Lectures 7, 9, 11 and 12:
Introduction to Bayesian Statistics models

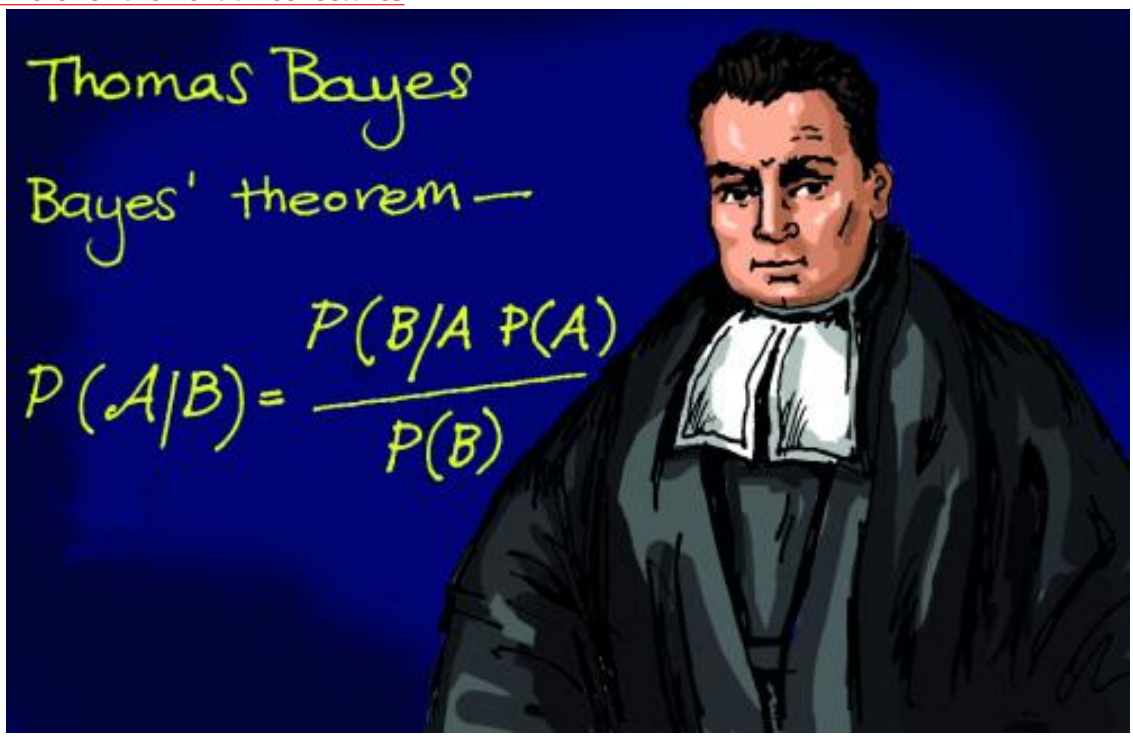
Data Science 2
CS 109b, Stat 121b, AC 209b, E-109b

Mark Glickman Pavlos Protopapas Chris Tanner

Bayesian statistics

- Foundations and philosophy of Bayesian statistics
- Recipe to performing a Bayesian analysis
- Advantages and disadvantages to the Bayesian paradigm
- Modern approach to Bayesian statistics
- Example analyses: Hierarchical Models

Your hero for the next three lectures



Are you a Bayesian?

In the following situations, what would you conclude about the (future) probability of success?

1. A music expert claims he can identify whether a piece of music was written by Mozart versus Haydn.

In 10 trials, the expert identifies the composer correctly 10 times.

2. A drunken friend claims he can predict the result of coin flips.

In 10 trials, he answers correctly 10 times.

3. A coffee connoisseur claims to be able to tell the difference between whether cream is poured in coffee first, or vice versa.

In 10 trials, the person answers correctly 10 times.

If you drew different conclusions in all three situations, then you are likely already Bayesian.

- One key feature of Bayesian statistics is that it is a "learning paradigm." You cannot (and should not) ignore knowledge you currently have in summarizing conclusions based on data.
- In the three example situations, we may have had strong prior beliefs that influenced our conclusions about the data. In classical settings, it is not straightforward to incorporate such beliefs.
- Bayesian statistics provides a formal way to incorporate prior information into an analysis.

Question

What is the underlying difference between classical and Bayesian statistics?

Answer

The definition of probability.

Example

Flip a fair coin; the probability the coin lands heads is $\frac{1}{2}$.

What does "the probability is $\frac{1}{2}$" mean?

Classical interpretation

On repeated flips of the coin, the coin lands heads $\frac{1}{2}$ of the time in the long run.

Frequency definition of probability Probability of an event is its **long-run frequency** of occurrence.

This definition is useful for describing the likelihood of potentially observable data.

Bayesian interpretation

There is an equal probability of heads and tails due to the symmetry of the coin, so that $\frac{1}{2}$ is an assessment of my belief that the coin will land heads.

Subjective definition of probability Probability of an event is one's **degree of belief** that the event will occur (1 means that it is certain to occur, and 0 means that it is certain not to occur).

This definition is useful for quantifying beliefs about non-data events (but also data events).

Classical statistics, which is founded on the frequency definition of probability, is usually called Frequentist statistics. Much of the theory developed by Jerzy Neyman in early 1900s.

Implications to a theory of statistical inference

- Statistics founded on the Frequentist definition of probability can only assign probabilities to future data or potentially observable quantities.
- Statistics founded on the subjective definition of probability can consider probabilities of values of unknown parameters (as well as values of potentially observable quantities).

Variation of coin flipping

Suppose yesterday, at 5:00pm, standing at home, I flipped a fair coin and recorded whether it landed heads.

What is the probability it landed heads?

Answer

If you are a Bayesian, the answer is unequivocally $\frac{1}{2}$.

If you are a Frequentist, you are in a bit of a bind because I am not asking about a long-run frequency – I am asking about what happened specifically at 5:00pm yesterday.

A Frequentist would need to say either

- that the probability is either 0 or 1, but would not know which is correct, or
- need to resort to frequencies of events across different universes within a multiverse!

Because the goal of statistical inference is to infer unknown quantities from data, Frequentists often find themselves in awkward philosophical quandaries.

Example Confidence intervals

Suppose $y_1, \ldots, y_n$ are a random sample of n values from $\text{N}(\mu, \sigma^2)$.

Formula for a 95% confidence interval is

$$(\bar{y} - t^*_{n-1}\frac{s}{\sqrt{n}},\ \bar{y} + t^*_{n-1}\frac{s}{\sqrt{n}})$$

where $\bar{y}$ is the sample mean, s is the sample standard deviation, and t^*_{n-1} is a critical value from the t-distribution on $n-1$ degrees of freedom.

Before observing data, 95% of the time these endpoints will contain μ, the value we are trying to infer.

- Notice that the Frequentist definition applies to characterizing the **procedure**. That is, if we apply the procedure for determining 95% confidence intervals according to the formula, we will be correct 95% of the time.
- Once data are observed, the coverage probability is either 0 or 1.

Here is a dialog I once heard between a Frequentist and his employer where the Frequentist tried to explain confidence intervals.

A Conversation between a Frequentist Consultant and his insightful Employer

Employer: So, do you have the results of the drug trial analyses?

Frequentist: Yes, I do.

Employer: Let's hear it.

Frequentist: Okay. I computed a 95% confidence interval for the average decrease in diastolic blood pressure from taking the medication, and I got $(10.18,\ 14.92)$ mm Hg.

Employer: Great! What does that mean?

Frequentist: It means we should be 95% "confident" that the true mean decrease is between 10.18 and 14.92.

Employer: Oh, you mean that there is a 95% probability that the mean decrease is between 10.18 and 14.92 mm Hg. Got it!

Frequentist: Well, not exactly.

Employer: Huh?

Frequentist: It means that if we were to compute "95% confidence intervals" from many data samples, about 95% of them would contain the true mean decrease in blood pressure.

Employer: Um... But what about **this** data set?

Frequentist: I don't know.

Employer: What do you mean you don't know?

Frequentist: All I know is that in 95% of the analyses I perform, my 95% confidence intervals contain the parameter I'm estimating.

Employer: But I've hired you to draw conclusions from this data set!

Frequentist: Well, you can hope that your data set is "typical."

Employer: What are you talking about? What does that have to do with 95%?

Frequentist: If your data set was one of the 95% "typical" data sets you would observe, then the confidence interval would contain the mean decrease in blood pressure.

Employer: But what does that tell me about the confidence interval you reported to me?

Frequentist: Nothing.

Employer: Nothing?

Frequentist: Can I get my consulting fee? I have to go.

Employer: No way! You're FIRED!!!

Bayesians compute and report intervals as well, but they have different meaning because they are based on subjective probability.

- For example, a Bayesian can compute an interval in which there is 95% probability (as a degree of belief) that μ is in the interval.
- Note that in this case the endpoints of the interval are values computed from data.

Here's a conversation between the same employer, but with a Bayesian.

A Conversation between a Bayesian Consultant and his insightful Employer

Employer: So, do you have the results of the drug trial analyses?

Bayesian: Yes, I do.

Employer: Let's hear it.

Bayesian: Okay. I computed a 95% credible interval for the average decrease in diastolic blood pressure from taking the medication, and I got $(10.04,\ 14.83)$ mm Hg.

Employer: Great! What does that mean?

Bayesian: It means we should be 95% certain that the true mean decrease is between 10.04 and 14.83 mm Hg.

Employer: Oh, you mean that there is a 95% probability that the mean decrease is between 10.04 and 14.83. Got it!

Bayesian: Exactly!

THE END!

A p-value example: (Lindley and Philips, 1976)

A coin has a probability θ of landing heads. Want to test whether the coin is fair (versus the alternative that the coin lands heads with greater probability), that is, whether $\theta = \frac{1}{2}$.

$\mathbf{H}_o : \theta = \frac{1}{2}$

$\mathbf{H}_a : \theta > \frac{1}{2}$

Will test at the $\alpha = 0.05$ significance level.

The coin is flipped 12 times independently, and we observe 9 heads and 3 tails.

H, T, H, H, H, H, H, H, T, H, H, T

Want to compute a p-value and compare to 0.05.

- If $p \leq 0.05$, then we reject $\mathbf{H}_o$ in favor of $\mathbf{H}_a$.
- If $p > 0.05$, then we do not have enough evidence to reject $\mathbf{H}_o$.

Out of 12 independent coin tosses, let X be the number times the coin lands heads. So X follows a Binomial distribution.

$$\begin{aligned} p\text{-value} &= \mathrm{P}\left(X \geq 9 \,\middle|\, \theta = \frac{1}{2}\right) \\ &= \sum_{x=9}^{12} \binom{12}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{12-x} = \boxed{0.075}. \end{aligned}$$

Because p-value > 0.05, we cannot reject $\mathbf{H}_o$ at the 0.05 significance level.

But there is something I didn't tell you...

The way in which I obtained 9 heads and 3 tails was that I flipped the coins until the third tail appeared.

H, T, H, H, H, H, H, H, T, H, H, T

Let Y be the number of times I flip the coin until the third tail appears. Then Y has a negative-binomial distribution with

$$\mathrm{P}(Y = y \mid \theta) = \binom{y-1}{2} \theta^{y-3} (1-\theta)^3.$$

The p-value assuming $\theta = \frac{1}{2}$ under $\mathbf{H}_o$ is

$$\begin{aligned} p\text{-value} &= \mathrm{P}\left(Y \geq 12 \,\middle|\, \theta = \frac{1}{2}\right) \\ &= \sum_{y=12}^{\infty} \binom{y-1}{2} \left(\frac{1}{2}\right)^{y-3} \left(1 - \frac{1}{2}\right)^{3} = \boxed{0.0325}. \end{aligned}$$

With this p-value, we can reject $\mathbf{H}_o$.

What just happened?

- In each case, even though the data and coin flip model were identical, we obtained different conclusions.
- The difference came about in what was considered "as or more extreme" than the observed data.
- Thus, in the classical approach, because the conclusions of a hypothesis test depend on data that we do not observe, strange results can occasionally occur.

  This does not happen in Bayesian statistical inference.

Simple but practical example: Back to the coffee connoisseur

- Suppose p is the probability the coffee connoisseur correctly guesses whether cream is poured in coffee first, or vice versa.
- We observed 10 correct guesses out of 10 tries.
- The classical point estimate of p is the sample proportion, that is $\hat{p} = 10/10 = 1$ (with a standard error of 0).

This means that our best guess is that the coffee connoisseur is 100% accurate.

How can that be?

Bayes theorem (or Bayes Rule)

Before introducing the Bayesian approach to statistics, it is worthwhile reviewing Bayes theorem for discrete events.

Suppose A and B are two events where A temporally or causally comes before B.

Example A is the event that you attend this lecture or watch the video, and B is the event that you obtain a perfect score on your Bayesian homework.

The use of Bayes Rule is to determine the probability of a "past" or "cause" event, given a current event has happened.

More formally, we use Bayes Rule to obtain $\mathrm{P}(A|B)$. This is a "detection" probability – given what we now know (B), what is the probability of a suspected event in the past (A)?

This is given by Bayes Rule:

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(A^c)\Pr(B|A^c)}$$

Usually $\Pr(A)$, $\Pr(B|A)$ and $\Pr(B|A^c)$ can be determined easily.

Example Two coins, one fair and one double-headed, are placed in a box. One coin is chosen at random, and flipped. If it lands heads, what is the probability it is the double-headed coin?

Answer Let

$$\begin{aligned} A &= \text{The double-headed coin is selected} \\ B &= \text{The coin selected lands heads} \end{aligned}$$

Can write down

$$\begin{aligned} \Pr(A) &= 1/2 \\ \Pr(B|A) &= 1 \\ \Pr(B|A^c) &= 1/2 \end{aligned}$$

So, by Bayes Rule,

$$\begin{aligned} \Pr(A|B) &= \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(A^c)\Pr(B|A^c)} \\ &= \frac{(1/2)(1)}{(1/2)(1) + (1/2)(1/2)} = \frac{2}{3} \end{aligned}$$

The probability that the double-headed coin was selected (the past event) given that the coin landed heads (the current event) is $\frac{2}{3}$.

Application of Bayes Rule: Naive Bayes Spam Filter

Goal Want to develop a way to assess the probability a new incoming e-mail is spam.

