# Basic Ideas of Bayesian Statistics

Probability and statistics provide tools to deal with uncertain events

- What is the probability that I have a particular disease?
- What is the probability that a student who has behaved in a particular way in school has attention-deficit disorder?
- How much more likely is it that a new medication will cure a disease, compared to the standard medication?
- How likely is it that Shakespeare actually wrote Hamlet?
- How likely is it that the Jets will win the Super Bowl next year?

Notice that classical statistical theory does not deal with these in a straightforward way.

- Any event that occurred in the past (e.g. writing Hamlet) is not subject to classical probability or statistical analysis; Shakespeare either did or did not write Hamlet. We just don't know which.
- Any single event, even a future one, cannot be described in terms of probability. So we can't talk about the probability that the Jets will win the Super Bowl. We only can talk about long run frequencies of similar events.
- When we compute a confidence interval, we can't say "There's a 95 percent chance that the new drug is better than the old"; we can only say that if the study were repeated many times, 95 percent of the time our interval will contain the population value.

Bayesian statistics is based on a very different
(and to most people more natural) principle:
We can assess our uncertainty about unknown quantities,
and use data to update our beliefs.

Our beliefs can be quantified
as a probability distribution,
and intervals can be interpreted
in terms of probability.

Bayesian statistics shows us how to

- quantify our current uncertainty or knowledge about events (prior distribution)
- assess the amount of information in new data (likelihood)
- revise our prior information using the new data (posterior distribution)
- make predictions about new events, when appropriate (predictive distribution)
- assess the costs and benefits of various outcomes of decisions
- combine our knowledge (probabilities) with costs and benefits to make decisions that are as good as possible given our limited information (decision theory)

To do Bayesian statistics, we need to know:

- How to quantify our current beliefs (or ignorance) in a probability distribution.
- How to choose a reasonable probability model for our data.
- How to derive the posterior distribution, how to interpret it, and how to use it to summarize our knowledge.
- How to form the predictive distribution, and how to use it to make predictions.
- How to assess costs and benefits, and how to combine these with our (posterior) beliefs to make decisions.

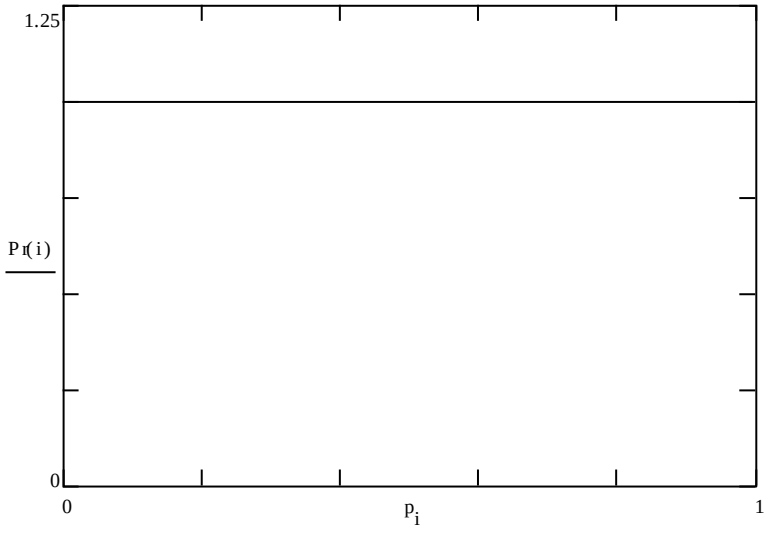
Philosophically, classical statistics uses deductive inference;
Bayesian statistics emphasizes inductive inference,
and shows how to use additional data to update our beliefs.

A simple example, without technical details:

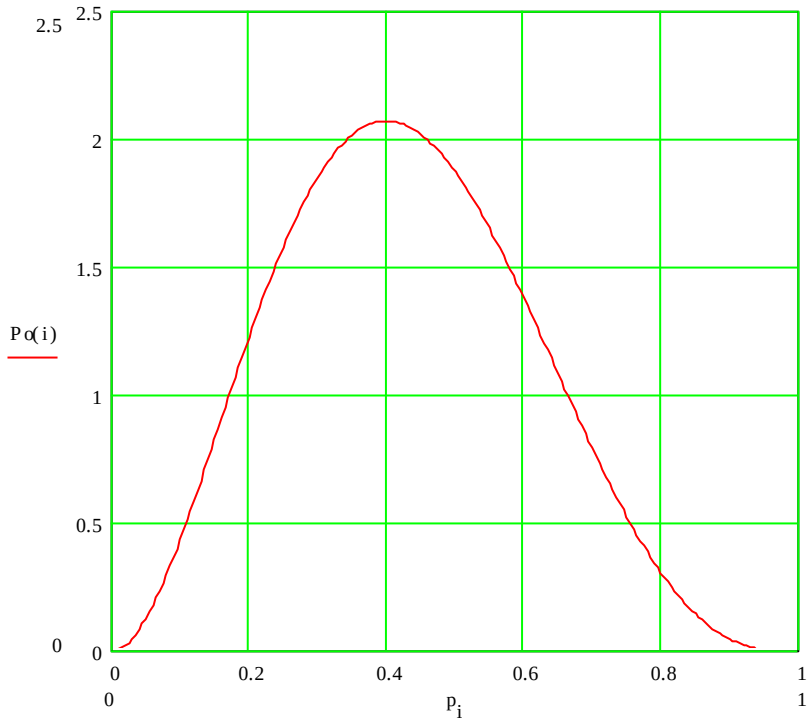
Suppose we have a thumbtack, and we want to be able to say something about the probability that, when tossed, it will land with its point up or its point down.


Assume that I have no experience with tossing thumbtacks, so I think that any probability of landing point up between 0 and 1 is equally likely.

This will be my prior distribution:
A flat line as below



Now I toss a thumbtack 5 times,
and it lands point up 2 out of the 5 times.
I can summarize the data in the form of a likelihood function:




This has the same form as a probability distribution
(except that the area under the curve may not be equal to one).
This tells you, for each possible value of the probability of landing point up,
how likely is the observed result of 2 out of 5 trials.

From the shape of this distribution (compare with the flat prior),
the data shows that extreme values of the probability
are unlikely to generate these data.

When we combine our prior information
(here, complete prior ignorance) with the data,
we get a posterior distribution
for the unknown parameter.

Because my prior distribution was uniform (flat),
my posterior will look the same as the likelihood.

We just have to make sure that we scale the height
so that the total area under the curve is 1,
like a probability distribution.
(I actually did that in the plot above.)

The plot summarizes my beliefs
about the plausibility of different values
being the correct probability
of landing with the point up.

It behaves like a probability density,
and like any other probability density,
areas under the curve represent probabilities.

These probabilities describe my beliefs
about this unknown quantity.

It looks like
the probability of landing with the point up,
which we will label $\pi$,
is somewhere between .2 and .6,
although it is possible that it is outside that range.

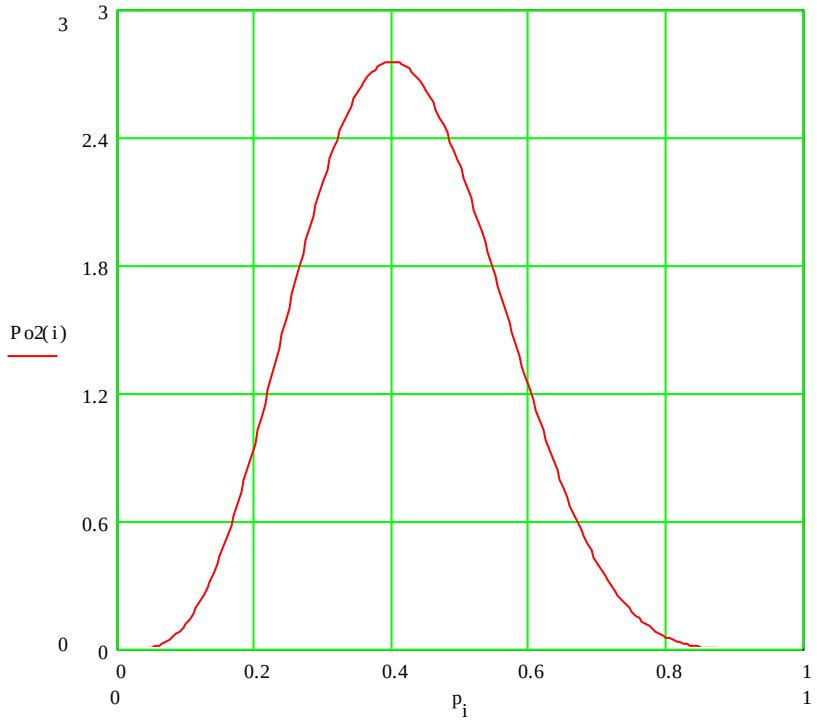
It is unlikely to be less than .1,
or greater than about .8.

We have narrowed our beliefs down somewhat,
but quite a bit of uncertainty remains.

Now suppose I gather more data;
I toss the thumbtack 5 more times.

I get another 2 points up out of 5 tosses.

Combined with the previous data,
my posterior now looks like this:




Now we are more certain that
$\pi$ is between .2 and .6,
and much more certain that
it is not less than .1 or greater than .8.

We can decide whether
this is enough information for our purposes,
or whether to gather more information
to further reduce our uncertainty.

# Simple Bayesian Example from Winkler

Description of situation:

- Quality control in manufacturing
- Process is either in control or not
- We have some long-run info on the process
- We gather some data from today's production run
- Goal 1: Make inferences about proportion of defectives
- Goal 2: Make predictions about future sample

# Possible Outcomes in Five Trials
## D = Defective, N=Non-defective

| (1) | (5) | (10) | (10) | (5) | (1) |
|---|---|---|---|---|---|
| NNNNN | DNNNN | DDNNN | DDDNN | DDDDN | DDDDD |
| | NDNNN | DNDNN | DDNDN | DDDND | |
| | NNDNN | DNNDN | DDNND | DDNDD | |
| | NNNDN | DNNND | DNDDN | DNDDD | |
| | NNNND | NDDNN | DNDND | NDDDD | |
| | | NDNDN | DNNDD | | |
| | | NDNND | NDDDN | | |
| | | NNDDN | NDDND | | |
| | | NNDND | NDNDD | | |
| | | NNNDD | NNDDD | | |

# Binomial Distribution

- How likely is any one of the outcomes?
- Let $\pi$ be the probability of a defective item
- Probability of any one of the posssible sequences of r defectives in sample of size n:
- Specific example:
  DNNDN has probability $\pi(1-\pi)(1-\pi)\pi(1-\pi)$
  $= \pi^2(1-\pi)^3$
- General form: $\pi^r(1-\pi)^{n-r}$
- For probability of r defectives, which could occur in any of a number of ways, multiply by number of ways it could occur
- Example: 2 defectives can occur in 10 ways, so probability of 2 defectives in 5 trials is
  $10\pi^2(1-\pi)^3$

# Probability vs Statistics (Likelihood)

| | .01 | .05 | .10 | .25 |
|---|---|---|---|---|
| 0 | 0.95 | 0.77 | 0.59 | 0.24 |
| 1 | 0.05 | 0.20 | 0.33 | 0.40 |
| 2 | 0.00 | 0.02 | 0.07 | 0.26 |
| 3 | 0.00 | 0.00 | 0.01 | 0.09 |
| 4 | 0.00 | 0.00 | 0.00 | 0.01 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 |

# Probability vs Statistics (Likelihood)

| | .01 | .05 | .10 | .25 |
|---|---|---|---|---|
| 0 | 0.95 | 0.77 | 0.59 | 0.24 |
| 1 | 0.05 | 0.20 | 0.33 | 0.40 |
| 2 | 0.00 | 0.02 | 0.07 | 0.26 |
| 3 | 0.00 | 0.00 | 0.01 | 0.09 |
| 4 | 0.00 | 0.00 | 0.00 | 0.01 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 |

