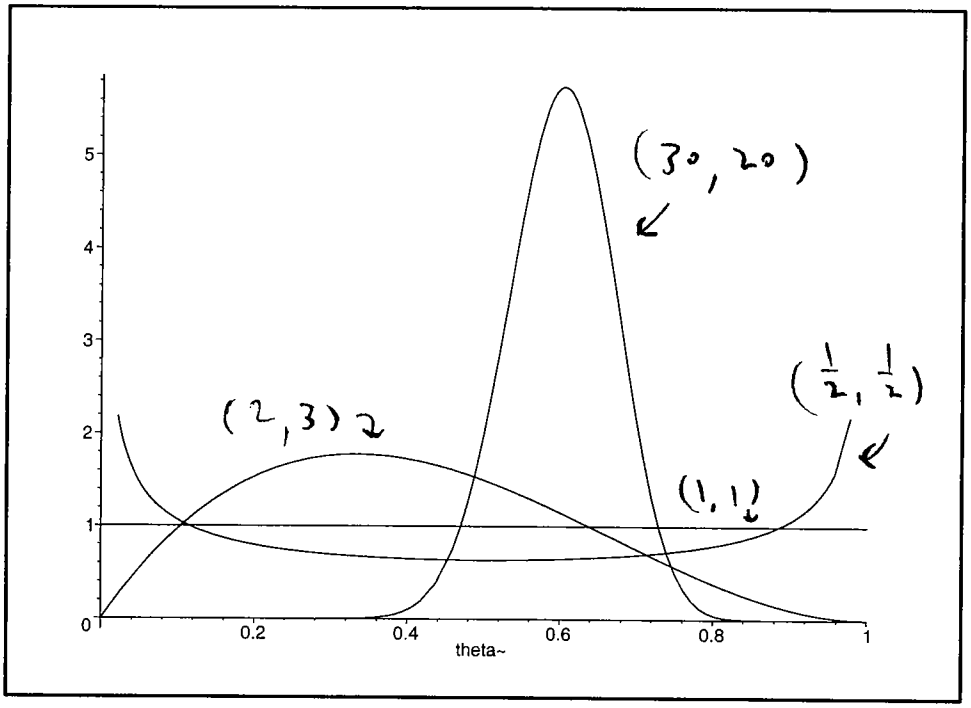
Conjugate Analysis (continued)



(2) As we saw above, the likelihood in this problem comes from the **Bernoulli** sampling distribution for the Y_i,

$$p(y_1, \ldots, y_n|\theta) = l(\theta|y) = \theta^s(1-\theta)^{n-s}, \qquad (41)$$

where s is the **sum** of the y_i.

Now Bayes' Theorem says that to get the posterior distribution $p(\theta|y)$ you **multiply** the prior $p(\theta)$ and the likelihood—in this case $\theta^s(1-\theta)^{n-s}$—and **renormalize** so that the product integrates to 1.

Rev. Bayes himself noticed back in the 1740s that if you take the prior to be of the form $c\,\theta^u\,(1-\theta)^v$, the product of the prior and the likelihood **will also be of this form**, which makes the **computations** more straightforward.

The beta family is said to be **conjugate** to the Bernoulli/binomial likelihood.

Conjugate Analysis (continued)

Conjugacy of a family of prior distributions to a given likelihood is a bit hard to define precisely, but the basic idea—given a particular likelihood function—is to try to find a family of prior distributions so that the product of members of this family with the likelihood function will also be in the family.

Conjugate analysis—finding conjugate priors for standard likelihoods and restricting attention to them on tractability grounds—is one of only two fairly general methods for getting closed-form answers in the Bayesian approach (the other is **asymptotic analysis**; see Bernardo and Smith, 1994).

Suppose we restrict attention (for now) to members of the beta family in trying to specify a **prior distribution** for θ in the AMI mortality example.