Assumptions

- You have a large collection of e-mails to develop your method. From these e-mails, you can easily characterize as "spam" versus "not spam."
- You have a program that can search for occurrences of words in your e-mails.

Suppose we identify 10 word phrases common in spam. For example,

- earn money, click to remove, visit our website, stay in shape, act now, free membership

Let W_j be the event that the j-th phrase appears in the e-mail. Let W_j^c (complement of W_j) denote the event that the j-th phrase is *not* in the e-mail.

Assume that an e-mail contains the 1st and 10th words, but none of the others. Want to compute or estimate

$$\Pr(\text{spam} \mid W_1, W_2^c, \ldots, W_9^c, W_{10}).$$

By Bayes Rule,

$$\Pr(\text{spam} \mid W_1, W_2^c, \ldots, W_9^c, W_{10}) =$$

$$\frac{\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}) \Pr(\text{spam})}{\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}) \Pr(\text{spam}) + \Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}^c) \Pr(\text{spam}^c)}.$$

We can empirically estimate $\Pr(\text{spam})$ from the sample proportion of spam you receive.

Also, $\Pr(\text{spam}^c) = 1 - \Pr(\text{spam})$.

Estimating

$$\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam})$$

and

$$\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}^c)$$

can also in principle be accomplished empirically by finding the sample proportion of e-mails in your development sample with the 1st and 10th phrase in the note, but none of the others.

Problem May be very few e-mails with exactly the 1st and 10th phrase in the note.

This problem is even worse if we considered, say, 100 spam phrases.

Possible solution: Naive Bayes

Use the following approximation:

$$\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}) \approx$$
$$\Pr(W_1|\text{spam}) \Pr(W_2^c|\text{spam}) \cdots \Pr(W_9^c|\text{spam}) \Pr(W_{10}|\text{spam})$$

$$\Pr(W_1, W_2^c, \ldots, W_9^c, W_{10} \mid \text{spam}^c) \approx$$
$$\Pr(W_1|\text{spam}^c) \Pr(W_2^c|\text{spam}^c) \cdots \Pr(W_9^c|\text{spam}^c) \Pr(W_{10}|\text{spam}^c)$$

This approximation assumes that the 10 phrase occurrences are *conditionally independent* given the spam status of each e-mail.

This is a huge simplification because it ignores the possibility that some spam phrases tend to travel together.

But this approach has the following advantages:

- Estimates of $\Pr(W_j|\text{spam})$ and $\Pr(W_j|\text{spam}^c)$ are likely to be accurate.
- Easy to compute.
- If the occurrences of pairs, triples, etc., of spam phrases are not very highly dependent, assuming conditional independence can give accurate estimates of combinations of phrases that are not observed in the development sample of e-mails.

See `http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf` to see how Naive Bayes can outperform state-of-the-art learner.

Can implement Naive Bayes using functions in the `sklearn.naive_bayes` module.

From Bayes Rule to Bayesian statistics

When we set up probability models for data, we specify

- Unknown model parameters
- How data are generated based on the model parameters

Can think of the model parameters coming first, and data coming second (based on the values of the model parameters).

Bayes theorem provides the tool to compute the probability of model parameter values based on the observed data.

Key idea In Bayesian statistics, all quantities (data, unknown parameters) have probability distributions to describe uncertainty.

- Parameters Uncertainty in model parameters are described through probability distributions $p(\theta)$. For instance, we may believe *a priori* that the mean adult male weight is uniformly distributed from 160 lbs to 200 lbs.
- Data Observations obtained from a data-generating mechanism are assumed to follow a probability distribution $p(y|\theta)$. For example, we may assume adult male weights are normally distributed around an unknown mean.

Note that assuming a probability distribution on a parameter does not mean that the parameter is "random." It is actually a fixed value, but our uncertainty about it is represented in the form of a probability distribution.

Summary of the Bayesian approach

A typical Bayesian analysis can be outlined in the following steps.

1. Formulate a probability model for the data.
2. Decide on a prior distribution for the unknown model parameters.
3. Observe the data, and construct the likelihood function based on the data.
4. Determine the posterior distribution.
5. Summarize important features of the posterior distribution, or calculate quantities of interest based on the posterior distribution.

Keeping your eye on the ball

The main goal of Bayesian inference is to obtain the **posterior distribution** of the unknown parameters. This is the probability distribution that describes the state of knowledge about the parameters once the data have been observed.

After the posterior distribution has been determined, just about any inferential question can be answered.

Simple example to illustrate steps in Bayesian analysis

Suppose we are examining patterns of 30-day mortality of heart attack patients at a local hospital.

- Among 5 randomly selected heart attack patients, 1 died and 4 survived beyond 30 days.
- Let θ be the probability a patient dies within 30 days after admission for a heart attack.
- We want to make inferences about θ from the data.

First step: Formulate the probability model

For $i = 1, \ldots, 5$, let

$$Y_i = \begin{cases} 0 & \text{if the } i\text{-th patient survives 30 days} \\ 1 & \text{if the } i\text{-th patient dies within 30 days} \end{cases}$$

Probability model for the Y_i:

$$\Pr(Y_i = y \mid \theta) = \begin{cases} \theta & \text{for } y = 1 \\ 1 - \theta & \text{for } y = 0 \end{cases}$$

More compactly,

$$\mathrm{P}(Y_i = y \mid \theta) = p(y|\theta) = \theta^y (1-\theta)^{1-y},$$

for $y = 0, 1$. This is a Bernoulli model for the data.

Second step: Choosing a prior distribution

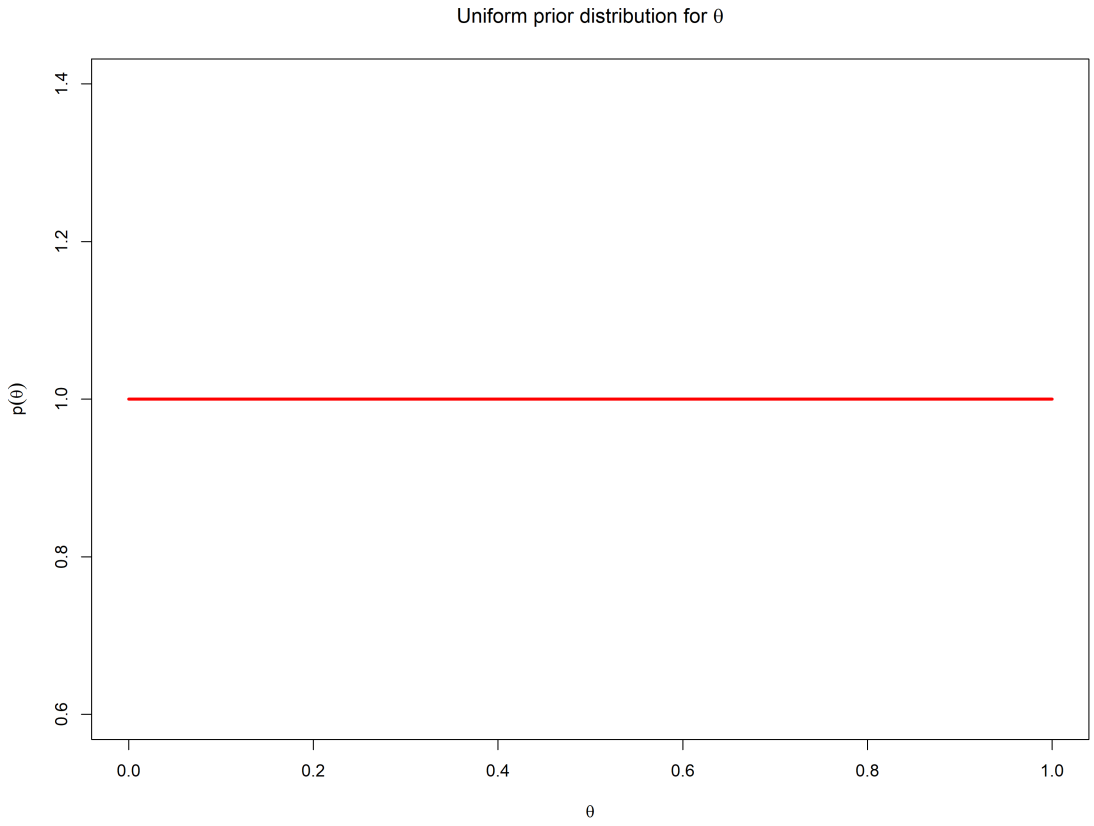
The prior distribution of the unknown parameter(s) represents the state of knowledge prior to observing the data.

For this setting, let us assume that all values of θ between 0 and 1 are equally believable.

Formally, this translates to

$$p(\theta) = 1$$

for $0 \leq \theta \leq 1$.



Third step: Construct the likelihood function

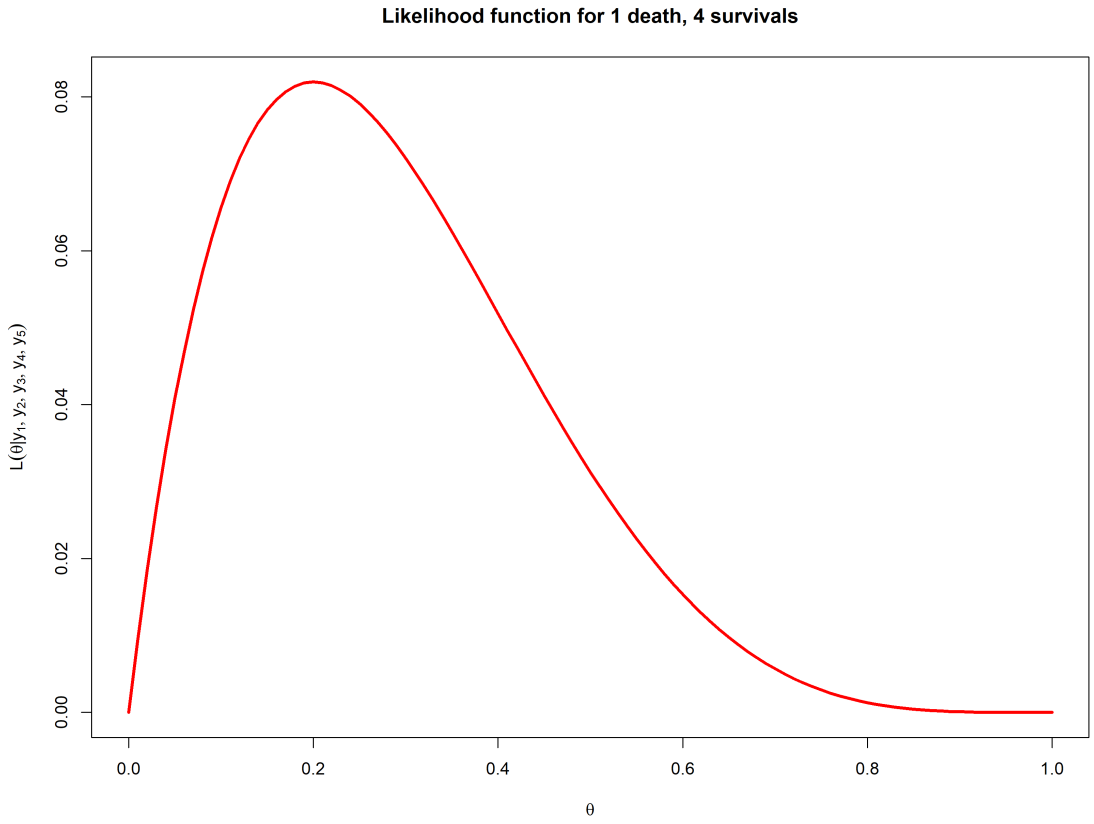
The likelihood function is the probability of the data (probability density, for continuous outcomes) conditional on the parameter(s), viewed as a function of the parameter(s).

$$p(y_1, y_2, y_3, y_4, y_5 \mid \theta) = \prod_{i=1}^{5} \theta^{y_i}(1-\theta)^{1-y_i} = \theta(1-\theta)^4.$$

Thus, the likelihood function is

$$L(\theta|\boldsymbol{y}) = \theta(1-\theta)^4,$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_5)$ are the five binary outcomes.



Fourth step: Determine the posterior distribution

The posterior distribution is the probability distribution of the unknown parameter(s) conditional on the observed data.

This is determined by Bayes rule:

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y} \mid \theta)}{\int p(\theta)p(\boldsymbol{y} \mid \theta)\, d\theta} = \frac{p(\theta)p(\boldsymbol{y} \mid \theta)}{p(\boldsymbol{y})}$$

$$\propto \quad p(\theta)p(\boldsymbol{y} \mid \theta) \propto p(\theta)L(\theta \mid \boldsymbol{y})$$

where "$\propto$" means "is proportional to."

The key tool in Bayesian statistics

$$\boxed{p(\theta \mid y) \propto p(\theta)L(\theta \mid y)}$$

Posterior $\propto$ Prior $\times$ Likelihood

Applying the key tool in our setting

We have

$$\begin{aligned} p(\theta) &= 1 \\ L(\theta \mid \boldsymbol{y}) &= \theta(1-\theta)^4 \end{aligned}$$

This means

$$p(\theta \mid \boldsymbol{y}) \propto p(\theta) L(\theta \mid \boldsymbol{y}) = (1)\theta(1-\theta)^4 = \theta(1-\theta)^4.$$

This in turn means that

$$p(\theta \mid \boldsymbol{y}) = c\theta(1-\theta)^4$$

for some constant c (that does not depend on θ).

