Introduction to Bayesian Analysis
Part I – Theory

Jean-Philippe Gauvin
Université de Montréal

April 24 2014

Goals for Today

- Why bother with bayesian?
- Who uses it? And for what?
- What is it? A look a the theory
- Where to go from there?
- Practical uses – Charles' presentation

What the frequentists do wrong

Some bayesian critiques of non-bayesian statistics:

- The assumptions make no sense!
  - Often, repeated sampling is illogical

What the frequentists do wrong

Some bayesian critiques of non-bayesian statistics:

- The assumptions make no sense!
  - Often, repeated sampling is illogical
- Frequentists act like Bayesians, but they shouldn't!
  - Interpretation of statistical significance is bad

What the frequentists do wrong

Some bayesian critiques of non-bayesian statistics:

- The assumptions make no sense!
  - Often, repeated sampling is illogical
- Frequentists act like Bayesians, but they shouldn't!
  - Interpretation of statistical significance is bad
- Frequentist statistics are software-driven
  - `reg y x1 x2 x3`
  - `m1 <- lm(y ~ x1 + x2 + x3, data=mydata)`

What the bayesians do right

Many advantages:

- The inference is simple and direct
  - Leads to a *posterior probability statement*. No extra step needed to test hypothesis.

What the bayesians do right

Many advantages:

- The inference is simple and direct
  - Leads to a *posterior probability statement*. No extra step needed to test hypothesis.
- Easy to investigate causal heterogeneity
  - θ parameter is always random. Easy to do hierarchical models.

What the bayesians do right

Many advantages:

- The inference is simple and direct
  - Leads to a *posterior probability statement*. No extra step needed to test hypothesis.
- Easy to investigate causal heterogeneity
  - θ parameter is always random. Easy to do hierarchical models.
- Easy to manage missing values
  - Can be treated a another random parameter!

What the bayesians do right

Many advantages:

- The inference is simple and direct
  - Leads to a *posterior probability statement*. No extra step needed to test hypothesis.
- Easy to investigate causal heterogeneity
  - θ parameter is always random. Easy to do hierarchical models.
- Easy to manage missing values
  - Can be treated a another random parameter!
- Let's you incorporate prior information
  - Prior evidence can be research or qualitative information, etc.

Who uses bayesian statistics?

Top 10 Best Rated (bayesian estimate) (Top 50)

#	title	rating	nb. votes
1	Monster (manga)	9.24	720
2	Berserk (manga)	9.24	1143
3	Nausicaä of the Valley of the Wind (manga)	9.22	590
4	Vinland Saga (manga)	9.11	193
5	Yokohama Kaidashi Kikou (manga)	9.11	336
6	20th Century Boys (manga)	9.04	561
7	Fullmetal Alchemist (manga)	8.95	1514
8	Akira (manga)	8.94	537
9	Yotsuba&! (manga)	8.94	729
10	Vagabond (manga)	8.93	309

This rating only includes titles that have at least 4 votes. The bayesian estimate is a statistical technique used to reduce the noise due to low sample counts. In effect, the less a title has votes, the more it is pulled towards the mean (7.806). In other words, these are the titles that **many** people agree are great. (formula)

bayesian rating = $(v \div (v+m)) \times R + (m \div (v+m)) \times C$

where:

R = average for the manga

v = number of votes for the manga

m = minimum votes required to be listed (currently 4)

C = the mean vote across the whole report (currently 7.806)

But also:

- Weather forecasting
- Health care/Clinical trials
- Criminal justice/Army

Objective vs. Subjective Probability

- Frequentists or classical statisticians see probability as objective and defined by repeated trials of the same process, such as flipping a coin a thousand times to find the probability of a Head on the next flip.

- Bayesians see probability as subjective, capturing the odds a person would place on an event occurring. In other terms: an individual's degree of belief in a statement. It can be influenced by personal beliefs, prior evidence, etc.