I want a member of this family which has **mean 0.15** and **95% central interval (0.05, 0.30)**.

```
> mean := integrate( theta * p( theta, alpha, beta ), theta = 0 .. 1 );
```

$$\text{mean} := \frac{\text{alpha}\tilde{}}{\text{alpha}\tilde{} + \text{beta}\tilde{}}$$

```
> variance :=simplify( integrate( ( theta - alpha /
    ( alpha + beta ) )^2 * p( theta, alpha, beta ), theta = 0 .. 1 ) );
```

$$\text{variance} := \frac{\text{alpha}\tilde{}\ \text{beta}\tilde{}}{(\text{alpha}\tilde{} + \text{beta}\tilde{})^2\ (\text{alpha}\tilde{} + \text{beta}\tilde{} + 1)}$$

Conjugate Analysis (continued)

As Maple has **demonstrated**,

$$\text{If } \theta \sim \text{Beta}(\alpha, \beta), \text{ then } E(\theta) = \frac{\alpha}{\alpha + \beta} \tag{42}$$

$$\text{and } V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

```
> solve( mean = 15 / 100, beta );

                                  17/3 alpha~

> solve( integrate( p( theta, alpha, 17 * alpha / 3 ),
    theta = 0.05 .. 0.30 ) = 0.95, alpha );

bytes used=3005456, alloc=1834672, time=0.82
bytes used=4006628, alloc=2293340, time=1.18
bytes used=5007408, alloc=2489912, time=1.58
```

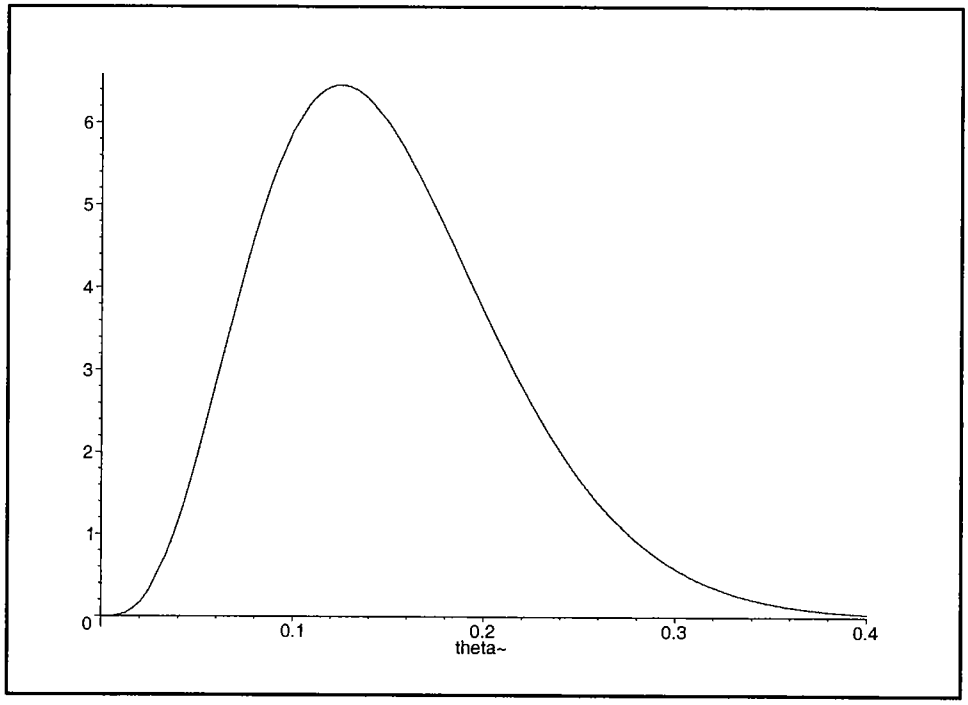
Maple can't solve this equation **symbolically** (and neither could you), but it can do so **numerically**:

```
> fsolve( integrate( p( theta, alpha, 17 * alpha / 3 ),
    theta = 0.05 .. 0.30 ) = 0.95, alpha );

bytes used=7083468, alloc=2686484, time=2.50

  (output suppressed)

bytes used=27099104, alloc=3538296, time=11.99

                                  4.506062414

> 17 * 4.506062414 / 3;

                                  25.53435368
```

Thus the beta distribution with $(\alpha, \beta) =$ **(4.5, 25.5)** meets my two prior specifications.