It turns out

- $p(\theta) = 30\ \theta(1-\theta)^4$ integrates to 1, so that $c = 30$ in this case.
- This is the density for a Beta distribution with parameters 2 and 5 (one more than the exponents of θ and $(1-\theta)$, respectively), abbreviated

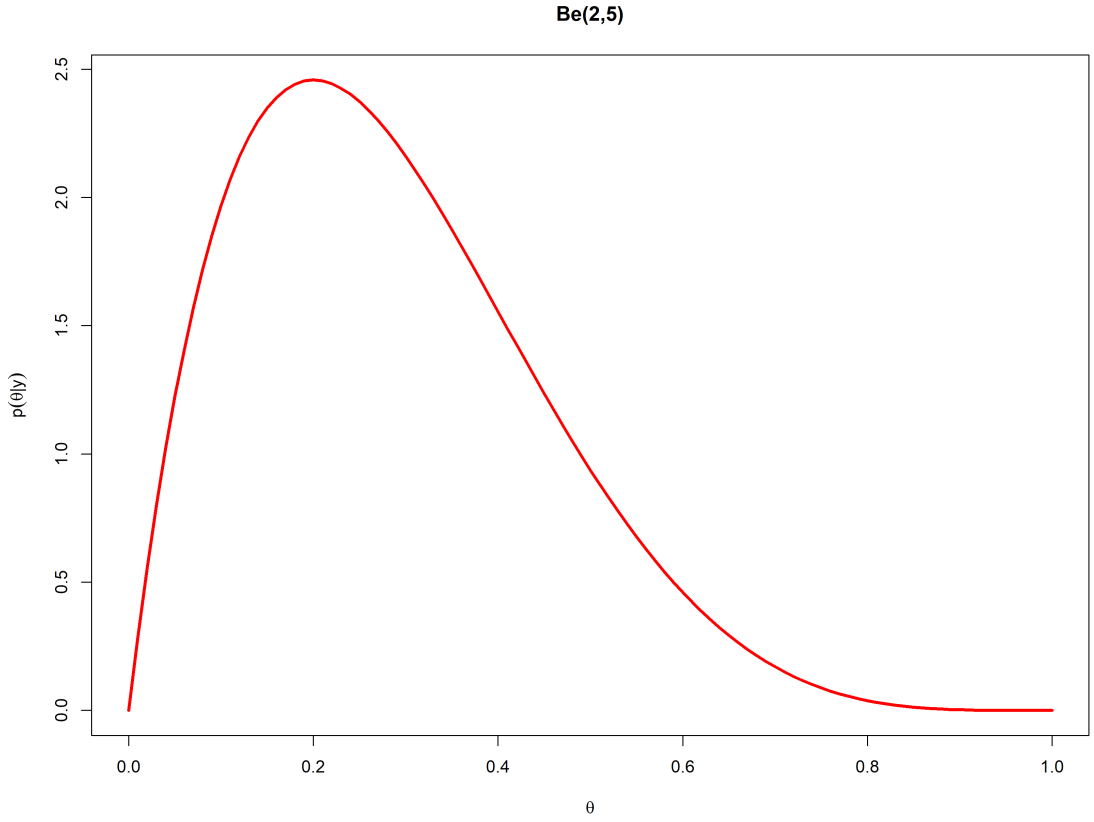
$$\theta \sim \text{Be}(2, 5).$$

This is a probability distribution used frequently by Bayesians.

For general Beta distributions, write $\text{Be}(\alpha, \beta)$.

Some properties of the Beta distribution

$$\begin{aligned} \text{E}(\theta|\alpha,\beta) &= \frac{\alpha}{\alpha+\beta} \\ \text{Var}(\theta|\alpha,\beta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ \text{Mode}(\theta|\alpha,\beta) &= \frac{\alpha-1}{\alpha+\beta-2} \end{aligned}$$



Fifth step: Summarizing the posterior distribution

- Posterior mean, $\mathrm{E}(\theta|\boldsymbol{y})$.
- Posterior mode (value of θ that maximizes $p(\theta \mid \boldsymbol{y})$).
- Central posterior interval for θ.
- Highest posterior density (HPD) region for θ – shortest interval with specified probability. Usually more difficult to compute than central intervals.

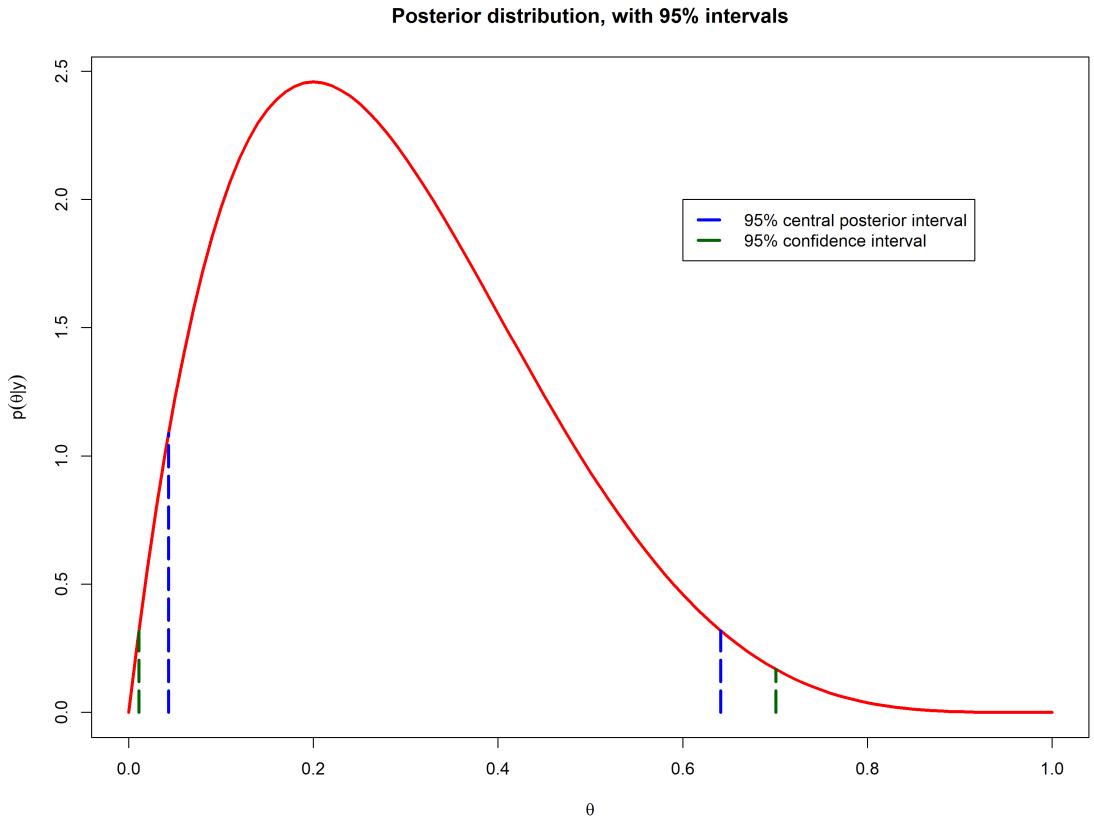
Posterior summaries for our setting

Based on properties of the Beta distribution

- $\mathrm{E}(\theta|\boldsymbol{y}) = \frac{2}{2+5} \approx 0.286$
- $\text{mode}(\theta|\boldsymbol{y}) = \frac{2-1}{2+5-2} = 0.2$
- 95% central posterior interval: Can compute the 2.5%-ile and 97.5%-ile numerically.

$$(0.0433,\ 0.641)$$

Compare to the Frequentist 95% confidence interval $(0.011,\ 0.701)$.



Comments

- The Bayesian approach recognizes the asymmetry in inferences about θ, whereas the standard Frequentist approach does not.
- The same approach outlined above applies identically to multi-parameter models.
- Frequentist approach can be improved, but need to use fancier machinery, e.g., the bootstrap, but even that is not often reliable with small samples.

Bayesian inference as a learning model

The Bayesian approach facilitates the idea that one learns about the state of the world as data are accumulated.

Sequential nature of Bayesian analysis

"Yesterday's posterior is today's prior"

Instead of performing a single analysis with the 5 observations, I could have performed (up to) 5 analyses in sequence:

Suppose $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$, and $y_5 = 0$.

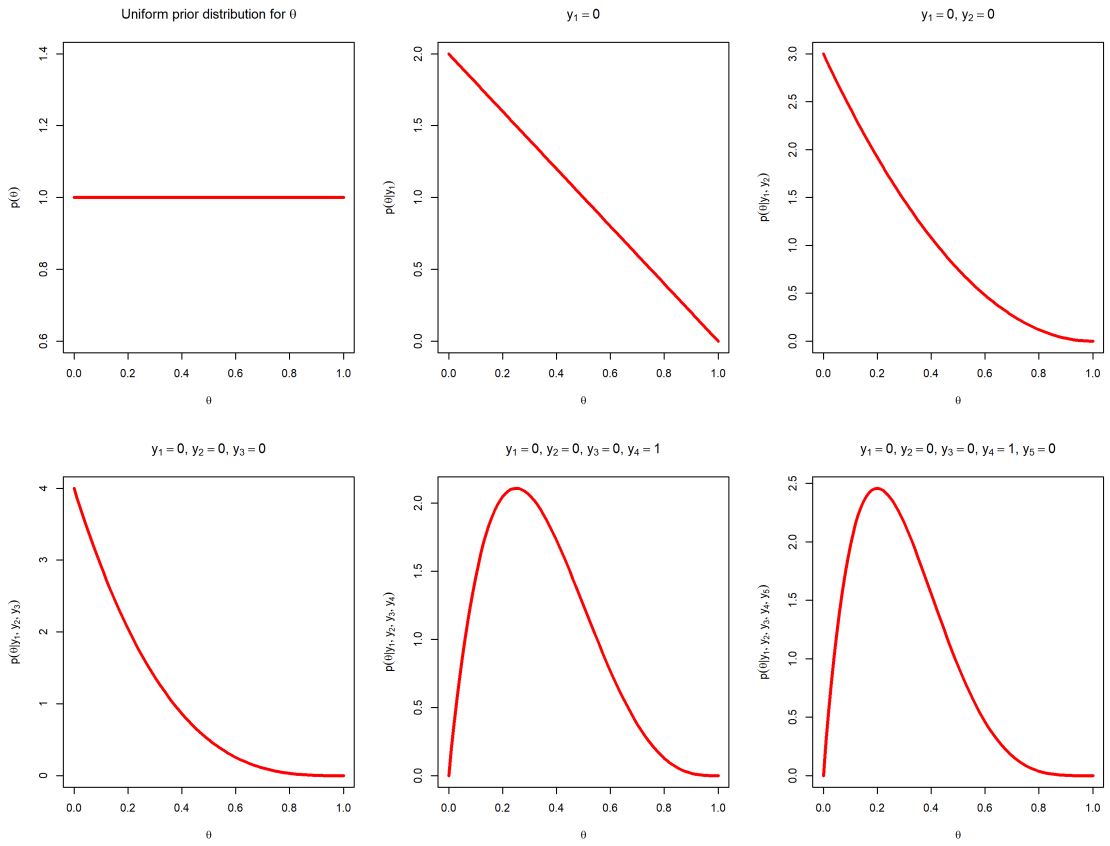
Start with $p(\theta)$.

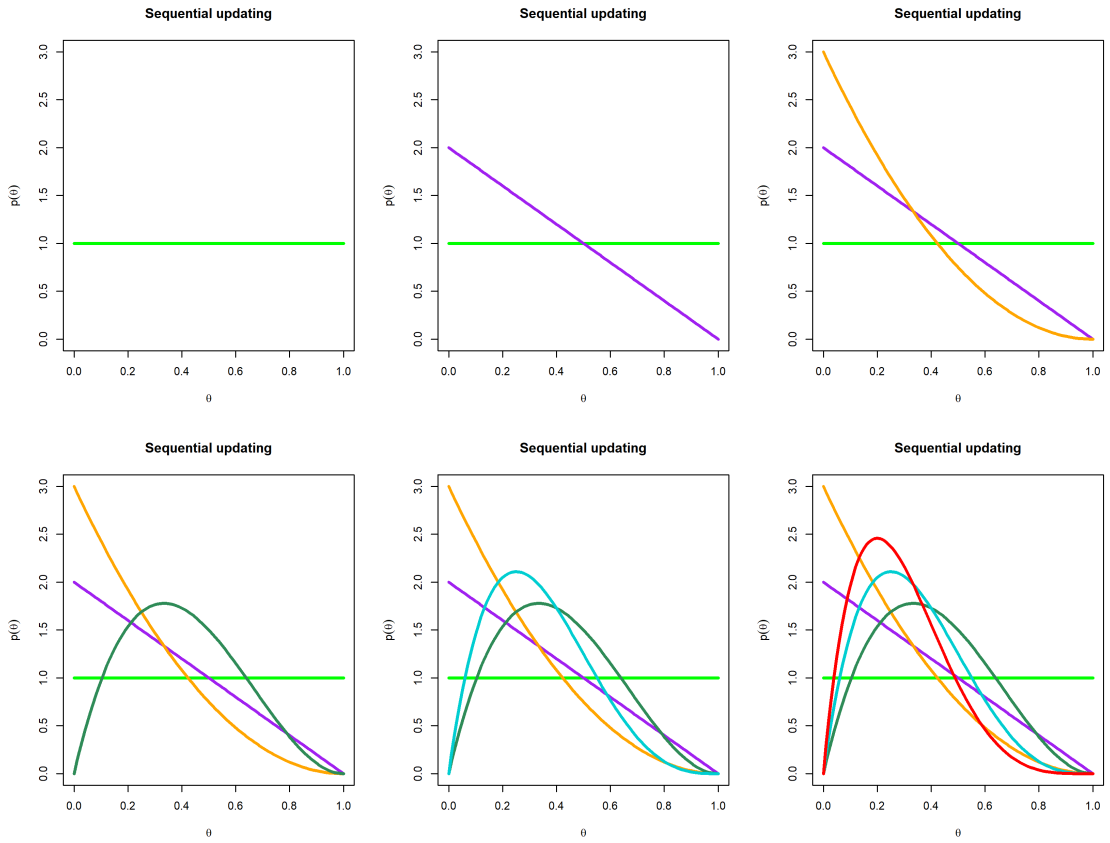
Observe $y_1 = 0$, and determine $p(\theta \mid y_1 = 0)$.

Now use $p(\theta \mid y_1)$ as the prior distribution for the second observation.

Observe $y_2 = 0$, and determine $p(\theta \mid y_1 = 0, y_2 = 0)$.

And so on! You will obtain the same posterior distribution as analyzing the data simultaneously.





Deeper discussion on prior distributions

A main difference operationally between Frequentist methods and Bayesian methods is the incorporation of a prior distribution.

What are the guidelines for choosing a prior distribution?

Two types of prior distributions: Informative versus non-informative

Informative prior distributions The statistician uses his/her knowledge about the substantive problem, along with elicited expert opinion if possible, to construct a prior distribution that properly reflects prior beliefs about the unknown parameters.

