Week 2: part 1, Basic Bayesian Methods

by Michael Escobar
January 12, 2014

1 Introduction

The first part of this handout contains some theoretical technical issues which were not covered in the main presentation slides. The second part of this handout contains computational issues.

The first part first contains some different technical points. The next section goes into more details on some probability terms. This is followed up with a more detailed explanation of the generalized t-distribution.

See the beginning of the second part for more details on how that part is organized.

2 Some Technical Jargon

Please note the following:

- *Probability measures, distributions, and random variables:* A probability measure is a function on sets of outcomes to a value between between zero and one. Technically, a random variable is a function which assigns a real number to an outcome. One connects a distribution to a probability measure by the cumulative distribution function which is defined by:

  $$P(X \leq a) = F_X(a).$$

  The cumulative distribution function is the "area under the curve". It describes the the probability that X is below an outcome a. Non-technically, consider a random variable like a persons height. The distribution provides how we weight our belief on the different values that a person's height might be. When we see a person's height, then we have a realization of the random variable X. Often we write the random variable as a capital letter and the realization as a small letter. So, X might be the random variable and x is the realization.

- *Density functions and probability function* The cumulative distribution function is sometimes not the easiest way to think about the distribution. Probably it is easier to think about how much weight different values have. The density function or the probability function is a way of defining a distribution which describes how much weight is on different values. For random variable which take on discrete values (for example, the random variable might only have a finite number of outcomes) then the probability function gives the probability of each outcome occurring. The sum of all possible outcomes equals one, because with probability one, one of the outcomes occurs. Later in the chapter, the binomial and Bernoulli distributions are described. These are examples of discrete distributions. For continuous random variables, the density function plays a similar role as the probability function plays for discrete random variables. The density function is like an idealized histogram of the random variable. In fact, the histogram is an estimate of the density function. If one draws a large independent sample from a continuous distribution and makes a histogram of this sample, then the histogram will look approximate the density function. To find the probability of observing a realization of the random variable in some interval, you would find the area under the density function above that interval. The area under the entire density function is one. Also, note, that when we use the term, "the integral of the density function over the set A", we mean the area under the curve over the set A. This integrated area defines the cumulative distribution function. Therefore, the density function is the derivative of the cumulative distribution function for continuous distributions. The normal distribution is an example of a continuous distribution.

- *Parameters, data, and random variables:* In more classical notion of probability and statistics, one has random variables from some distribution. These distributions often have parameters which are considered to be fixed constants. Data are sampled from a distribution and as such are realizations of a random variable. In the Bayesian paradigm, the parameters to the models are themselves random variables from some distribution. Parameters of a model are random when we are uncertain as to their value. Even if the parameter is some physical constant, it can still be described as a probability if we are uncertain as the value. It is a probability because it quantifies the uncertainty of our belief in the value of the physical constant.
- *Parameters and the question of interest:* The purpose of a statistical analysis is to learn about some value of interest. When this is translated into a mathematical model, this value of interest is usually converted into some parameter of a probability distribution.
- *Data, events, and outcomes:* Events and outcomes are things which are observed. When talking about a mathematical model of the situation, we will usually refer to these observations as data. Data are the realization of some random variable.
- *Conditional Distribution* Most distributions that we will use are conditional distributions. They are our belief of the likelihood of certain events happening if we know or assume we know some other events. That is, they express our belief about some value if we know something else. So, the data is a sample of a conditional distribution given knowledge of the unknown parameter. What we are interested in is finding our belief of the value of the parameter given knowledge of the data that we observe. So, we are interested in the conditional distribution of parameter given the data.

3 Some Basic Probability Issues

3.1 Independence

Let X and Y be two random variables and let A and B be two sets of possible outcomes for X and Y. The probability that X is in A and Y is in B can be calculated by the formula: $P(X \in A \& Y \in B) = P(X \in A|Y \in B)P(Y \in B)$. Equivalently, independence can be defined by the property that the probability that X is in A given Y is in B is equal to the probability that X is in A. Mathematically this can be expressed as $P(X \in A|Y \in B) = P(X \in A)$.

Although classical statistics often uses independence when describing several random variables, this is not ideal. To see this, consider the example of the efficacy of an anti-cancer drug. In the Bayesian set up, we do not want the individual subjects to be independent. That is, we actually do want the probability of one subject to change when we know something about another subject. So, if there are 21 subjects in the study, then the probability that the 21st subject showing an improvement changes when we have knowledge on the first 20 subjects. Let us see how this could change by looking at this example in more detail. Let X_i be 1 if the i-th patient shows an improvement under the drug treatment. Let θ be the probability that X_i equals 1 given θ. Let θ have the flat uniform prior. Then without knowledge of the other patients the probability that X_{21} will show an improvement is 0.5. Remember this is the marginal probability that X_i will show an improvement. That[1] is:

$$p(X = 1) = E(P(X = 1|\theta)) = E(E(X|\theta)) = E(\theta) = 1/2$$

Now if we see 5 out of 20 patients showing an improvement, then the probability that $X_{21} = 1$ changes. Let D be the information that there was 5 improved subjects out of 20. Then, by Bayes theorem last week, we showed that the posterior distribution of θ given D is a Beta(6,16) distribution which has expected value (mean) of 6/22 which is approximately 0.27. So, we have:

$$P(X = 1|D) = E(P(X = 1|\theta, D)|D) = E(\theta|D) = 6/22 = .27.$$

[1] If this looks like complete magic, tell me and I can go over the simple rules which makes this work.

So, the probability that X_{21} changes depending on whether we know the results of $X_1, \ldots, X_{20}$. Therefore, X_{21} is not independent of the previous X's.

After looking at that example a little bit, we quickly realize that one does not want the subjects to be independent. The whole purpose of doing the experiment is to learn about the next subject from the previous subjects. If our belief did not change, then the experiment would not have been useful. What we do have is that given θ, then the subjects are conditionally independent. Previous subjects give us information on the value of θ. It is this information about θ which changes. If we knew θ, then the previous subjects would not tell us more about the next subject. Note that using conditional independence breaks the model into pieces. This is an important technique. It used in later chapters to break complex models into smaller, linked submodels. This leads to models which are called conditionally independent hierarchical models. More on these models later.

3.2 Exchangeability

Exchangeability is a weaker condition than independence and it is a useful way develop Bayesian models. Two random variables, say X_1 and X_2 are exchangeable if they have the same outcomes with the same probability outcomes and if the conditional distribution of X_1 given X_2 is equal to the conditional distribution of X_2 given X_1. Notationally this means that $f(X_1) = f(X_2)$ and $f(X_1|X_2) = f(X_2|X_1)$. This means that X_1 are X_2 or interchangeable. A sequence of random variables, $X_1, \ldots, X_n$, is exchangeable if each X_i has the same set of possible outcomes and the joint probabilities of the vector of random variables, $(X_1, \ldots, X_n)$, is the same no matter what the order of the X's. That is, if $f(X_1, \ldots, X_n) = f(X_{(1)}, x_{(2)}, \ldots, X_{(n)})$ where the vector $((X_{(1)}, x_{(2)}, \ldots, X_{(n)})$ is a permutation of the vector $(X_1, \ldots, X_n)$. An infinite sequence of random variables, $(X_1, \ldots, X_n, \ldots)$, is exchangeable if every finite sequence is exchangeable.

NOTE: the rest of this subsection is just for enrichment. I won't be examining you on this paragraph. Having an infinite, exchangeable sequence can be used to define conditional independence and long run frequency. This is expressed by the "representation theorems". For example suppose that we have an infinite exchangeable sequence of random variables $X_1, X_2, \ldots$ where the X_i's take on values of 0 or 1. Then, the following is from Bernardo and Smith (1994)[2]:

> Then, this sequence can be viewed as *if*:
>
> 1. The X_i are judged to be independent, Bernoulli random quantities conditional on a random quantity θ;
> 2. θ is itself assigned a probability measure Q;
> 3. by the strong law of large numbers, $\theta = \lim_{n\to\infty}(\sum_{i=1}^{n} X_i/n)$, so that Q may be interpreted as "beliefs about the limiting relative frequency of 1's".