# Choosing a Prior Distribution

- Simplification: Choose 4 possible values of $\pi$
  proportion of defective items
  (Makes it easy to see calculations)
- Suppose only possible values of $\pi$ are .01, .05, .10, and .25
- From experience, we may know the prior distribution:

| prob | 0.01 | 0.05 | 0.10 | 0.25 |
|---|---|---|---|---|
| prior | 0.60 | 0.30 | 0.08 | 0.02 |

# Data and Rule for Posterior

- We take a sample of 5 items,
  and find 1 defective
- Likelihood from binomial:
  $p(r|\pi) = 5\pi^1(1-\pi)^4$
- Example: If $\pi = .1$, $p(r = 1|\pi = .1) = 5(.1)(.9)^4 = .32805$
- Combining prior and data: $p(\pi|r) \propto p(r|\pi)p(\pi)$

# Posterior Distribution

| prob | prior | like | post.un | post |
|---|---|---|---|---|
| 0.01 | 0.60 | 0.05 | 0.03 | 0.23 |
| 0.05 | 0.30 | 0.20 | 0.06 | 0.49 |
| 0.10 | 0.08 | 0.33 | 0.03 | 0.21 |
| 0.25 | 0.02 | 0.40 | 0.01 | 0.06 |
| Sum | 1.00 | | .124 | 1.00 |

# Predictive distribution for new sample

| post: | .23 | .49 | .21 | .06 | |
|---|---|---|---|---|---|
| y | pred01 | pred05 | pred10 | pred25 | aver |
| 0 | 0.95 | 0.77 | 0.59 | 0.24 | 0.57 |
| 1 | 0.05 | 0.20 | 0.33 | 0.40 | 0.22 |
| 2 | 0.00 | 0.02 | 0.07 | 0.26 | 0.08 |
| 3 | 0.00 | 0.00 | 0.01 | 0.09 | 0.04 |
| 4 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

- Prediction is weighted sum: $\sum f(y|\pi)f(\pi|x)$
- E.g. $Prob(y = 0) =$
  $.23(.95) + .49(.77) + .21(.59) + .06(.24) = .57$
- Interpretation: We don't know $\pi$, so we average the probabilities of (e.g.) no defectives over the possible values of $\pi$, weighting by how likely each is to be true.

# Analyzing Count Data (Poisson distribution)

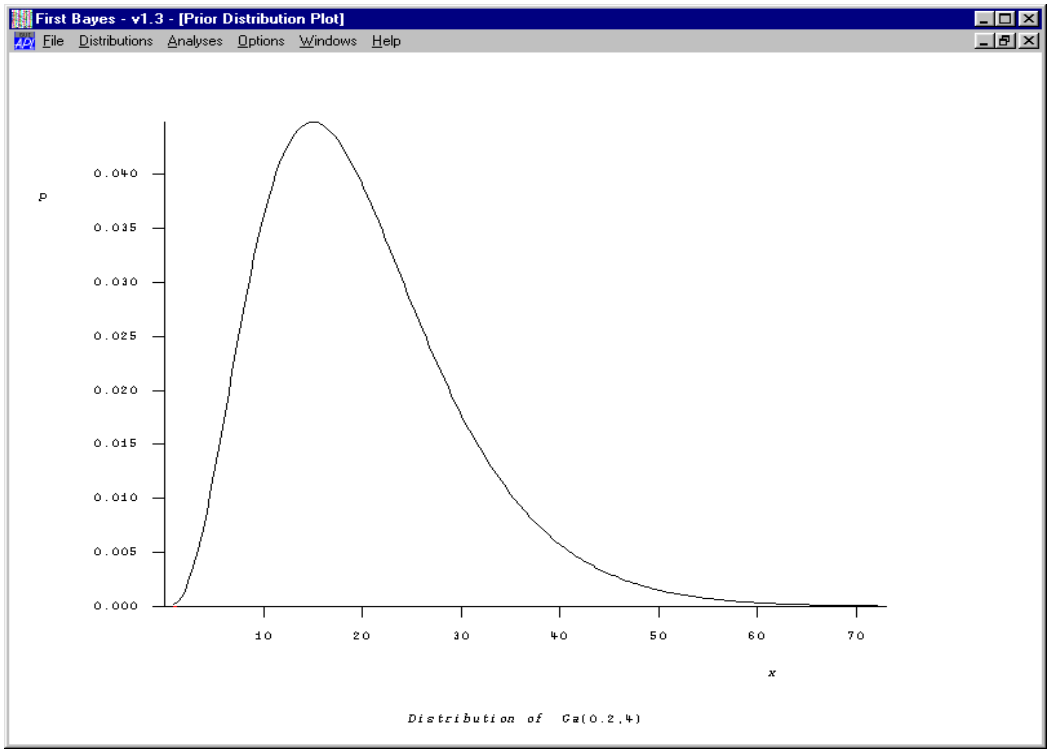
**Problem:** Suppose I want to know how many students to expect in next year's entering class. I have data from the past 10 years, during which the numbers were:

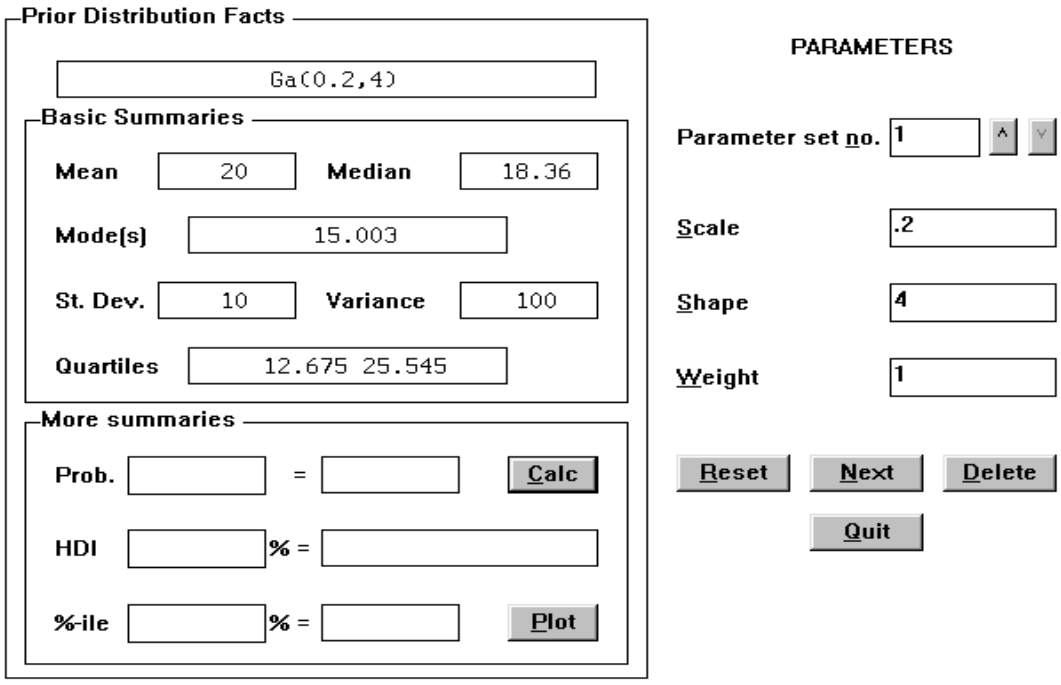
15, 13, 24, 21, 18, 17, 12, 26, 14, 20

The data are in the form of counts (non-negative integers), for which the Poisson distribution is sometimes useful. As we will see later, there is one parameter, which we will denote $\lambda$. This parameter represents the expected number (average) of entering students in a particular year. We will assume that $\lambda$ is a constant; that is, it is not changing over years. (One aspect of model checking is to determine whether this is plausible.)

**Prior Distribution:** Even before looking at the data, I have a feeling about the relative plausibility of possible values of $\lambda$. I know we always have more than 10 students, and never more than 30, so the expected number must be well within that range. So, to be somewhat conservative, I will choose a prior that puts most of the probability between those numbers.

The gamma is a convenient prior distributional form for $\lambda$.
Using First Bayes, we produce a plot of the prior distribution:






First, notice that the mean, standard deviation, and variance are all what we wanted. Next, notice that this appears to be a fairly informative prior distribution

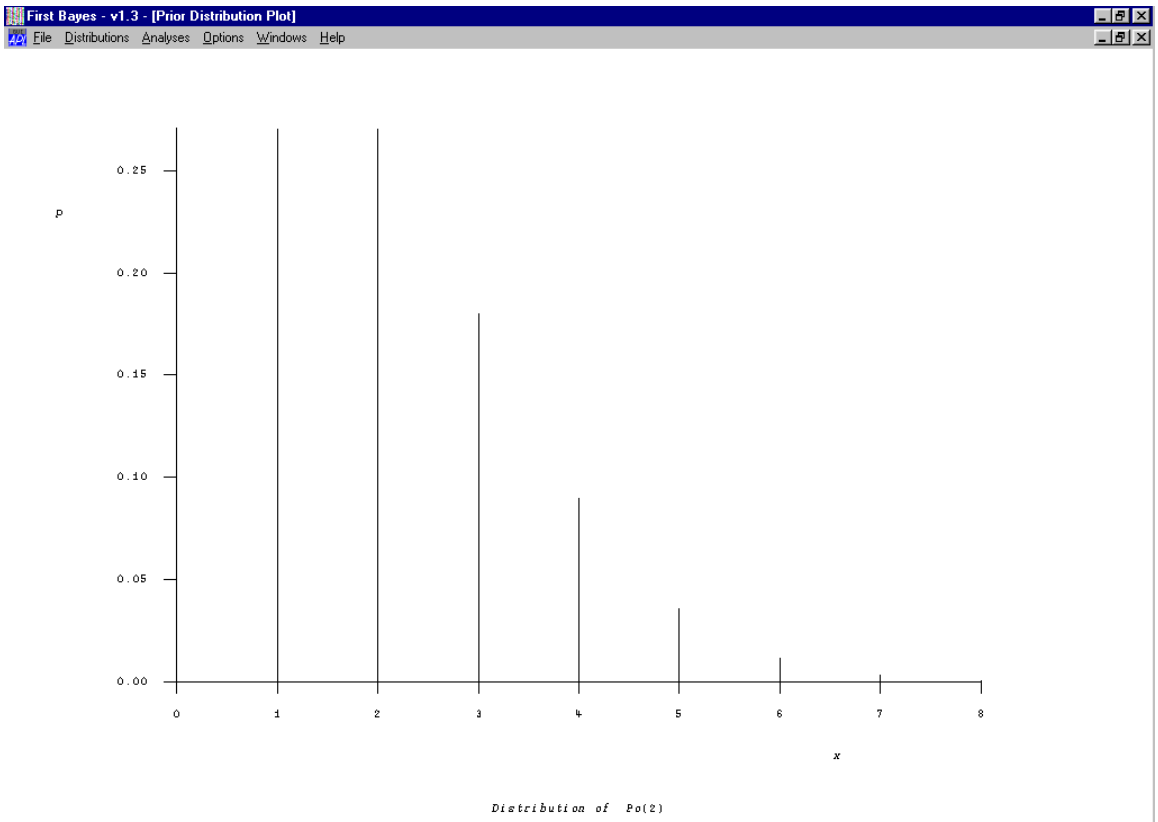
**Data Model**: A beginning place for most count data is the Poisson distribution. While models can be quite complex, taking into account many explanatory variables (much as in ANOVA and regression), we will start with a simple model that assumes constancy across years.