Non-informative prior distributions The statistician chooses a distribution in an attempt to be objective, acting as though no prior knowledge about the parameters exists before observing the data.

When one can construct a defensible informative prior distribution, this is the best and most scientific approach.

Criticisms of assuming an informative prior distribution

1. Two different Bayesians could end up using two different informative prior distributions, and would therefore obtain two different posterior distributions.

2. Some argue informative prior distributions are not "objective" or "scientific."

Noninformative prior distributions

Also termed "vague," "diffuse," and "objective."

It is often desirable to have prior distributions that

- formally express ignorance, and
- can be viewed as default choices when no prior knowledge is available.

Main desired property of choosing an objective prior distribution

- Assigns "equal probability" (or at least approximately equal) to all values of the parameters.

That is, the prior density should be fairly flat, and that the likelihood should overwhelm the prior density with even small amounts of data.

Problems with objective prior distributions

- For most problems, there is no unique non-informative prior distribution.
- Any method for constructing a non-informative prior distribution should be invariant to the scale of the parameter.

  This is a difficult property to satisfy in practice.
- Common methods for constructing non-informative prior distributions result in "improper" probability distributions.

So being Bayesian is not necessarily the perfect solution.

Bayesian models as generative models

A generative model is a probabilitistic specification to generate data. Model ingredients:

- The parameters, θ.
- The features (or predictors or covariates), x.
- The labels (or the outcome variable), y.

For the Bayesian framework, we assume

$$\begin{aligned} \theta &\sim p(\theta) && \text{Prior distribution} \\ y|\theta, x &\sim p(y|\theta, x) && \text{Data probability model} \end{aligned}$$

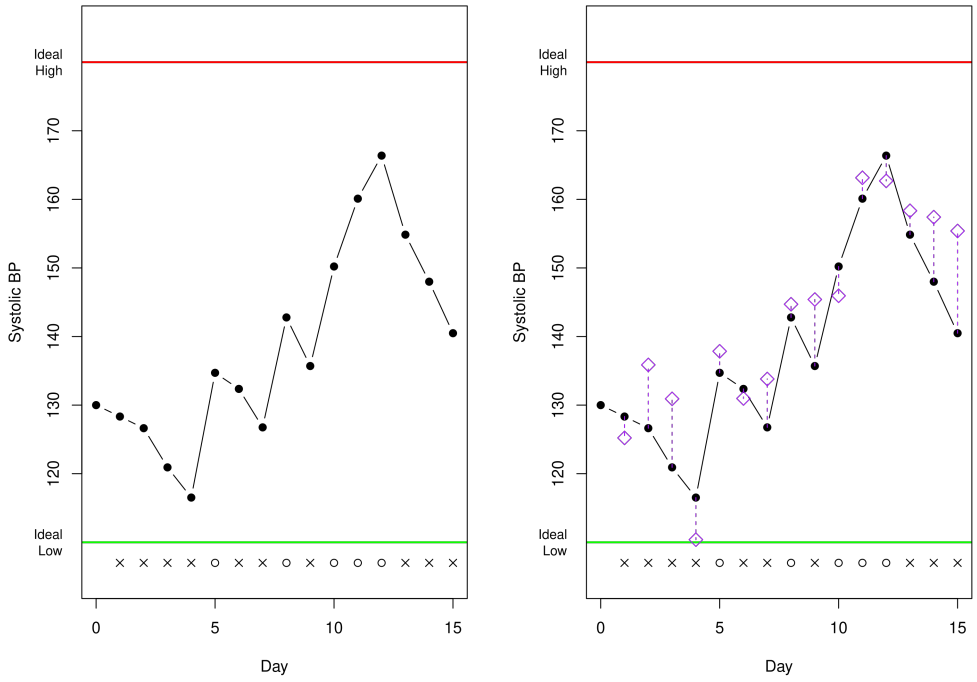
To generate or simulate an outcome value,

- Generate a value θ from the distribution $p(\theta)$.
- Given the simulated value θ, and the known covariates x, now simulate an outcome value y from the distribution $p(y|\theta, x)$.

Examples of generative models

- Hierarchical (linear) models – to be seen shortly
- Generative Adversarial Networks (GANs) – to be seen not so shortly
- Hidden Markov models (HMMs) – next slide!

Hidden Markov model for Systolic blood pressure over time



Towards a modern approach to Bayesian data analyses

The previous discussion laid out the foundations of the Bayesian approach, but more complex modeling situations can present great challenges.

- The posterior density for most real-life models can be complex expressions.
- More often than not, the posterior density cannot be written exactly because the normalizing constant cannot be evaluated.
- Even if one can write down the posterior density exactly, it is almost always the case that it is difficult to summarize using standard analytic tools.

Early computational approaches

For awhile (through the 1980s), the approach that seemed most promising was quadrature methods. E.g., see Naylor & Smith, 1982, *Applied Statistics*.

- Basic idea: Replace $p(\theta|\boldsymbol{y})$, a continuous density, with an approximating discrete mass function $\tilde{p}(\theta|\boldsymbol{y})$. Obtain posterior summaries through the approximating discrete distribution.
- A common particular version is Gauss-Hermite quadrature. Choose the spacing of the values of θ according to a normal distribution along with relevant weights.
  - Requires computation to approximate the normal mean and variance.
  - For multivariate parameter θ, the quadrature spacing is typically accomplished one variable at a time.

Problem with quadrature approaches

Can be quite limited in its applicability.

- The quadrature grid of values of θ may not adequately represent the true distribution of θ.
- The real problem is with multivariate θ. Very difficult to obtain reasonable representation of discrete values in multivariate space (curse of dimensionality!) without making the computation unwieldly.

This approach never got very serious traction.

Monte Carlo simulation

The posterior distribution is a probability distribution, and the goal of a Bayesian analysis is to study the posterior distribution, or derive summaries from the posterior distribution.

Summarizing computer-simulated values from the posterior distribution is a legitimate alternative to analytic summaries.

Monte Carlo summaries of probability distributions

- Rather than go through laborious calculations to obtain the mean, median, variance, percentiles, etc., of a probability distribution, one can calculate the sample-version of the distribution from the simulated sample.

- Usually need to simulate a very large sample in order to precisely approximate the generating distribution.

Example

Summarizing Be$(2, 5)$ distribution:

Rather than report the analytically-determined summaries of the Beta posterior distribution for the 30-day mortality example, do the following:

1. Simulate 10,000 values from Be$(2, 5)$.
2. Report *sample* summaries from the distribution of 10,000 simulated values.

Code

```
# simulate from Be(2,5)
theta = np.random.beta(2,5,10000) # generate 10,000 values
print(theta[0:9]) # first ten simulated values
print(np.mean(theta)) # sample mean
# 2.5% and 97.5% of empirical distn
print(np.percentile(theta,[2.5,97.5]))
```

```
[0.12471433 0.38933883 0.1296898  0.23946542 0.55924411
 0.23182827 0.13178348 0.29691004 0.27515282 0.31043925]

0.28478820646628294

[0.04471399 0.63485071]
```

Recall that E$(\theta|\alpha = 2, \beta = 5) = 0.286$ and the 95% central posterior interval was $(0.0433,\ 0.641)$.

Predictive inference

One of the real strengths of the Bayesian setting is the ability to produce model-based predictions that accounts for the uncertainty in parameter inferences.

Goal Want to determine $p(\tilde{y} \mid \boldsymbol{y})$, that is, the probability distribution for a new value $\tilde{y}$ given the data we already analyzed. This distribution is called the **posterior predictive distribution**.

Notice that θ is no longer in the expression – the posterior predictive distribution "averages out" the uncertainty about θ.

Analytic solution to determining $p(\tilde{y} \mid \boldsymbol{y})$:

$$p(\tilde{y} \mid \boldsymbol{y}) = \int p(\tilde{y} \mid \theta) p(\theta|\boldsymbol{y})\, d\theta.$$

For our setting, we would compute

$$\begin{aligned} p(\tilde{y} \mid \boldsymbol{y}) &= \int \left(\theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}}\right)\ \left(30\ \theta(1-\theta)^4\right)\ d\theta \\ &= 30 \int \theta^{\tilde{y}+1}(1-\theta)^{5-\tilde{y}}\ d\theta \end{aligned}$$

This simplifies to

$$p(\tilde{y} \mid \boldsymbol{y}) = \begin{cases} 5/7 \approx 0.7143 & \text{if } \tilde{y} = 0 \\ 2/7 \approx 0.2857 & \text{if } \tilde{y} = 1 \end{cases}$$

Thus we can conclude that there is a $2/7$ chance that a new (randomly selected) patient entering the hospital with a heart attack would die within 30 days.

Computing $p(\tilde{y} \mid \boldsymbol{y})$ via simulation

This method highlights the power of simulation-based inference.

1. Generate (say) 10,000 simulated parameter values from the posterior distribution. Call these $\theta^{(1)}, \ldots, \theta^{(10000)}$.
2. For each $j = 1, \ldots, 10000$ separately, generate a value of $\tilde{y}^{(j)}$ from the probability model with parameter value $\theta^{(j)}$.
3. Summarize features of the 10,000 values $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(10000)}$ for predictions.

Example with 30-day mortality

The posterior distribution is $\theta \sim \text{Be}(2, 5)$. Also, $\Pr(\tilde{y} = 1|\theta) = \theta$.

```
# simulate y from posterior predictive distribution
theta = np.random.beta(2,5,10000) # generate 10,000 values
print(theta[0:9]) # first ten simulated values
# generate 10,000 binary values with different probabilities
y = np.random.binomial(1,theta,10000)
print(y[0:9]) # first ten values
print(np.bincount(y)/10000)  # frequency of 0 and 1
```

```
[0.36309086 0.05705801 0.05409083 0.5388881  0.49666649
 0.30191721 0.3444435  0.08902907 0.08247285 0.21758449]

[0 0 0 1 0 0 1 1 0 0]

[0.7083 0.2917]
```

Indirect Monte Carlo methods

It is extremely rare in practice that the posterior density will be of a convenient form that permits direct simulation.

Bayesians started to rely on several indirect sampling approaches, including

1. Rejection sampling
2. Weighted Bootstrap

We will examine both and their relevance in Bayesian analysis. An easy-to-read review can be found in Smith and Gelfand, 1992, *The American Statistician*.

Rejection sampling

- Suppose we are interested in simulating values from density $h(\boldsymbol{\theta})$ but cannot do so directly.
- Density $h(\boldsymbol{\theta})$ may be unnormalized; that is, $\int h(\boldsymbol{\theta})d\boldsymbol{\theta} = c \neq 1$.
- Suppose also that we can simulate directly from density $g(\boldsymbol{\theta})$.
- Need to require that there exists constant $M > 0$ such that $h(\boldsymbol{\theta})/g(\boldsymbol{\theta}) \leq M$ for all $\boldsymbol{\theta}$.

Then to obtain a simulated value of $\boldsymbol{\theta}$ from $h(\boldsymbol{\theta})$, do the following.

Rejection sampling (cont'd)

1. Simulate $\boldsymbol{\theta}$ from $g(\boldsymbol{\theta})$.
2. Simulate U from a uniform distribution on $(0, 1)$.
3. If
$$U \leq \frac{h(\boldsymbol{\theta})}{Mg(\boldsymbol{\theta})},$$
then accept $\boldsymbol{\theta}$. Otherwise, repeat steps $1 - 3$.

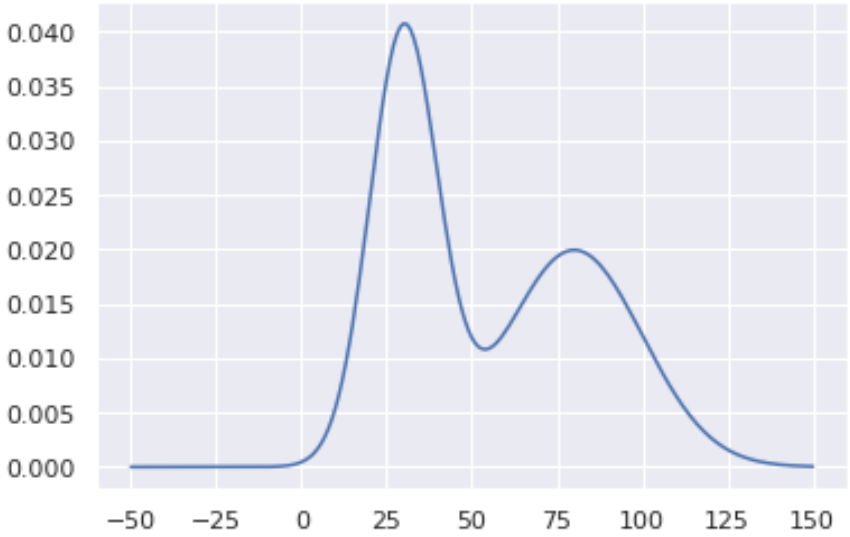
The resulting $\boldsymbol{\theta}$ is a simulated value from $h(\boldsymbol{\theta})$, the desired distribution.

The easy proof is in the monograph by Ripley (1986).