The general idea is that if a sequence of random variables $(X_1, \ldots, X_n)$ is a random sample from a probability model with some parameter θ, then the X's are conditionally independent given the parameter θ. Also, the sequence of X's are necessarily exchangeable. Also in a general way[3], the converse is true and exchangeability can be used to define the concept of a probability model and a parameter. Let $X_1, X_2, \ldots, X_n, \ldots$ be an infinite exchangeable sequence. Then any finite subsequence $X_1, \ldots, X_n$ can be represented as a random sample from some probability model $f(X|\theta)$ with parameter θ, with a probability distribution on the parameter θ. Also, the parameter θ can be defined as the limit of some function on the sequence of X_i's. So, the simple concept of exchangeability gives rise to the notion of a probability model, a parameter, and how the parameter is related to the limit of a function of the data (much like a "long run" frequency).

[2] This statement is from Bernardo and Smith, 1994, Bayes Theory, New York: Wiley and Sons, page 173. Also, note that this theorem is originally from De Finetti (1930).

[3] Things get a bit complicated since one is considering general spaces and such. So, at this point, let us not try and pin this done too much

3.3 Types of distributions: multivariate

When considering two more more variables, one is interested in the joint distribution, the marginal distributions, and the conditional distributions. Without lost of generality, let us assume that there are two random variables, X and Y.

- *joint distribution* This describes the probability weight for the points (X, Y).
- *conditional distribution* This is the distribution of a random variable if you know some information. If X is the random variable and Y is the know information, then we write: $p(X|Y)$ and say, "the probability of X *given* Y."
- *marginal distribution* This is the distribution of a random variable averaged out over the other variables.
- *predictive distribution* It we have an exchangeable sequence $X_1, \ldots, X_n, X_{n+1}$, then the predictive distribution is the conditional distribution of X_{n+1} given the random variables $X_1, \ldots, X_n$.

To illustrate these distributions, let us assume that the variables X and Y each take on the values $\{1, 2, 3\}$. Also, let the probability of the different, joint values of X and Y be given by the following table:

			Y	
		1	2	3
X	1	.20	.05	.05
	2	.10	.10	.30
	3	.05	.10	.05

So, according to the above table, $P(X = 1, Y = 1) = .20$ and $P(X = 1, Y = 2) = .05$, etc. This defines the joint distribution with probability mass function, $f_{X,Y}(1, 1) = .20$ and $f_{X,Y}(1, 2) = .05$, etc.

To fine the marginal distribution, for example, the probability that $X = 1$, then one needs to sum over all the values of Y where $X = 1$, so $f_X(1) = P(X = 1) = .20 + .05 + .05 = .30$. So, this gives the following marginal distribution for X and for Y:

	$f_Y(\cdot)$	
1	2	3
.35	.25	.40

	$f_X(\cdot)$	
1	2	3
.30	.50	.20

The marginal distribution for X and Y are the row and column sums of the original table.

The conditional distribution is the distribution for one random variable if you know something about the other. So, if we know the value of X then the distribution of Y given the value of X is given as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

So, for the above example table, we can find the conditional distribution of Y given a particular value of X:

		Y	
	1	2	3
$f_{Y\|X}(y\|X = 1)$:	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$
$f_{Y\|X}(y\|X = 2)$:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{3}{5}$
$f_{Y\|X}(y\|X = 3)$:	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Note that each row sums to one since each row is a probability distribution in its own right.

From the conditional distribution and the marginal distribution, one can obtain the joint distribution. As one can see from the equation above which defines the conditional distribution, one can rewrite that equation to be:

$$f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y),$$

which can be written as:

$$f(Y|X)f(X) = f(X,Y).$$

This means that one can generate sample from a joint distribution by first, say, generating a sample from the marginal distribution of X and then sampling Y given the sampled value from X.

The concept of starting with a marginal distribution and then sampling from a cascade of conditional distribution is used in hierarchical models. Hierarchical models include models like multilevel models, growth curves, and random effect models.

3.4 The Predictive Distributions

The predictive distribution is the distribution of a new observation after we have seen some other observations.

If we know that the observations $X_1, \ldots, X_n$ given parameter θ have the likelihood defined by $f(X_1, \ldots, X_n|\theta)$ and θ has a prior $f(\theta)$, then using Bayes theorem, we our knowledge of θ can be expressed by the posterior distribution $f(\theta|X_1, \ldots, X_n)$. Then, to sample from the distribution for a new X_{n+1}, one can first sample from the posterior of θ and then use this sampled value to sample from the conditional distribution $f(X_{n+1}|\theta)$.

Technically, we can define the predictive distribution by the following: If we have an exchangeable sequence $X_1, \ldots, X_n, X_{n+1}$, then the predictive distribution is the conditional distribution of X_{n+1} given the random variables $X_1, \ldots, X_n$.

Consider the following:

- Suppose that $X_1, X_2, \ldots, X_n, X_{n+1}$ are all sampled from an identical distribution F with a common unknown parameter θ. Furthermore, assume that the prior distribution of θ is $H(\theta)$.

- So, for example, we could consider the following way of sampling X_{n+1}. First, one can sample θ from $H(\theta)$ and then sample X_{n+1} from $F(\theta)$.

- However, if we know the sampled values of $X_1, \ldots, X_n$, then the sample rule for X_{n+1} would be different based on this new information. This is because, if we know the sequence $X_1, \ldots, X_n$, then we should not sample θ from the prior $H(\theta)$ but θ should now be sample from the posterior distribution (which we call $H(\theta|X_1, \ldots, X_n)$) based on this new information.

- So, if the sampled values of $X_1, \ldots, X_n$ are known, then one first samples θ from the posterior distribution $H(\theta|X_1, \ldots, X_n)$.

This new distribution of X_{n+1} given the previous sampled values of X is called the predictive distribution.

Consider the following example of a predictive distribution. Let us revisit first example involving sampling some small boxes from a large box. In that example, there were 10 boxes which each contained 1 red ball and 5 blue balls and 10 boxes which contained 3 red balls and 3 blue balls. We called the boxes with 1 red ball a type I box and the boxes with 3 red balls a type III box. We picked one of the small boxes out and sampled a ball from the box and founded out that it was red.

Now, suppose we put the red ball back in that box, shook up the box and drew another ball out of the box. This probability is represented by the predictive distribution.

To figure this out, first consider the posterior distribution of whether the box was a type I or III box. From previous work in the lecture hand out, we know that:

$$\begin{aligned} P(\text{Box Type III }|\text{Red ball}) &= \frac{3}{4} \\ P(\text{Box Type I}|\text{Red ball}) &= \frac{1}{4} \end{aligned}$$

So, the probability of drawing a second red ball is calculated by combining the conditional distribution of the drawing a red ball given the box type with the posterior probability of the different box types. That is:

$$\begin{aligned} P(\text{Drawing a second red ball}|\text{First ball was red}) &= \left(\frac{1}{2}\right)\left(\frac{3}{4}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{4}\right) \\ &= \frac{5}{12} \end{aligned}$$

So, the predictive probability of a future red ball given the first was red is 5/12.

Now, let us consider the clinical trial example. In that case, we found 5 improvements in 20 trials. If we use a uniform distribution (a beta(1,1)) for the prior, then the posterior distribution is a beta(6,16). So, with this information, if we were to run 20 more trials, what is the distributions of the number of improvements expected?

The name for this distribution is a beta-binomial. (If you are interested in the exact form of this distribution, you can google it up if you wish.) Instead of looking it up, one can easily sample from this distribution. Here are the steps to do this:

- Step 1: Generate θ^* from the correct posterior distribution. In this case it is a Beta(6,16).
- Step 2: Generate the new Y from the binomial distribution with $n = 20$ and $p = \theta^*$ where θ^* is the value sampled in the first step.

Note, this is not the same as generating a new Y from a binomial distribution with a fixed value for p. If you use a fixed value of p estmated from the data (like the posterior mean or the MLE), you will not get enough "spread" in the distribution. This is because plugging-in a single estimated value will ignore the uncertainty in the estimate.

For example, figure 1 shows the probability mass function for the beta-binomial compared to the binomial which uses the posterior mean as a plug in estimate for the binomial. In that plot one sees that the beta-binomial is more spread out then the binomial. This is because the beta-binomial incorporates uncertainty in the value of θ.