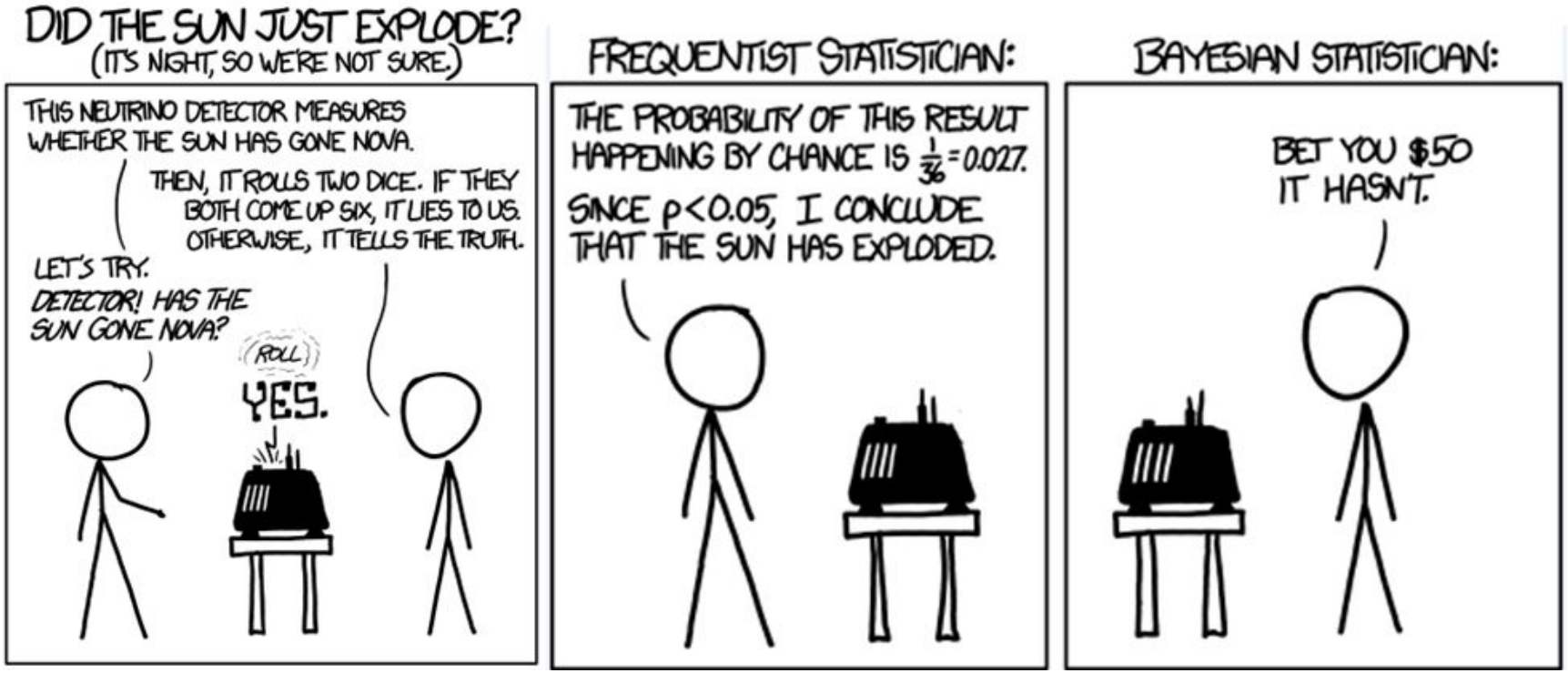
Objective vs. Subjective Probability

- Frequentists or classical statisticians see probability as objective and defined by repeated trials of the same process, such as flipping a coin a thousand times to find the probability of a Head on the next flip.

- Bayesians see probability as subjective, capturing the odds a person would place on an event occurring. In other terms: an individual's degree of belief in a statement. It can be influenced by personal beliefs, prior evidence, etc.

Leads to three views of statistics (frequentist, likelihood and bayesian). Bayesian statistics do not require repeated sampling or large-N assumptions.