The probability density function for a Poisson distribution is:

$$f(x) = \exp(-\lambda)\,\lambda^{x} / x!$$

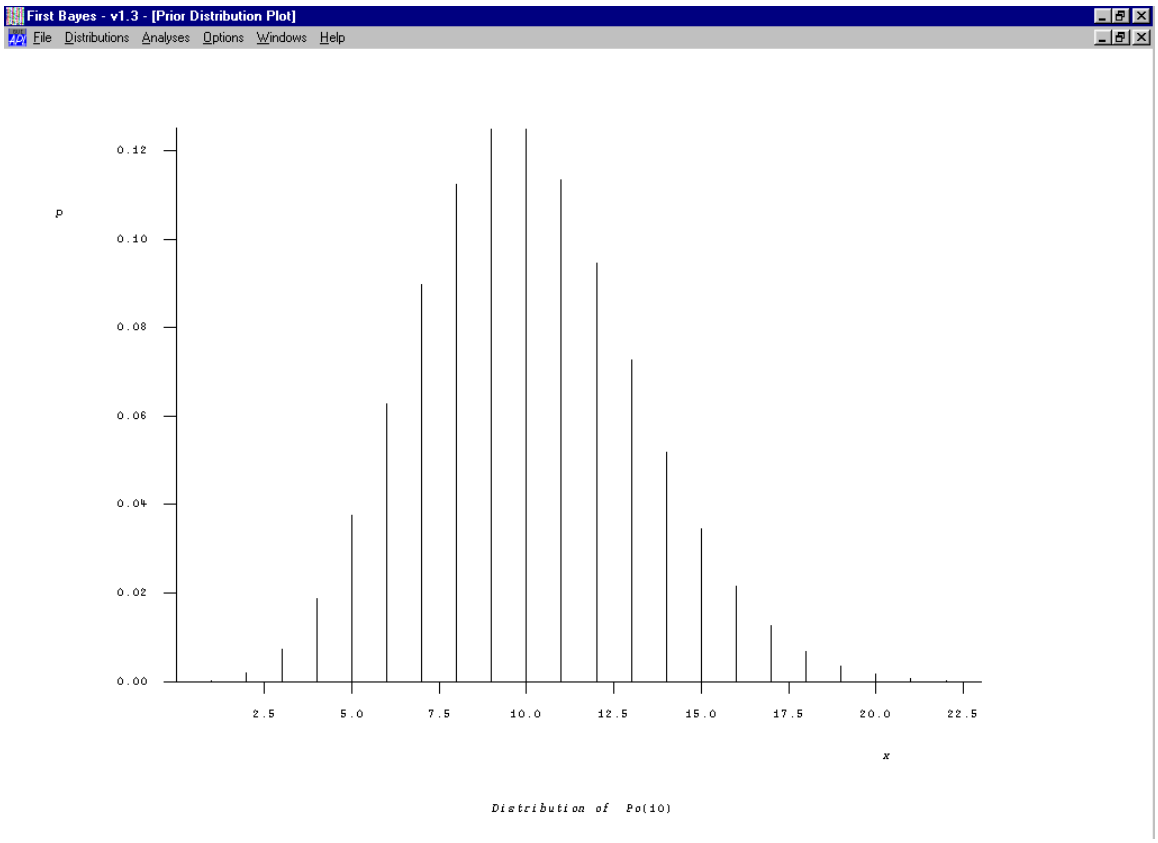
where x takes values 0, 1, 2, …. and λ is a parameter that can have any positive value. The parameter λ is the mean of the distribution (and also the variance).

Here is a plot of a Poisson distribution for which λ = 2:




(The probability that X=0 is .135, which doesn't show up clearly on the plot.)
The plot shows that for a process with an average of 2 events occurring during a particular time (or in a particular area), you might expect to see anywhere from 0 up to 5 events. There is quite a bit of variability in outcomes that can occur from this process.

The next plot shows the Poisson for a larger mean; here λ=10




Again, we see that many possible outcomes (here from about 4 to 16) would not be unusual. The distribution looks closer to being symmetric than the previous one, and not too far from a normal (except that the Poisson only considers discrete outcomes, whereas the normal is continuous). As you can probably imagine, when the mean is relatively large, we can use the normal distribution as an approximation to the Poisson.

**Likelihood**: Our data represent a sample of 10 observations, each assumed to be from the same Poisson process. That is, the value of λ is assumed to be the same for each observation. The likelihood is therefore a product of 10 terms, each with the same form, but with the value of X varying.

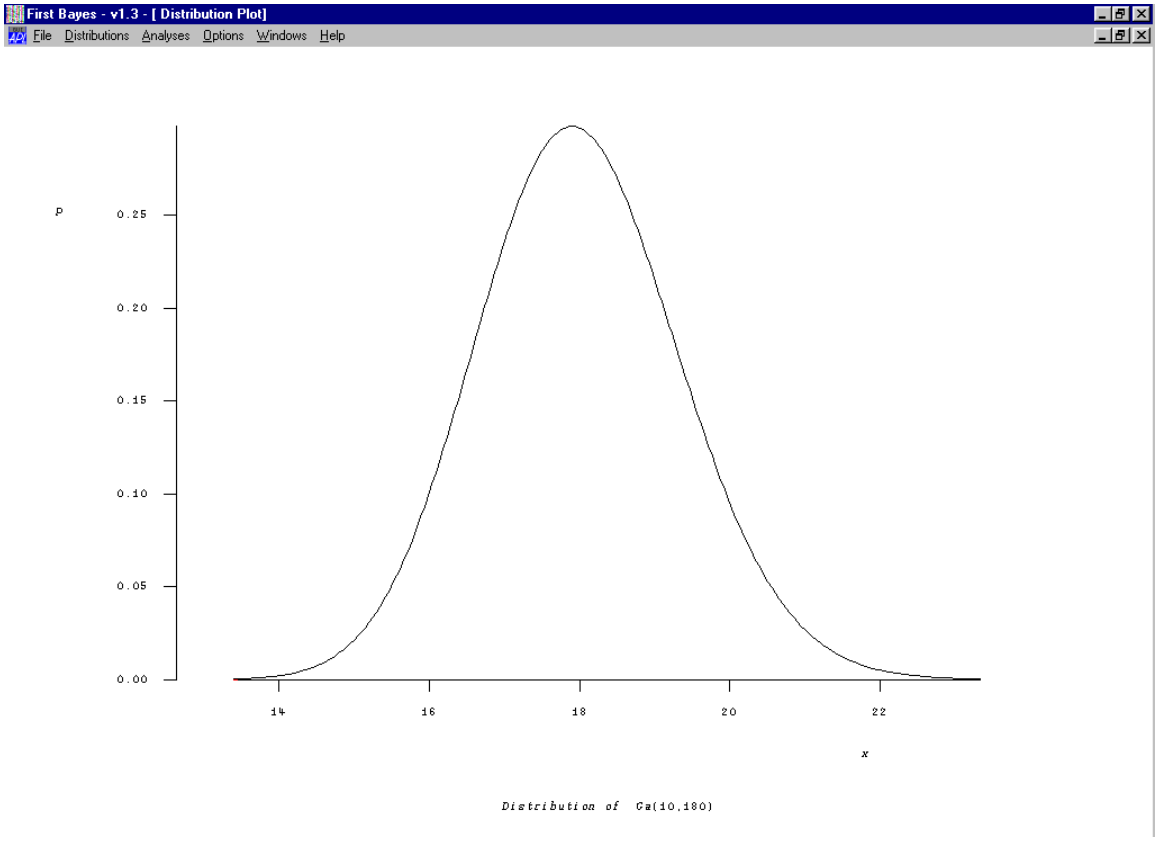
$$Lik(\lambda \mid X) \propto \{\exp(-\lambda)\}^{10} \lambda^{x_1} \lambda^{x_2} \cdots \lambda^{x_{10}}$$
$$= \{\exp(-10\lambda)\} \lambda^{\Sigma x_i}$$

Notice that all that is left of the individual observations is their sum; the exact values of the 10 observations are not needed. The sum is therefore a sufficient statistic.

We are now in a position to give a better meaning to the parameters of the gamma distribution that we are using as a prior (and posterior) distribution. The sample size in the likelihood corresponds to the place where the parameter t appears in the prior distribution; therefore t may represent the equivalent sample size of the prior information.

The sum of the observations appears where r-1 appears in the prior; therefore r is (almost) the sum, and r/t would represent the mean (as we already know).

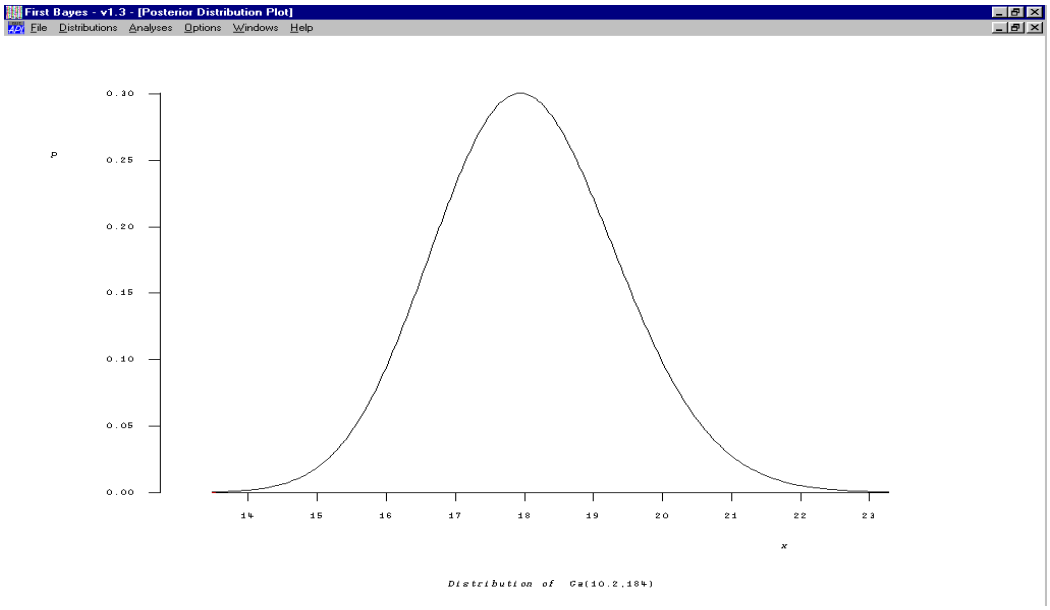
A picture of the likelihood appears below; it was actually produced by plotting a gamma(10,180) distribution, where 10 is the sample size, and 180 is the sum of the observations.



