Bayesian Data Analysis

2020

Week 2: technical necessities – summarizing probability distributions, Monte Carlo Methods

Jarno.Vanhatalo@helsinki.fi

The previous lecture

Two types of uncertainty

- Aleatory (stochastic) uncertainty, which originates from randomness
- Epistemic (knowledge) uncertainty, which originates from the lack of knowledge
- Bayes theorem

likelihood

prior

$$p(\theta|y,n,I) = \frac{p(y|\theta,n,I)p(\theta|I)}{p(y|I)}$$

posterior

Normalization term (marginal likelihood)

Aims of the week

- Summarizing probability distributions
- Monte Carlo method
- Common probability distributions: Poisson and Binomial
- Examples
  - Estimating population parameters
  - Estimating population size

Discrete random variables

Let X be a discrete random variable.

- Probability mass function (pmf) for X is

$$\mathrm{p}(x) = P(X = x)$$

- $\sum_k p(x_k) = 1$
- Cumulative distribution function (cdf)

$$F(x) = P(X \leq x) = \sum_{k:x_k \leq x} p(x_k)$$

Continuous random variables

- A random variable X has a continuous distribution with probability density function (pdf) p(x) if

$$P(a \leq X \leq b) = \int_a^b f(x)dx, a, b \in R, a < b.$$

- Hence, if F(x) is the cumulative distribution function (cdf) of X, then

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- $\int_{-\infty}^{\infty} f(x)dx = 1$

Discrete vs. continuous

- If θ is <u>discrete</u>
  - $P(\theta = 1)$ is the <u>probability</u> that θ has value 1
  - $P(0 \leq \theta \leq 2) = P(\theta = 0) + P(\theta = 1) + P(\theta = 2)$
- If θ is <u>continuous</u>
  - $p(\theta = 1)$ is the <u>probability density</u> of θ at value 1
  - $P(0 \leq \theta \leq 2) = \int_1^2 p(\theta)\, d\theta$ is the probability that θ is between 1 and 2
  - $P(\theta = x_0) = 0,$ for all $x_0 \in \mathbb{R}$ i.e. probability of single value is zero
- <u>We will denote both the probability and probability density function with lower case p!</u>

Expectation/mean, variance, sd and p-fractiles

- $\mathrm{E}(\mathrm{X}) = \sum_k x_k \cdot p(x_k)$
- E(X)=$\int_{-\infty}^{\infty} x \cdot p(x)dx$
- E(g(X))=$\int_{-\infty}^{\infty} g(x) \cdot p(x)dx$
- Variance Var(X)=E[(X-E(X))^2] =E(X^2)-[E(X)]^2
- Standard deviation sd(X)=$\sqrt{\mathrm{Var}(X)}$
- Let $0<p<1$. A p-fractile for X is a number x_p to which applies $P(X \leq x_p) \geq p$ and $P(X \geq x_p) \geq 1-p$.
- If p=0.5 we have median.

Cumulative distribution function

- How much of cumulative probability mass is below a certain value.
- If $-\infty < \theta < \infty$
  - continuous

