Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Lecture 26: Bayesian Statistics I

Jason Mezey

April 30, 2024 (T) 8:40-9:55

Announcements

- For those in NYC (!!) Thurs lecture (May 2) WILL BE BY ZOOM, i.e., we do not have a room (Ithaca classroom available as always)
- Reminder: last computer lab this week is optional (!!) we will show you examples of an EM algorithm for mixed models and MCMC algorithm for Bayesian inference (see lecture Thurs)!
- Last work for the class: Project and Final
  - For project (due by 11:59PM, May 7!)
  - Final will be same format as midterm (available May 11 and due by 11:59PM, May 18!) and **you will do a GWAS analysis with a linear regression with and without covariates AND a logistic regression with and without covariates (!!)**

Quantitative Genomics and Genetics - Spring 2024
BIOCB 4830/6830; PBSB 5201.01

Final exam available: Sat., May 11

Final exam due: 11:59PM, Sat., May 18

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. **YOU ARE TO COMPLETE THIS EXAM ALONE!** The exam is open book, so you are allowed to use any books or information available online (even ChatGPT or similar!), your own notes and your previously constructed code, etc. **HOWEVER <u>YOU ARE NOT ALLOWED</u> TO COMMUNICATE OR IN ANY <u>WAY ASK ANYONE FOR ASSISTANCE WITH</u> THIS EXAM IN ANY FORM e.g., DO NOT POST PUBLIC MESSAGES ON ED DISCUSSION!** (the only exceptions are Beulah, Sam, and Dr. Mezey, e.g., you MAY send us a private message on Canvas). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell's honor code if you act otherwise.

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to your advantage to attempt every part of every question.

3. A complete answer to this exam will include R code answers, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). Note there will be penalties for scripts that fail to compile (!!). Also, as always, you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!).

4. The exam must be uploaded on Canvas before 11:59PM (!!) (ET) Sat, May 18. It is your responsibility to make sure that it is in uploaded by then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to upload early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

Summary of lecture 26: Bayesian Statistics I

- Last lecture, we completed our (brief) discussion of mixed models (and EM algorithms)
- Today, we will continue our (very brief) introduction to Bayesian Statistics!

Review: Introduction to Bayesian analysis 1

- Up to this point, we have considered statistical analysis (and inference) using a Frequentist formalism
- There is an alternative formalism called Bayesian that we will now introduce in a very brief manner
- Note that there is an important conceptual split between statisticians who consider themselves Frequentist of Bayesian but for GWAS analysis (and for most applications where we are concerned with analyzing data) we do not have a preference, i.e. we only care about getting the "right" biological answer so any (or both) frameworks that get us to this goal are useful
- In GWAS (and mapping) analysis, you will see both frequentist (i.e. the framework we have built up to this point!) and Bayesian approaches applied

Review: Intro to Bayesian analysis II

- In both frequentist and Bayesian analyses, we have the same probabilistic framework (sample spaces, random variables, probability models, etc.) and when assuming our probability model falls in a family of parameterized distributions, we assume that a single fixed parameter value(s) describes the true model that produced our sample
- However, in a Bayesian framework, we now allow the parameter to have it's own probability distribution (we DO NOT do this in a frequentist analysis), such that we treat it as a random variable
- This may seem strange - how can we consider a parameter to have a probability distribution if it is fixed?
- However, we can if we have some prior assumptions about what values the parameter value will take for our system compared to others and we can make this prior assumption rigorous by assuming there is a probability distribution associated with the parameter
- It turns out, this assumption produces major differences between the two analysis procedures (in how they consider probability, how they perform inference, etc.

