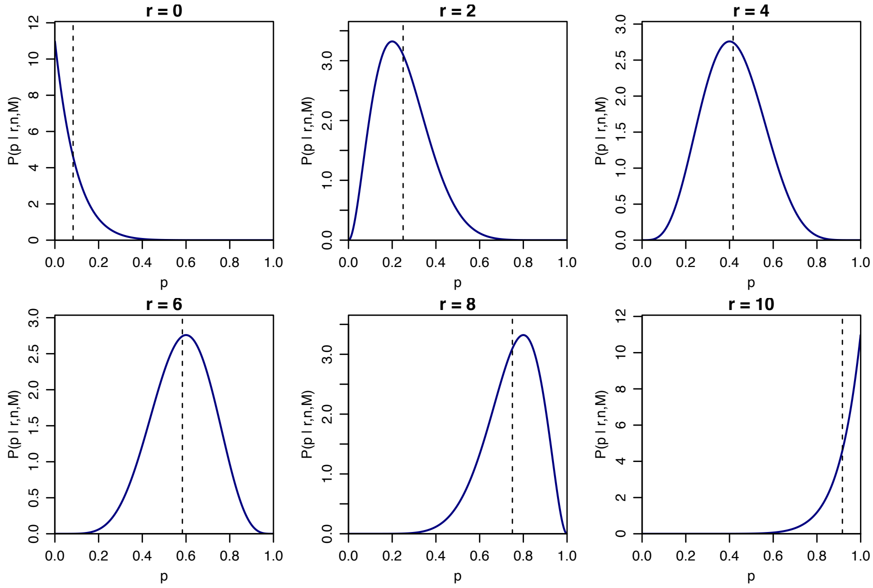
$$Z = \int_0^1 P^*\left(\pi \mid r,n,H\right)\,d\pi \approx \sum_j P^*\left(\pi_j \mid r,n,H\right)\Delta\pi_j$$

- 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,H)")
  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

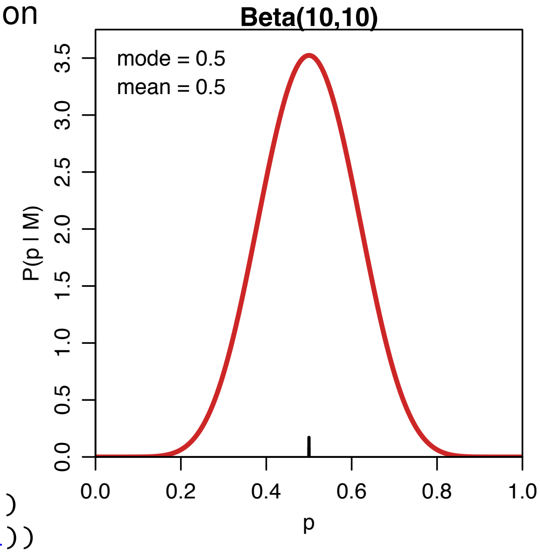
- given a random coin, we may believe the coin is fair, or close to fair
- an appropriate probability density function is the Beta distribution

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

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␣|␣H)"),
     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 π in the constant term Z, we get

$$
\begin{aligned}
P\left(\pi \,\middle|\, r, n, H\right) &= \frac{1}{Z}\, \pi^{r}(1-\pi)^{n-r} \times \pi^{\alpha-1}(1-\pi)^{\beta-1} \\
&= \frac{1}{Z}\, \pi^{r+\alpha-1}(1-\pi)^{n-r+\beta-1}
\end{aligned}
$$