4 Comments on standardized and generalized t-Distribution

4.1 Introduction

In the lecture 1 slide presentation, it is mentioned that the μ parameter in the normal distribution will have a generalized t-distribution under certain conditions. However, currently, that slide package does not mentioned how to work with this distribution. This handout will discuss some details.

4.2 The basic normal model setup

Here the following model is assumed. Assume that $X_1, X_2, \ldots, X_n$ given μ and τ has a normal distribution (the prior) with mean μ and precision τ. Assume that μ given μ_0, τ, and θ has a normal distribution with mean μ_0 and precision $\theta\tau$. The parameter τ given α and β is assumed to have a gamma distribution with shape parameter α and rate parameter β.

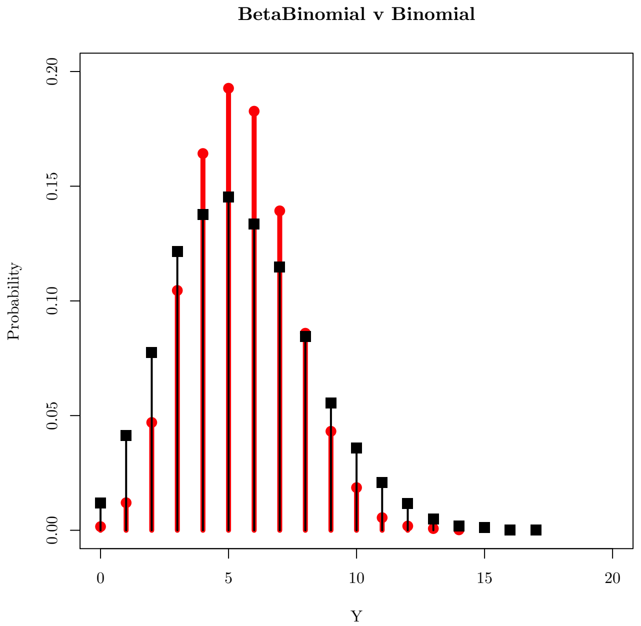

Figure 1: Probability mass function estimated from a sample. The wide red bars are for a binomial distribution with n=20 and p=6/22. The thin black bars are for a beta-binomial formed from a binomial with n=20 and with p sampled from a beta(6,16).

Following the discussion in the lecture notes, given the above model specification, one can get the joint posterior distribution of (μ, τ). This is the product of the posterior distribution of τ and the posterior distribution of μ conditionally knowing τ.

The posterior distribution of τ is a gamma distribution with parameters α' and β'. The parameter $\alpha' = \alpha + n/2$ and the parameter β' is defined as:

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(X - \bar{X})^2 + \frac{n\theta(\bar{X} - \mu_0)^2}{\theta + n}.$$

The posterior distribution of μ given τ in a normal distribution with precision $(\theta + n)\tau$ and mean μ_0' where μ_0' is defined as:

$$\mu_0' = \frac{n\bar{X} + \theta\mu_0}{n + \theta}$$

Note 1: in the above, $\bar{X}$ is the mean of the observed X_i values. Note 2: in some of the below, it might be helpful to consider $\theta + n$ as θ'.

4.3 The marginal posterior distribution of μ

To get the marginal posterior distribution of μ, one needs to integrate out the τ parameter from the joint posterior distribution. This results in a generalized t distribution. As the slides state, the term:

$$(\mu - \mu_0') \Big/ \sqrt{\frac{\beta'}{\theta'\alpha'}}\;.$$

has a regular t-distribution with $2\alpha'$ degrees of freedom.

Note: 1) in the above expression, θ' is substituted for $\theta + n$ as it appears in the lecture note 1 slide package. 2) the parameters α', β', and μ_0' are defined above and are all "fixed" values based on the prior parameters or the observed data.

To work with this distribution it is useful to remember the following:

- One can consider the above equation as a "standardized form" and use this standardized t in a similar way as one would use a standardized normal. That is, it is helpful to look at the regular t distribution and then back transform to get information about the posterior distribution of μ.
- For the standardized normal, the variance of the normal is 1. However, the variance for the standard t is not 1 but depends on the degrees of freedom. For a standard t distribution with ν degrees of freedom, the variance is $\frac{\nu}{\nu-2}$ when $\nu > 2$. (If $\nu \in (1, 2]$, then variance is infinite and if $\nu < 1$ then variance is undefined.)
- From the above two points, one can then calculate the variance for the marginal posterior distribution of μ. Let $T(\nu)$ have a random t distribution with ν degrees of freedom. Also, in the below, assume that the Var function is the variance under the marginal posterior distribution. Then, we have the following:

$$
\begin{aligned}
&\mathrm{Var}(\mu) \\
&= \mathrm{Var}\left[\left(\frac{\mu - \mu_0'}{\sqrt{\beta'/(\theta'\alpha'}}\right)\sqrt{\beta'/(\theta'\alpha')} + \mu_0'\right] \\
&= \mathrm{Var}\left[T(2\alpha')\sqrt{\beta'/(\theta'\alpha')} + \mu_0'\right] \\
&= \left(\frac{\beta'}{\theta'\alpha'}\right)\mathrm{Var}\left[T(2\alpha')\right] \\
&= \left(\frac{\beta'}{\theta'\alpha'}\right)\left(\frac{2\alpha'}{2\alpha' - 2}\right)
\end{aligned}
$$

- If one wants to find quantile functions, then one first can find the quantiles for the standard t distribution with $2\alpha'$ degrees of freedom. Then, one would multiple the value by $\sqrt{\beta'/(\theta'\alpha')}$ and then add μ_0'.

4.4 Examples

To see how to apply these, lets consider an example. Let n equal 4, $\alpha = 2$ (so $\alpha' = 4$), $\beta' = 100$, $\theta = 1$ (so $\theta' = 5$), and $\mu_0' = 50$.

The above leads to the following R commands and outputs:

```
> nbig=100000
> ap=4;bp=100; mu0p=50; thp=5
> tquant=qt(c(.025,.05,.25,.5,.75,.95,.975),2*ap)
> tquant*sqrt(bp/ap/thp)+mu0p
[1] 44.84362 45.84192 48.42047 50.00000 51.57953 54.15808 55.15638
> (bp/ap/thp)*(2*ap/(2*ap-2))
[1] 6.666667
```

Compare this to generating a sample from the joint distribution. That is, generate 100,000 values from the posterior distribution for τ and then use these values to generate values of μ from the posterior given τ. This leads to the following R commands and output:

```
> nbig=100000
> ap=4;bp=100; mu0p=50; thp=5
> t1=rgamma(nbig, ap, bp)
> m1=rnorm(nbig, mu0p, 1/sqrt(thp*t1))
> var(m1)
[1] 6.587882
> quantile(m1, probs=c(.025, .05, .25, .5,.75, .95, .975))
    2.5%       5%      25%      50%      75%      95%    97.5%
44.85468 45.87521 48.43976 49.99962 51.57357 54.11968 55.09158
```

Note that the quantiles are about the same. The variance from the simulation is 6.587 and the variance from direct calculations from the t distribution is 6.667.

Week 2: Part 2, Some Computing notes

5 Introduction

Since we are using the computer to substitute for technical calculations, it is very important to be able to gain some skill in statistical computing. Later in the course, we will be using WinBugs, a freeware computer package, to do the specialized calculations we need for complex Bayesian calculations. However, this package is somewhat limiting, so we need to use another statistical package to perform other calculations. Possible packages include SAS, Splus, R, or Matlab. The package R is a fairly complete and powerful package and it is also freeware. It is also very similar to Splus with basically the same commands. So, most of the general computing commands that I will be demonstrating in this course will be in either R or Splus.

In this handout, I will be showing example of some of that calculations which I have used in the early part of the course.

Here is an outline of the sections in this handout:

Section 5 Introduction. This is this section.

Section 6 Methods of using R for several basic distributions are discussed. There are subsections on the Binomial Distribution, the beta distribution, generating basic graphs for them, and a final subsection on some of the other distributions.