Review Intro to Bayesian analysis III

- To introduce Bayesian statistics, we need to begin by introducing Bayes theorem
- Consider a set of events (remember events!?) $\mathcal{A} = \mathcal{A}_1 ... \mathcal{A}_k$ of a sample space Ω (where k may be infinite), which form a partition of the sample space, i.e. $\bigcup_i^k \mathcal{A}_i = \Omega \text{ and } \mathcal{A}_i \cap \mathcal{A}_j = \emptyset \text{ for all } i \neq j$
- For another event $\mathcal{B} \subset \Omega$ (which may be Ω itself) define the Law of total probability:

$$Pr(\mathcal{B}) = \sum_{i=1}^{k} Pr(\mathcal{B} \cap \mathcal{A}_i) = \sum_{i=1}^{k} Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)$$

- Now we can state Bayes theorem:

$$Pr(\mathcal{A}_i|\mathcal{B}) = \frac{Pr(\mathcal{A}_i \cap \mathcal{B})}{Pr(\mathcal{B})} = \frac{Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)}{Pr(\mathcal{B})} = \frac{Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A})}{\sum_{i=1}^{k} Pr(\mathcal{B}|\mathcal{A}_i) Pr(\mathcal{A}_i)}$$

Introduction to Bayesian analysis IV

- Remember that in a Bayesian (not frequentist!) framework, our parameter(s) have a probability distribution associated with them that reflects our belief in the values that might be the true value of the parameter

- Since we are treating the parameter as a random variable, we can consider the joint distribution of the parameter AND a sample **Y** produced under a probability model:

$$Pr(\theta \cap \mathbf{Y})$$

- Fo inference, we are interested in the probability the parameter takes a certain value given a sample:

$$Pr(\theta|\mathbf{y})$$

- Using Bayes theorem, we can write:

$$Pr(\theta|\mathbf{y}) = \frac{Pr(\mathbf{y}|\theta)Pr(\theta)}{Pr(\mathbf{y})}$$

- Also note that since the sample is fixed (i.e. we are considering a single sample) $Pr(\mathbf{y}) = c,$ we can rewrite this as follows:

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

Introduction to Bayesian analysis V

- Let's consider the structure of our main equation in Bayesian statistics:

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

- Note that the left hand side is called the posterior probability:

$$Pr(\theta|\mathbf{y})$$

- The first term of the right hand side is something we have seen before, i.e. the likelihood (!!):

$$Pr(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$$

- The second term of the right hand side is new and is called the prior:

$$Pr(\theta)$$

- Note that the prior is how we incorporate our assumptions concerning the values the true parameter value may take

- In a Bayesian framework, we are making two assumptions (unlike a frequentist where we make one assumption): 1. the probability distribution that generated the sample, 2. the probability distribution of the parameter

Probability in a Bayesian framework

- By allowing for the parameter to have an prior probability distribution, we produce a change in how we consider probability in a Bayesian versus Frequentist perspective
- For example, consider a coin flip, with Bern(p)
  - In a Frequentist framework, we consider a conception of probability that we use for inference to reflect the outcomes as if we flipped the coin an infinite number of times, i.e. if we flipped the coin 100 times and it was "heads" each time, we do not use this information to change how we consider a new experiment with this same coin if we flipped it again
  - In a Bayesian framework, we consider a conception of probability can incorporate previous observations, i.e. if we flipped a coin 100 times and it was "heads" each time, we might want to incorporate this information in to our inferences from a new experiment with this same coin if we flipped it again
- Note that this philosophic distinction is very deep (=we have only scratched the surface with this one example)

Debating the Frequentist versus Bayesian frameworks

- Frequentists often argue that because they "do not" take previous experience into account when performing their inference concerning the value of a parameter, such that they do not introduce biases into their inference framework
- In response, Bayesians often argue:
  - Previous experience is used to specify the probability model in the first place
  - By not incorporating previous experience in the inference procedure, prior assumptions are still being used (which can introduce logical inconsistencies!)
  - The idea of considering an infinite number of observations is not particular realistic (and can be a non-sensical abstraction for the real world)
  - The impact of prior assumptions in Bayesian inference disappear as the sample size goes to infinite
- Again, note that we have only scratched the surface of this debate!