Conjugate Analysis (continued)

```
> plot( p( theta, 4.5, 25.5 ), theta = 0 .. 0.4 );
```



This prior distribution looks just like I want it to: it has a **long right-hand tail** and is **quite spread out**: the prior SD with this choice of (α, β) is $\sqrt{\frac{(4.5)(25.5)}{(4.5+25.5)^2(4.5+25.5+1)}} \doteq 0.064$, i.e., my prior says that I think the underlying AMI mortality rate at the DH is around **15%**, give or take about **6 or 7%**.

In the usual jargon α and β are called **hyperparameters** since they're parameters of the prior distribution.

Written **hierarchically** the model we've arrived at is

$$\begin{array}{rcll} (\alpha, \beta) & = & (4.5, 25.5) & \text{(hyperparameters)} \\ (\theta|\alpha, \beta) & \sim & \text{Beta}(\alpha, \beta) & \text{(prior)} \\ (Y_1, \ldots, Y_n|\theta) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta) & \text{(likelihood)} \end{array} \qquad (43)$$

Conjugate Analysis (continued)

(43) suggests what to do if you're not sure about the specifications that led to $(\alpha, \beta) = (4.5, 25.5)$: **hierarchically expand** the model by placing a distribution on (α, β) centered at $(4.5, 25.5)$.

This is an important Bayesian modeling tool: if the model is inadequate in some way, **expand it hierarchically** in directions suggested by the nature of its inadequacy (I'll give more examples of this later).

Q: Doesn't this set up the possibility of an **infinite regress**, i.e., how do you know **when to stop** adding layers to the hierarchy?

A: (1) In practice people stop when they run out of (time, money), after having made sure that the final model passes **diagnostic checks**, and comfort may be taken from the empirical fact that (2) there tends to be a kind of **diminishing returns** principle: the farther a given layer in the hierarchy is from the likelihood (data) layer, the less it tends to affect the answer.

The conjugacy of the prior leads to a **simple closed form** for the posterior here: with y as the vector of observed $Y_i, i = 1, \ldots, n$ and s as the sum of the y_i (a **sufficient statistic** for θ with the Bernoulli likelihood),

$$\begin{aligned} p(\theta|y, \alpha, \beta) &= c\, l(\theta|y)\, p(\theta|\alpha, \beta) \\ &= c\, \theta^s (1-\theta)^{n-s}\, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= c\, \theta^{(s+\alpha)-1} (1-\theta)^{(n-s+\beta)-1}, \end{aligned} \tag{44}$$

i.e., the posterior for θ is Beta$(\alpha + s, \beta + n - s)$.

This gives the hyperparameters a nice interpretation in terms of **effective information content of the prior**: it's as if the data (Beta$(s+1, n-s+1)$) were worth $(s+1) + (n-s+1) \doteq n$ observations and the prior (Beta(α, β)) were worth $(\alpha + \beta)$ observations.

This can be used to judge whether the prior is **"too informative"**—here it's equivalent to $(4.5 + 25.5) = 30$ binary observables with a mean of 0.15.