Section 7 How to learn about distributions from a sample of the distribution. Here there is a discussion of the underlying general theory which allows us to use sampled values as a substitute for doing analytical calculations.

Section 8 How to do more some more advanced procedures such as plotting two dimensional plots.

Section 9 Getting the values for the beta and gamma distribution.

Section 10 Looking at the mean and variance of the beta distribution. When justifying the Jeffery's prior, there was a discuss about unequal variances. This section discusses this in some more detail. *Note: this is just for enrichment. Some students ask me about this every year, so this is just for their benefit.*

6 Some ways to use R for calculations on distributions

6.1 The Binomial distribution

The four commands for the binomial distribution are:

- `dbinom(x,size, prob)`; This is the probability (density) function. This function gives the probability that a random variable from a binomial distribution with the given "size" and "prob" will be "x". That is, it calculates $P(X = x|\text{size}, \text{prob})$. In the course notes, "size" corresponds to the parameter "n" and "prob" corresponds to the parameter "θ". In general, for other distributions, this is the density or the probability function. (For example, below, we have the function "dbeta" which is the density function for the beta distribution).

- `pbinom(q,size,prob)`: This (cumulative) distribution function. That is, it calculates the value of $P(X \leq q|\text{size}, \text{prob})$. The value of "size" and "prob" have the same meaning as above (and will have the same meaning in the next two functions). In general, for other distributions, this is the "area under the curve" function.

- `qbinom(p, size, prob)`: This is the quantile function. It is the inverse of the function "pbinom". This function gives the value "a" such that $P(X \leq a|\text{size}, \text{prob}) = p$. That is, it gives the value "a" such that the area under the curve up to the value "a" is equal to "q".
- `rbinom(n,size, prob)`: This command generates random variables. It will generate a vector which has "n" random variable in it.

Here are some comments about using binomial distributions with R:

- The `help` command: If you want to know more about any of these function, use the `help` function. For example, type:

  ```
  help(rbinom)
  ```

  will get you more information about the above four commands. In general, the help function is quite useful when you quickly need to learn some of the details of any of the R functions.
- Example: Suppose one wanted to generate 10 samples from a binomial with size 6 and probability .5. The following shows the screen from the command window from R. Note, R generates the ">" before the commands you type:

  ```
  > rbinom(10, 6,.5)
   [1] 2 4 4 2 5 3 4 3 5 5
  ```

6.2 Beta Distribution

First, an aside: the mean and the variance of a beta distribution are: The mean is $\alpha/(\alpha+\beta)$ *and the variance is* $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.

The four main functions for the beta distribution are similar to the binomial functions. They are:

- `dbeta(x, shape1, shape2)`
- `pbeta(q, shape1, shape2)`
- `qbeta(p, shape1, shape2)`
- `rbeta(n, shape1, shape2)`

These four functions are the density function, the distribution function, the quantile function and the generator function. The parameters α and β which we discussed in class are the parameters "shape1" and "shape2" respectively in the above functions.

Here are some examples where R is used to calculate values for the beta distribution.

- In the class hand out, we had an experiment with 5 successes out of 20 trials and where the prior was uniform (or a Beta(1,1)). So the posterior distribution was Beta(6, 16). How do we find out the summary values in this situation? Well, we know (from above aside) that for a Beta(a,b) the mean is $a/(a+b)$ and the variance is $ab/[(a+b)^2(a+b+1)]$. So, we can use R to do the calculations. Also, to put things on one line, I'll use the `c(...)` function in R. This will just put both the mean and the standard deviation on one line.

  Doing this, I get the following output:

  ```
  > a<-6
  > b<-16
  > c(  a/(a+b),   sqrt(  a*b/((a+b)^2*(a+b+1))))
  [1] 0.27272727 0.09286435
  >
  ```

So, the mean is 0.27 and the standard deviation is 0.09.

We can also get a 95% credible region (equal tail error) using the `qbeta` function:

```
> qbeta(c(.025,.975),6,16)
[1] 0.1128094 0.4716598
>
```

Note that I used the `c(...)` function to ask for two values. So, the credible region is (0.11, 0.47).

- One can also get approximate answers by sampling values from the distribution of interest. These answers will not be exact, but they can be fairly accurate if you take large enough samples. In this course, it is acceptable to give these approximate answers instead of the exact answers obtained through analytically methods. So, to redo the above example, we first generate a large sample from the Beta(6,16) and get the answers of interest. For this you will probably want the functions: `mean`, `sd`, and `quantiles`. Below is the R window for these commands:

```
> x<-rbeta(100000,6,16)
> c(mean(x), sd(x))
[1] 0.27276785 0.09300739
> quantile(x, probs=c(.025,.975))
     2.5%     97.5%
0.1123894 0.4732656
```

  As you can see, the above answers are accurate to the second decimal place. The accuracy is determined by the number of random samples taken.

6.3 Pictures

The program R is fairly good at making high quality pictures. In fact many plots which are in statistical journals and books come from R (or its sister package Splus). The following you might find useful:

- `plot`: This command produces the basic X-Y type of plot.
- `hist`: This produces a histogram
- `lines` and `points`: These commands will produce a plot which will overlay a previously drawn plot using the `plot` or `hist` commands.
- `par`: This command is used to adjust plotting parameters for you graphic plots. For example, the command: `par(mfrow=c(3,3))` will put a 3x3 grid of plots on one page.
- `postscript` and `jpeg`: These commands will redirect the output to a graphic file. (Either a postscript file of a jpeg file). Please note, that you need to shut off the redirection with the command `dev.off()` so R can finish off the file.
- Making density plots. The simplest way is to first define the x values as a sequence of about 100 or 200 points along the range you are plotting. Then, simple use the plot function with the option `type="l"`. The following will plot the density for a Beta(10,10):

```
theta<-seq(.001, .999, length=200)
plot(theta, dbeta(theta,10,10),type="l")
```

  Alternatively, one could plot the histogram of a large sample from this distribution. The more adventuresome might consider plotting a kernel density estimate of this random sample.

- Example: The following commands are used to plot figure 2. It is written to a postscript file which is then inserted into this Latex file for this document. If this document was written in MS Word, I might have sent the output to a jpeg file instead.

  Here is the command:

  ```
  #  This file produces a 3x3 set of plots of different Beta densities
  #

  theta<-seq(.001, .999, length=200)

  postscript(file="c:/mike workstation/bayescourse4/chap1BetaTable.eps",
  width = 7.0, height = 7.0, horizontal = FALSE, onefile = FALSE,
  paper = "special",
  pagecentre=TRUE,
  family = "ComputerModern")

  par(mfrow=c(3,3))

  plot(theta, dbeta(theta,.5,.5),type="l", main="Beta(1/2, 1/2)",
     ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,.5,1),type="l", main="Beta(1/2, 1)",
     ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,.5,10),type="l", main="Beta(1/2, 10)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,1,.5),type="l", main="Beta(1, 1/2)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,1,1),type="l", main="Beta(1,1)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,1,10),type="l", main="Beta(1,10)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,10,.5),type="l", main="Beta(10, 1/2)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,10,1),type="l", main="Beta(10,1)",
      ylab=" ",xlab=" ")
  plot(theta, dbeta(theta,10,10),type="l", main="Beta(10,10)",
      ylab=" ",xlab=" ")

  par(mfrow=c(1,1))
  dev.off()
  ```

6.4 Other distributions

For normal distribution material, you might wish to use the normal distribution, the t-distribution and/or the gamma distribution. To see these commands, you can look at these you can use the R `help` function. Type the following in R:

- `help(dnorm)`
- `help(dt)`

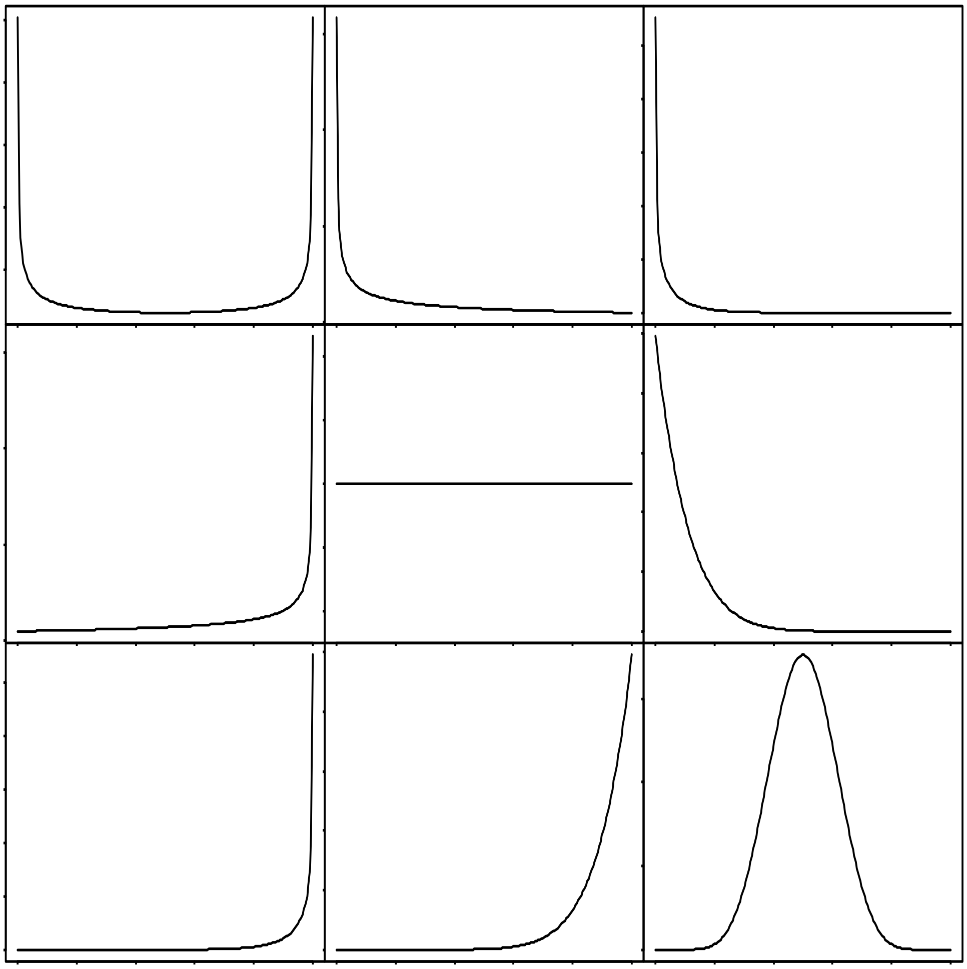

Figure 2: The beta distribution for different values of α and β. The rows (t-b) are for $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$. The columns (l-r) are for $\beta < 1$, $\beta = 1$, $\beta > 1$.

- `help(dgamma)`

Please note that for the gamma distribution, I talk about the parameters α and β. In R these are called the "shape" and "rate" parameter. So, to get 10 random samples from a gamma(2, 20) type `rgamma(10, 2, 20)`.

7 Learning about a Distribution from a Sample

Given a sampled values from a distribution allows one to learn about the distribution. The basic methods that are used here are techniques like averages, histograms and the like. First, I should comment on the fact that it might look like we are abandoning Bayesian methods and turning to frequentist methods at this point. So, this needs to be commented on. Then, we need to consider the type of information that we might be interested in getting from the analysis, and how to obtain this information.

I do admit that the techniques below have a very "frequentist flavour" about them. So, at first glance here, it appears that we first state that Bayesian methods are better than frequentist methods. Then, we sample from the posterior distribution and then to figure out what the posterior distribution is, we use very frequentist methods. To be more "pure", I guess we could put yet another prior. However, as we have seen, if we start with a prior which is has "low information" and we have a very large amount of data, then the usual frequentist procedures are close to the low informative Bayesian procedures. The other thing to note is that there is more to the differences between frequentist and Bayesian methods.

Anyway, in this section, methods for calculating important features of a distribution from a sample of the distribution are discussed. The basic idea is to appeal to the law of large numbers and the central limit theorem. That is, sampled averages are an approximation to the expected value of a random number. This also applies to functions of the expected values. Similarly, one can use the same technique to get the probability of sets. One needs a little bit more machinery to estimates density functions and percentiles, but the same general principles apply. These methods are first discussed for independent samples and then for samples from a Markov chain sampler.

As an aside, please note that these methods parallel the methods used for the numerical technique know as Monte Carlo integration. In Monte Carlo integration, integration is approximated by noting that getting an expected value is a type of integration. So, one can approximate integration by getting computer simulated draws from a random variable or a pseudo-random variable. These values are then averaged in some way to obtain an approximation to the integration.

7.1 Getting the Means and other Moments

Let X be a random variable from a distribution F, and let $X_1, X_2, \ldots, X_n$ be samples from distribution F. Let the $(1/n)\sum_{i=1}^{n} X$ be $\bar{X}$. The Law of Large Numbers states that given certain conditions, then as n goes to infinite,

$$\bar{X} \to \mathrm{E}(X).$$

There are many different "Law of Large Numbers". Each one has different "certain conditions" (and types of convergence). The sufficient conditions are generally fairly weak. The simplest version requires that the samples X_i be independent and that the variance of X be finite. The statement about the path averages converging when the X_i's are sampled from a Markov chain is another Law of Large Number.

Also, there are Central Limit Theorems. Let the variance of X be σ, then the Central limit theorem states that under certain conditions as n goes to infinite,

$$\frac{\bar{x} - \mathrm{E}(X)}{\sqrt{\sigma^2/n}} \xrightarrow{\mathcal{L}} \mathrm{Normal}(0, 1).$$

Again, there are many different types of Central Limit Theorems which depend on the conditions. One of the simplest statements is that the X_i's be independent samples and that the variance, σ^2, be finite. There also exist versions when the X_i's are sampled from a Markov Chain.

So, for our purposes, when the X_i's are sampled from a Markov Chain, we behave as if sufficient conditions are meet and that there at least exist a Law of Large Numbers. Therefore, we can estimate the mean of the distribution F by simply taking the path averages. That is, if X_i is the sampled value at the i-th iteration, then one can use $\bar{X}$ as an approximation to $\mathrm{E}(X)$.

Besides getting an approximation to $\mathrm{E}(X)$, one can use the above method to also find an approximation to $\mathrm{E}[g(X)]$ for some function $g(X)$. Again, under fairly weak conditions, there is a Law of Large Numbers and a Central Limit Theorem which apply and one can use the approximation:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \approx \mathrm{E}[g(X)].$$

The above can be used to find the various moments of the distribution F. This of course includes the variance, skewness, and kurtosis.

Also this method can be used to approximate the cumulative distribution function. The cumulative distribution function, $F(a)$, is the probability that a random X is less than the number a. To do this, define a function $g_a(z)$ to equal one if z is less than a and zero otherwise. Then, $F(a) = \mathrm{E}[g_a(X)]$. So, this can be easily approximated by the above. Also, please note that as a function of a random variable X, $g_a(X)$ is a random variable with a Bernoulli distribution. So, the variance of the average of $g_a(X)$ is $F(a)[1 - F(a)]/n$. So, if one wants to know the value of $F(a)$ to the second decimal place, then one needs to sample about 10,000 values. (To see this, see the statement of the central limit theorem above.)

It takes a little bit more work, but it can be shown that one can estimate the population percentiles (the percentiles of F) by the sample percentiles (the percentiles of the data values $X_1, X_2, \ldots, X_n$. Basically, one can show that their exist Law of Large Numbers and Central Limit Theorems for these statistics.

7.2 Estimating Densities

Some methods:

- Histogram
- Kernel density estimation: Suppose you wish to estimate the value of a density, $f(\cdot)$ at the point x_*. One way is to estimate it by a kernel density estimate defined by the following:

  $$f(x_*) \approx \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i - x_*}{h_n}\right),$$

  where $K(\cdot)$ is the kernel function and h_n defines the window. Examples of the kernel function include the standard normal density. The window controls the amount of smoothing that is done in the estimate and it is usually some function of n. As n gets larger, then h_n gets smaller at some rate. Given certain conditions, (including the kernel function and the rate that the window decreases as a function of n), then the kernel function converges to the density function for all values of x_*.