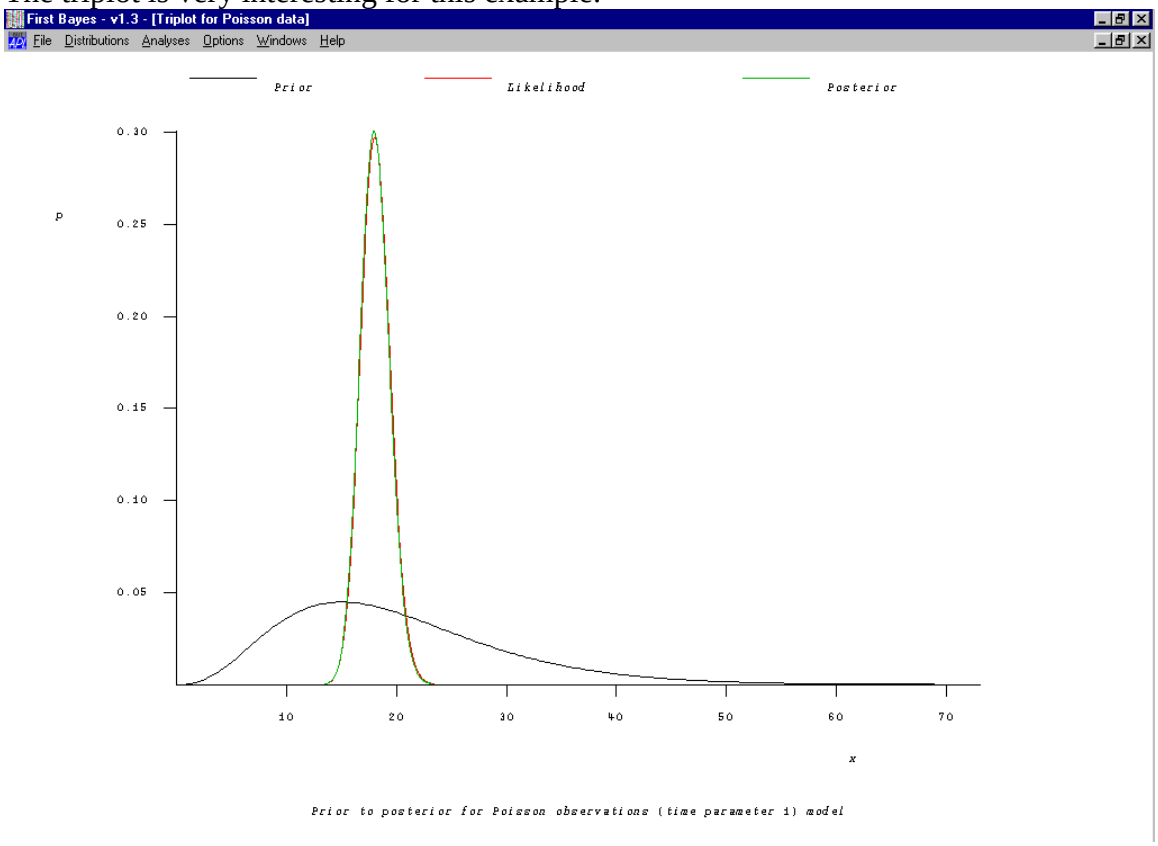

With a flat prior, the posterior would be proportional to the likelihood. Examining the likelihood from that point of view leads us to conclude that $\lambda$ is almost certainly between 14 and 22, and probably between 15 and 21.

**Posterior Distribution of $\lambda$.** As you might assume from the discussion above, because the parameters of the prior can be interpreted in terms of prior number of observations and sum of the observations, the posterior is easy to find. Merely add the number of observations in the prior to the number in the data to get t" (i.e., t for the posterior), and add the prior sum of observations (r') to the sum of observations in the data to get r" (i.e., r for the posterior).

In our case, we had chosen r' = 4 and t' = .2 for the prior; the data have N=10 and $\Sigma X$ = 180. Therefore, the posterior is a gamma distribution with parameters r" = 4 + 180 = 184 for the shape, and t" = .2 + 10 = 10.2 for the scale. A plot of this distribution is:

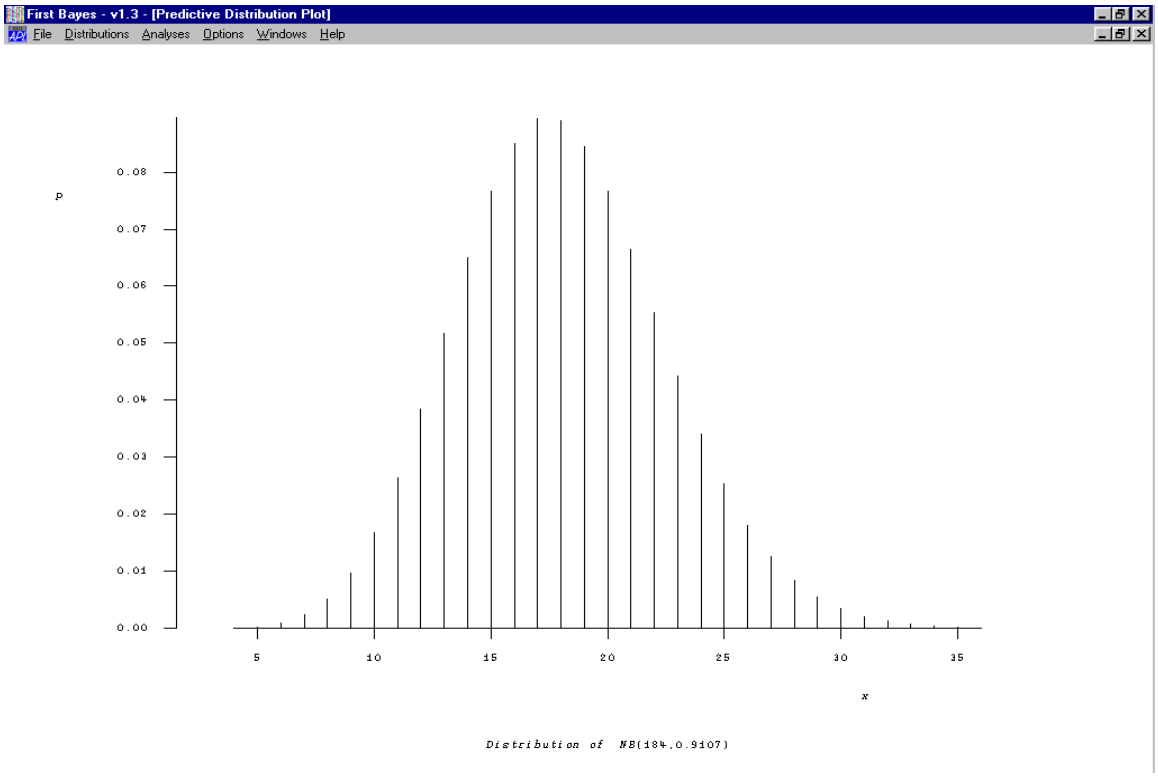

First Bayes - v1.3 - [Posterior Distribution Plot]
File Distributions Analyses Options Windows Help
P
0.30
0.25
0.20
0.15
0.10
0.05
0.00
14
15
16
17
18
19
20
21
22
23
x
Distribution of Ga(10.2,184)

The triplot is very interesting for this example:

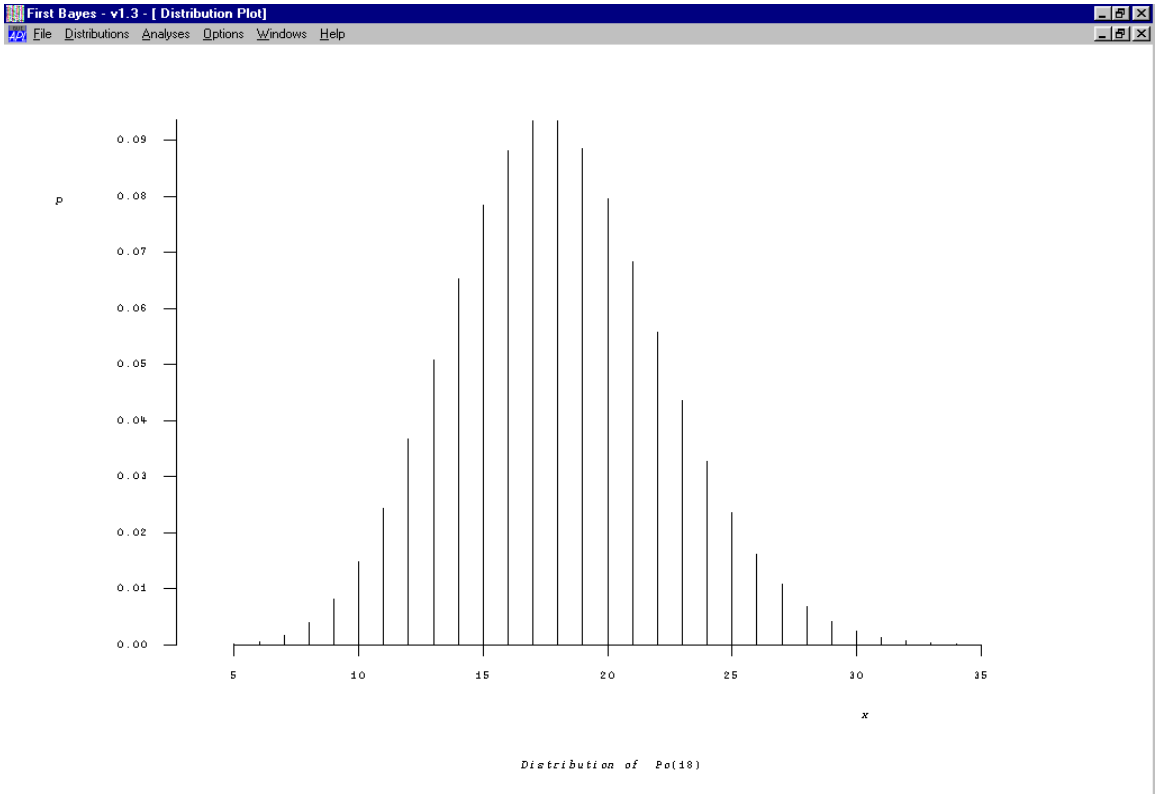



In spite of using what we thought was a fairly informative prior, the posterior is very similar to the likelihood. In fact, in this plot they are not distinguishable. The answer to this apparent paradox is to notice what is happening to the prior in the area where the likelihood is large, that is, for values of $\lambda$ between about 15 and 21. In this area, the prior distribution is relatively flat, in spite of it being distinctly not flat in other areas. Because the prior is relatively flat where the likelihood is appreciable, the posterior will look very similar to the likelihood. In Bayesian statistics, this is sometimes known as the principle of stable estimation: Informative priors will have relatively little impact on the posterior if the prior is relatively flat where the likelihood is large. In this area, the prior acts as if it were relatively noninformative.

**Predictive Distribution**. Finding the posterior distribution is not the final step in our process. We still need to think about the original problem: Predicting what might happen next year. If the posterior distribution had a very small variance, so that we were fairly certain about the value of $\lambda$, then we could use that value in a Poisson distribution to find the probability of any outcome next year. However, a range of values of $\lambda$ are still plausible, even with 10 years' data. Each plausible value of $\lambda$ produces a Poisson to predict the outcome if that were the correct value of $\lambda$. What we must do is to average all the plausible Poisson distributions, weighting by how likely each is to be the correct one. The predictive distribution is therefore a mixture of Poissons, where the mixing distribution is a gamma. The resulting mixture is called a negative binomial distribution. A plot of the posterior is shown below:




Compare this to a Poisson distribution with λ=18:



While it is difficult to make an accurate visual comparison, the variance of the Poisson is 18 (the same as the mean), while for the NB(184, .9107) predictive distribution the variance is 19.81, somewhat larger. In this case, using a point estimate of the parameter would not have resulted in predictions that were very different from using the (correct) predictive distribution.

What can we say about the number of students likely to enter next year? First, that 18, the observed average, is nearly equal to the expected number. But there is great uncertainty about how many will come; a 90 percent credible interval is from 11 to 25, while a 95 percent interval is from 10 to 27. So any value from 10 to 27 entering students would be consistent with our model and data. This is quite a bit more variation than most people would expect to occur by chance. Therefore, when they see enrollments higher in one year, the faculty may think they did something special to attract students; when they see lower enrollments, the faculty may wonder what they should do to recruit additional good students. However, a large amount of such fluctuation is likely to be purely random, due completely to chance.