Conjugate Analysis (continued)

(44) can be **summarized** by saying

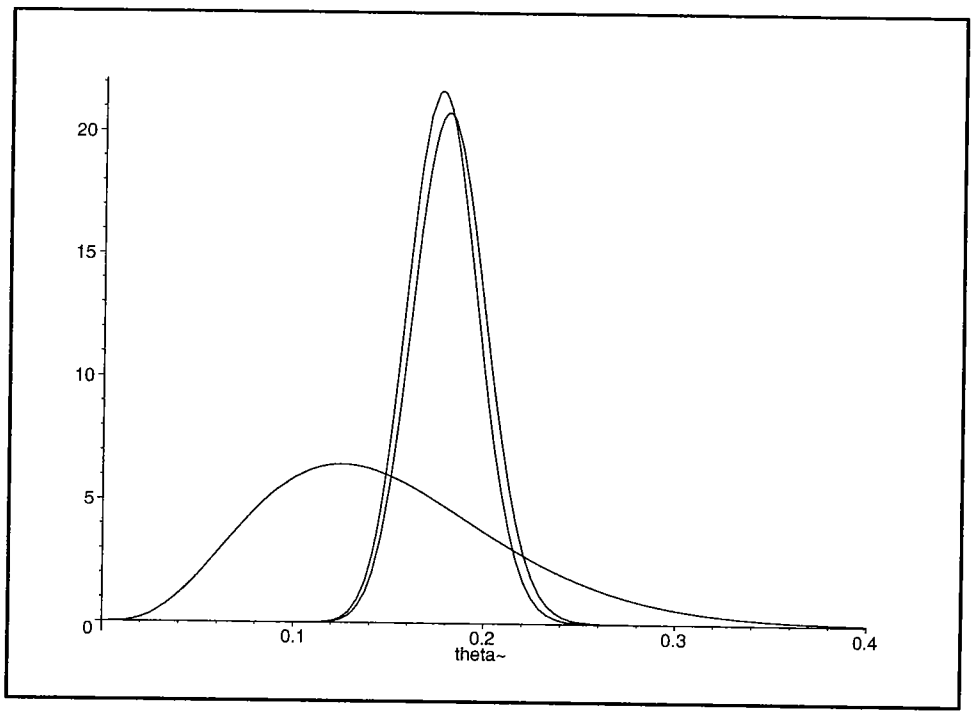
$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\theta|y) \sim \text{Beta}(\alpha + s, \beta + n - s), \tag{45}$$

where $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^{n} y_i$.

Suppose the $n = 400$ **observed mortality indicators** consist of $s = 72$ 1s and $(n - s) = 328$ 0s.

Then the **prior** is Beta(4.5,25.5), the **likelihood** is Beta(73,329), the **posterior** for θ is Beta(76.5,353.5), and the three densities plotted on the same graph come out as follows:

```
> plot( { p( theta, 4.5, 25.5 ), p( theta, 73.0, 329.0 ),
    p( theta, 76.5, 353.5 ) }, theta = 0 .. 0.4, color = black );
```



In this case the posterior and the likelihood nearly coincide, because the **data information** outweighs the **prior information** by $\frac{400}{30}$ = more than 13 to 1.

Comparison with Frequentist Modeling

The mean of a Beta(α, β) distribution is $\frac{\alpha}{\alpha+\beta}$; with this in mind the posterior mean has a nice expression as a weighted average of the prior mean and data mean, with weights determined by the **effective sample size** of the prior, $(\alpha+\beta)$, and the **data sample size** n:

$$\frac{\alpha+s}{\alpha+\beta+n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{s}{n}\right)$$

$$\begin{array}{ccccccccc} \text{posterior} & & \left(\begin{array}{c}\text{prior}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{prior}\\ \text{mean}\end{array}\right) & & \left(\begin{array}{c}\text{data}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{data}\\ \text{mean}\end{array}\right) \\ \text{mean} & = & & & + & & \\ .178 & = & (.070) & (.15) & + & (.93) & (.18)\ . \end{array}$$

Another way to put this is that the data mean, $\bar{y} = \frac{s}{n} = \frac{72}{400} = .18$, has been **shrunk** toward the prior mean .15 by (in this case) a modest amount: the posterior mean is about .178, and the **shrinkage factor** is $\frac{30}{30+400}$ = about .07.

