UPPSALA
UNIVERSITET

Bayesian Statistics and Data Analysis
Lecture 1

Måns Magnusson
Department of Statistics, Uppsala University

Thanks to Aki Vehtari, Aalto University

UPPSALA UNIVERSITET

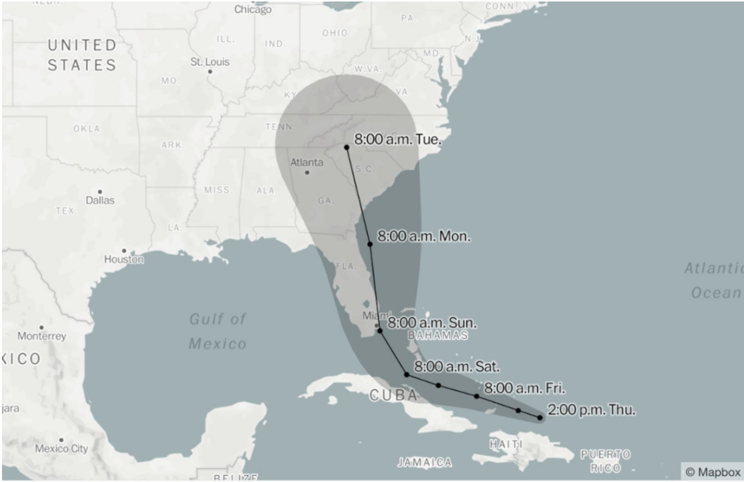
Section 1

Introduction

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Decision making in case of uncertainties



UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Bayesian Analysis

- Bayesian probability theory
  - uncertainty is presented with probabilities
  - probabilities are updated based on new information
- Thomas Bayes (170?–1761)
  - English nonconformist, Presbyterian minister, mathematician
  - considered the problem of *inverse probability*
    - significant part of the Bayesian theory
- Bayes did not invent all, but was first to solve problem of inverse probability in special case
- Modern Bayesian theory with rigorous proofs developed in 20th century

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Term Bayesian used first time in mid 20th century

- Earlier there was just "probability theory"
  - concept of the probability was not strictly defined, although it was close to modern Bayesian interpretation
  - in the end of 19th century there were increasing demand for more strict definition of probability (mathematical and philosophical problem)
- In the beginning of 20th century frequentist view gained popularity
  - accepts definition of probabilities only through frequencies
  - does not accept inverse probability or use of prior
  - gained popularity due to apparent objectivity and "cook book" like reference books
- R. A. Fisher used in 1950 first time term "Bayesian" to emphasize the difference to general term "probability theory"
  - term became quickly popular, because alternative descriptions were longer

Uncertainty and probabilistic modeling

- Two types of uncertainty: aleatoric and epistemic
- Representing uncertainty with probabilities
- Updating uncertainty

Two types of uncertainty

- Aleatoric uncertainty due to randomness
  - we are not able to obtain observations which could reduce this uncertainty
- Epistemic uncertainty due to lack of knowledge
  - we are able to obtain observations which can reduce this uncertainty
  - two observers may have different epistemic uncertainty

UPPSALA UNIVERSITET

Section 2

Bayesian Statistical Inference

Probability in Bayesian Statistics

- Uncertainty and knowledge is expressed as probability distributions, why?
  - Analogy to physical randomness
  - Knowledge as coherence of bets (the Dutch book)
- Everyone can have their own 'subjective' uncertainty, e.g. P(rain tomorrow), P(Magdalena Andersson will be the next primeminister)
- Frequency arguments can be difficult in some situations: P(other life in the universe)
- Bayesian epistemiology
  - State of knowledge is a probability distribution,
  - e.g. unlike Popperian aproaches, we can talk about P(black swan)
  - Research in Philisophy of Science

UPPSALA UNIVERSITET

Statistical inference

- Draw conclusions of unobserved entities, based on data
- Different types of unobserved entities
  - potentially observed: future observations, treatment effects
  - parameters: data-generating process (e.g. regression coefficients)

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Parameters, data, and predictions

- Model parameters: $\theta = (\theta_1, ..., \theta_p)$
- Observed data: $y = (y_1, ..., y_n)$
  - y_i can be a vector and is assumed to be random
- Potentially observed data: $\tilde{y} = (\tilde{y}_1, ..., \tilde{y}_m)$
- Observed (known) covariates: x
- We assume *exchangeability* of the observations:

$$p(y_1, ..., y_n)$$

Notation

- $p(\cdot|\cdot)$: conditional pdf/pmf
- $p(\cdot)$: marginal pdf/pmf
- $P(\cdot)$ or $Pr(\cdot)$: probability, e.g. $P(\theta > 0) = \int_0^{\infty} p(\theta) d\theta$
- Random variable: $\theta \sim N(\mu, \sigma)$
- pdf/pmf: $p(\theta) = N(\theta|\mu, \sigma)$

The basic steps of Bayesian inference

1. Setting up a full probability model $p(y|\theta) \cdot p(\theta) = p(y, \theta)$
2. Conditioning on observed data y to calculate the posterior distribution $p(\theta|y)$
3. Evaluate the model. If not satisfied, go back to 1.