- 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

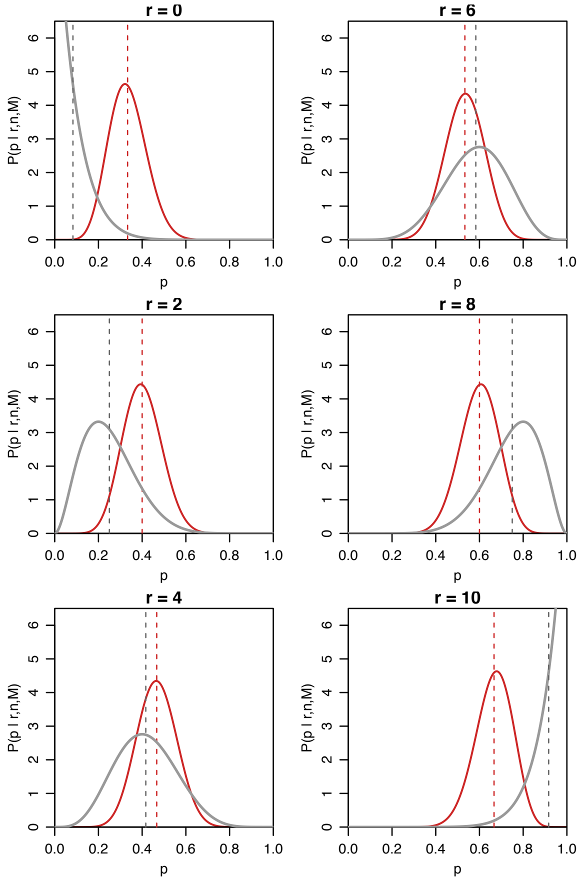
$$\text{mean} = \frac{\alpha_p + r}{\alpha_p + \beta_p + n} \quad \text{and} \quad \text{mode} = \frac{\alpha_p + r - 1}{\alpha_p + \beta_p + n - 2}$$

- 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

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

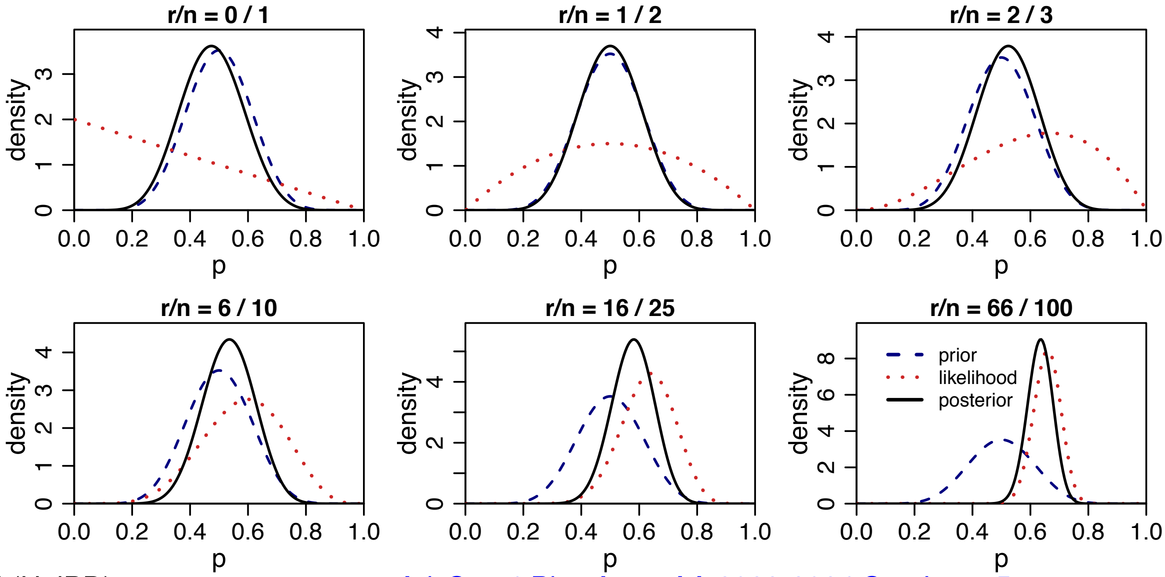
Beta Prior vs Uniform Prior

```
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)
}
```



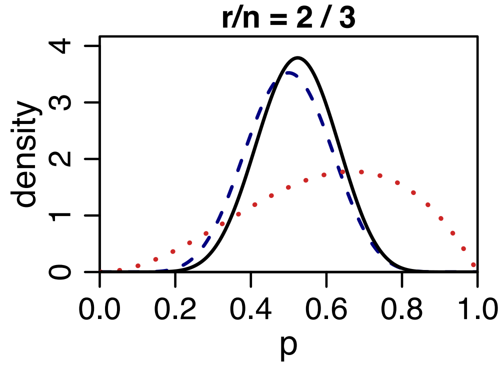
Posterior evolution with data size

- 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 bought 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

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

- 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

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

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

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

with A, a normalization constant

Parameters best estimates and reliability

- we have approximated our posterior pdf by a Gaussian distribution

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

- 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

$$P\left(\left|\theta - \theta_\circ\right| \le \sigma \mid DH\right) \sim 0.67$$

$$P\left(\left|\theta - \theta_\circ\right| \le 2\sigma \mid DH\right) \sim 0.95$$