As we saw back on pp. 9–10, to analyze these data as a frequentist you would appeal to the **Central Limit Theorem:** $n = 400$ is big enough so that the sampling distribution of $\bar{Y}$ is approximately $\mathrm{N}\left[\theta, \frac{\theta(1-\theta)}{n}\right]$, so an approximate **95% confidence interval** for θ would be centered at $\hat{\theta} = \bar{y} = 0.18$, with an estimated standard error of $\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.0192$, and would run roughly from 0.142 to 0.218.

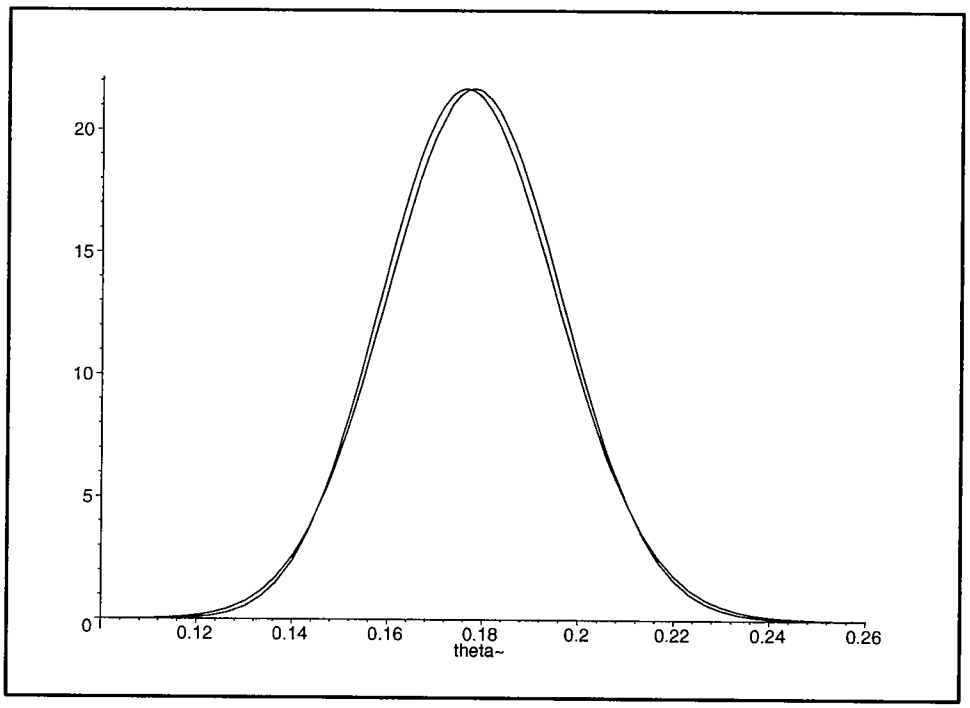
By contrast the posterior for θ is also **approximately Gaussian** (see the graph on the next page), with a mean of 0.178 and an SD of $\sqrt{\frac{\alpha^*\beta^*}{(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)}} = 0.0184$, where α^* and β^* are the parameters of the beta posterior distribution; a **95% central posterior interval** for θ would then run from about $0.178 - (1.96)(0.0184) = 0.142$ to $0.178 + (1.96)(0.0184) = 0.215$.

Comparison with Frequentist Modeling

```
> g := ( theta, mu, sigma ) -> exp( - ( theta - mu )^2 /
    ( 2 * sigma^2 ) ) / ( sigma * sqrt( 2 * Pi ) );
```

$$g := (\theta, \mu, \sigma) \to \frac{\exp\left(-\frac{1}{2}\frac{(\theta-\mu)^2}{\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

```
> plot( { p( theta, 76.5, 353.5 ), g( theta, 0.178, 0.0184 ) },
    theta = 0.10 .. 0.26, color = black );
```



The Bayesian analysis here is equivalent to one in which a dataset consisting of $(0.15)(30) = 4.5$ 1s and $(1 - 0.15)(30) = 25.5$ 0s is appended to the observed data, and **a frequentist analysis is carried out on this merged dataset**.