**Homework:** The following data are taken from a famous example illustrating the Poisson distribution. The Prussian army kept track of how many soldiers in each of its units died from being kicked by horses (this was back in the 19$^{th}$ Century, of course). Corps XI had the following numbers killed each year, from 1875 to 1894:

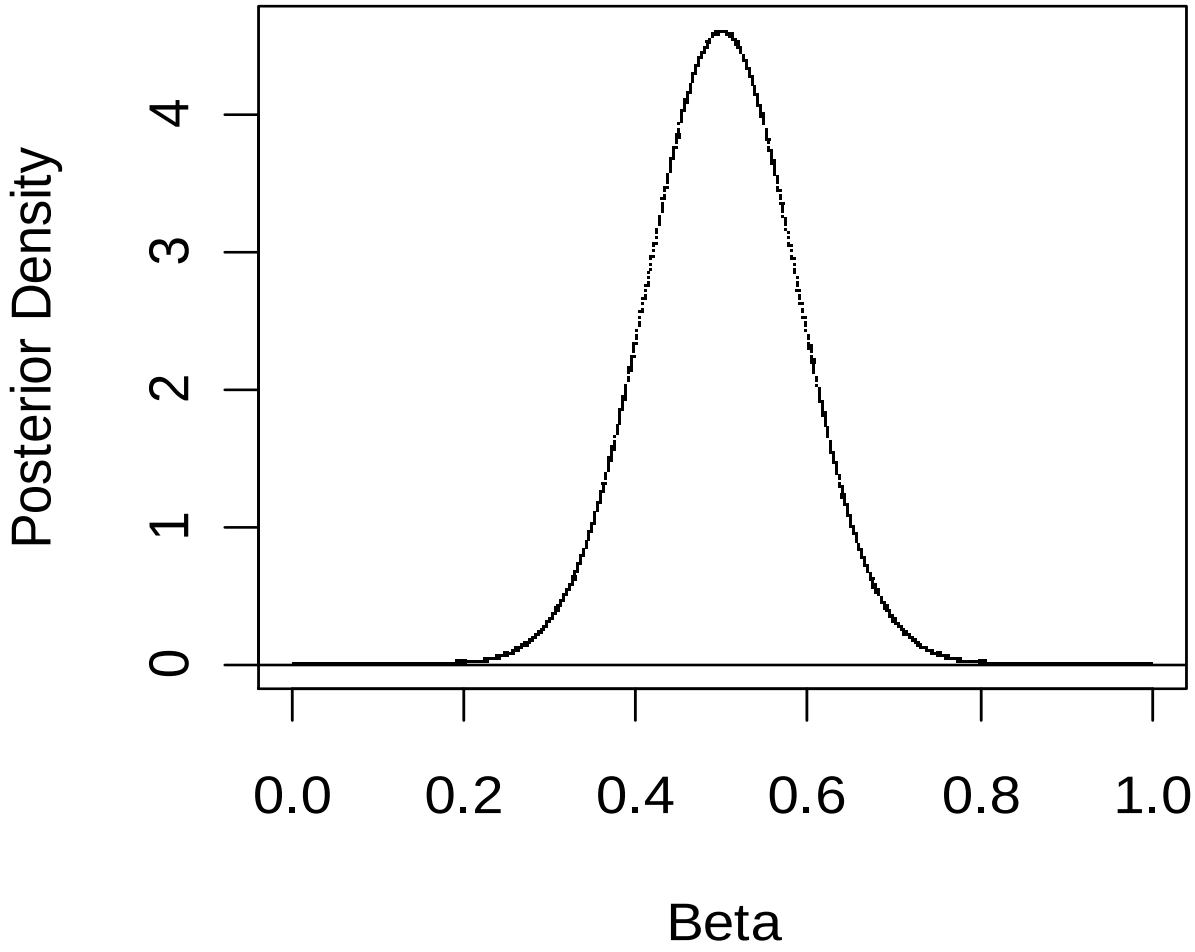
0, 0, 0, 0, 2, 4, 0, 1, 3, 0, 0, 1, 1, 1, 2, 1, 3, 1, 3, 1.

Assume that the true rate is a constant across years (this can be tested, but we will not do it). Conduct a Bayesian analysis of these data, and write up the results. Try various plausible prior distributions, and see how much effect the choice of prior has on the outcome.

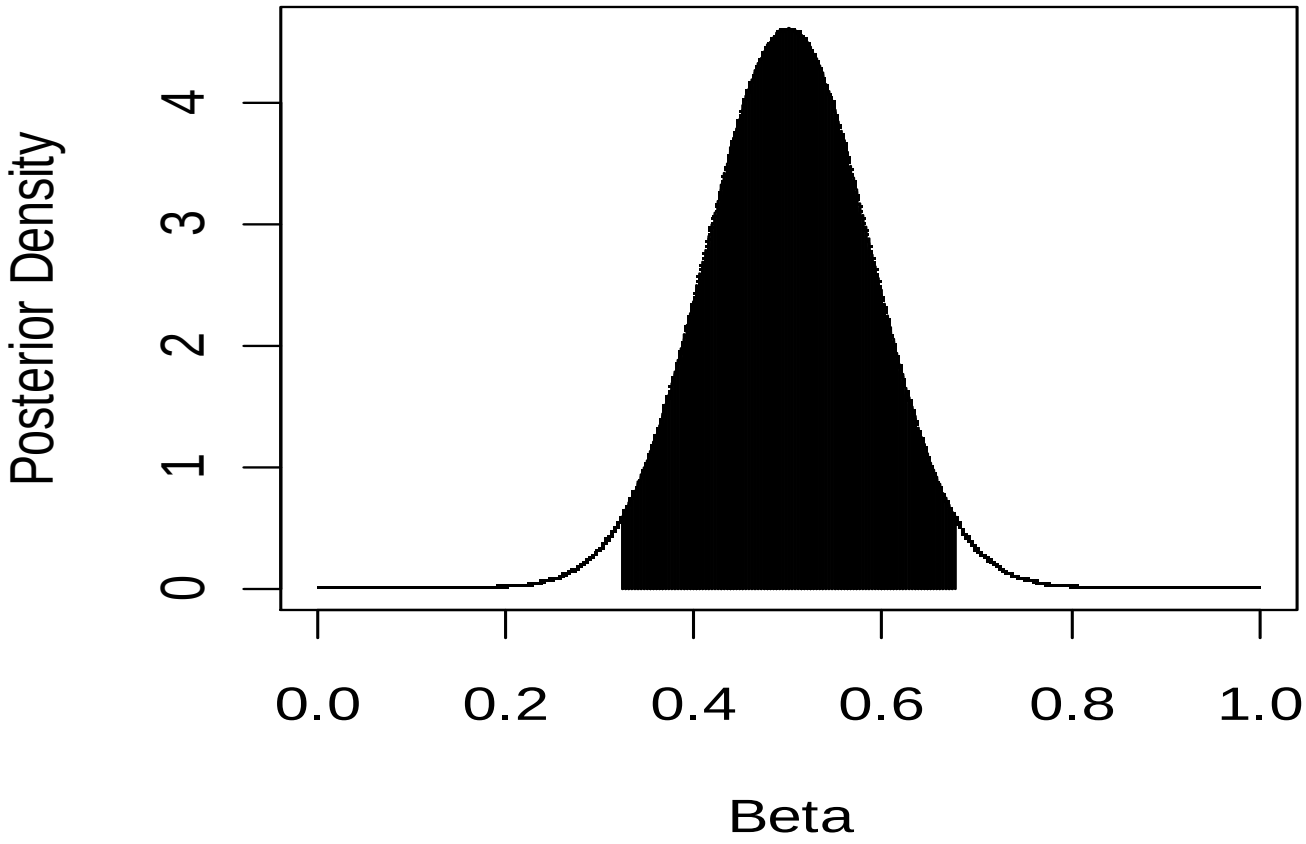
# Simple Bayesian Regression: OLS with new interpretation

- Suppose r(xy) = .5, and n = 102
- Beta.hat = r = .5
- Then $r^2$ = .25, and
- std.err = sqrt(.75)/10 = .087
- Posterior for beta is t, but nearly normal, with mean = .5 and std.dev = .087
- Same as OLS, but new interpretation:

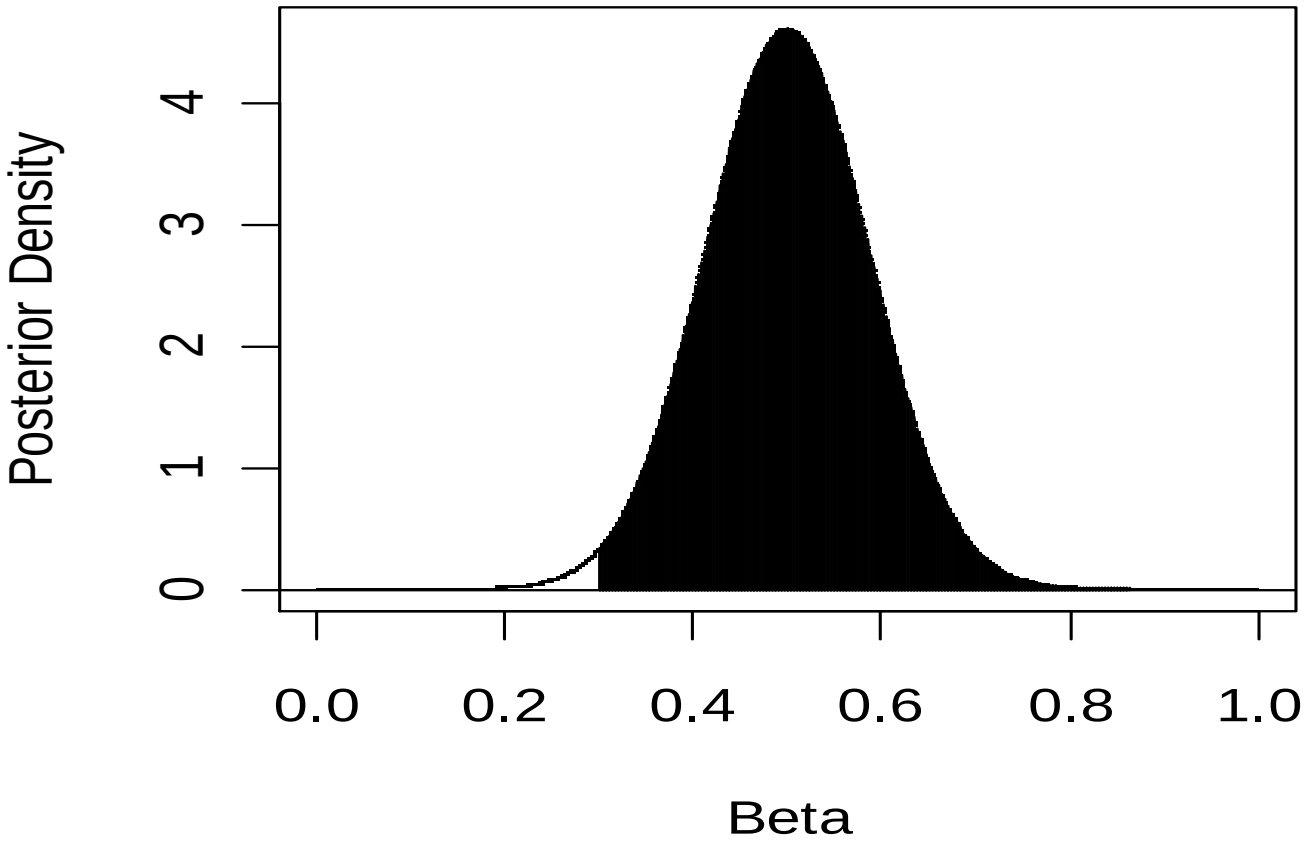
# Posterior distribution of standardized regression coefficient beta



# Probability is .95 that beta is in the interval (.32, .68)



# Probability that beta > .3 (at least medium effect) is .99+



# U836 S03 Week 12
# Hierarchical Bayesian Models

Here we will learn what to do with data that are obtained from a number of independent sources, but are "similar" in some respects. The specific example concerns death rates in operations in each of 12 hospitals. Even though the hospitals are unrelated, it would seem that the death rate in any hospital would be similar to that in other hospitals, and that such information could be used to improve estimates in each hospital. This example is taken from the BUGS manual, but with some alterations. (The original is in Arial font.)