Parameters estimates, coin example, Uniform Prior

- the Posterior is

$$P\left(\pi \,\middle|\, r, n, H\right) \propto \pi^r (1-\pi)^{n-r}$$

- taking the natural logarithm

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

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

- from the request of a maximum

$$\frac{dL}{d\pi} = 0 \quad \Rightarrow \quad \pi_\circ = \frac{r}{n}$$

- the reliability is given by the second derivative

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

- therefore

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

Parameters estimates, coin example, Beta Prior

- the Posterior is

$$P\left(\pi \,\middle|\, r, n, H\right) \propto \pi^{r+\alpha-1} (1-\pi)^{n-r+\beta-1}$$

- taking the natural logarithm

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

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

- from the request of a maximum

$$\frac{dL}{d\pi} = 0 \quad \Rightarrow \quad \pi_\circ = \frac{r+\alpha-1}{n+\alpha+\beta-2}$$

- the reliability is given by the second derivative

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

- therefore

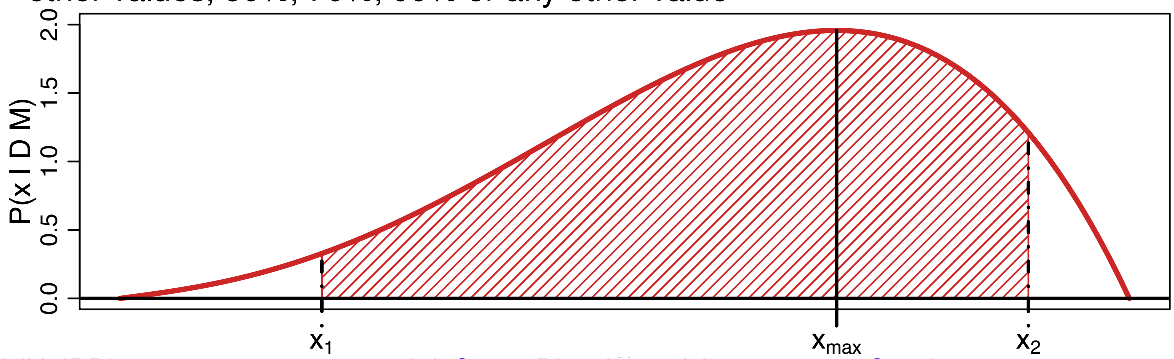
$$\sigma = \left(- \left.\frac{d^2L}{d\theta^2}\right|_{\theta_\circ} \right)^{-1/2} = \frac{1}{\alpha+\beta+n-2}\sqrt{\frac{\alpha+r-1}{\alpha+r}}$$

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

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

- 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

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

Exercise : a survey for the next Uni elections

The Problem

- In proximity of the elections for student's representatives in some University board, Anna, Chris and Maggie decide to perform a survey among their classmates to check how strong is their candidate friend
- the aim is to infere the probability that she gets elected

Step 1: choosing the Priors

- Before starting the interviews, they have different opinions about the results of the elections:

- Anna thinks that there will be a 20% chanches that their friend will be elected, and moreover, the probability has a standard deviation of 0.08.
  She therefore assumes a Beta prior such that:

$$E[x] = \frac{a}{a+b} = 0.2 \quad 1 - E[x] = \frac{b}{a+b} = 0.8 \qquad \frac{0.2 \times 0.8}{a+b+1} = 0.08^2 \ ,$$