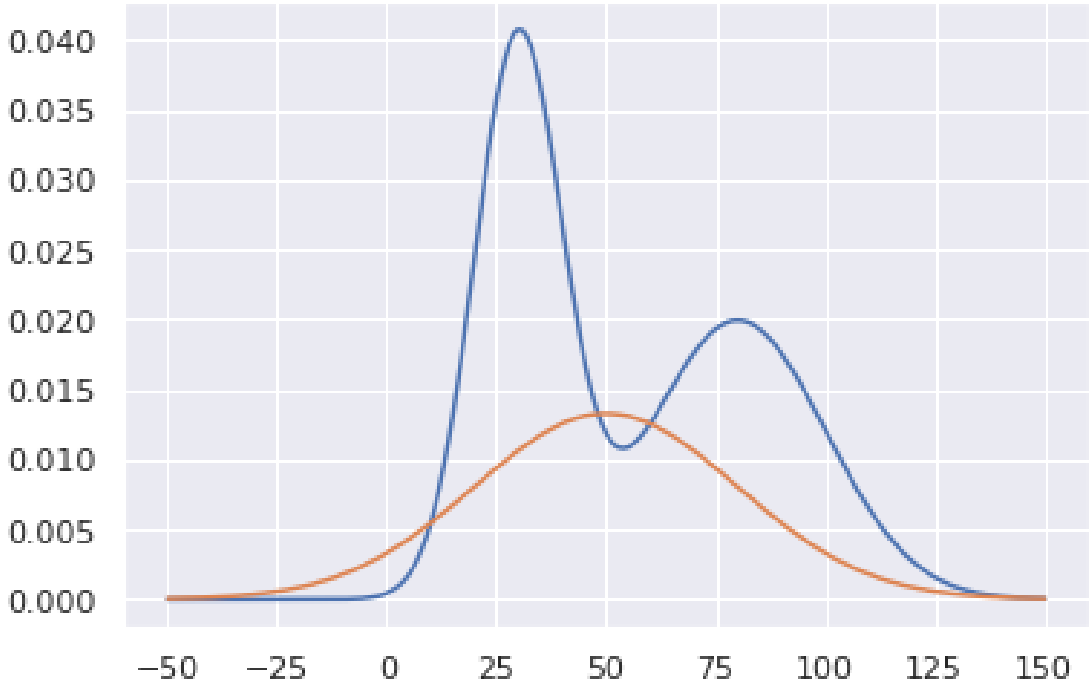
Example: Sampling from a mixture of normal distributions

Let

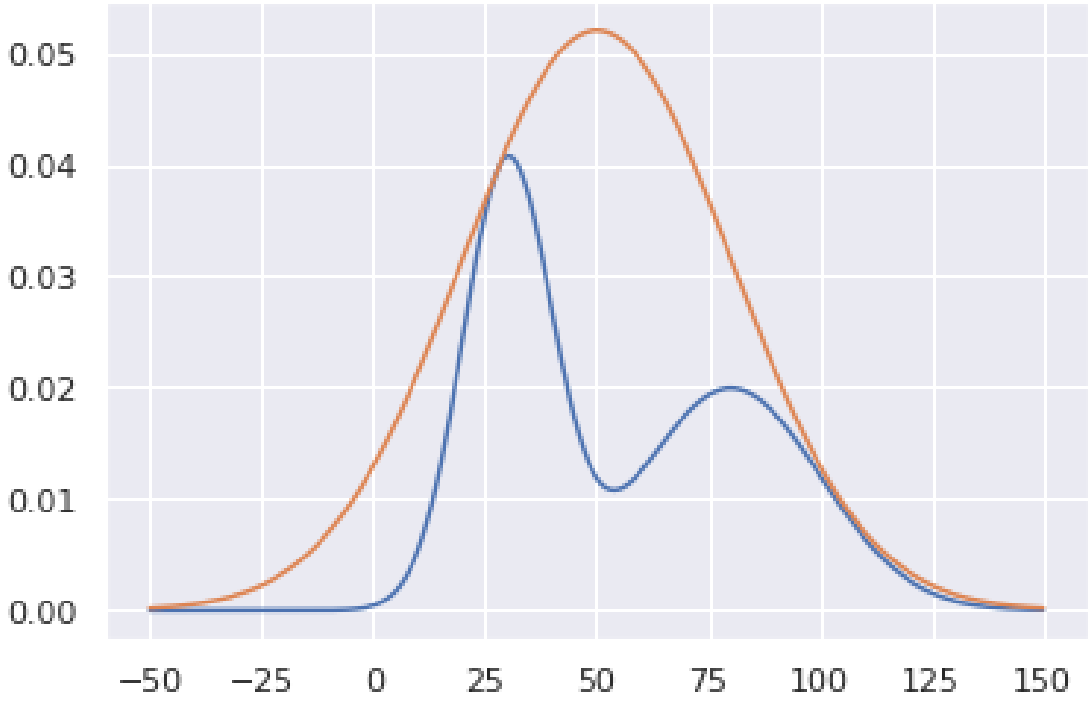
$$\begin{aligned} h(\theta) &= \frac{1}{\sqrt{2\pi(10^2)}} \exp\left(\frac{-1}{2(10^2)}(\theta - 30)^2\right) + \frac{1}{\sqrt{2\pi(20^2)}} \exp\left(\frac{-1}{2(20^2)}(\theta - 80)^2\right) \\ g(\theta) &= \frac{1}{\sqrt{2\pi(30^2)}} \exp\left(\frac{-1}{2(30^2)}(\theta - 50)^2\right) \end{aligned}$$



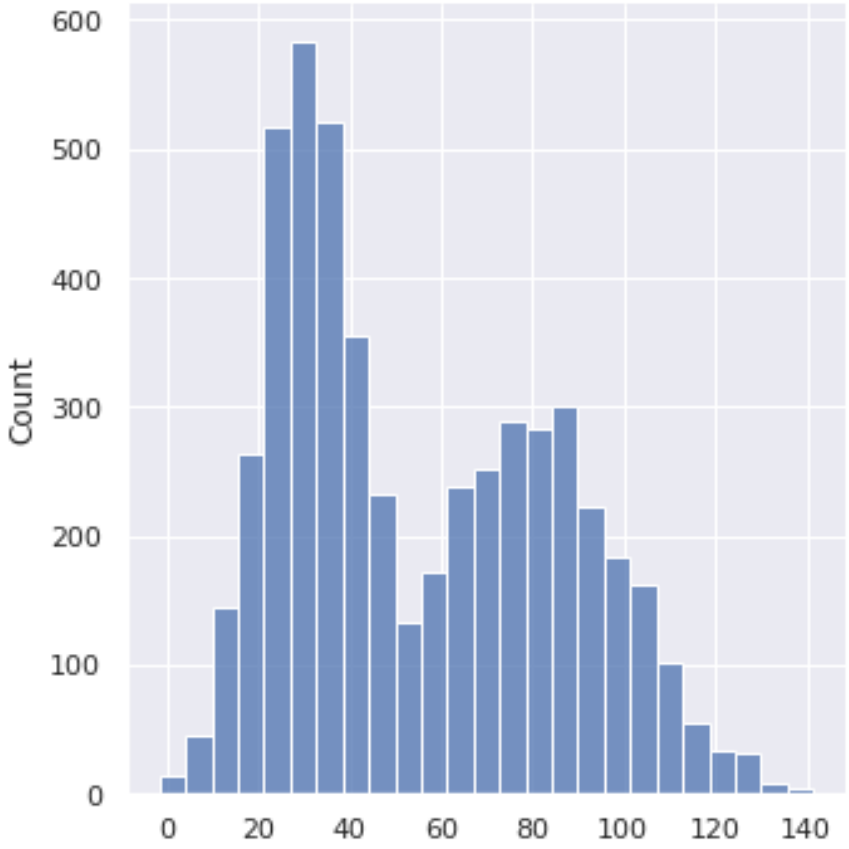
Superimposing proposal normal density



Multiplicative increase to become envelope



Results of rejection sampling with 10000 iterations



Weighted Bootstrap

Consider the same setup as rejection sampling:

- Want to simulate from $h(\boldsymbol{\theta})$ but cannot directly.
- Density $h(\boldsymbol{\theta})$ may be unnormalized; that is, $\int h(\boldsymbol{\theta})d\boldsymbol{\theta} = c \neq 1$.
- We can simulate directly from $g(\boldsymbol{\theta})$.

However, this time suppose we cannot readily produce an M such that $h(\boldsymbol{\theta})/g(\boldsymbol{\theta}) \leq M$ for all $\boldsymbol{\theta}$.

Can use the following procedure.

Weighted Bootstrap (cont'd)

1. Simulate K draws $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ from $g(\boldsymbol{\theta})$.
2. For each $k = 1, \ldots, K$, compute $w_k = h(\boldsymbol{\theta}_k)/g(\boldsymbol{\theta}_k)$, and let

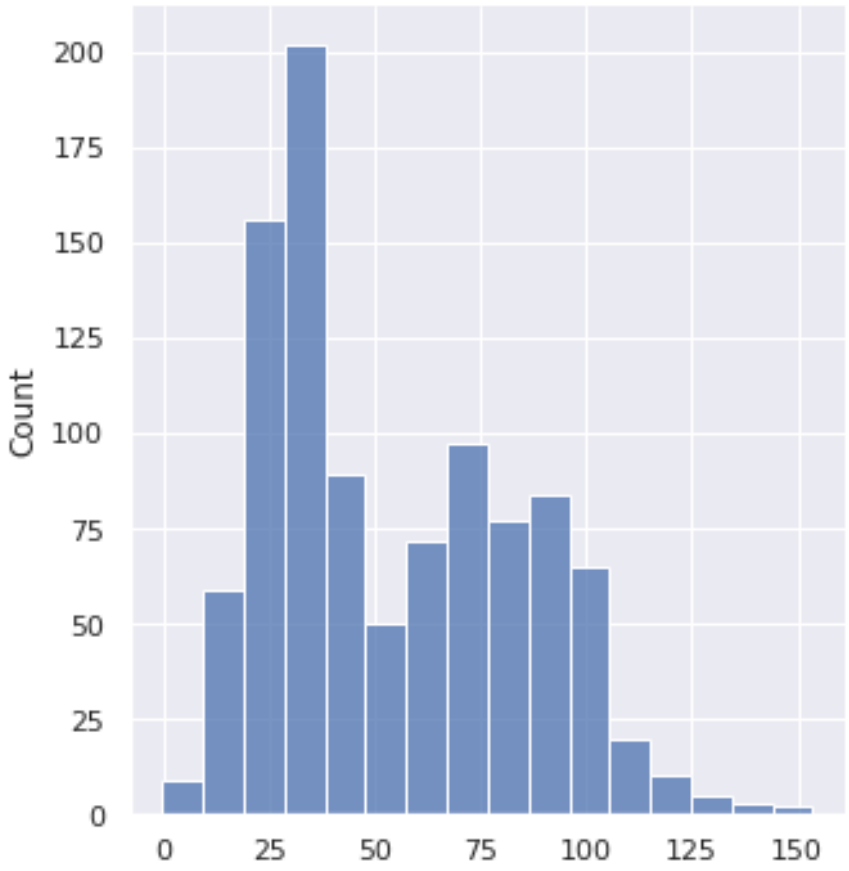
$$q_k = \frac{w_k}{\sum_{j=1}^{K} w_j},$$

   the normalized values of the w_k.
3. Now simulate $\boldsymbol{\theta}$ from the discrete distribution of values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ with probabilities $\{q_1, \ldots, q_K\}$.

The resulting value of $\boldsymbol{\theta}$ is *approximately* distributed according to $h(\boldsymbol{\theta})$, with the approximation improving as K increases.

This procedure is sometimes known as SIR (sampling/importance resampling).

Results of weighted bootstrap with 1000 iterations



Application to Bayesian analysis

One clever way to use rejection sampling or the weighted bootstrap for simulating from a posterior density is to do the following.

With prior density $p(\boldsymbol{\theta})$, likelihood $L(\boldsymbol{\theta}|\boldsymbol{y})$, and posterior density $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{y})$, suppose

- one can simulate values directly from $p(\boldsymbol{\theta})$,
- we want to simulate values from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Bayesian rejection sampling

In the rejection sampling algorithm, let $h(\boldsymbol{\theta}) = p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{y})$ (which is a constant factor of the posterior density), and $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

Let M be the minimum value for which $h(\boldsymbol{\theta})/g(\boldsymbol{\theta}) \leq M$ for all $\boldsymbol{\theta}$. That is,

$$\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})} = L(\boldsymbol{\theta}|\boldsymbol{y}) \leq M.$$

Thus choose $M = L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ where $\hat{\boldsymbol{\theta}}$ is the MLE.

Then

1. Simulate $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$.
2. Generate U from a uniform distribution over $(0, 1)$.
3. If
$$U \leq \frac{h(\boldsymbol{\theta})}{Mg(\boldsymbol{\theta})} = \frac{L(\boldsymbol{\theta}|\boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})},$$
then accept $\boldsymbol{\theta}$ as a simulated value from $p(\boldsymbol{\theta}|\boldsymbol{y})$; otherwise repeat.

Bayesian weighted bootstrap

In the weighted bootstrap algorithm, again let $h(\boldsymbol{\theta}) = p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{y})$ and $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

Simulate values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ from $p(\boldsymbol{\theta})$, and evaluate for each $k = 1, \ldots, K$
$$w_k = \frac{h(\boldsymbol{\theta}_k)}{g(\boldsymbol{\theta}_k)} = L(\boldsymbol{\theta}_k|\boldsymbol{y}),$$
and compute the normalized values
$$q_k = \frac{w_k}{\sum_{j=1}^K w_j} = \frac{L(\boldsymbol{\theta}_k|\boldsymbol{y})}{\sum_{j=1}^K L(\boldsymbol{\theta}_j|\boldsymbol{y})}.$$

Now simulate a value $\boldsymbol{\theta}$ from the discrete distribution with support on $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ and with probabilities $\{q_1, \ldots, q_K\}$.

The resulting $\boldsymbol{\theta}$ is approximately from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Problems with these approaches

The main problem is computational efficiency.

- Both algorithms work well when $h(\boldsymbol{\theta})/g(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$ is nearly constant with respect to $\boldsymbol{\theta}$.
- When $h(\boldsymbol{\theta})/g(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$ is not constant, then
  - It could take ages in rejection sampling before a $\boldsymbol{\theta}$ is accepted!
  - The probabilities $\{q_1, \ldots, q_K\}$ in the weighted bootstrap could be terribly skewed so that the discrete approximation to $p(\boldsymbol{\theta}|\boldsymbol{y})$ is horrible!

Another problem is that if we assume an improper prior distribution, then we cannot simulate from $p(\boldsymbol{\theta})$.

For these reasons, these two algorithms are not often used.

The Modern World

In the late 1980s, a few statisticians (Alan Gelfand, Adrian Smith, and others) stumbled on an idea that revolutionized Bayesian statistics.

Rather than directly or indirectly simulate from $p(\boldsymbol{\theta}|\boldsymbol{y})$, simulate a random walk in the space of $\boldsymbol{\theta}$ which converges to $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Markov chain Monte Carlo (MCMC) simulation

- The key idea is to create a Markov chain whose stationary distribution is $p(\boldsymbol{\theta}|\boldsymbol{y})$. Values are then computer-simulated from the Markov chain.
- Once the Markov chain is run long enough, simulated values can be treated as coming from the posterior distribution.