7.3 Rao-Blackwellization

(ignore this till week 3 and 4)

For samples from the Markov chain, sometimes the estimate can be improved by using the Rao-Blackwell theorem. The best example is in its use to estimate density functions. Assume that in the MCMC, we are sampling parameters θ_1 and θ_2 and we have data vector Y. So, to do the Gibbs sampler, we need to sample from the following densities:

$$f_{\theta_1}(\theta_1|\theta_2, Y)$$
$$f_{\theta_2}(\theta_2|\theta_1, Y)$$

Now suppose you wish to estimate the marginal posterior density of θ_1. One could simply collect the sampled values of θ_1 from the simulation and then do either a histogram or a kernel density estimate of θ_1. However, there is a more efferent way using the sampled values of θ_2.

Here is how to do it. The estimation is on grid of points. Suppose that we have $\theta_2^{(m)}$ which is the m-th sample of θ_2 and there are M total samples. So, lets look at estimating the density at the point where θ_1 equals t_*. Use the following estimate of $f_{\theta_1}(t_*|Y)$:

$$\hat{f}_{\theta_1}(t_*|Y) = \frac{1}{n}\sum_{m=1}^{M} f_{\theta_1}(t_*|\theta_2 = \theta_2^{(m)}, Y).$$

Now consider the heights problem example. If one wants to estimate the posterior distribution of the mean μ, then one could get the sampled value $\mu^{(m)}$ and then get a histogram estimate or a density estimate or one could get the Rao-Blackwellized estimate. Let us look at the Rao-Blackwellized estimate. For this, we would only consider the sampled values $\tau^{(m)}$. Suppose one wanted to get the estimate of the density value where μ equal 66 and where one received the following sampled values of $\tau^{(m)} = (0.1322, 0.0607, 0.0398)$. These value of τ correspond to the following values of τ_0: $(0.645, 1.322, 0.607)$. (See page 45 of the second set of lecture/overhead notes.) Then, one would have:

$$\begin{aligned}
\hat{f}_{\theta_1}(t_*|Y) &= \frac{1}{n}\sum_{m=1}^{M} f_{\theta_1}(t_*|\theta_2 = \theta_2^{(m)}, Y) \\
\hat{f}_{\mu}(66|Y) &= \frac{1}{3}\left(\phi\left(0.645(66-66.80)\right) + \phi\left(1.322(66-66.80)\right) + \phi\left(0.607(66-66.80)\right)\right) \\
&= (-0.5160 - 1.0576 - 0.4856)/3 \\
&= 0.3106,
\end{aligned}$$

where the function ϕ is the density function for the standard normal.

8 Some helpful ways to use R to make calculations

During the lectures during the first weeks, there were several calculations which were performed which you might find a bit tricky. Below, some of these computer calculations are discussed. These include calculating the two dimensional contour plots and finding the parameters of the beta and gamma distributions which have certain moment or percentile properties. This section ends with a an example of a MCMC algorithm for normal models.

8.1 Getting density plots and probability mass plots

Getting a distribution by sampling and then using a kernel density estimate or a histrogram is an important technique in this course. The following code produces figures 3 and 4:

```
xxx=rgamma(10000,2,4)
ii=seq(.001,3,length=100)

postscript(
    file="c:/mike workstation/bayescourse10/LectureNoteSlide/gamHistTrue.eps",
    width = 7.0, height = 7.0, horizontal = FALSE, onefile = FALSE,
paper = "special",
pagecentre=TRUE,
family = "ComputerModern")
hist(xxx,freq=F,nclass=50,main="Histogram with true density",
   xlab="gamma(2,4)")
lines(ii,dgamma(ii,2,4),lwd=2)
dev.off()

postscript(
    file="c:/mike workstation/bayescourse10/LectureNoteSlide/gamHistEst.eps",
    width = 7.0, height = 7.0, horizontal = FALSE, onefile = FALSE,
paper = "special",
pagecentre=TRUE,
family = "ComputerModern")
hist(xxx,freq=F,nclass=50,main="Histogram with estimated density",
   xlab="gamma(2,4)")
lines(density(xxx),lwd=2)
dev.off()
```

Note that one usually plots estimates the a vector of values with the command `plot(density(xxx))`.

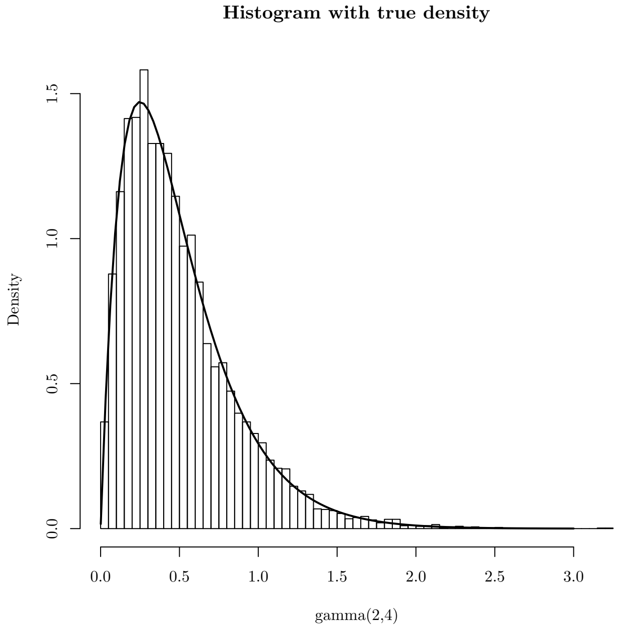

Figure 3: An histogram estimate of the density of a gamma(2,4) compared to the true density function.

In figure 1, the predictive distribution of future trial outcomes was estimated via sampling and plotting. Here is the commands that produced that plot:

```
nbig=20000
bb=rbeta(nbig, 6, 16)
betaval=(table(rbinom(nbig,20,bb))/nbig)
binval=table(rbinom(nbig,20,6/22))/nbig
lbin=length(binval)
```

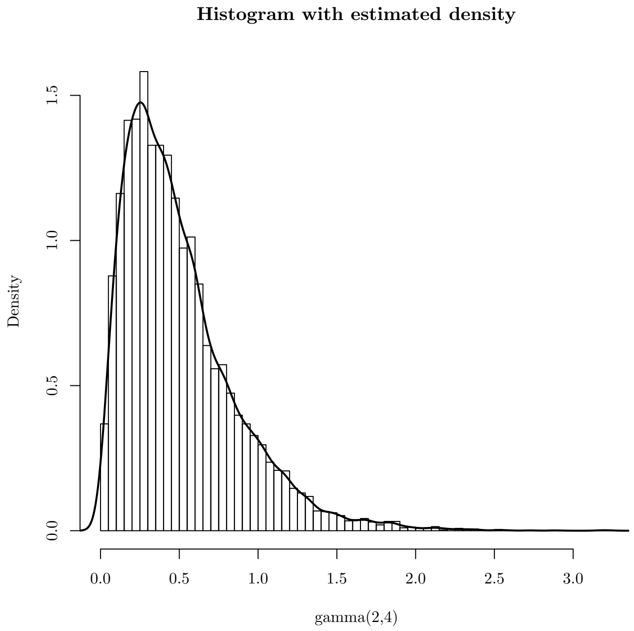

Figure 4: An histogram estimate of the density of a gamma(2,4) compared to the kernel density estimate.

```
lbeta=length(betaval)


postscript(
    file="c:/mike workstation/bayescourse11/LectureNoteSlide/betabin.eps",
    width = 7.0, height = 7.0, horizontal = FALSE, onefile = FALSE,
paper = "special",
pagecentre=TRUE,
family = "ComputerModern")

plot(c(0,20),c(0,.20),type="n", main="BetaBinomial v Binomial",
  xlab="Y", ylab="Probability")
#lines(table(rbinom(nbig,20,6/22))/nbig,col="red",lwd=5)
#lines(table(rbinom(nbig,20,bb))/nbig)
lines(binval, col="red", lwd=5)
points( 0:(lbin-1),binval, pch=16, col="red", cex=1.5)
lines(betaval)
points( 0:(lbeta-1), betaval, pch=15, cex=1.5)
dev.off()
```

8.2 Getting contour plots in R and Splus

There are slightly different commands to produce plots for the joint distribution of two random variables from a sample. Here are some of the important functions:

- `contour(...)`: This is the basic function to produce contour plots. These plots are similar to a topographical map of the probability distribution. For this plot, you need to input the value of the two dimensional function over a grid of X-Y values. So, you cannot simply input the data points, but you need to pre-process the data and estimate a histogram or a two dimensional density plot from the data points. For Splus, one can get the two dimensional histogram from the function `hist2d`. This function also exist in R, but the two dimensional density estimator `kde2d` gives a smoother looking plot. Note: besides the function `contour` you might want to try the functions `image` or `persp` to plot the joint distribution function. These two functions provide "heat plots" or "perspective plots" of the joint distribution.