Types of priors in Bayesian analysis

- Up to this point, we have discussed priors in an abstract manner
- To start making this concept more clear, let's consider one of our original examples where we are interested in the knowing the mean human height in the US (what are the components of the statistical framework for this example!? Note the basic components are the same in Frequentist / Bayesian!)
- If we assume a normal probability model of human height (what parameter are we interested in inferring in this case and why?) in a Bayesian framework, we will at least need to define a prior:

$$Pr(\mu)$$

- One possible approach is to make the probability of each possible value of the parameter the same (what distribution are we assuming and what is a problem with this approach), which defines an improper prior:

$$Pr(\mu) = c$$

- Another possible approach is to incorporate our previous observations that heights are seldom infinite, etc. where one choice for incorporating this observations is my defining a prior that has the same distribution as our probability model, which defines a conjugate prior (which is also a proper prior):

$$Pr(\mu) \sim N(\kappa, \phi^2)$$

Constructing the posterior probability

- Let's put this all together for our "heights in the US" example
- First recall that our assumption is the probability model is normal (so what is the form of the likelihood?):

$$Y \sim N(\mu, \sigma^2)$$

- Second, assume a normal prior for the parameter we are interested in:

$$Pr(\mu) \sim N(\kappa, \phi^2)$$

- From the Bayesian equation, we can now put this together as follows:

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

$$Pr(\mu|\mathbf{y}) \propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}} \right) \frac{1}{\sqrt{2\pi\phi^2}} e^{\frac{-(\mu-\kappa)^2}{2\phi^2}}$$

- Note that with a little rearrangement, this can be written in the following form:

$$Pr(\mu|\mathbf{y}) \sim N\left(\frac{(\frac{\kappa}{\sigma^2} + \frac{\sum_i^n y_i}{\sigma^2})}{(\frac{1}{\phi^2} + \frac{n}{\sigma^2})}, (\frac{1}{\phi^2} + \frac{n}{\sigma^2})^{-1} \right)$$

Bayesian inference: estimation I

- Inference in a Bayesian framework differs from a frequentist framework in both estimation and hypothesis testing
- For example, for estimation in a Bayesian framework, we always construct estimators using the posterior probability distribution, for example:

$$\hat{\theta} = mean(\theta|\mathbf{y}) = \int \theta Pr(\theta|\mathbf{y}) d\theta \quad \text{or} \quad \hat{\theta} = median(\theta|\mathbf{y})$$

- Estimates in a Bayesian framework can be different than in a likelihood (Frequentist) framework since estimator construction is fundamentally different (!!)

Bayesian inference: estimation II

- For example, for estimation in a Bayesian framework, we always construct estimators using the posterior probability distribution, for example:

$$\hat{\theta} = mean(\theta|\mathbf{y}) = \int \theta Pr(\theta|\mathbf{y}) d\theta \quad \text{or} \quad \hat{\theta} = median(\theta|\mathbf{y})$$

- For example, in our "heights in the US" example our estimator is:

$$\hat{\mu} = median(\mu|\mathbf{y}) = mean(\mu|\mathbf{y}) = \frac{(\frac{\kappa}{\sigma^2} + \frac{n\bar{y}}{\sigma^2})}{(\frac{1}{\phi^2} + \frac{n}{\sigma^2})}$$

- Notice that the impact of the prior disappears as the sample size goes to infinite (=same as MLE under this condition):

$$\frac{(\frac{\kappa}{\sigma^2} + \frac{n\bar{y}}{\sigma^2})}{(\frac{1}{\phi^2} + \frac{n}{\sigma^2})} \approx \frac{(\frac{n\bar{y}}{\sigma^2})}{(\frac{n}{\sigma^2})} \approx \bar{y}$$

Bayesian inference: hypothesis testing

- For hypothesis testing in a Bayesian analysis, we use the same null and alternative hypothesis framework:

$$H_0 : \theta \in \Theta_0$$

$$H_A : \theta \in \Theta_A$$

- However, the approach to hypothesis testing is completely different than in a frequentist framework, where we use a *Bayes factor* to indicate the relative support for one hypothesis versus the other:

$$Bayes = \frac{\int_{\theta \in \Theta_0} Pr(\mathbf{y}|\theta) Pr(\theta) d\theta}{\int_{\theta \in \Theta_A} Pr(\mathbf{y}|\theta) Pr(\theta) d\theta}$$