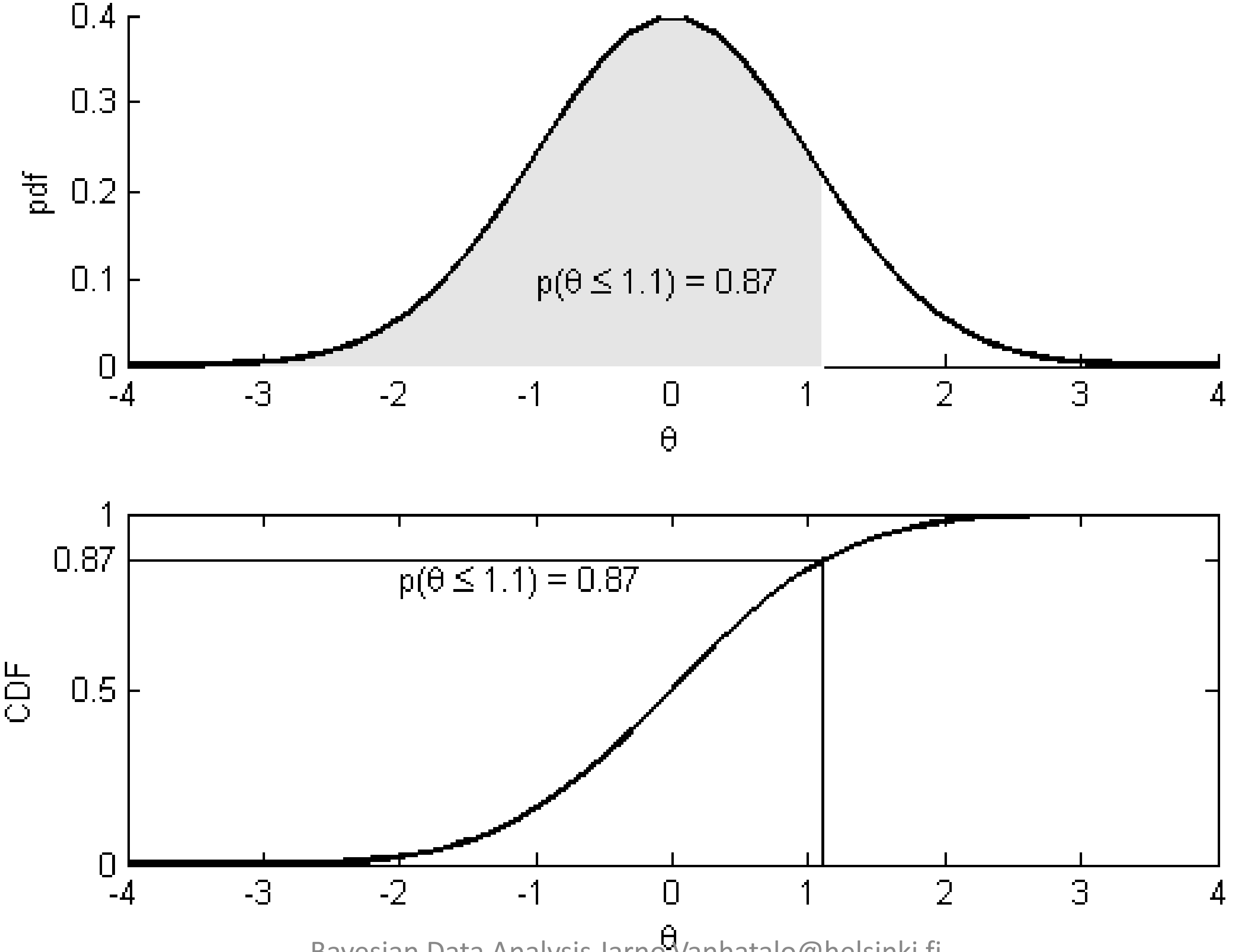
$$\mathrm{F}(a) = P(\theta \leq a) = \int_{-\infty}^{a} p(\theta) d\theta$$

  - discrete

$$= \sum_{0=-\infty}^{a} p(\theta_i)$$

- Can be defined only for one dimensional distributions
- Ready made functions for stardard distributions

Cumulative distribution function (CDF)



Discrete vs. continous

- Compare
  - Posterior predictive distribution, discrete

$$p(\tilde{y}|I,y) = \sum p(\tilde{y}|\theta_i, I)p(\theta_i|y,I)$$

  - Posterior predictive distribution, continuous

$$p(\tilde{y}|I,y) = \int p(\tilde{y}|\theta, I)p(\theta|y,I)d\theta$$

Discrete vs. continuous

$$p(\theta|y,n,I) = \frac{p(y|\theta,n,I)p(\theta|I)}{p(y|I)}$$

- Marginal likelihood in Bayes theorem
  - $p(y|I) = \begin{cases} \int p(y|\theta,n,I)p(\theta|I)\,d\theta, & \text{if } \theta \text{ is continuous} \\ \sum_{i=1}^{d} p(y|\theta_i,n,I)p(\theta_i|I), & \text{if } \theta \text{ is discrete} \end{cases}$
- Integral $\int$ is a generalization of a sum $\sum$ for continuous variables
- From now on we will use notation $\int$ for both continues and discrete variables