There are actually many ways to set up such a Markov chain. A common way to construct a Markov chain is to perform **Random Walk Metropolis** (RWM) sampling.

Random Walk Metropolis (RWM) sampling

Let $p(\boldsymbol{\theta}|\boldsymbol{y})$ be the posterior density of d-dimensional $\boldsymbol{\theta}$. We want to sample vectors of $\boldsymbol{\theta}$ from this distribution.

- Step 1: Pick arbitrary starting vector $\boldsymbol{\theta}_1$.
- Step 2: For $j = 1, 2, \ldots,$
  - (a) simulate d-dimensional vector Δ_j (the proposed "jump") from pre-specified distribution $\tilde{p}(\Delta)$.
  - (b) evaluate the proposal acceptance probability

$$\alpha(\boldsymbol{\theta}_j, \Delta_j) = \min\left(1, \frac{p(\boldsymbol{\theta}_j + \Delta_j|\boldsymbol{y})}{p(\boldsymbol{\theta}_j|\boldsymbol{y})}\right).$$

  - (c) let

$$\boldsymbol{\theta}_{j+1} = \begin{cases} \boldsymbol{\theta}_j & \text{with probability } 1 - \alpha(\boldsymbol{\theta}_j, \Delta_j) \\ \boldsymbol{\theta}_j + \Delta_j & \text{with probability } \alpha(\boldsymbol{\theta}_j, \Delta_j) \end{cases}$$

As $j \to \infty$, $\boldsymbol{\theta}_j$ is a simulated value from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Interpreting RWM sampling

- Uphill proposals (ones that take the Markov chain to a local maximum) are always accepted.
- Downhill proposals (ones that move away from a local maximum) are accepted with probability equal to the relative heights of the posterior density at the proposed and current values.

MCMC Random Walk Metropolis Demo

Implementing an MCMC sampler to obtain simulated parameter values

1. Run several parallel MCMC samplers with different starting values (preferably widely-dispersed)
2. Simulate values from the Markov chains for a "burn-in" period (before the Markov chains have converged to the stationary distribution), and discard the burn-in simulations
3. Save simulated values after burn-in period. These will be the simulated values on which to perform inferential summaries.

If I were you, I would be left with a **big headache** because a lot of work still seems necessary to implement a good MCMC sampler.

- Still need to pick a good proposal distribution (for RWM sampling) where the jumps are not too big or too small.
- Still need to simulate from the probability distributions
- Still need to write code the performs the Markov chain simulation, checks for convergence, etc.

Fortunately, many packages now exist that do the hard work for you.

The two most well-known packages for implementing Bayesian analyses using MCMC are

- WinBUGS, OpenBUGS, and JAGS (BUGS = **B**ayesian inference **U**sing **G**ibbs **S**ampling).
- Stan, which is a lot like WinBUGS, but uses (by default) a different type of Markov chain sampler.
- pymc3 library in python.

You will get exposure to pymc3 in this course.

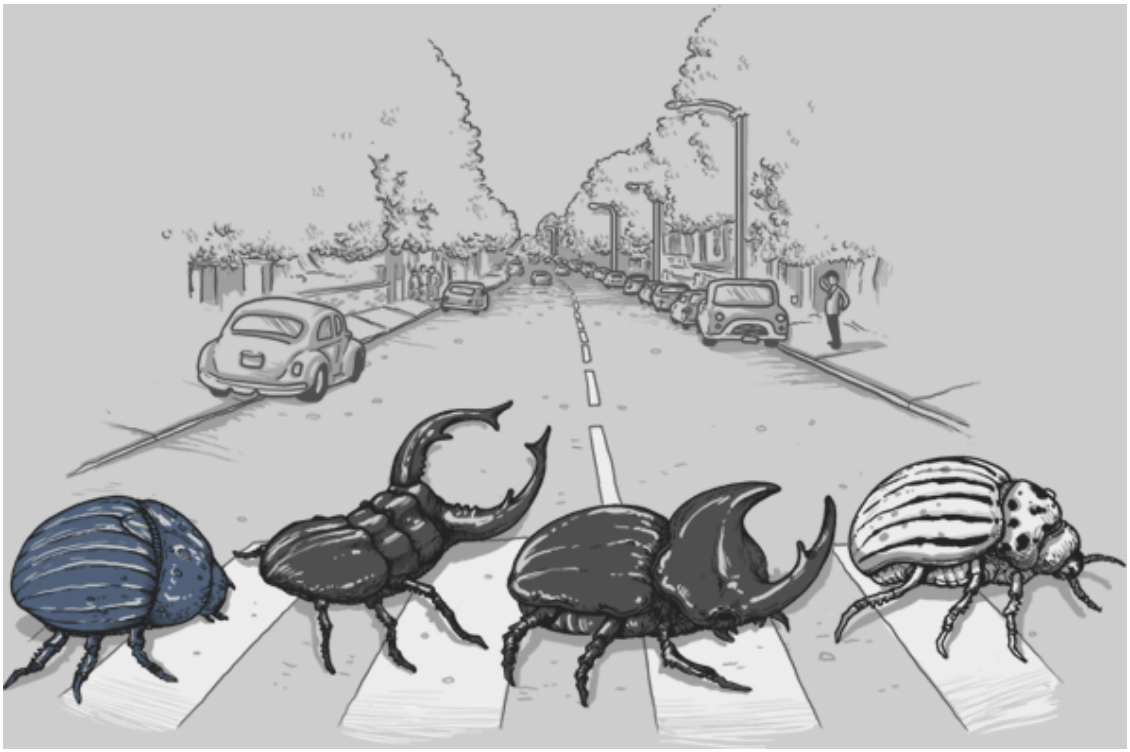
These programs actually figure out the necessary distributional assumptions and automatically sample from them!

That's right – they automatically determine and implement the Markov chain!

All you need to do typically is

- specify the model for data, and
- specify logistical issues concerning MCMC simulation (starting values, burn-in period, how many iterations to run after burn-in)

Example: The Beetles!



Bayesian logistic regression

A dose-response study was performed that counted the number of beetles killed after 5-hour exposure to carbon disulphide. The data are shown below

Concentration of carbon disulphide (x_i)	Number of beetles exposed (n_i)	Number killed (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	52
1.8610	62	61
1.8839	60	60

Model for probability of beetle death

For $i = 1, \ldots, N$, where $N = 8$ observations,

$$\begin{aligned} y_i &\sim \text{Bin}(n_i,\ p_i) \\ \text{logit}\ p_i &= \alpha + \beta x_i \end{aligned}$$

Not necessary, but slightly better computationally to center the x_i:

$$\text{logit}\ p_i = \alpha^* + \beta(x_i - \bar{x})$$

where

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \alpha^* &= \alpha + \beta\bar{x} \end{aligned}$$

Prior distribution

Assume non-informative but proper prior distribution that has independent components

$$\begin{aligned} \alpha^* &\sim \mathrm{N}(0, 10000) \\ \beta &\sim \mathrm{N}(0, 10000) \end{aligned}$$

By Bayes rule, the posterior density for α^* and β can be written as

$$\mathsf{p}(\alpha^*, \beta|\boldsymbol{y}) = c \cdot \mathrm{N}(\alpha^*|0, 10000)\mathrm{N}(\beta|0, 10000)\prod_{i=1}^{8} p_i^{y_i}(1-p_i)^{n_i - y_i}$$

where c is a normalizing constant, and the $\mathrm{N}(\cdot|\cdot,\cdot)$ are normal densities with the given mean and variance.

pymc3 code Written as a generative model

```
with pm.Model() as beetle_model:
    alpha_star = pm.Normal('alpha*', mu=0, sigma=100)
    beta = pm.Normal('beta', mu=0, sigma=100)
    p_i = pm.Deterministic('$P_i$',
      pm.math.invlogit(alpha_star + beta*beetles_x))
    deaths = pm.Binomial('obs_deaths',
      n=beetles_n, p=p_i, observed=beetles_y)
    trace = pm.sample(2000, tune=2000, target_accept=0.9)
```

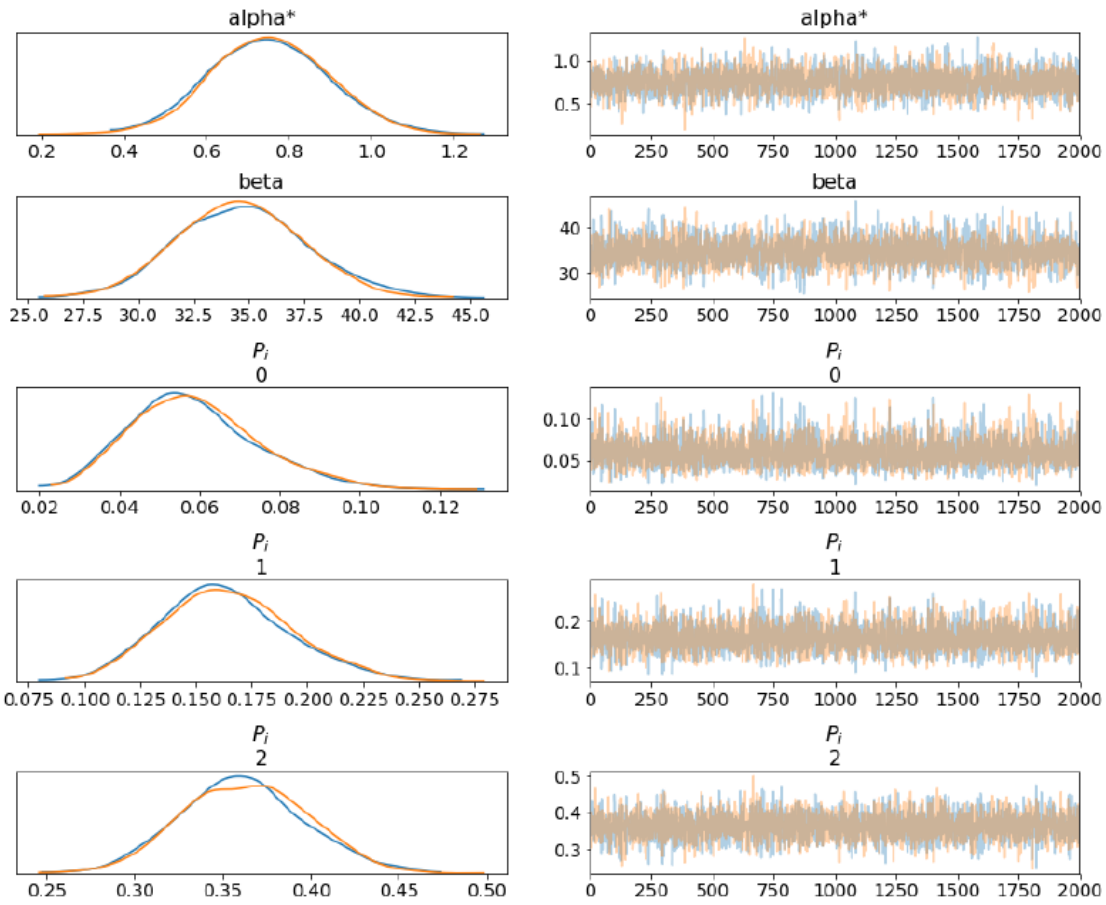
By default, `pymc3` implements four parallel Markov chain samplers. Running three or four chains is typical.

Diagnosing convergence

Run parallel MCMC chains with different starting points.

After the burn-in period, can plot the parameter draws as a function of iteration for each parameter.

Want to look for whether the chains coalesce.



R_hat statistic

The `R_hat` statistic is a numerical measure that indicates whether the Markov chain was run long enough to reach convergence.

- Values near 1.0 indicate convergence.
- Large values (say 1.3 or greater) indicate either that the procedure has not been run long enough, or that the parameter itself may be difficult to obtain reasonable samples given strong autocorrelation in the sampler.

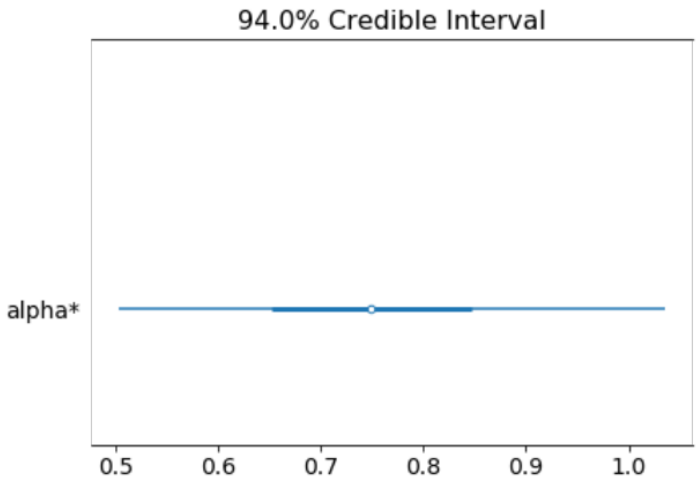
The statistic essentially computes a ratio of between-chain variance and within-chain variance.