## Surgical: Institutional ranking

This example considers mortality rates in 12 hospitals performing cardiac surgery in babies. The data are shown below.

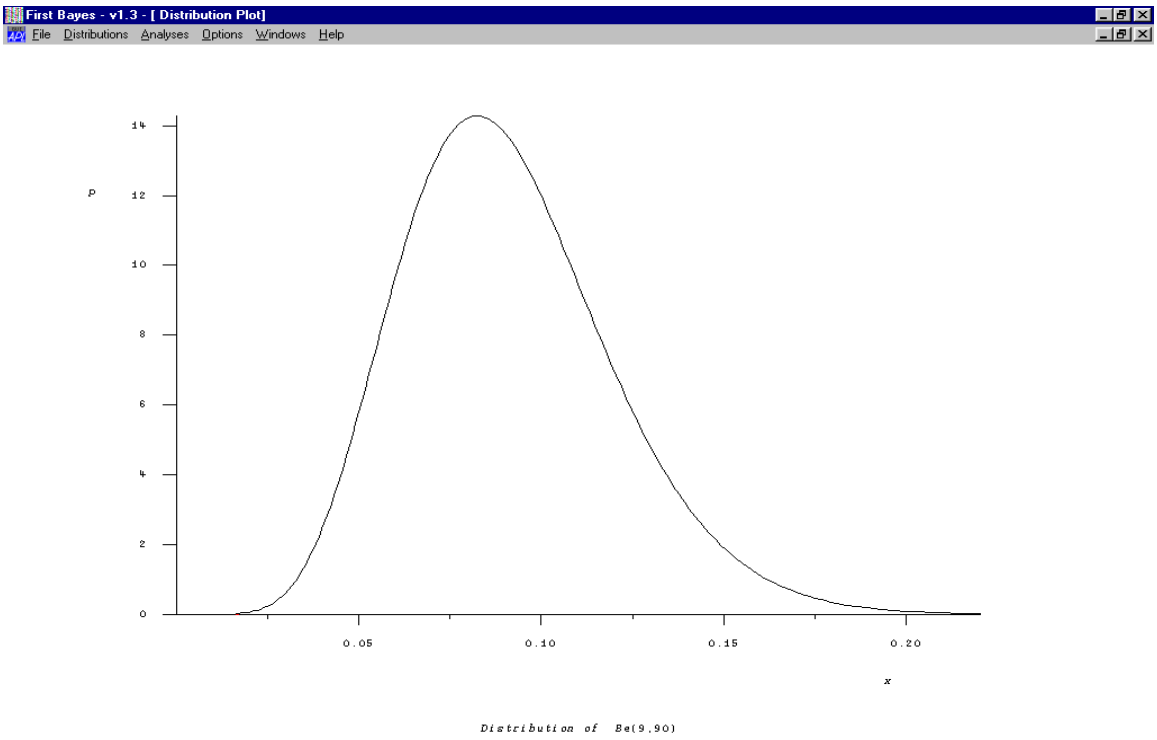
| Hospital | No of ops | No of deaths |
|---|---|---|
| A | 47 | 0 |
| B | 148 | 18 |
| C | 119 | 8 |
| D | 810 | 46 |
| E | 211 | 8 |
| F | 196 | 13 |
| G | 148 | 9 |
| H | 215 | 31 |
| I | 207 | 14 |
| J | 97 | 8 |
| K | 256 | 29 |
| L | 360 | 24 |

Here are the data again, with p(death) computed, and the data sorted in increasing order of p(death):

**BUGS Surgical Data**
**Sorted by p(death)**

| | CASE | HOSPITAL | DEATH | N | P.DEATH |
|---|---|---|---|---|---|
| 1 | 1 | A | 0 | 47 | .00 |
| 2 | 5 | E | 8 | 211 | .04 |
| 3 | 4 | D | 46 | 810 | .06 |
| 4 | 7 | G | 9 | 148 | .06 |
| 5 | 6 | F | 13 | 196 | .07 |
| 6 | 12 | L | 24 | 360 | .07 |
| 7 | 3 | C | 8 | 119 | .07 |
| 8 | 9 | I | 14 | 207 | .07 |
| 9 | 10 | J | 8 | 97 | .08 |
| 10 | 11 | K | 29 | 256 | .11 |
| 11 | 2 | B | 18 | 148 | .12 |
| 12 | 8 | H | 31 | 215 | .14 |

Before considering the use of BUGS for the analyses of these data, think about what you would do if you were the statistician for one of these hospitals, and had only the data from that hospital to analyze. Consider hospital J, which had 8 deaths in 97 operations. We analyze the data using First Bayes, as we have done for similar problems in the past. We will use a prior Be(1,1), giving a posterior that is Be(9,90), pictured below:




The mode of the posterior is .082, and the 95 percent HDI is (.039, .148).

Here is a BUGS version, in which the data for all 12 hospitals are analyzed at once. In this analysis, each hospital's data is analyzed separately; i.e., the results for one hospital are not used to help model the outcome for any other hospital.

The number of deaths $r_i$ for hospital *i* are modelled as a binary response variable with `true' failure probability $p_i$:

$$r_i \sim \text{Binomial}(p_i, n_i)$$

We first assume that the true failure probabilities are *independent* (*i.e.*fixed effects) for each hospital. This is equivalent to assuming a standard non-informative prior distribution for the $p_i$'s, namely:

$$p_i \sim \text{Beta}(1.0, 1.0)$$

BUGS *language for fixed effects surgical model:*

```
model
{
  for( i in 1 : N ) {
     p[i] ~ dbeta(1.0, 1.0)
          r[i] ~ dbin(p[i], n[i])
  }
  }
```

## Data

```
list(n = c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360),
    r  = c( 0,   18,    8,   46,   8,   13,    9,   31,  14,   8,   29,   24),
    N = 12)
```

## Inits

```
list(p = c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1))
```

| node | mean | sd | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|
| p[1] | 0.01999 | 0.0198 | 5.961E-4 | 0.01406 | 0.07174 | 5001 | 5000 |
| p[2] | 0.1268 | 0.02718 | 0.07853 | 0.125 | 0.1859 | 5001 | 5000 |
| p[3] | 0.07446 | 0.02366 | 0.03479 | 0.07206 | 0.1259 | 5001 | 5000 |
| p[4] | 0.0579 | 0.008362 | 0.04265 | 0.05756 | 0.07512 | 5001 | 5000 |
| p[5] | 0.04243 | 0.014 | 0.01949 | 0.04081 | 0.0742 | 5001 | 5000 |
| p[6] | 0.0709 | 0.01823 | 0.03922 | 0.06942 | 0.1105 | 5001 | 5000 |
| p[7] | 0.06694 | 0.02032 | 0.03223 | 0.06471 | 0.1117 | 5001 | 5000 |
| p[8] | 0.1473 | 0.02375 | 0.1048 | 0.1458 | 0.1978 | 5001 | 5000 |
| p[9] | 0.0718 | 0.01808 | 0.04069 | 0.07044 | 0.1107 | 5001 | 5000 |
| p[10] | 0.09079 | 0.02875 | 0.04298 | 0.08805 | 0.1525 | 5001 | 5000 |
| p[11] | 0.1168 | 0.01997 | 0.08034 | 0.1157 | 0.1592 | 5001 | 5000 |
| p[12] | 0.06906 | 0.01352 | 0.04533 | 0.06824 | 0.09758 | 5001 | 5000 |

Hospital J is number 10. The modal estimate of the death rate is .088, and a 95 percent posterior interval (not necessarily HDI, but close) is (.043, .153). These are not far from the exact results provided by First Bayes.

## Hierarchical Bayes Model

Next we consider how to make use of the fact that each of the 12 hospitals are estimating the same type of quantity. While their death rates may truly differ (because some are better than others, or some treat more serious cases than others, for example), we still expect some similarities among the rates. How can we incorporate this idea into the analysis?

Let us suppose that we believe each hospital to have a death rate, pi(i), where the subscript i refers to hospitals. The pi(i) for the 12 hospitals will have some distribution; a simple possibility is that they have a beta distribution. If this distribution has a small variance, then hospitals all have similar death rates from this operation, and we can "borrow information" from all hospitals to help estimate the death rate in each individual hospital. That is, to estimate the true death rate in hospital A, we can use not only the information from that hospital, but also the information from hospitals B, C, …, L. In a worst case scenario, suppose we had no information from hospital A; if we know that hospitals all have similar death rates, we could get a reasonably good estimate of the death rate in hospital A by using the data from the other hospitals.

Remember that we do not observe the pi(i); we observe the numbers of deaths and operations in each hospital. The observed death rate in each hospital has sampling error; the large the number of operations performed in that hospital, the smaller the sampling error. The observed death rates therefore vary for two reasons: first, the true rates vary from hospital to hospital, and second, the observed rate varies from the true rate due to sampling error. This defines the hierarchy in the model: true variation among hospitals, and sampling variability within hospitals.

How should we alter our statistical model to reflect this perspective? What we need is to specify a distribution for the unknown pi(i), and then specify a vague prior for the parameters of this distribution. Because proportions vary from 0 to 1, the beta is a natural distributional form (we will see later a different possibility). A beta has two parameters, each of which is restricted to be positive. What kind of prior distribution might we choose for them? We have seen a similar situation in the analysis of a Poisson variable: The mean of the Poisson was restricted to be positive; we used a gamma distribution for the prior (which also became the form of the posterior). We discovered that a Ga(.01,.01), or any other with very small parameters, is a prior form that is nearly uninformative. BUGS cannot use a completely noninformative prior here, because it is not a proper distribution (that is, the area under the curve would not be finite). So the modified BUGS code is:

```
# surgical data 2a

# random effects, but in original scale of probabilities
# use beta as distribution of effects (random variable)

model
        {
        for( i in 1 : N ) {
             p[i] ~ dbeta(a,b)
             r[i] ~ dbin(p[i], n[i])
            }
        a ~ dgamma(.01,.01)   # hyperpopulation parameters
        b ~ dgamma(.01,.01)

        mean <- a/(a+b)                              #  better characterizations of
        sd <- sqrt(a*b/((a+b)*(a+b)*(a+b+1)))        #   posterior
        mode <- (a-1)/(a+b-2)
        }
```

```
# data

    list(n = c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360),
         r  = c(0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24),
         N = 12)

# initialization of parameters
              list(a=2,b=2)
```

The only change in the first part of the code is that the proportion of deaths for each hospital is drawn from the same distribution: The parameters a and b are used, and these have no subscripts, so they are the same for each hospital. The *a* and *b* values have gamma prior distributions.

The last part of the model code creates some new values that are useful in describing the posterior distribution of the proportions. The original parameters are difficult to interpret, so we create a mean, mode, and standard deviation to help visualize the shape of the beta posterior distribution.