Representing probability

- The posterior probability distribution contains all the current information about the parameter θ
- Ideally one reports the entire distribution
- Often this is compressed to location (e.g. mean/median) and width (e.g. variance) parameters or quantiles
  - The posterior in female birth example: Beta(438; 544)
  - The statistics of Beta(α,β)

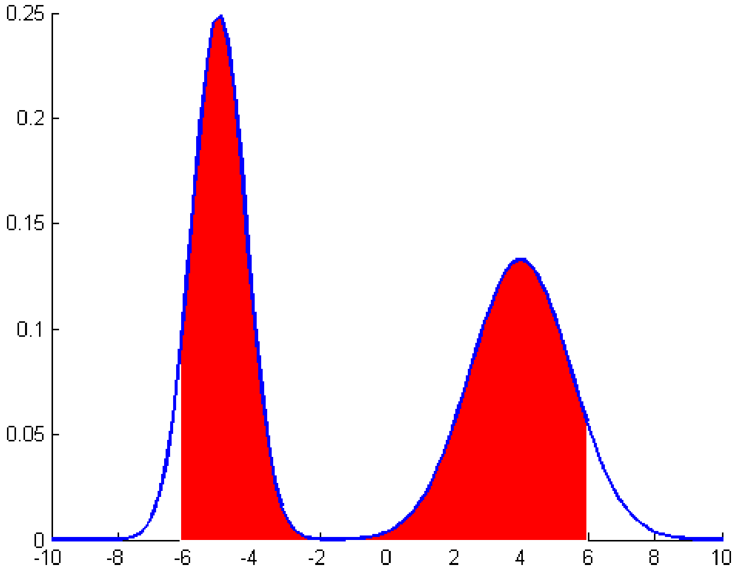
$$E[\theta] = \frac{\alpha}{\beta + \alpha} \approx 0.446$$

$$\mathrm{Var}[\theta] = \frac{\alpha\beta}{(\beta + \alpha)^2\,(\beta + \alpha + 1)} \approx 0.016^2$$

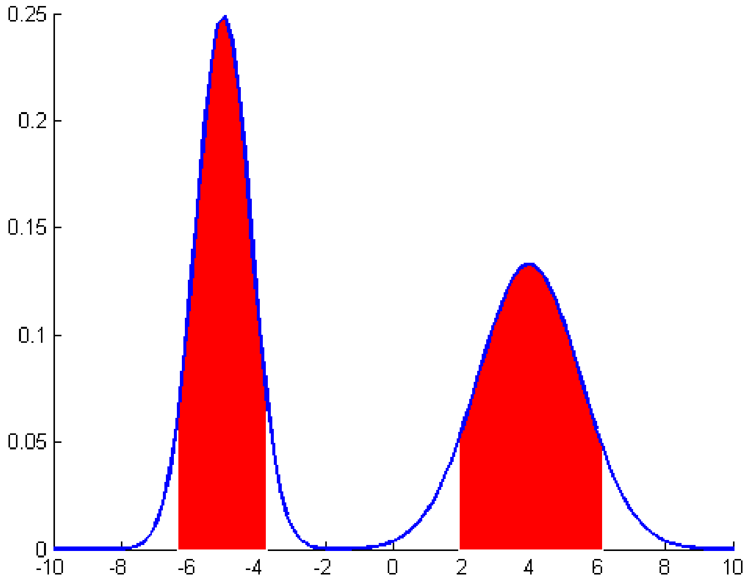
Representing probability

- Posterior intervals can be used to describe both location and width
- Posterior interval is called also credible interval (or Bayesian confidence interval)
  - different from frequentist confidence interval (even though in some special cases they match)
- Posterior interval contains a certain portion of the probability mass (e.g. 95%)
  - the interval is not uniquely defined
- Most common options are
  - central posterior interval – equal amount of probability mass below and above the interval
  - highest posterior density (HPD) interval – shortest possible interval