UPPSALA UNIVERSITET

Bayesian Inference I

- Bayesian conclusions about θ or $\tilde{y}$ are made using probabilities, conditional on data y
- We state our uncertainty about θ or $\tilde{y}$ as distributions
  - potentially observed: $p(\tilde{y}|y)$
  - parameters: $p(\theta|y)$
- We implicitly condition on x, i.e. $p(\theta|y) = p(\theta|y, x)$

Bayesian Inference: Setting up the model

- Start out with a model: a *joint distribution* for data and parameters:

$$p(y, \theta) = p(y|\theta)p(\theta)$$

- $p(y|\theta)$ is our data model, and when conditioned on y, the *likelihood*
- $p(\theta)$ is our prior distribution for our parameters

UPPSALA UNIVERSITET

Bayesian Inference: Computing the posterior

- Conditioning on data y, using *Bayes theorem*, we can the compute the *posterior distribution*

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

where

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

- $p(y)$ is the *marginal likelihood*
- $p(\theta|y)$ summarize our knowledge about θ
- Bayesian statistics obey the *likelihood principle*: data only affects $p(\theta|y)$ through the likelihood $p(y|\theta)$

Predictions

- How to do inference on an unknown (potential) observable $\tilde{y}$? E.g. a future observation
- We use our *data* model and our posterior and *marginalize* over the uncertainty

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

- The *posterior predictive distribution*
- 'An average of conditional predictions over the posterior distribution'

Example: Updating uncertainty

UPPSALA UNIVERSITET

- Probability of red $\frac{\#\text{red}}{\#\text{red}+\#\text{yellow}} = \theta$
- $p(y = \text{red}|\theta) = \theta$ aleatoric uncertainty
- $p(\theta)$ epistemic uncertainty
- Picking many chips updates our uncertainty about the proportion
- $p(\theta|\text{y} = \text{red}, \text{yellow}, \text{red}, \text{red}, \ldots) =?$
- Bayes rule $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$

Model vs. Likelihood

- Bayes rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

- Data model: $p(y|\theta)$ as a function of y given fixed θ describes the aleatoric uncertainty
- Likelihood: $p(y|\theta) = L(\theta|y)$ as a function of θ given fixed y provides information about epistemic uncertainty, but is not a probability distribution, why?
- Bayes rule combines the likelihood with prior uncertainty $p(\theta)$ and transforms them to updated posterior uncertainty $p(\theta|y)$

Example application: Effects on roaches

- Question: Effect of treamtments on captured roaches
- Outcome (y): The number of roaches
- Coeffcients (x): Treatment, Senior home, Pre-treatment number of roaches
- Data model:

$$p(y|\theta) = Po(\lambda) = \frac{1}{y!}\lambda^{y}\exp(\lambda)$$

where

$$\lambda = \exp(\alpha + \beta X)$$

- Prior:

$$p(\theta) = p(\alpha) = p(\beta_k) = N(\theta|\mu, \sigma)$$

- Posterior:

$$p(\alpha, \beta|y) \propto N(\alpha|\mu, \sigma)\prod^{K} N(\beta_k|\mu, \sigma)\prod_{i=1}^{N} Po(\exp(\alpha + \beta x))$$

- Predictive distribution:

$$p(\tilde{y}|y, \tilde{x}) = \int Po(\exp(\alpha + \beta\tilde{x}))p(\alpha, \beta|y)d\alpha d\beta$$

UPPSALA UNIVERSITET

Section 3

Probabilistic Modeling

The art of probabilistic modeling

- Subjectivity: we need to specify both $p(\theta)$ and $p(y|\theta)$
- The art of probabilistic modeling is to describe in a mathematical form (model and prior distributions) what we already know and what we don't know
- "Easy" part is to use Bayes rule to update the uncertainties
  - computational challenges
- Other parts of the art of probabilistic modeling are, for example,
  - model checking/assesment: is data in conflict with our prior knowledge?
  - model choice: which model should we use?
  - presentation: presenting the model and the results to the application experts