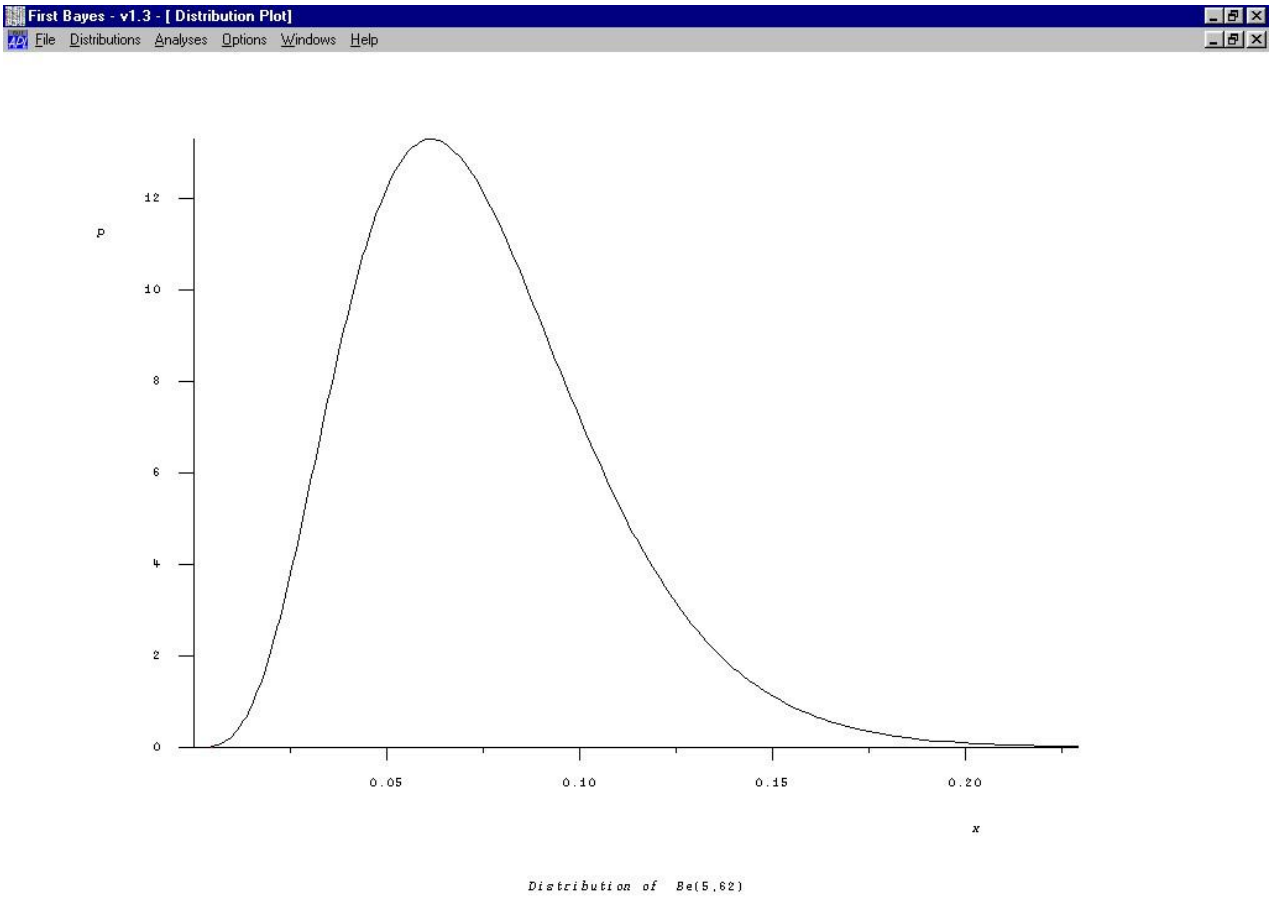
The data are in lists of number of operations and numbers of deaths, and the number of hospitals.

Because the prior distributions for the parameters of the beta prior are so diffuse (noninformative), we must give guesses for these parameters in order for the program to find reasonable estimates. These don't have to be really accurate, just in the ballpark. One way to check the convergence of the procedure is to try different starting values and run the model again.

Here are some of the results for this model:

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| a | 5.982 | 3.453 | 0.2929 | 1.539 | 5.207 | 14.59 | 5001 | 5000 |
| b | 71.87 | 41.89 | 3.566 | 17.54 | 62.39 | 174.8 | 5001 | 5000 |
| mean | 0.07809 | 0.01209 | 2.092E-4 | 0.05755 | 0.07696 | 0.1056 | 5001 | 5000 |
| mode | 0.06183 | 0.01532 | 7.294E-4 | 0.02722 | 0.06343 | 0.0871 | 5001 | 5000 |
| p[1] | 0.04504 | 0.02195 | 8.477E-4 | 0.008348 | 0.04312 | 0.09298 | 5001 | 5000 |
| p[2] | 0.1077 | 0.02228 | 5.856E-4 | 0.06926 | 0.1064 | 0.1556 | 5001 | 5000 |
| p[3] | 0.07084 | 0.01888 | 3.076E-4 | 0.03822 | 0.06958 | 0.1124 | 5001 | 5000 |
| p[4] | 0.05847 | 0.008054 | 1.432E-4 | 0.04369 | 0.05793 | 0.07511 | 5001 | 5000 |
| p[5] | 0.04804 | 0.0135 | 3.756E-4 | 0.0246 | 0.04709 | 0.07675 | 5001 | 5000 |
| p[6] | 0.06923 | 0.01588 | 2.468E-4 | 0.04138 | 0.06819 | 0.1038 | 5001 | 5000 |
| p[7] | 0.06623 | 0.01708 | 2.576E-4 | 0.03593 | 0.06502 | 0.1031 | 5001 | 5000 |
| p[8] | 0.1276 | 0.02098 | 6.535E-4 | 0.09017 | 0.1269 | 0.1726 | 5001 | 5000 |
| p[9] | 0.07023 | 0.01562 | 2.176E-4 | 0.04252 | 0.06926 | 0.1034 | 5001 | 5000 |
| p[10] | 0.08014 | 0.02145 | 3.098E-4 | 0.044 | 0.07835 | 0.1264 | 5001 | 5000 |
| p[11] | 0.1056 | 0.01747 | 4.084E-4 | 0.07481 | 0.1045 | 0.1424 | 5001 | 5000 |
| p[12] | 0.06843 | 0.01249 | 1.965E-4 | 0.04636 | 0.06763 | 0.09466 | 5001 | 5000 |
| sd | 0.03427 | 0.01183 | 8.216E-4 | 0.01881 | 0.03198 | 0.06372 | 5001 | 5000 |

The posterior distribution of the true hospital death rates is a beta with parameters estimated to be approximately 5 and 62. This corresponds to an average of about .078, and a standard deviation of .034. The rates for individual hospitals should be pulled towards .078; those with smaller sample sizes will be pulled more strongly towards that number, while those with larger sample sizes will not be pulled so strongly. The posterior is pictured below:



From the picture we see that most hospitals should have death rates between about .02 and .15. (The actual 95 percent HDI is (.02, .14).) Therefore, if we were told that there is a 13th hospital from this group, but we have no data on that hospital, we could say with some confidence that it should have a death rate in that interval. More precisely, we would look at the predictive distribution, since there is some uncertainty here about exactly what the parameters a and b should be. To do this in BUGS, we actually create a 13th hospital with missing data, so the data becomes:

```
list(n = c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360, 0),
           r  = c( 0,  18,    8,   46,   8,   13,   9,  31,  14,    8,   29,  24, NA),
           N = 13)
```

The statistics for the "phantom" 13th hospital form the predictive distribution:

| node | mean | sd | 2.5% | 5.0% | 10.0% | 25.0% | median | 75.0% | 90.0% | 95.0% | 97.5% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| p[13] | 0.07816 | 0.0391 | 0.01974 | 0.02601 | 0.03561 | 0.05237 | 0.07258 | 0.09664 | 0.1257 | 0.1477 | 0.1713 |

For another hospital with no data, we would predict a death rate of .078, and a 95 percent interval estimate would be (.02, .17). As indicated, we can make reasonably accurate inferences with no data, other than knowing that this hospital was drawn from the same population as the others.

Finally, we compare the results for hospital 10 (J) with what we calculated previously. Here, the mean is .080, the median estimate is .078, and the 95 percent interval is (.044, .126). The mean is pulled down toward the overall mean of .078, and the interval is much narrower. The narrower interval means that we have more certainty about the estimate for this hospital when we use information from other, similar, hospitals than when we only use information from that hospital alone.

# Simple example of simulation of parameters and functions of them from posterior

- Ex. from Ioannis Ntzoufras, *Bayesian Modeling Using WinBUGS*, p. 33-35
- Case-control study: {case/control} {exposed/not}
- Model: pr(case | group) ~ binomial
- We'd like to estimate effect using one of three standard methods:
  - Attributable risk (difference)
  - Relative risk (ratio)
  - Odds ratio

# Observed frequencies and row proportions

| | case | control |
|---|---|---|
| exposed | 25 | 300 |
| not exp | 30 | 900 |

| | case | control |
|---|---|---|
| exposed | 0.08 | 0.92 |
| not exp | 0.03 | 0.97 |

# Notation and problem statement

```
#  Nzoutfras p 34

#  y0 and y1 are non-exposed and exposed cases,
#  n0 and n1 are non-exposed and exposed controls.

#  simulate from posterior;
#  find posterior of derived quantities:
#    AR : attributable risk (diff in probabilities)
#    RR : relative risk (ratio of probabilities)
#    OR : odds ratio
```

# Specify priors and data; draw parameters from posterior

```
a <- 1
a0 <- a1 <- b0 <- b1 <- a

y1 <- 25; y0 <- 30; n1 <- 300; n0 <- 900
p0 <- rbeta(10000, y0+a0, n0+b0)
p1 <- rbeta(10000, y1+a1, n1+b1)
```

# Calculate derived quantities, describe their posterior

```
AR <- p1-p0;  RR <- p1/p0
OR <- p1*(1-p0)/(p0*(1-p1))

mean(AR);  mean(RR);  mean(OR)
sd(AR);    sd(RR);    sd(OR)

quantile(AR,c(.025,.975))
quantile(RR,c(.025,.975))
quantile(OR,c(.025,.975))
```

# Display posterior means and standard deviations (similar to standard error)

```
> mean(AR);  mean(RR);  mean(OR)
[1] 0.04635321
[1] 2.471100
[1] 2.606942

> sd(AR);    sd(RR);    sd(OR)
[1] 0.01601468
[1] 0.6512437
[1] 0.7311651
```

# Display credible intervals (like confidence intervals)

```
> quantile(AR,c(.025,.975))
      2.5%      97.5%
0.01736515 0.07969480

> quantile(RR,c(.025,.975))
    2.5%    97.5%
1.446549 3.949577

> quantile(OR,c(.025,.975))
    2.5%    97.5%
1.476032 4.283846
>
```

# Prettier output using xtable

| | average | std.err | 2.5% | 97.5% |
|---|---|---|---|---|
| Risk Diff | 0.05 | 0.02 | 0.02 | 0.08 |
| Rel Risk | 2.47 | 0.65 | 1.45 | 3.95 |
| Odds Ratio | 2.61 | 0.73 | 1.48 | 4.28 |

# Plot smoothed posterior distributions

```
plot(density(AR), main=' ', xlab='Attributable Risk',
  ylab='Posterior Density')

plot(density(RR), main=' ', xlab='Relative Risk',
  ylab='Posterior Density')

plot(density(OR), main=' ', xlab='Odds Ratio',
  ylab='Posterior Density')
```