Representing probability



Central posterior interval



highest posterior region

Representing probability

- Posterior probability, Bayesian p-value (different from the frequentist p-value)
  - The amount of probability mass in certain area

$$p(a < \theta < b|y, I) = \int_a^b p(\theta|y, I)\, d\theta$$

Representing probability

- For most of the standard distributions mean, median and standard deviation can be evaluated analytically
  - Also quantiles and intervals are easily available from cumulative density
  - However, usually these are not trivial to solve for general posterior distribution
- In general case we need approximations
  - e.g. (Markov chain) Monte Carlo

Example: estimating female birth ratio

Example: Estimating population size with mark-recapture

See exercise 2 of week 2

Example: mark-recapture study

- Mark-recapture is a method to estimate the size of a population *N*
  - mark M = 25 animals
  - Let the marked animals mix with rest of the population
  - capture C = 20 animals
  - Count the number of recaptured animals $R =$?
- Assumptions, e.g.,
  - Time between consecutive captures enough for ”perfect mixing”
  - The behaviour and capture probability do not change due to marking
  - The population is closed between the captures
    - Animals do not die and no births either
    - No immigration / emigration
  - Marks are not lost

Example: mark-recapture study

- Estimate the total number of balls in the bag
  - M=25 marked balls
  - C= 20 Number of drawn balls at the second time
  - R = 3 recaptured balls
- Observation model, (lets approximate with binomial)

$$\mathrm{p}(R|M,N,C) = \mathrm{Bin}(R|M/N,C)$$

- Prior for the number of balls

$$\mathrm{p}(N) = ?$$

- The posterior probability of the number of balls

$$\mathrm{p}(N|R,M,C) \propto \mathrm{p}(R|M,N,C) \times \mathrm{p}(N)$$

Example: mark-recapture

- The mark-recapture model was clearly wrong. E.g.
  - Observations were not Binomially distributed since we did not replace the balls
    - Right model would have been Hyper-geometric distribution
    - In this case the difference between these two distributions is negligible
  - Prior distribution did not encode our prior thoughts perfectly
    - Mathematical description (almost) always simplifies
- Despite the model deficiencies we obtained an estimate that contains the most essential uncertainties
  - Next we could improve the model
  - > model validation and comparison in later lectures

Monte Carlo methods

- ”Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.”

- Basic idea is to draw simulation samples from the distribution
- These samples are observations from the distribution
- With the samples we can
  - evaluate expectations and standard deviations
  - evaluate quantiles
  - draw histograms
  - marginalize
  - etc.

Monte Carlo

- Racall Strong law of large numbers: let $X_1, \dots, X_n$ be a sequence of independent random variables having a common distribution p(x) and let E(X_i)= μ. Then with probability 1,

$$\frac{X_1 + \cdots + X_n}{n} \rightarrow \mu \; as \; n \rightarrow \infty.$$

- Hence one way to estimate $E[\theta|y]$ is
    - $\theta^i \sim p(\theta|y)$ (θ^i is sampled from the posterior)
    - Expected value of the unknown parameter

$$E[\theta|y] = \int \theta p(\theta|y) d\theta \approx \frac{1}{S} \sum_{i=1}^{S} \theta^i$$

Monte Carlo

- Posterior probability

$$\mathrm{p}(\mathrm{a} < \theta < b|y) \approx \frac{1}{S}\sum_{i=1}^{S} I\left(\mathrm{a} < \theta^{i} < b\right)$$