Example applications

- Galaxy clusters for cosmology
- Coagulation of blood
- Gene regulation
- Pharmacokinetics and -dynamics
- Decision support
- Effects of nutrition for diabetes
- Evolutionary anthropology
- Clinical trial designs
- Daily demand for gas
- Brain structure trees
- School enrollment
- Sports
- Product demand
- Cocoa bean fermentation
- Marine propulsion power
- Alcohol consumption trends
- Flood probability
- Instantaneous heart rate distributions
- Drug dosing regimens in pediatrics
- Human T stem cell memory cells
- Fairness in university admission policies
- Destruction of bacteria and bacterial spores under heat

Benefits and problems when going Bayesian

UPPSALA UNIVERSITET

- Benefits of Bayesian approach
  - integrate over uncertainties to focus to interesting parts
  - straight-forward predictive distributions
  - use relevant prior information
  - hierarchical models
  - model checking and evaluation
  - easier interpretation of uncertainty intervals
- Complications of Bayesian approach
  - most models does not have nice analytical posteriors
  - we need to *approximate* our posterior
  - can be computationally costly

Section 4

Bayesian Computation

UPPSALA UNIVERSITET

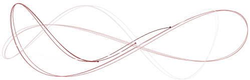
Computation

We need to be able to compute expectations with respect to posterior distribution $p(\theta|y)$

$$\mathrm{E}_{\theta|y}\left[g(\theta)\right] = \int p(\theta|y)g(\theta)d\theta$$

- Analytic
  - only for very simple models
- Monte Carlo, Markov chain Monte Carlo
  - generic
- Distributional approximations
  - e.g. Laplace, variational inference
  - less generic, but can be much faster with sufficient accuracy

Probabilistic programming



Enables agile workflow for developing probabilistic models

language – automated inference – diagnostics

mc-stan.org

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Binomial model for treatment/control comparison

- Two groups of patients: treatment and control
  - Binary outcome
  - Is the treatment useful?

Binomial model for treatment/control comparison

```
data {
  int<lower=0> N1;
  int<lower=0> y1;
  int<lower=0> N2;
  int<lower=0> y2;
}
parameters {
  real<lower=0,upper=1> theta1;
  real<lower=0,upper=1> theta2;
}
model {
  theta1 ~ beta(1,1);
  theta2 ~ beta(1,1);
  y1 ~ binomial(N1, theta1);
  y2 ~ binomial(N2, theta2);
}
generated quantities {
  real oddsratio;
  oddsratio = (theta2/(1-theta2))/(theta1/(1-theta1));
}
```

Binomial model for treatment/control comparison

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

RStanARM

```
fit_bin2 <- stan_glm(y/N ~ grp2, family = binomial(),
                     data = d_bin2, weights = N)
```

Modeling nature

- Drop a ball from different heights and measure time
  - Newton
  - air resistance, air pressure, shape and surface structure of the ball
  - relativity
- Taking into account the accuracy of the measurements, how accurate model is needed?
  - often simple models are adequate and useful
  - *All models are wrong, but some of them are useful*, George P. Box

UPPSALA UNIVERSITET

Recap: Uncertainty and probabilistic modeling

- Two types of uncertainty: aleatoric and epistemic
- Representing uncertainty with probabilities
- Updating uncertainty

Rest of the course

UPPSALA UNIVERSITET

- Basic models which can be used as building blocks
- Basic computation of posterior distributions
- Typical simple scientific data analysis cases
  - e.g. comparison of treatments
- Presentation of the results

UPPSALA UNIVERSITET

- Introduction
- Bayesian Statistical Inference
- Probabilistic Modeling
- Bayesian Computation

Ambiguous notation in statistics

In $p(y|\theta)$

- y can be variable or (observed) value

  we could clarify by using $p(Y|\theta)$ or $p(y|\theta)$
- θ can be variable or value

  we could clarify by using $p(y|\Theta)$ or $p(y|\theta)$
- p can be a discrete or continuous function of y or θ

  we could clarify by using P_Y, P_Θ, p_Y or p_Θ
- $P_Y(Y|\Theta = \theta)$ is a probability mass function, sampling distribution, observation model
- $P(Y = y|\Theta = \theta)$ is a probability
- $P_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)
- $p_Y(Y|\Theta = \theta)$ is a probability density function, sampling distribution, observation model
- $p(Y = y|\Theta = \theta)$ is a density
- $p_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)