- `hist2d(...)`: One inputs two dimensional data into this function and the output is a list with three output variables that provide the counts on the three dimensional grid of points on an X-Y grid. If you are using the program R, then you need to first load in the package "gplots". If you are using R in windows, then you simple go to the "packages" tab in the top menu of R and click on the "load packages..." option.

- `kde2d(...)`: This function accepts the two dimensional data as input and outputs a grid of X-Y values and an estimate of the joint probability function at the grid value using a two dimensional kernel density estimator. I believe that this function is only available in the R software package and not in Splus. One needs to load the "MASS" package in R to run this program.

So, for the example in the overhead, we found that the posterior distribution for μ give τ is normal with mean 66.80 and precision $10 * \tau$ and the marginal posterior distribution of τ is gamma with parameters 4 and 71.787. So, in Splus, we can get a contour plot for the joint distribution of μ and τ or of mu and sd where $sd = 1/\sqrt{\tau}$ with the following commands:

```
tau<-rgamma(50000,4,71.787)
mu<-rnorm(50000, mean=66.80, sd=1/sqrt(10*tau))

#  for the joint of mu and sd:
contour(hist2d(list(x=mu,
    y=1/sqrt(tau)),nxbin=30, nybin=30), xlab='mu',
    ylab="sd", labex=0, nlevels=19)
iind<-seq(1,1000)
points (mu[iind],1/sqrt(tau[iind]))

# for the joint between mu and tau:
contour(hist2d(list(x=mu,
    y=tau),nxbin=30, nybin=30), xlab='mu',
    ylab="tau", labex=0, nlevels=19)
iind<-seq(1,1000)
points (mu[iind],tau[iind])
```

The commands in R are similar. First make sure that the package "MASS" is loaded, then type the following commands:

```
tau<-rgamma(50000,4,71.787)
mu<-rnorm(50000, mean=66.80, sd=1/sqrt(10*tau))

#  for the joint of mu and sd:
contour(kde2d(mu,1/sqrt(tau)), xlab="mu",ylab="sd", labex=0, nlevels=19)
iind<-seq(1,200)
```

```
points (mu[iind],1/sqrt(tau[iind]))

# for the joint between mu and tau:
contour(kde2d(mu,tau), xlab="mu",ylab="tau", labex=0, nlevels=19)
iind<-seq(1,200)
points (mu[iind],tau[iind])
```

9 Parameter values for Beta's and Gamma's

The Beta distribution and the Gamma distribution are often used to model one's prior belief for binomial and normal data, respectively. So, one needs to find the parameter values for these distributions. Some strategies one might try are:

1. If ones belief can be expressed by two simple functions of the distribution, then maybe only simple algebra might be needed.
2. Sometimes one can "guess" the parameter value if one can express ones belief as one simple function and then one can quickly find the answer by checking a grid of points.
3. One can formally run an optimization program to find the answer such as the `optim` function in R.

9.1 Using simple functions

Features of these distribution such as the mean, variance, and "effective prior sample size" are simple functions of the parameters of the beta and gamma distributions. For example, if one wishes a beta prior distribution which has mean .3 and effective prior sample size of 2, then $\alpha + \beta = 2$ and $\alpha/(\alpha + \beta) = .3$, so simple algebra gives $\alpha = .6$ and $\beta = 1.4$.

9.2 Guessing

If one of the features of the distribution can be expressed as a simple function, then sometimes it is not too painful to simply guess on a grid of points to get the distribution you want. This is basically doing a bisection algorithm my hand.

For example, suppose one is interested in a beta distribution which has 95% mass between .1 and .9. Also, assume that the belief is symmetric about 0.5. Then, we know that the parameters α and β are equal. Then, we just need to figure out the what to set the α parameter so that 2.5%-tile equals .1. So, perhaps we might try values of α of 1, 5, 10, and 20. Here is the results from R:

```
> a=c(1, 5, 10, 20)
> cbind(a, qbeta(.025,a,a))
      a
[1,]  1 0.0250000
[2,]  5 0.2120085
[3,] 10 0.2886432
[4,] 20 0.3478022
```

In the above, `cbind` is used to put the value of `a` in a column along side the value of `qbeta(.025,a,a)`.

From the above, we see that the value of α that we want is between 1 and 5. So, we can use the `seq` command to get all the α values between 1 and 5 in increments of .2. This gives the following result:

```
> a=seq(1,5,by=.2)
> cbind(a, qbeta(.025,a,a))
        a
```

```
 [1,] 1.0 0.02500000
 [2,] 1.2 0.03910378
 [3,] 1.4 0.05364196
 [4,] 1.6 0.06788860
 [5,] 1.8 0.08148970
 [6,] 2.0 0.09429932
 [7,] 2.2 0.10628153
 [8,] 2.4 0.11745590
 [9,] 2.6 0.12786809
[10,] 2.8 0.13757425
[11,] 3.0 0.14663280
[12,] 3.2 0.15510038
[13,] 3.4 0.16302990
[14,] 3.6 0.17046978
[15,] 3.8 0.17746387
[16,] 4.0 0.18405157
[17,] 4.2 0.19026820
[18,] 4.4 0.19614537
[19,] 4.6 0.20171136
[20,] 4.8 0.20699150
[21,] 5.0 0.21200851
>
```

In the above, we see that the value we want is between 2.0 and 2.2. We might stop here and use, say, a beta(2,2). This would give use a 95% interval between the value (.094, .906) which is close to the interval (.1, .9). If one wanted to go further, one could look at a series of points between 2 and 2.2.

9.3 Finding parameters of a Beta distribution using the optim function

A more complicated way of finding the parameters of the beta distribution is to use a formal R program to find the answer.

In the handout, I found the values of a beta distribution which corresponded to a mean of .30 and a 97.5 percentile of .70. There is actually a fairly easy to obtain the values of the beta distribution with a simple call the the `optim` function in R[4]. The `optim` function will find the minimum value of function. (For the details, type `help(optim)` in R and it will pop up the help file for the function.) *(Also, note that this is a general optimization program. Sometimes it does a bad job and other times it will fail altogether. Be careful with blackboxes!)*

This can be demonstrated in the following output from R:

```
>
> crit1=function(x,y) abs(0.30-x/(x+y) )
> crit2=function(x,y) abs(qbeta(.975,x,y) -0.7)
> v=optim(c(.5,.5), function(x) crit1(x[1],x[2])+crit2(x[1],x[2]))$par
>
> a=v[1]
> b=v[2]
> c(a,b)
[1] 1.638575 3.823341
> c(crit1(a,b),crit2(a,b))
[1] 2.096173e-09 4.365976e-09
```

[4] I'm grateful to a student, Apostolos Dimitromanolakis, for showing me this trick.

```
>
> c( a/(a+b), qbeta(.975,a,b))
[1] 0.3 0.7
```

The first two commands define the criteria functions. The parameters x and y will be used for the parameters of the beta distribution. So, the first function `crit1` is minimized and equal to zero when the mean, $x/(x + y)$, is equal to 0.30. The second function `crit2` is minimized and equal to zero when the 97.5%-tile is equal to 0.7.

The `optim` command needs two basic parameters (and could have more–see the help file for more information). The first is a vector of starting values for the function to be optimized. The second command is a function. In the above output, the expression:

```
function(x) crit1(x[1],x[2])+crit2(x[1],x[2])
```

is used to define the function inside the call to `optim`. The function we are optimizing here is

```
crit1(x[1],x[2])+crit2(x[1],x[2])
```

The output to the `optim` function is a list, and in the third line above, the `par` values of the output are assigned to the variable `v`.