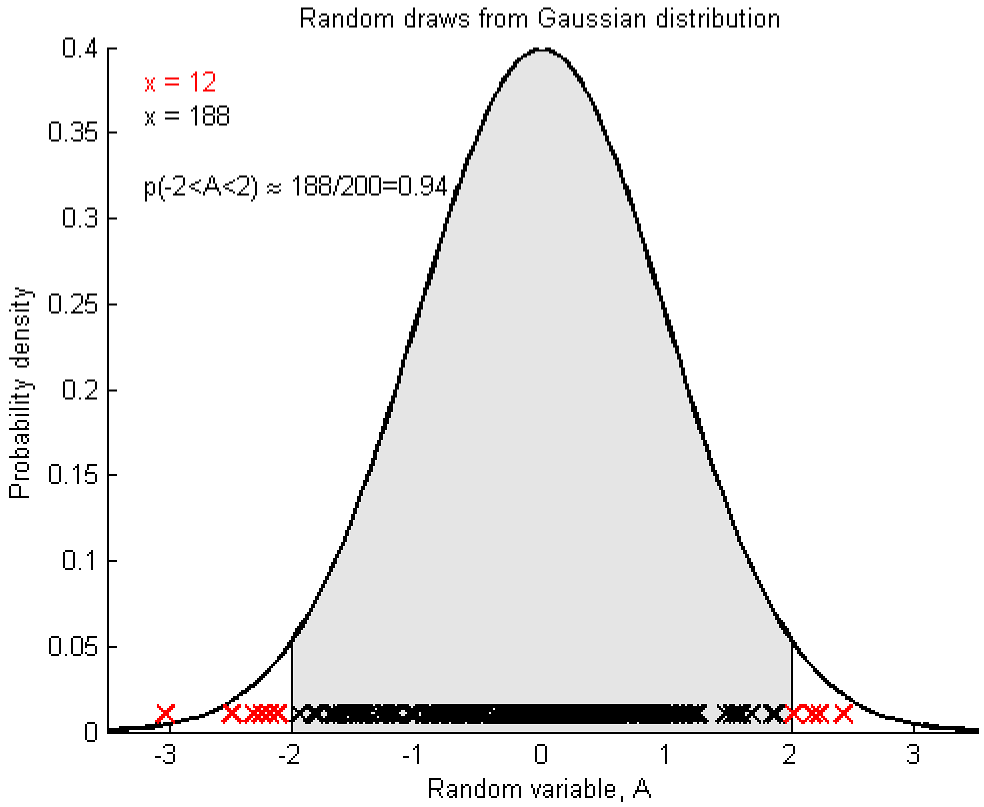
  - $I(\mathrm{a} < \theta^{i} < b)$=1, if $\mathrm{a} < \theta^{i} < b$
  - $I(\mathrm{a} < \theta^{i} < b)$=0, otherwise

- In general, fewer simulations are needed to estimate e.g. posterior medians or probabilities near 0.5 than e.g. extreme quantiles, posterior means or probabilities for rare events.