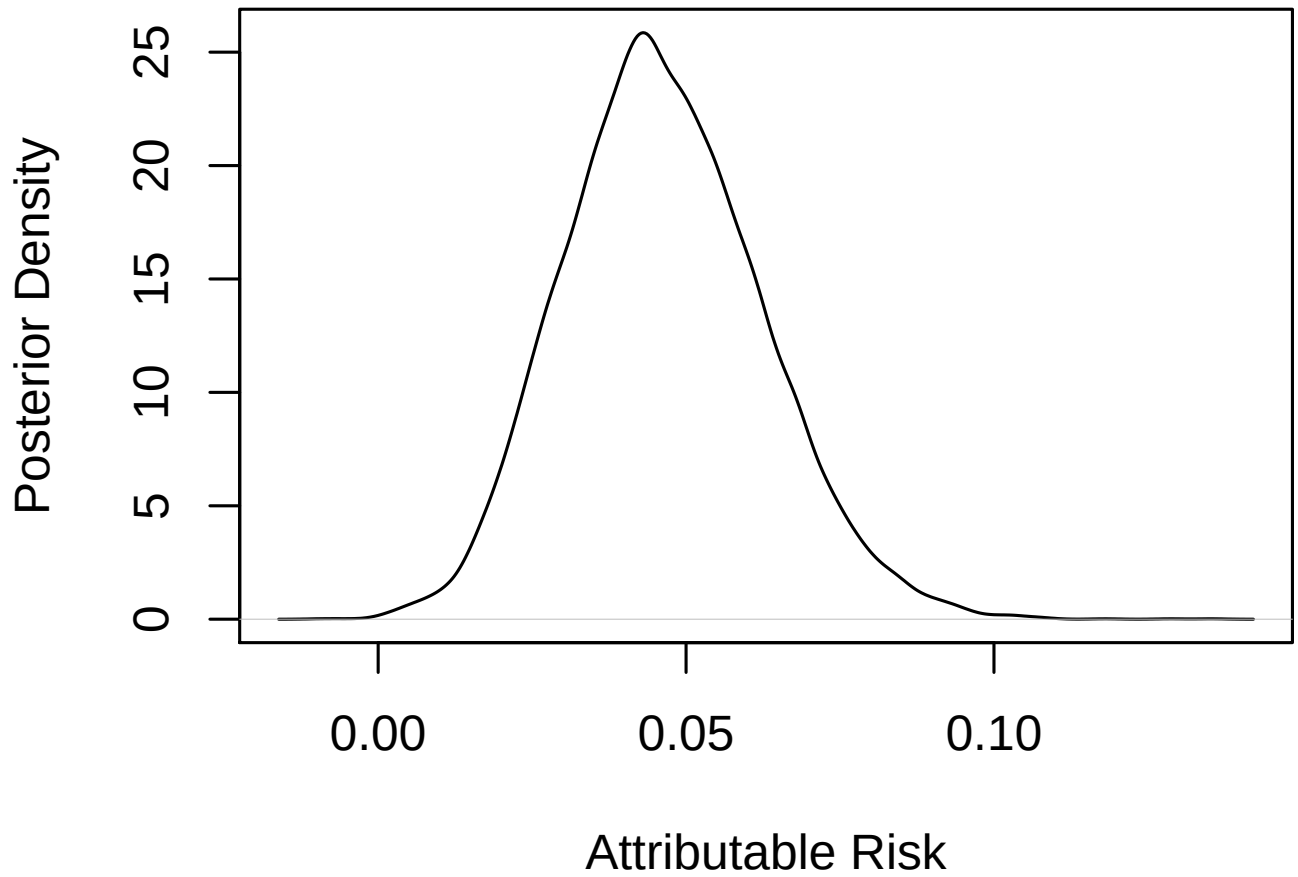



Figure: *Posterior distribution of attributable risk (risk difference)*

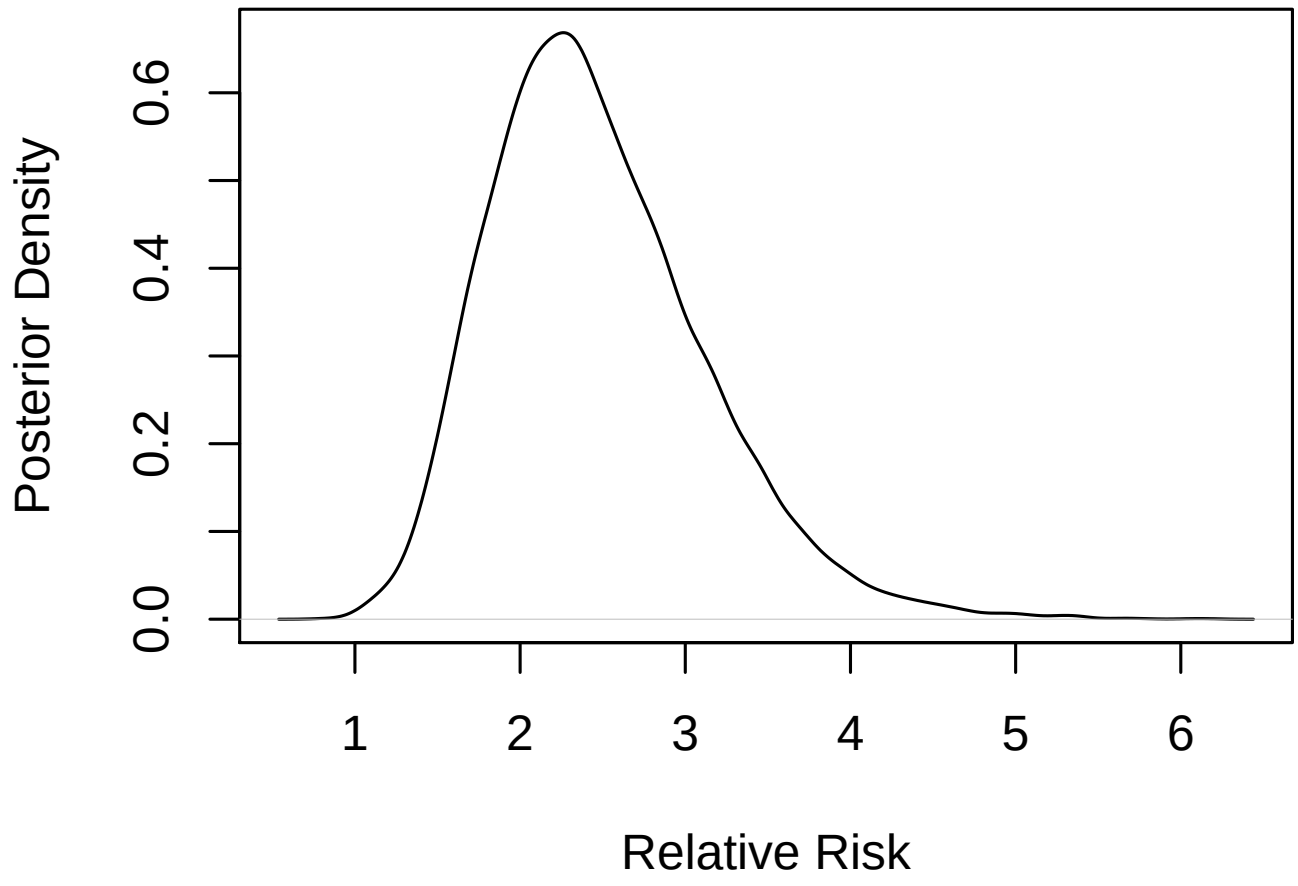



Figure: *Posterior distribution of relative risk (risk ratio)*

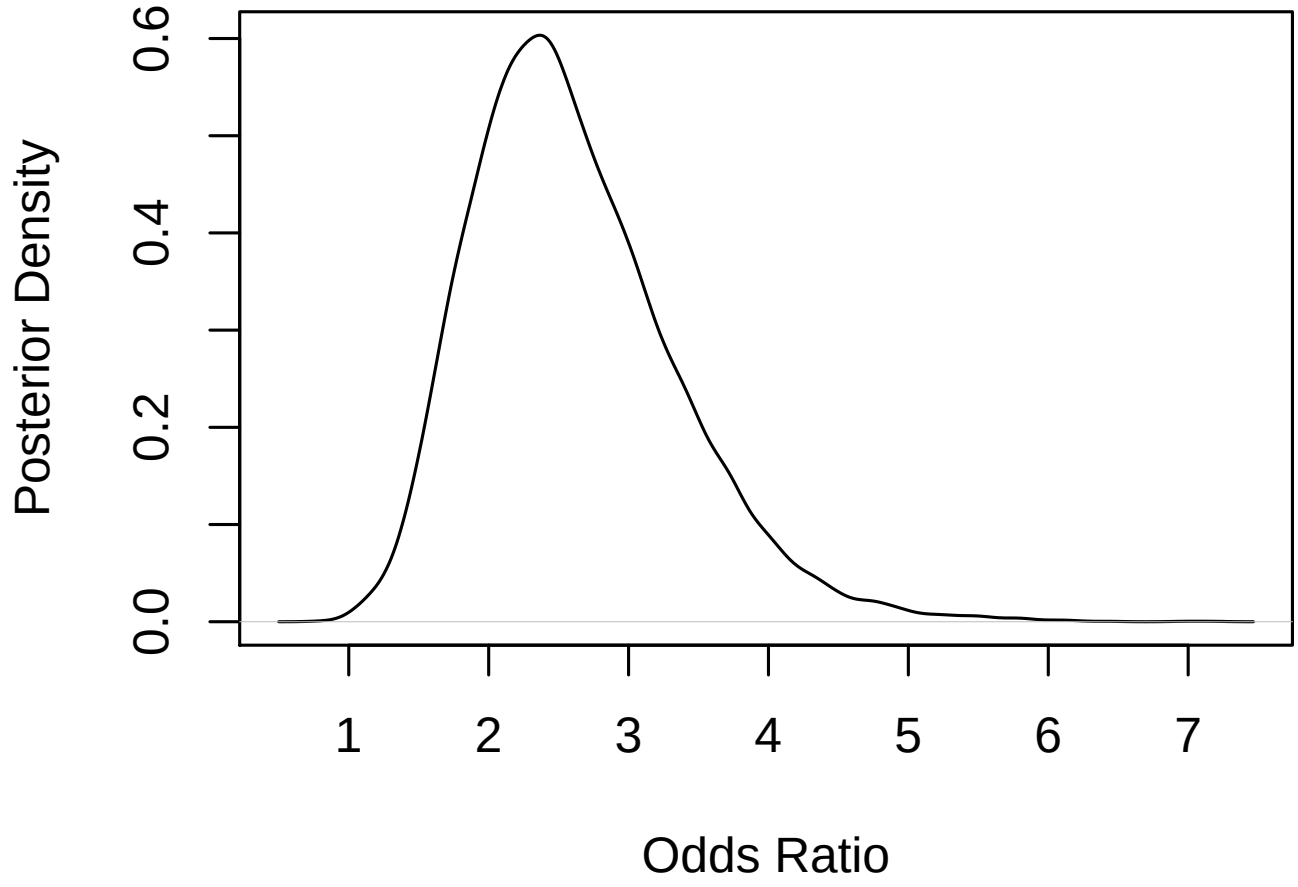



Figure: *Posterior distribution of odds ratio*

# Computer programs (free on the web)

- First Bayes: `http://www.tonyohagan.co.uk/1b/`
  Note that installation is a bit complicated.
- BUGS: `http://www.mrc-bsu.cam.ac.uk/bugs/`
- R: `http://www.r-project.org/`

# Introductory articles

- In praise of Bayes. *Economist* (2000).
  `http://www.cs.berkeley.edu/~murphyk/Bayes/economist.html`
- Hively, W. (1996). The mathematics of making up your mind. *Discover*.
  `http://discovermagazine.com/1996/may/themathematicsof760`
- Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches.
  Pages 319-332 in L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors,
  *What if there were no significance tests?*
  Mahwah, NJ: Lawrence Erlbaum Associates.

# Recommended books I

- Anthony O'Hagan, Bryan R. Luce. *Bayesian Statistics in Health Economics and Outcomes Research*. Bayesian Initiative in Health Economics and Outcomes Research, Centre for Bayesian Statistics in Health Economics
  `http://www.shef.ac.uk/content/1/c6/02/55/92/primer.pdf`

  Free intro booklet; good place to start
- Winkler, R., (2003). *An Introduction to Bayesian Inference and Decision*. Gainesville: Probabilistic Pub.
  `http://www.decisions-books.com/`

# Recommended books II

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D.B. (2003). *Bayesian Data Analysis*, Second Edition. Boca Raton: Chapman & Hall/CRC.

  Likely to become a classic for applied Bayesian data analysis. See review in
  `http://polmeth.wustl.edu/tpm/tpm_v11_n2.pdf`