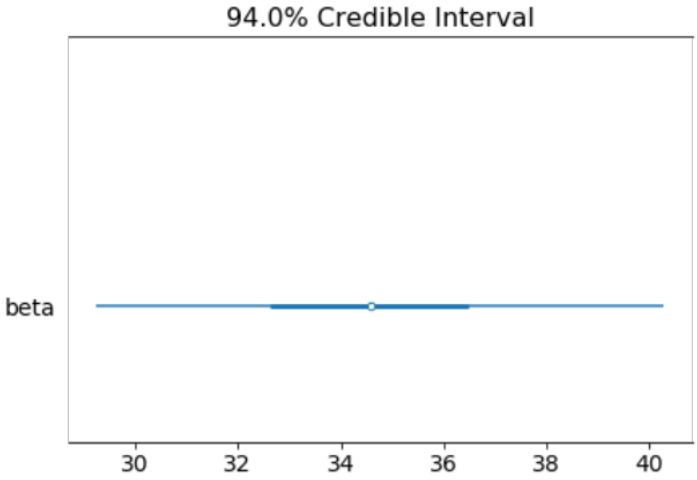
	mean	sd	hpd_3%	hpd_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_ha
alpha*	0.751	0.142	0.503	1.035	0.003	0.002	2665.0	2665.0	2665.0	2345.0	1.(
beta	34.617	2.932	29.249	40.277	0.058	0.041	2574.0	2550.0	2591.0	2500.0	1.(
P_i[0]	0.059	0.016	0.032	0.090	0.000	0.000	3302.0	3211.0	3261.0	2703.0	1.(
P_i[1]	0.164	0.029	0.112	0.220	0.000	0.000	3520.0	3502.0	3495.0	2802.0	1.(
P_i[2]	0.362	0.035	0.298	0.427	0.001	0.000	3612.0	3612.0	3596.0	2665.0	1.(
P_i[3]	0.606	0.032	0.542	0.664	0.001	0.000	2917.0	2917.0	2924.0	2393.0	1.(
P_i[4]	0.796	0.026	0.745	0.844	0.001	0.000	2369.0	2369.0	2367.0	2337.0	1.(
P_i[5]	0.904	0.019	0.868	0.937	0.000	0.000	2253.0	2253.0	2254.0	2097.0	1.(
P_i[6]	0.955	0.012	0.933	0.975	0.000	0.000	2262.0	2262.0	2255.0	2012.0	1.(
P_i[7]	0.979	0.007	0.966	0.991	0.000	0.000	2294.0	2294.0	2281.0	2258.0	1.(

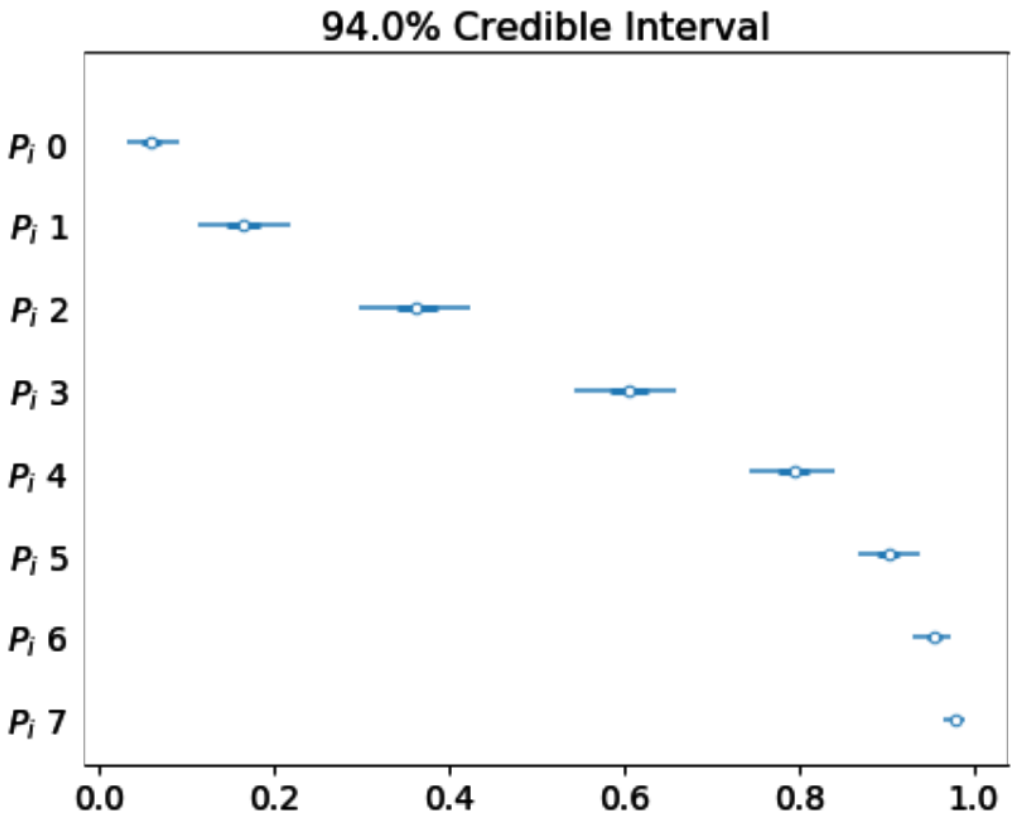
Summarizing model results

Once convergence of the MCMC sampler has been established, the following can be done with the simulated values across all chains after the burn-in iterations:

- Approximate posterior mean and standard deviations of parameters by their sample counterparts.
- Can compute (say) 95% central posterior intervals from the 2.5%-ile and 97.5%-ile of the sample of simulated values.


94.0% Credible Interval
alpha*
0.5
0.6
0.7
0.8
0.9
1.0


94.0% Credible Interval
beta
30
32
34
36
38
40



Numerical summaries

	mean	sd	2.5%	25.0%	50.0%	75.0%	97.5%
alpha*	0.751390	0.141753	0.478711	0.652996	0.748475	0.846896	1.033710
beta	34.616909	2.931295	28.936764	32.611365	34.572128	36.488580	40.633339
P_i [0]	0.059134	0.016350	0.032743	0.047590	0.057312	0.068486	0.096447
P_i [1]	0.163829	0.028815	0.112463	0.143730	0.161691	0.182162	0.224477
P_i [2]	0.361563	0.035229	0.294835	0.337852	0.360971	0.384664	0.430386
P_i [3]	0.605892	0.032435	0.541279	0.583828	0.605457	0.627823	0.668789
P_i [4]	0.796114	0.026486	0.742425	0.778172	0.796806	0.814523	0.846053
P_i [5]	0.903587	0.018611	0.863879	0.891767	0.904673	0.917122	0.936220
P_i [6]	0.955078	0.011669	0.929487	0.947880	0.956265	0.963560	0.974797
P_i [7]	0.978768	0.006921	0.962938	0.974709	0.979674	0.983812	0.989860

Hierarchical linear models

A class of models that is particularly well-served by the Bayesian framework is hierarchical mod-

els.

Typical setup Observe response data and predictor variables as in the usual regression setting.

Suppose now that the samples are clustered by a grouping variable.

Examples

- Obtain a sample of patients seeing their doctor for chest pain. Want to measure the number of days until pain goes away as a function of patient characteristics and treatment given.

  The data may be clustered by clinic, and we might expect that the effects of the patient characteristics differ by clinic.

- Obtain a sample of children taking a standardized achievement test. Want to measure the relationship between test score and background variables about each child.

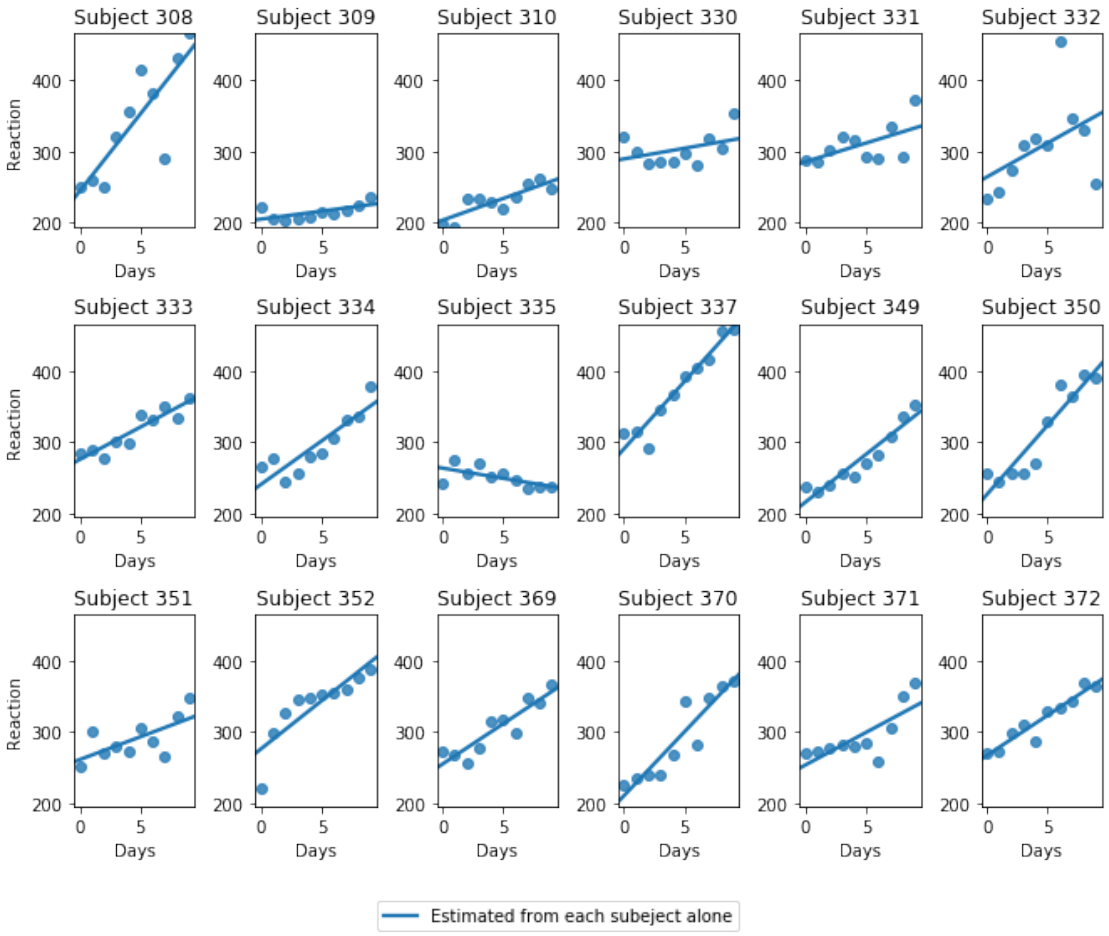
  The data may be clustered by school, and we might expect that the relationship between child characteristics and test score might depend on the school.

Example to implement: Reaction times and sleep deprivation

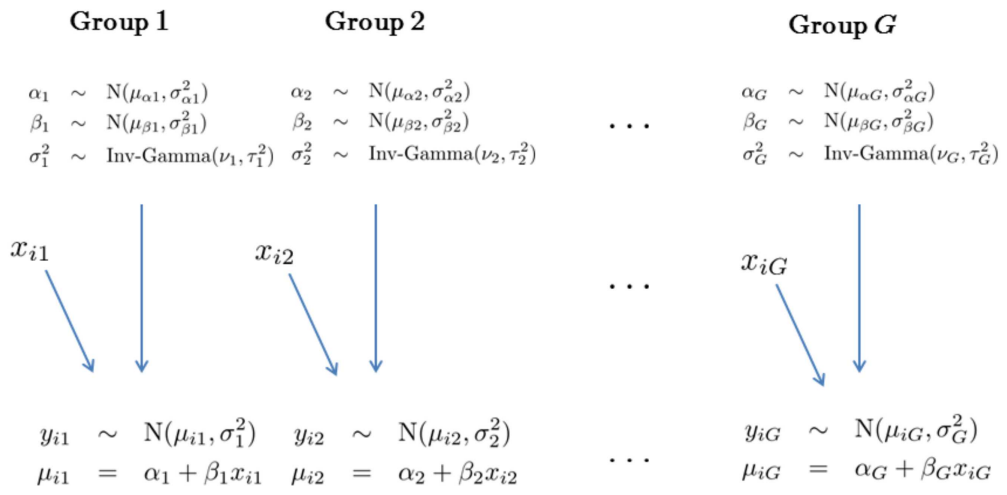
- A study published in a 2003 issue of the *Journal of Sleep Research* measured the reaction time per day for 18 subjects in a sleep deprivation study.
- On day 0, each subject had a normal amount of sleep. Starting that night, each subject was restricted to 3 hours of sleep.
- The response represents the average reaction time on a series of tests given to each subject across 10 days.

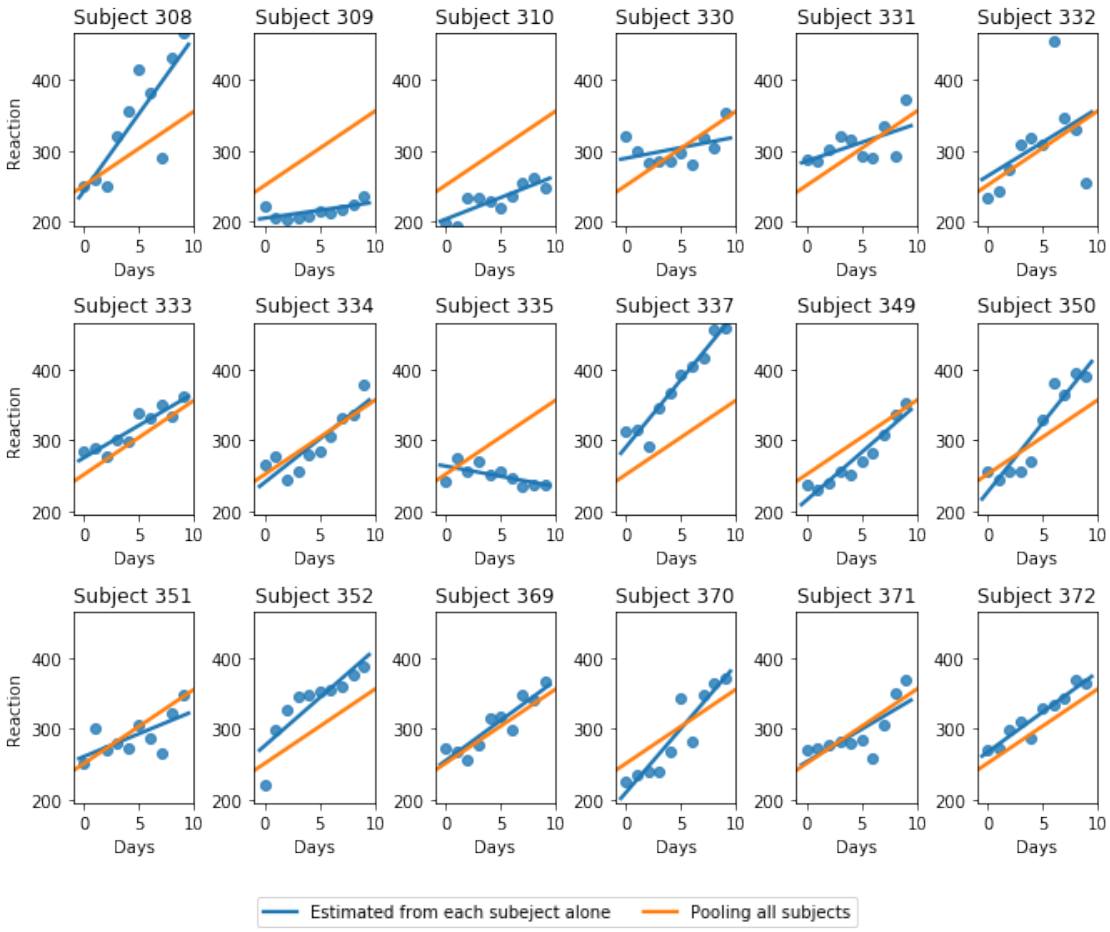
Consider the following two extremes:

- For each subject, fit a separate least-squares regression.
  - Good: Acknowledges that the relationship between x and y may differ within each group.
  - Bad: Very little data on which to estimate effects.
- Pool all the data together and fit a single least-squares regression.
  - Good: Makes full use of the entire data.
  - Bad: Does not recognize differences in effects within groups.