- Note that a downside to using a Bayes factor to assess hypotheses is that it can be difficult to assign priors for hypotheses that have completely different ranges of support (e.g. the null is a point and alternative is a range of values)

- As a consequence, people often use an alternative "psuedo-Bayesian" approach to hypothesis testing that makes use of *credible intervals* (which is what we will use in this course)

Bayesian credible intervals (versus frequentist confidence intervals)

- Recall that in a Frequentist framework that we can estimate a confidence interval at some level (say 0.95), which is an interval that will include the value of the parameter 0.95 of the times we performed the experiment an infinite number of times, calculating the confidence interval each time (note: a strange definition...)
- In a Bayesian interval, the parallel concept is a credible interval that has a completely different interpretation: *this interval has a given probability of including the parameter value* (!!)
- The definition of a credible interval is as follows:

$$c.i.(\theta) = \int_{-c_\alpha}^{c_\alpha} Pr(\theta|\mathbf{y})d\theta = 1 - \alpha$$

- Note that we can assess a null hypothesis using a credible interval by determining if this interval includes the value of the parameter under the null hypothesis (!!)

Bayesian inference: genetic model I

- We are now ready to tackle Bayesian inference for our genetic model (note that we will focus on the linear regression model but we can perform Bayesian inference for any GLM!):

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

- Recall for a sample generated under this model, we can write:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

$$\epsilon \sim multiN(0, \mathbf{I}\sigma^2_\epsilon)$$

- In this case, we are interested in the following hypotheses:

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \qquad H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- We are therefore interested in the *marginal posterior probability* of these two parameters

Bayesian inference: genetic model II

- To calculate these probabilities, we need to assign a joint probability distribution for the prior

$$Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon)$$

- One possible choice is as follows (are these proper or improper!?):

$$Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) = Pr(\beta_\mu)Pr(\beta_a)Pr(\beta_d)Pr(\sigma^2_\epsilon)$$

$$Pr(\beta_\mu) = Pr(\beta_a) = Pr(\beta_d) = c$$

$$Pr(\sigma^2_\epsilon) = c$$

- Under this prior the complete posterior distribution is multivariate normal (!!):

$$Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon|\mathbf{y}) \propto Pr(\mathbf{y}|\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon)$$

$$Pr(\theta|\mathbf{y}) \propto (\sigma^2_\epsilon)^{-\frac{n}{2}} e^{\frac{(\mathbf{y}-\mathbf{x}\beta)^{\mathrm{T}}(\mathbf{y}-\mathbf{x}\beta)}{2\sigma^2_\epsilon}}$$

Bayesian inference: genetic model III

- For the linear model with sample:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

$$\epsilon \sim multiN(0, \mathbf{I}\sigma_\epsilon^2)$$

- The complete posterior probability for the genetic model is:

$$Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2|\mathbf{y}) \propto Pr(\mathbf{y}|\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2)Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2)$$

- With a uniform prior is:

$$Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2|\mathbf{y}) \propto Pr(\mathbf{y}|\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2)$$

- The marginal posterior probability of the parameters we are interested in is:

$$Pr(\beta_a, \beta_d|\mathbf{y}) = \int_0^\infty \int_{-\infty}^\infty Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2|\mathbf{y}) d\beta_\mu d\sigma_\epsilon^2$$

Bayesian inference: genetic model IV

- Assuming uniform (improper!) priors, the marginal distribution is:

$$Pr(\beta_a, \beta_d|\mathbf{y}) = \int_{-\infty}^{\infty}\int_{0}^{\infty} Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon|\mathbf{y})d\beta_\mu d\sigma^2_\epsilon \sim multi\text{-}t\text{-}distribution$$

- With the following parameter values:

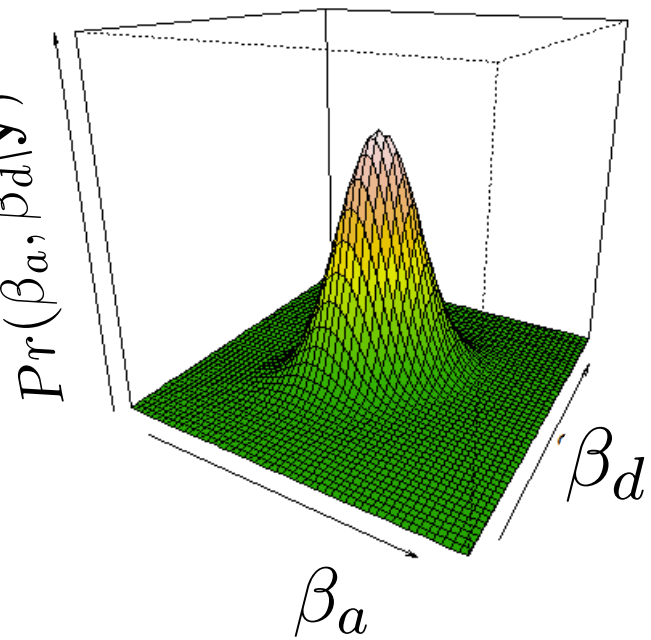
$$mean(Pr(\beta_a, \beta_d|\mathbf{y})) = \left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}} = \mathbf{C}^{-1}\left[\mathbf{X}_a, \mathbf{X}_d\right]^{\mathrm{T}}\mathbf{y} \qquad \mathbf{C} = \begin{bmatrix}\mathbf{X}_a^{\mathrm{T}}\mathbf{X}_a & \mathbf{X}_a^{\mathrm{T}}\mathbf{X}_d \\ \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_a & \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_d\end{bmatrix}$$

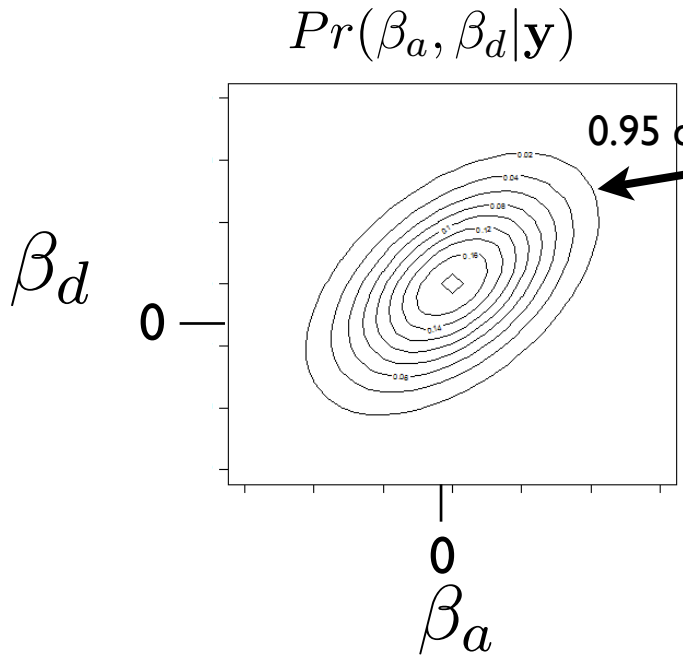
$$cov = \frac{(\mathbf{y} - [\mathbf{X}_a, \mathbf{X}_d]\left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}})^{\mathrm{T}}(\mathbf{y} - [\mathbf{X}_a, \mathbf{X}_d]\left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}})}{n-6}\mathbf{C}^{-1}$$