Example
- Monte Carlo

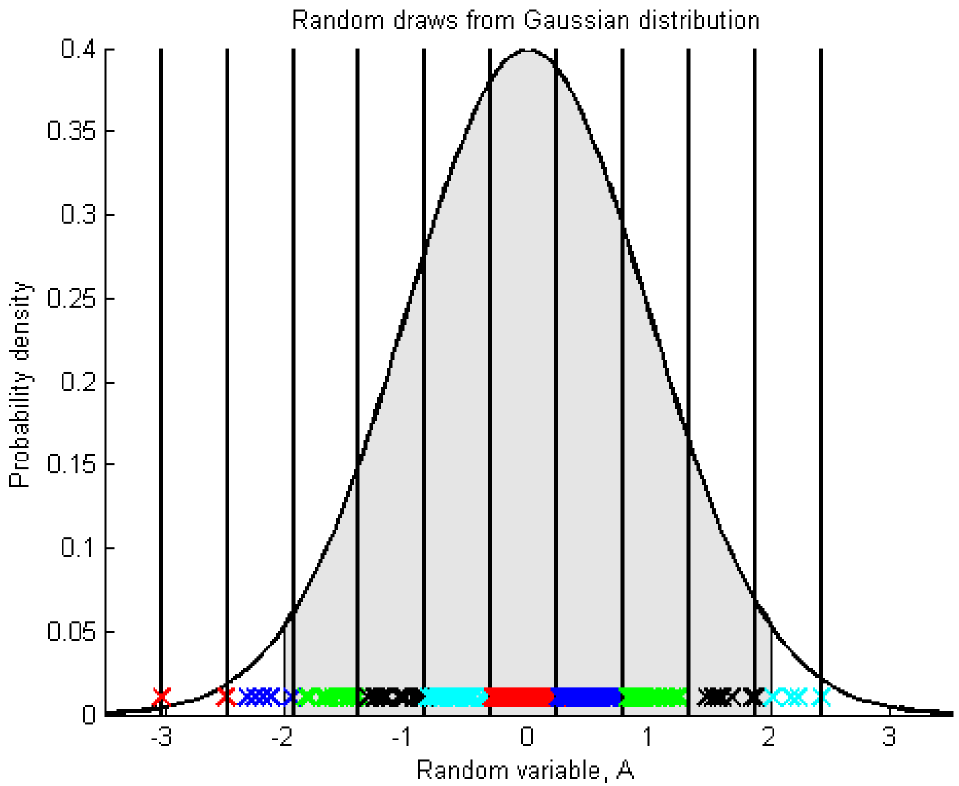
Example: Monte Carlo

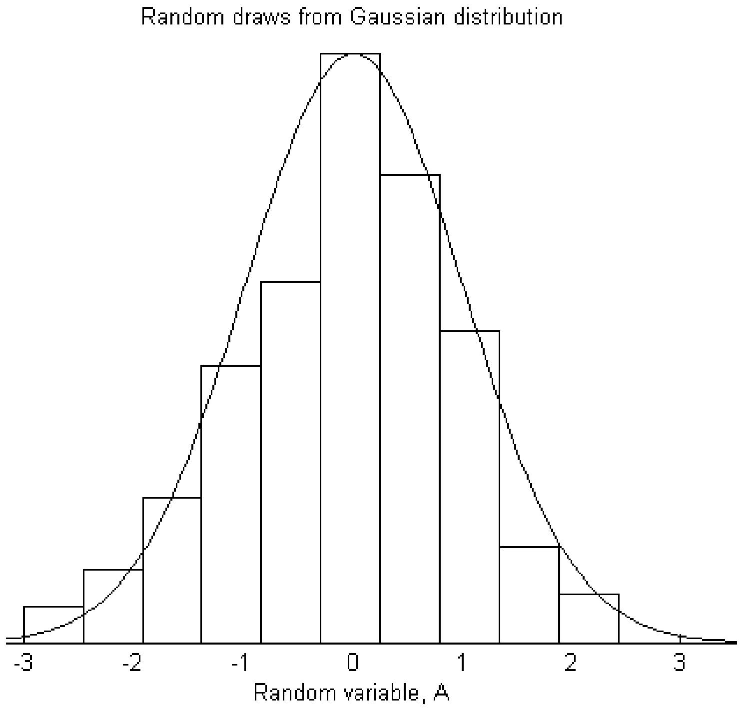
- Random draws from a distribution and approximating probability



Example: Monte Carlo

- Approximating the distribution with a histogram





Direct simulation

- Direct simulation from a distribution produces independent samples
- Build in functions available for standard distributions
  - e.g. Gaussian, Binomial, Beta,... (In R rnorm, rbinom, rbeta,…)
  - In practice computer programs produce pseudo random numbers
    - problem only in very specific situations -> Good enough in practice

Monte Carlo and change of variable

- Consider a case that you are interested in posterior distribution of parameter $\mathrm{u} = \mathrm{f}(\theta)$ where $\mathrm{f}(.)$ is a function With Monte Carlo you can
  - Sample $\theta^i \sim p(\theta|y)$
    (θ^i is sampled from the posterior)
  - For each θ^i calculate $u^i = f(\theta^i)$
  - Then u^i will be distributed as

$$u^i \sim p(u|y)$$

Marginalization with Monte Carlo

- Consider a joint distribution of two variables

$$p(\alpha, \beta)$$

The marginal distributions of α and β are

$$p(\alpha) = \int p(\alpha, \beta) d\beta$$

$$p(\beta) = \int p(\alpha, \beta) d\alpha$$

These are often hard or impossible to calculate in closed form. However, given a sample from the joint distribution, we can extract a sample from each of the marginals by simply taking only the samples of the corresponding variables.