Subject 308
Subject 309
Subject 310
Subject 330
Subject 331
Subject 332
Subject 333
Subject 334
Subject 335
Subject 337
Subject 349
Subject 350
Subject 351
Subject 352
Subject 369
Subject 370
Subject 371
Subject 372
Reaction
Days
Estimated from each subeject alone

Separate least-squares regressions as a generative model:





Hierarchical modeling

A compromise between separate regressions and one overall regression.

Suppose for group $g = 1, \ldots, G$,

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta}_g + \varepsilon_{ig} = \beta_{0g} + \beta_{1g}x_{i1g} + \cdots + \beta_{Jg}x_{iJg} + \varepsilon_{ig}$$

with $\varepsilon_{ig} \sim \mathrm{N}(0, \sigma^2)$.

Notice that this model assumes a different $\boldsymbol{\beta}_g$ for each group g.

Now introduce the model component

$$\boldsymbol{\beta}_g \sim \mathrm{N}(\mu_\beta, \Sigma_\beta).$$

We will also assume a prior distribution on μ_β and Σ_β.

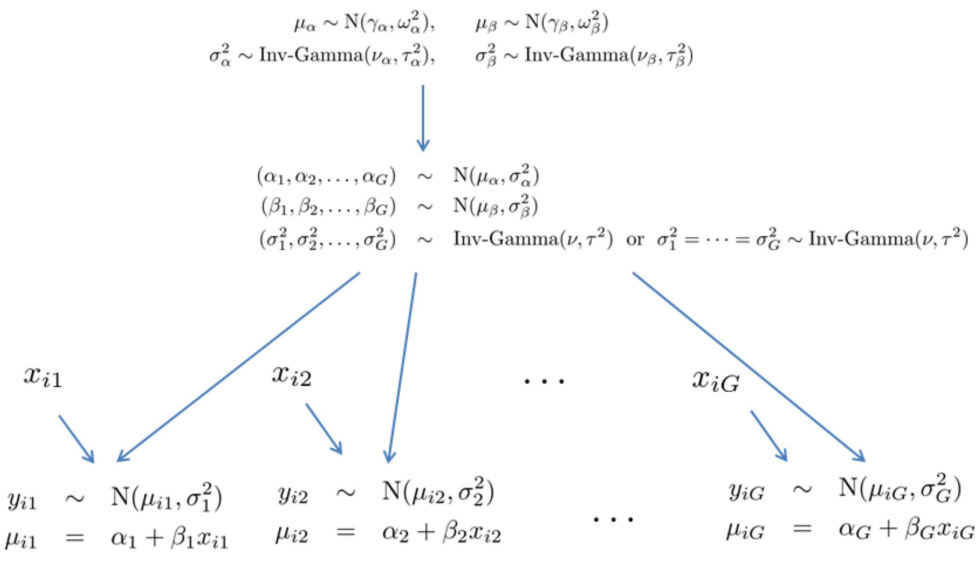
This normal "random effects" distribution $N(\mu_\beta, \Sigma_\beta)$ has an important role:

- The $\boldsymbol{\beta}_g$ are assumed distinct, but come from the same "family" (i.e., $N(\mu_\beta, \Sigma_\beta)$).

- The data across all groups inform the values of μ_β and Σ_β, which in turn means that the $\boldsymbol{\beta}_g$ are informed from other groups besides the data in group g. Sometimes called "partial pooling" of data across groups.
- Can think of the normal distribution acting like a rubber band around the $\boldsymbol{\beta}_g$. They can vary, but the random effects distribution keeps them from being too far apart.

- The random effects distribution "shrinks" the $\boldsymbol{\beta}_g$ to a common population mean.
- Shrinkage tends to be particularly noticeable when some groups have small numbers of observations relative to others.
- Assuming a population distribution on the the $\boldsymbol{\beta}_g$ with unknown covariance Σ_β can be viewed as a form of regularization.

Hierarchical least-squares regressions as a generative model:



pymc3 code

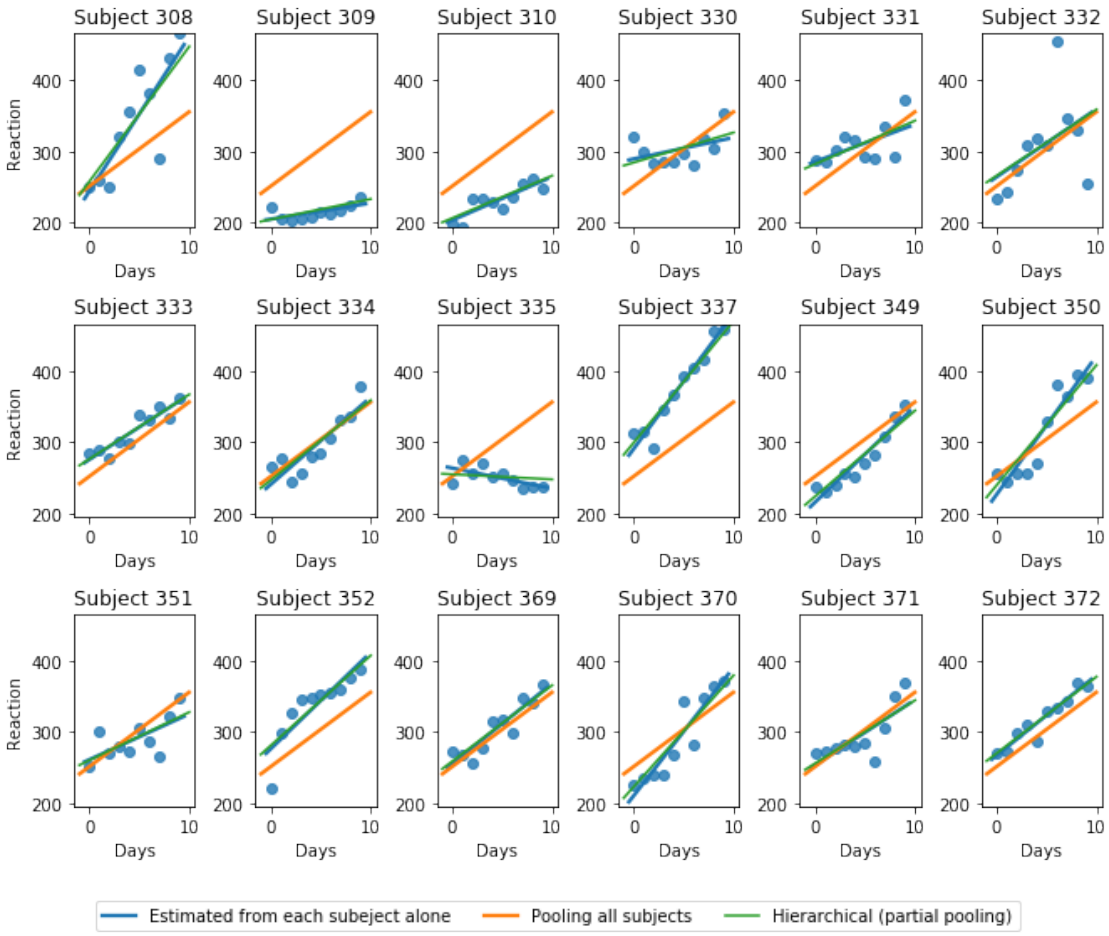
```
with pm.Model() as sleep_model:
    tau_alpha = pm.Gamma('tau_alpha', alpha=.001, beta=.001)
    tau_beta = pm.Gamma('tau_beta', alpha=.001, beta=.001)
    alpha = pm.Normal('alpha', mu=300, tau=tau_alpha,
       shape=len(raw_ids))
    beta = pm.Normal('beta', mu=10, tau=tau_beta,
```

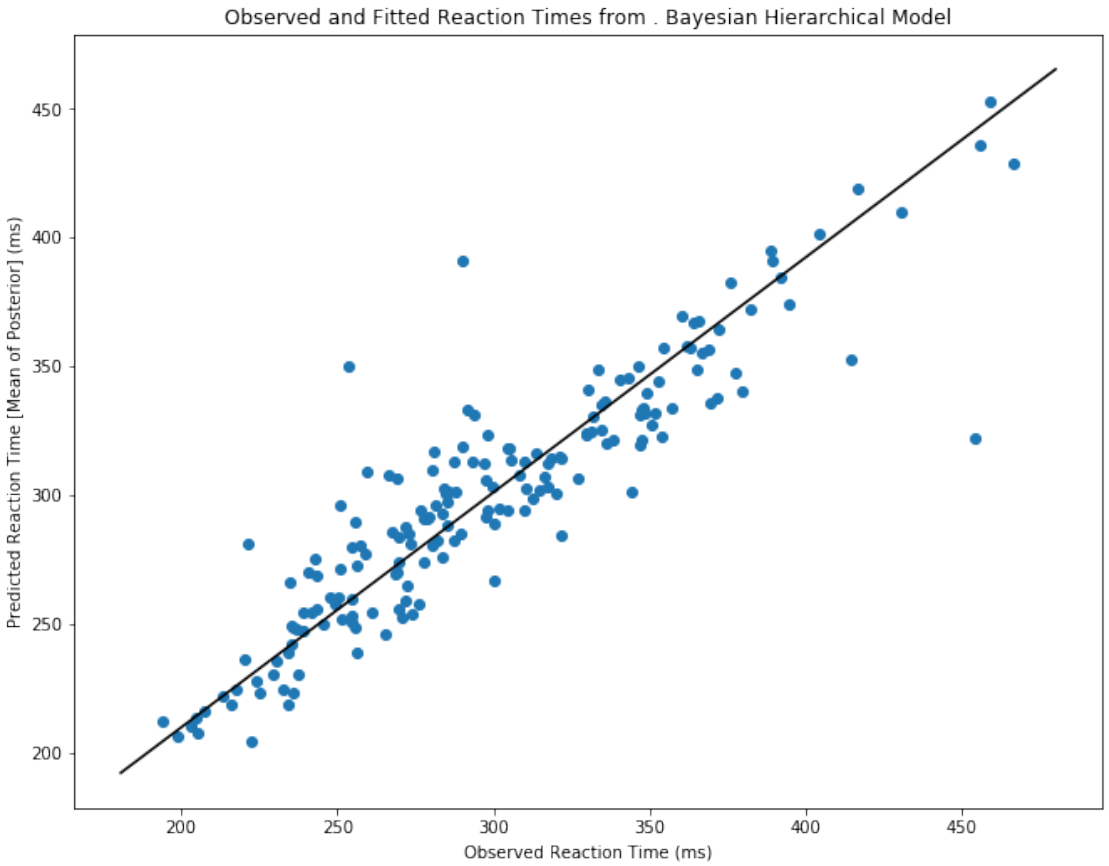
```
    shape=len(raw_ids))
intercepts = alpha[sleepstudy['SeqSubject']]
slopes = beta[sleepstudy['SeqSubject']]
mu_i = pm.Deterministic('mu_i', intercepts +
    slopes*sleepstudy['Days'])
tau_obs = pm.Gamma('tau_obs', 0.001, 0.001)
obs = pm.Normal('observed', mu=mu_i, tau=tau_obs,
   observed=sleepstudy['Reaction'])

trace = pm.sample(2000, tune=2000, target_accept=0.9)
```

Numerical summaries

	mean	sd	hpd_3%	hpd_97%	mcse_mean	mcse_sd	ess_mean	ess_sd
tau_alpha	0.000	0.000	0.000	0.001	0.000	0.000	4962.0	4841.0
tau_beta	0.033	0.015	0.010	0.060	0.000	0.000	3944.0	3156.0
alpha[0]	257.610	14.278	230.512	282.679	0.191	0.135	5581.0	5581.0
alpha[1]	204.767	14.188	179.835	233.325	0.195	0.138	5302.0	5261.0
alpha[2]	206.309	13.938	179.206	231.732	0.187	0.132	5577.0	5571.0
alpha[3]	284.217	13.814	257.316	308.996	0.182	0.129	5783.0	5720.0
...								
beta[0]	18.997	2.668	13.982	23.852	0.036	0.026	5595.0	5175.0
beta[1]	2.823	2.582	-2.183	7.532	0.033	0.030	6111.0	3708.0
beta[2]	5.993	2.525	1.093	10.622	0.034	0.027	5590.0	4462.0
beta[3]	4.326	2.578	-0.702	8.992	0.033	0.026	6240.0	4928.0
...								
tau_obs	0.002	0.000	0.001	0.002	0.000	0.000	3644.0	3601.0


Subject 308
Subject 309
Subject 310
Subject 330
Subject 331
Subject 332
Subject 333
Subject 334
Subject 335
Subject 337
Subject 349
Subject 350
Subject 351
Subject 352
Subject 369
Subject 370
Subject 371
Subject 372
Reaction
Days
Estimated from each subeject alone
Pooling all subjects
Hierarchical (partial pooling)



Final thoughts about Bayesian statistics

- Precise modeling leads to statistical inferences through a leak-proof route
- Given a probability model, Bayesian analysis makes full use of all the data
- Statistical inferences that are unacceptable must come from inappropriate modeling assumptions, not a problem in the underlying inferential mechanism

- Less attention need to be given to mathematical convenience of models, and more attention on scientific merit
- Awkward problems that Frequentists face (e.g., choice of estimators, adjustments for certain types of data) do not arise for Bayesians
- Modern computational methods (MCMC sampling from the posterior distribution) enables fitting even complex data models.