  which means $a = 4.8$ and $b = 19.2$

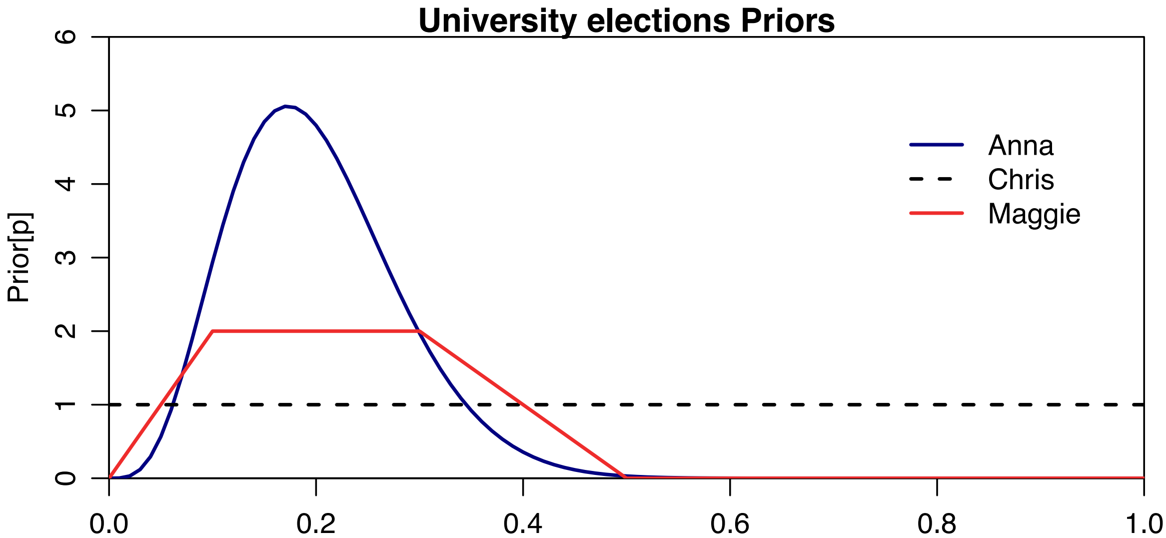
- Chris is a new student and he does not have any feeling how popular their candidate is, therefore he assumes a Uniform prior distribution. For him $a = b = 1$

Exercise : a survey for the next Uni elections (2)

Step 1: choosing the Priors (cont'd)

Before starting the interviews, they have different opinions about the results of the elections:

- Maggie thinks that the probability distribution is flat, but not over the whole domain. Therefore she assumes a trapezoidal distribution which is flat between 0.1 and 0.3, and goes to zero outside that domain



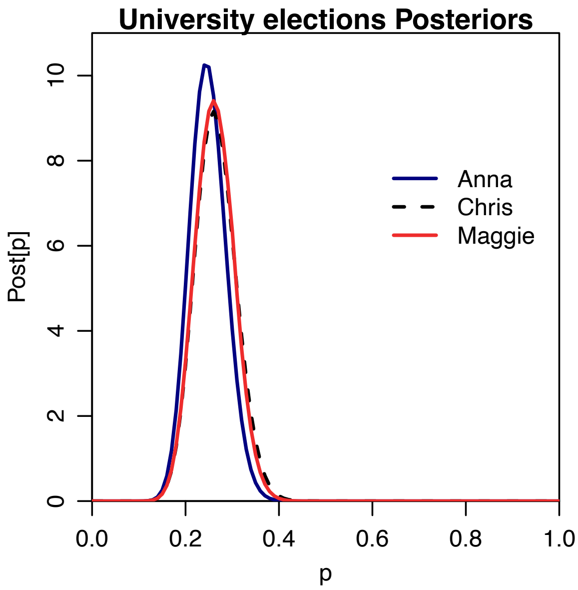
Exercise : a survey for the next Uni elections (3)

Step 2: getting the data

- now they start the survey and decide to interview $n = 100$ students regularly coming to the University canteen but they do not personally know
- out of the interviewed students, $x = 26$ claim they will support and vote the candidate

Step 3: computing the Posterior

- Anna and Chris use a Beta prior ➜ they get a conjugate prior $\mathrm{Beta}(\alpha = a + x, \beta = b + n - x)$
- Anna has $\mathrm{Beta}(\alpha = 4.8 + 26, \beta = 19.2 + 74)$
- Chris gets $\mathrm{Beta}(\alpha = 1 + 26, \beta = 1 + 74)$
- Maggie has to perform a numerical computation of the posterior, given her user-defined Prior



Exercise : a survey for the next Uni elections (4)

Step 4: computing Credibility Intervals

- given the Posterior distributions, we can compute the mean value and the variance
- by integrating the Posterior distribution, it is possible to compute the Credibility Interval, 95%, as the area between the 2.5% and 97.5%
- Maggie's estimate must be done by numerical integration

	Post(α, β)	mean	sigma	95% Cr. Int.
Anna	Beta$(\alpha = 30.8, \beta = 93.2)$	0.248	0.039	0.177 - 0.328
Chris	Beta$(\alpha = 27, \beta = 75)$	0.265	0.043	0.184 - 0.354
Maggie	numerical	0.262	0.042	0.183 - 0.346