For example consider a matrix A of samples from the above joint distribution

$$A = \begin{bmatrix} \alpha^1, \beta^1 \\ \vdots \\ \alpha^m, \beta^m \end{bmatrix} \quad \text{so that} \quad (\alpha^i, \beta^i) \sim p(\alpha, \beta) \text{ for all } i$$

Then $A_{\cdot,1} = \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^m \end{bmatrix}$ where $\alpha^i \sim p(\alpha)$ for all i

Sampling from joint distribution through conditionals

- Consider a joint distribution of two variables

$$p(\alpha, \beta) = p(\alpha|\beta)p(\beta)$$

  Direct sampling from this joint distribution is often hard whereas sampling from the marginal and conditional might be easier. In that case we can sample from the joint as follows.
  - Repeat for $i = 1, \dots, m$
    - Sample $\beta^i \sim p(\beta)$
    - Sample $\alpha^i \sim p\left(\alpha|\beta^i\right)$
    - Set $A_{i,\cdot} = [\alpha^i, \beta^i]$
  - After this A is a matrix where each row $A_{i,\cdot} \sim p(\alpha, \beta)$ and each column corresponds to samples from marginal of either α or β
- Note only the row-wise samples are from the joint but for example if $i \neq j$ then $\left(A_{i,1}, A_{j,2}\right) = (\alpha^i, \beta^j)$ is not sample from $p(\alpha, \beta)$!

Prediction with Monte Carlo

- Posterior predictive distribution

$$p(\tilde{y}|I, y) = \int p(\tilde{y}|\theta, I)\, p(\theta\,|\,y, I)\, d\theta$$

- With Monte Carlo.
  - Repeat for *i=1,…,m*
    - Sample $\theta^{i} \sim p(\theta\,|\,y, I)$
    - Sample $\tilde{y}^{i} \sim p(\tilde{y}|\theta^{i}, I)$
  - Use $\tilde{y}^{1}, \ldots \tilde{y}^{i}$ as an approximation for $p(\tilde{y}|I, y)$

Probability distributions of the week

- **Beta** y~Beta(α, β)

For 0 ≤ y ≤ 1 and shape parameters α, β > 0. A conjugate prior for Binomial disteribution.

$$p(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

- **Poisson** y~Poisson(θ)

Event can occur 0,1,2,… times in an interval. Let θ be the average number of events in an interval (rate parameter). The probability of observing y events in an interval is

$$p(y|\theta) = e^{-\theta}\frac{\theta^y}{y!}$$

- **Gamma** y~Gamma(α, β)

For y > 0 with shape parameter α > 0 and inverse scale parameter , β > 0. A conjugate prior for Poisson rate

$$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

About distributions and normalization

- The posterior is often written as

$$p(\theta|y,n,I) = \frac{p(y|\theta,n,I)p(\theta|I)}{p(y|I)}$$

$$p(\theta|y,n,I) \propto p(y|\theta,n,I)p(\theta|I)$$

- Unnormalized distributions are often used since
  - The normalization can be calculated afterwards
  - One uses computational methods that work for unnormalized distributions
- Naming convention for the distribution
  - If $\int \pi(\theta)d\theta = \infty$, $\pi(\theta)$ is improper
  - If $\int q(\theta)d\theta = Z \neq 1$, $q(\theta)$ is unnormalized
  - If $\int p(\theta)d\theta = 1$, $p(\theta)$ is proper and normalized

Integrals in Bayesian analysis

- The normalization term

$$p(y|I) = \int p(y|\theta,n,I)p(\theta|I)\,d\theta$$

- Posterior predictive distribution

$$p(\tilde{y}|I,y) = \int p(\tilde{y}|\theta,I)p(\theta|y,I)d\theta$$

- Marginalization (compare to normalization term)

$$p(\theta_1|y) = \int p(\theta_1,\theta_2|y)d\theta_2$$

- How to solve these in practice?

The Bayes theorem

likelihood

prior

$$p(\theta|y,n,I) = \frac{p(y|\theta,n,I)\,p(\theta|I)}{p(y|I)}$$

posterior

Normalization term (marginal likelihood)

- The normalization term $p(y|I) = \int p(y|\theta,n,I)\,p(\theta|I)\,d\theta$ is often hard to evaluate
- The unnormalized posterior $p(\theta|y,n,I) \propto p(y|\theta,n,I)\,p(\theta|I)$ is often enough for computations

This week

- Practicalities with R
- Calculations with discrete variable model (Mark recapture)
- Calculations with analytically tractable model (female birth analysis with binomial model)
- Monte Carlo method

Next week

- Markov chain Monte Carlo methods
- First steps with Stan