The rest of the command are to print the values and to check the results.

So, from the above R output, we see that a beta$(1.64, 3.82)$ will approximately have the values that we want.

9.4 Finding Parameters of a Gamma distribution using the optim function

The gamma distribution is a two parameter distribution. Given two restrictions on the gamma distribution, then one would expect that one could unique define the parameters of the gamma distribution. As with the beta distribution, one can use the `optim` command in R to solve this problem.

In the handout, the parameters of the gamma distribution is found when the quartiles of the distribution were specified. Here, a method is demonstrated which finds the parameters of the gamma distribution when the quartiles are specified. In this case, we want to find the values of the gamma parameters α and β such that the interquartiles (the 25%-tile and the 75%-tile) have the values 0.025 and 160 respectively.

```
>
> crit1=function(x,y) abs(0.025-qgamma(.025,x,y) )
> crit2=function(x,y) abs(160 - qgamma(.975,x,y) )
> v=optim(c(1, 1), function(x) crit1(x[1],x[2])+crit2(x[1],x[2]))$par
There were 14 warnings (use warnings() to see them)
>
> a=v[1]
> b=v[2]
> c(a,b)
[1] 0.50934769 0.01585891
> c(crit1(a,b),crit2(a,b))
[1] 0.010637979 0.001112534
>
> qgamma(c(.025, .975), a,b)
[1]   0.03563798 159.99888747
>
```

So, in the above output, we see that a gamma$(.51, .016)$ will be a gamma distribution which has approximately the values that we want.

10 Unequal variance of the beta distribution

**** This section is purely for enrichment *****

In the lecture notes, I stated that for the binomial model, when one moves from a success rate of .95 to a .99 it seems harder then going from .50 to .54. I stated that this is related to the variance of the beta distribution. In this section, this is explained a little more clearly.

Basically, the mean of a beta distribution is related to the variance. Note that for a beta(α, β) the mean is $\alpha/(\alpha+\beta)$ and the variance is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1)$. Let us reparametrize the model by letting $\mu = \alpha/(\alpha+\beta)$ and $M = \alpha+\beta$. Then, the mean is μ and the variance is $\mu(1-\mu)/(M+1)$. So, with this reparametrization, we can clearly see that the variance is a function of the mean.

In the notes, it is claimed that transforming θ to $\sin^{-1}(\sqrt{\theta})$ results in a random variable where the variance is a constant with respect to the mean. To show this, one can do this by a one term Taylor expansion and some calculus. However, let us show this via a simulation. The simulation does not a proof per se, but it does show that we are in the right direction. The below code does the simulation for the figure 5

```
ApB = 200   #### this is (a+b+1) parameter
nx  = 100  ### number of mu’s I will look at...
nn = 50000   ### number of random theta’s sampled for the simulation
####

mux = seq(.001, .999,length=nx)
varVal= rep(0, nx)
varTVal=rep(0,nx)
meanTVal=rep(0,nx)
for(i in 1:nx){
  aa= mux[i] *ApB
  bb= ApB-aa
  theta=rbeta(nn,aa,bb)
  varVal[i] = var(theta)
  Ttheta= asin(sqrt(theta))
  varTVal[i] = var(Ttheta)
  meanTVal[i] =mean(Ttheta)
  }

postscript(
    file="c:/mike workstation/bayescourse11/LectureNoteSlide/meanvar.eps",
    width = 4, height = 7.0, horizontal = FALSE, onefile = FALSE,
paper = "special",
pagecentre=TRUE,
family = "ComputerModern")

par(mfcol=c(2,1))
plot(mux, varVal, type="b", main="plot of mean versus variance", sub="Untransformed", xlab="E(theta)
#junk=locator(1)

plot(meanTVal, varTVal, type="b", main="plot of mean versus variance", sub="Transformed", xlab="E(h(
```

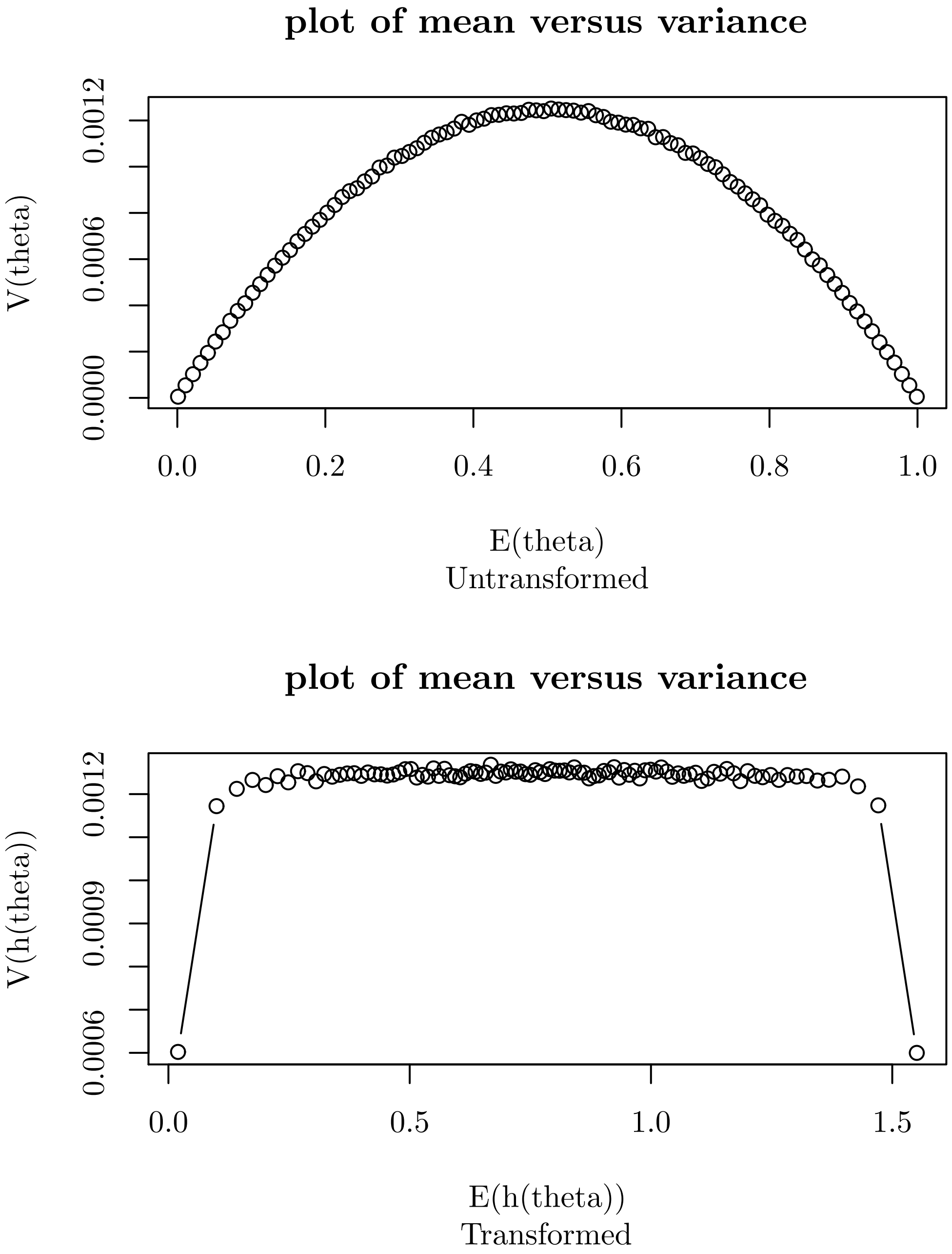

Figure 5: These two plots look at the relationship between the mean and variance of different beta distributions with the same “effective sample size”. The top plot is the untransformed θ and therefore these are samples from the beta distribution. The bottom plot is for the mean and variance of the transformed random variable: $\sin^{-1}(\sqrt{\theta})$. Note that for the untransformed, the relationship is a quadratic while for the transformed variable the relationship is a a line (except for the extreme points).

```
dev.off()
```