The two approaches (frequentist based only on the sample, Bayesian based on the sample and the prior I'm using) give **almost the same** answers in this case, a result that's typical of situations with fairly large n and relatively **diffuse** prior information.

Comparison (continued)

Note, however, that the **interpretation** of the two analyses differs somewhat:

- In the frequentist approach θ **is fixed but unknown and** $\bar{Y}$ **is random**, with the analysis based on imagining what would happen if the hypothetical random sampling were repeated, and appealing to the fact that across these repetitions $(\bar{Y} - \theta) \dot{\sim} \text{N}(0, .019^2)$; whereas

- In the Bayesian approach $\bar{y}$ **is fixed at its observed value and** θ **is treated as random**, as a means of quantifying your posterior uncertainty about it: $(\theta - \bar{y}|\bar{y}) \dot{\sim} \text{N}(0, .018^2)$.

This means among other things that, while it's **not legitimate** with the frequentist approach to say that $P_f(.14 \leq \theta \leq .22) \doteq .95$, which is what many users of confidence intervals would like them to mean, the corresponding statement $P_B(.14 \leq \theta \leq .22|y, \text{diffuse prior information}) \doteq .95$ is a **natural consequence** of the Bayesian approach.

In the case of diffuse prior information this justifies the fairly common practice of **computing inferential summaries in a frequentist way and then interpreting them Bayesianly**.

When **nondiffuse** prior information is available and you use it, your answer will **differ** from a frequentist analysis based on the same likelihood.

If your prior is retrospectively seen to have been **well-calibrated** you will get a better answer than with the frequentist approach; if poorly calibrated, a worse answer (Samaniego and Reneau, 1994):

"bad" Bayesian $\leq$ frequentist $\leq$ "good" Bayesian

What you make of this depends on your **risk-aversion**: Is it better to try to land on the right in this box, running some risk of landing on the left, or to steer a middle course?

(**NB** I'll give several examples later in which a Bayesian analysis is better **even with diffuse prior information**.)

Bernoulli Prediction

The **predictive distribution** for future Y_i in the Bernoulli model was shown back on p. 33 (equation (35)) to be

$$p(Y_{m+1} = y_{m+1}, \ldots, Y_n = y_n | y_1, \ldots, y_m) = \tag{46}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\, p(\theta | y_1, \ldots, y_m)\, d\theta$$

We've seen that if the **prior** is taken to be Beta(α, β) the **posterior** $p(\theta | y_1, \ldots, y_m)$ in this expression is Beta(α^*, β^*), where $\alpha^* = \alpha + s$ and $\beta^* = \beta + (n - s)$.

As an example of an **explicit calculation** with (46) in this case, suppose that we've observed n of the Y_i, obtaining data vector $y = (y_1, \ldots, y_n)$, and we want to predict Y_{n+1}.

Obviously $p(Y_{n+1} = y_{n+1} | y)$ has to be a **Bernoulli**(θ^*) distribution for some θ^*, and intuition says that θ^* should just be the **mean** $\frac{\alpha^*}{\alpha^* + \beta^*}$ of the posterior distribution for θ given y.

(46) and an **application of (39)** in this case give for $p(Y_{n+1} = y_{n+1} | y)$ the expressions

$$\int_0^1 \theta^{y_{n+1}}(1-\theta)^{1-y_{n+1}} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}\, d\theta \tag{47}$$
$$= \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^{\alpha^* + y_{n+1} - 1}(1-\theta)^{(\beta^* - y_{n+1} + 1) - 1}\, d\theta$$
$$= \left[\frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)}\right] \left[\frac{\Gamma(\beta^* - y_{n+1} + 1)}{\Gamma(\beta^*)}\right] \left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right]$$

Now it's a **fact about the Gamma function**, which you can verify with `Maple`, that for any real number x, $\frac{\Gamma(x+1)}{\Gamma(x)} = x$:

Bernoulli Prediction (continued)

```
rosalind 175> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> assume( x, real );

> simplify( GAMMA( x + 1 ) / GAMMA( x ) );

                                x~
```

So **(47)**, for example in the case $y_{n+1} = 1$, becomes

$$\begin{aligned} p(Y_{n+1} = 1|y) &= \left[\frac{\Gamma(\alpha^* + 1)}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right] \\ &= \frac{\alpha^*}{\alpha^* + \beta^*}, \end{aligned} \tag{48}$$

confirming intuition.

For example, with $(\alpha, \beta) = (4.5, 25.5)$ and $n = 400$ with $s = 72$, we saw earlier that the **posterior** for θ was Beta$(76.5, 353.5)$, and this posterior distribution has mean $\frac{\alpha^*}{\alpha^*+\beta^*} = 0.178$.

In this situation you would expect the next AMI patient who comes along to die within 30 days of admission with probability **0.178**, so the predictive distribution above **makes good sense**.

The Binomial Distribution

We've seen that a **sufficient statistic** for θ with a Bernoulli likelihood is the **sum** $s = \sum_{i=1}^{n} y_i$ of the 1s and 0s.

This means that if you buy into the model $(Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$ **you don't care** whether you observe the entire data vector $Y = (Y_1, \ldots, Y_n)$ or its sum $S = \sum_{i=1}^{n} Y_i$.

The distribution of S in repeated sampling has a **familiar form**: it's just the **binomial** distribution $\text{Binomial}(n, \theta)$, which counts the number of successes in a series of IID success/failure trials.

Recall that if $S \sim \text{Binomial}(n, \theta)$ then S has **discrete density**

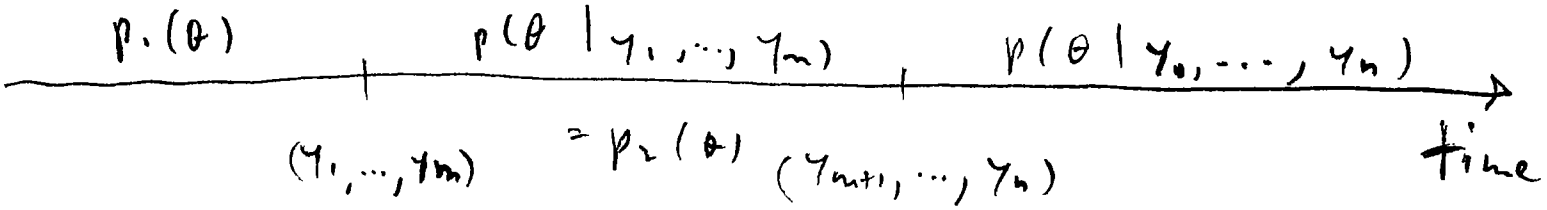
$$p(S = s|\theta) = \left\{ \begin{array}{cc} \binom{n}{s} \theta^s (1-\theta)^{n-s} & \text{if } s = 0, \ldots, n \\ 0 & \text{otherwise} \end{array} \right\}.$$

Thus we've learned **another conjugate updating rule** in simple Bayesian modeling, more or less for free: if the data set just consists of a single draw S from a binomial distribution, then the conjugate prior for the success probability θ is $\text{Beta}(\alpha, \beta)$, and the updating rule, which follows directly from (45), is

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (S|\theta) \sim \text{Binomial}(n, \theta) \end{array} \right\} \rightarrow (\theta|s) \sim \text{Beta}(\alpha + s, \beta + n - s). \quad (49)$$

Two Important General Points

1 (the **sequential** nature of **Bayesian learning**) Suppose you and I are observing data $(y_1, \ldots, y_n)$ to **learn** about a **parameter** θ, and we have no reason throughout this observation process to **change** (the sampling distribution/likelihood part of) our **model**.