$$df(multi{-}t) = n - 4$$

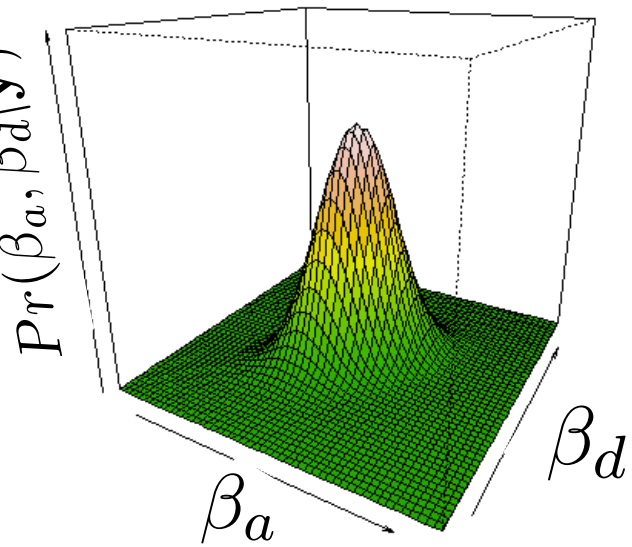
- With these estimates (equations) we can now construct a credible interval for our genetic null hypothesis and test a marker for a phenotype association and we can perform a GWAS by doing this for each marker (!!)

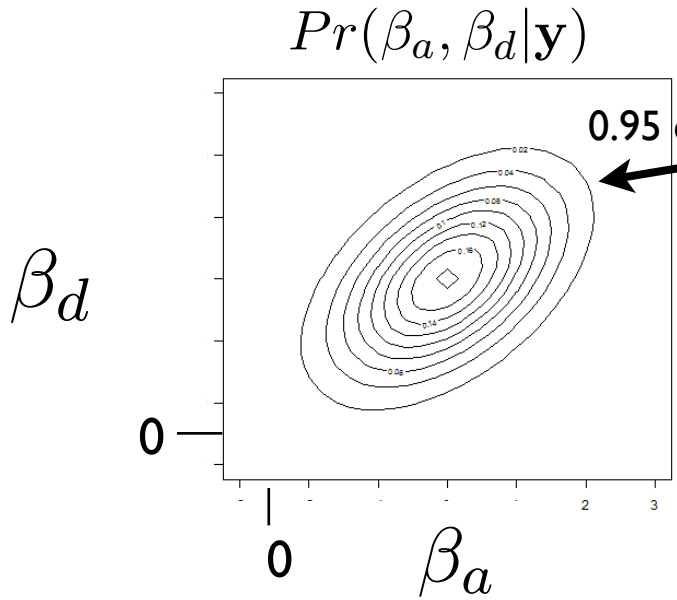
Bayesian inference: genetic model V





Cannot reject H0!





Reject H0!

Bayesian inference for more "complex" posterior distributions

- For a linear regression, with a simple (uniform) prior, we have a simple closed form of the overall posterior
- This is not always (=often not the case), since we may often choose to put together more complex priors with our likelihood or consider a more complicated likelihood equation (e.g. for a logistic regression!)
- To perform hypothesis testing with these more complex cases, we still need to determine the credible interval from the posterior (or marginal) probability distribution so we need to determine the form of this distribution
- To do this we will need an algorithm and we will introduce the Markov chain Monte Carlo (MCMC) algorithm for this purpose

Stochastic processes

- To introduce the MCMC algorithm for our purpose, we need to consider models from another branch of probability (remember, probability is a field much larger than the components that we use for statistics / inference!): *Stochastic processes*
- **Stochastic process** (intuitive def) - a collection of random vectors (variables) with defined conditional relationships, often indexed by an ordered set *t*
- We will be interested in one particular class of models within this probability sub-field: *Markov processes* (or more specifically *Markov chains*)
- Our MCMC will be a Markov chain (probability model)

Markov processes

- A *Markov chain* can be thought of as a random vector (or more accurately, a set of random vectors), which we will index with *t*:

$$X_t, X_{t+1}, X_{t+2}, ..., X_{t+k}$$

$$X_t, X_{t-1}, X_{t-2}, ..., X_{t-k}$$

- **Markov chain -** a stochastic process that satisfies the Markov property:

$$Pr(X_t, |X_{t-1}, X_{t-2}, ..., X_{t-k}) = Pr(X_t, |X_{t-1})$$

- While we often assume each of the random variables in a Markov chain are in the same class of random variables (e.g. Bernoulli, normal, etc.) we allow the parameters of these random variables to be different, e.g. at time *t* and *t*+1

- How does this differ from a random vector of an iid sample!?

That's it for today

- Next OPTIONAL lectures: Bayesian Statistics II (!!)