We both start with the **same prior** $p_1(\theta)$ before any of the data arrive, but we adopt what appear to be **different analytic strategies**:

- You wait until the whole data set $(y_1, \ldots, y_n)$ has been observed and **update** $p_1(\theta)$ **directly** to the posterior distribution $p(\theta|y_1, \ldots, y_n)$, whereas

- I **stop** after seeing $(y_1, \ldots, y_m)$ for some $m < n$, update $p_1(\theta)$ to an **intermediate** posterior distribution $p(\theta|y_1, \ldots, y_m)$, and then I want to go on from there, observing $(y_{m+1}, \ldots, y_n)$ and finally updating to a posterior on θ that takes account of the **whole data set** $(y_1, \ldots, y_n)$.

Q$_1$ What should I use for my **intermediate prior distribution** $p_2(\theta)$?

A$_1$ Naturally enough, the **right thing to do** is to set $p_2(\theta) = p(\theta|y_1, \ldots, y_m)$.

The informal way people refer to this is to say that **yesterday's posterior distribution is today's prior distribution**.

Q$_2$ If I use the posterior in **A**$_1$, do you and I get the **same answer** for $p(\theta|y_1, \ldots, y_n)$ in the end?

A$_2$ **Yes** (you can check this).

Two Important Points (continued)

2 (the generality of **conjugate analysis**) Having seen **conjugate priors** used with binary outcomes, you can see that **conjugate analysis** has a variety of **advantages**:

- It's **mathematically straightforward**;
- The **posterior mean** turns out to be a **weighted average** of the **prior** and **data means**; and
- You get the nice interpretation of the prior as an information source that's **equivalent to a data set**, and it's easy to figure out the **prior sample size**.

It's natural to wonder, though, what's **lost** in addition to what's **gained** by adopting a conjugate prior.

The main **disadvantage** of conjugate priors is that in their simplest form they're **not flexible enough** to express **all possible forms** of prior information.

For example, in the AMI mortality case study, what if you wanted to combine a **bimodal** prior distribution with the Bernoulli likelihood?

This isn't possible when using a **single member** of the Beta(α, β) family.

However, it's possible to **prove** the following:

Theorem (Diaconis and Ylvisaker 1985). Given a particular likelihood that's a member of the **exponential family** (this will be covered in Section 2.9 below), any prior distribution can be expressed as a **mixture** of priors that are conjugate to that likelihood.

For example, in the **AMI case study** the model could be

$$\begin{array}{rcl} J & \sim & p(J) \\ (\theta|J) & \sim & \text{Beta}(\alpha_J, \beta_J) \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & \text{B}(\theta), \; i = 1, \ldots, n, \end{array} \tag{50}$$

for some distribution $p(J)$ on the positive integers—this is **completely general** but loses some of the advantages of simple conjugate analysis (e.g., **closed-form computations** are no longer possible).

2.7 The Exponential Family

In our first (and only, so far) example of **conjugate** analysis, with the Bernoulli/binomial likelihood (41), we worked out the form of the **conjugate prior** just by **looking** at the likelihood function.

This works in **simple** problems, but it would be nice to have a **general** way of figuring out what the conjugate prior has to be (if it exists) once the likelihood is **specified**.

It was noticed a long time ago that many of the **standard sampling distributions** that you're likely to want to use in constructing likelihood functions in parametric Bayesian modeling have the **same general form**, which is referred to as the **exponential family**.

I bring this up here because there's a **simple theorem** which specifies the **conjugate prior** for likelihoods that belong to the **exponential family**.

With the Bernoulli likelihood (41) in the hospital mortality case study, the unknown quantity θ in the likelihood function was a **scalar** (**1–dimensional**; **univariate**), but this will not always be true: more generally and more usually θ is a **vector** of length (say) k.

We'll begin to look at problems with **multivariate** θ ($k > 1$) in Section 2.9, but for continuity with the later material I'm going to give the **definition** of the exponential family for vector θ.