Objective vs. Subjective Probability

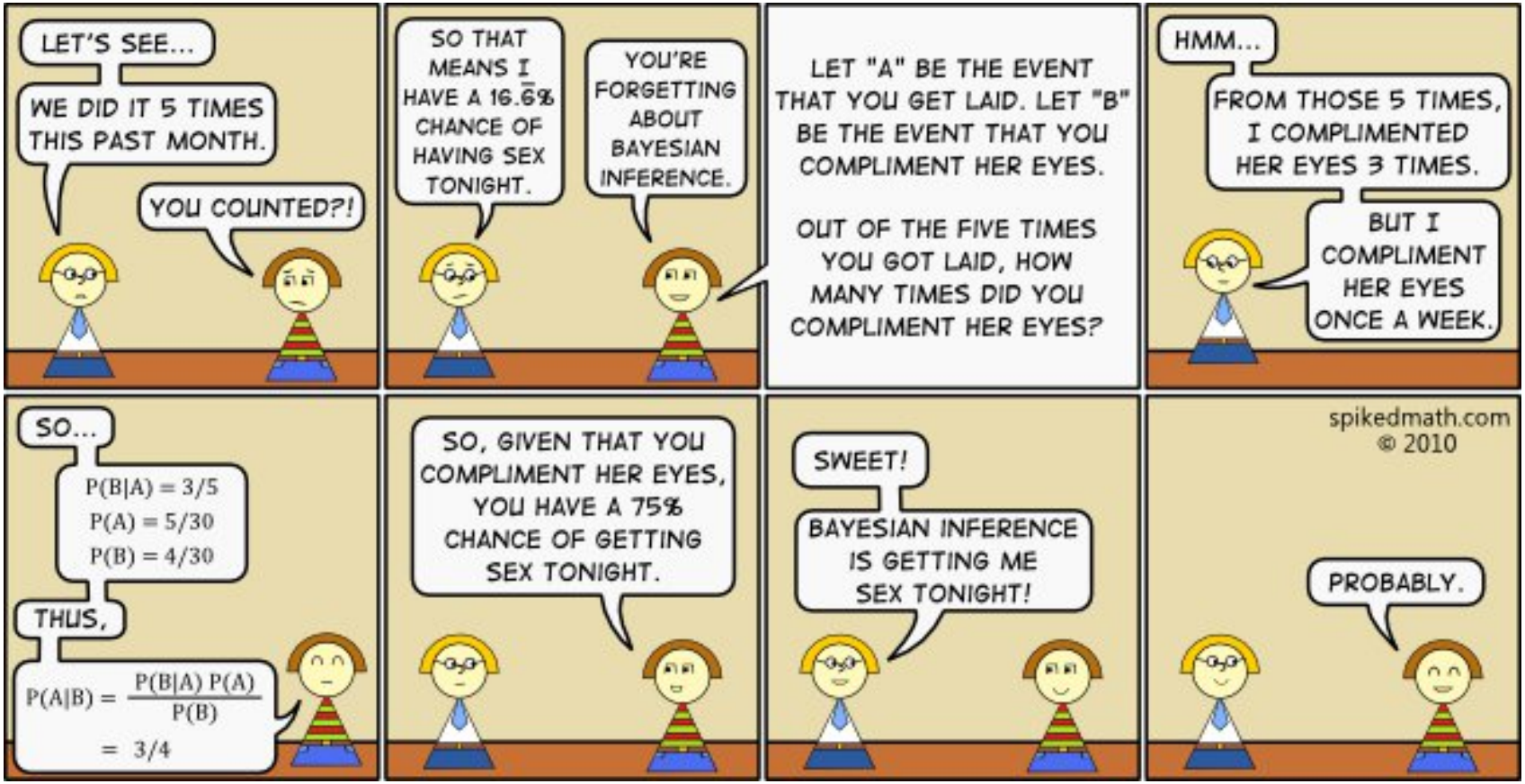


Bayes' Theorem

Very uncontroversial, easily derived from conditional probability:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Practical Uses for the Theorem?



Simple Mecanics of the Theorem

Do we condition on the model or the hypothesis?

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

Measuring $p(\theta)$ is the object of a debate between likelihoodists and bayesians.

For likelihoodists, $p(\theta)/p(y)$ is simply a constant that can be dropped. Hence,

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

In other words, there is a fixed value of θ and we maximize the likelihood to estimate θ and make assumptions to generate uncertainty about the estimate.

How the bayesians do it

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

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

$$\propto p(\theta)L(\theta|y)$$

The Posterior probability $\propto$ Prior Probability X Likelihood function

How the bayesians do it

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

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

$$\propto p(\theta)L(\theta|y)$$

The Posterior probability $\propto$ Prior Probability X Likelihood function

- We have a **prior** distribution: $p(\theta)$
- We update the prior with **observed data**: $p(y|\theta) = L(\theta|y)$
- We get a **posterior** distribution for θ: $p(\theta|y)$

Fixed vs Random Variable

Frequentists see θ as a fixed parameter. Bayesians see it as stochastic.

θ as a fixed parameter:

- We estimate θ with a measure of uncertainty (SE, CIs)
- Given repeated sampling, θ should fall within the CI 95% of the time

Fixed vs Random Variable

Frequentists see θ as a fixed parameter. Bayesians see it as stochastic.

θ as a fixed parameter:

- We estimate θ with a measure of uncertainty (SE, CIs)
- Given repeated sampling, θ should fall within the CI 95% of the time

θ as random parameter:

- We want to model the uncertainty around θ. Thus, we find its posterior distribution.
- We have statistics useful for probability statements (posterior mean, posterior SD, posterior credibility interval, etc.)

The Three Steps of Bayesian Analysis

- Specify a probability model for unknown parameter values (θ) that include some **prior knowledge** about the parameters if available.
- Update knowledge about the unknown parameters (θ) by **conditioning** this probability model **on the observed data** (y).
- Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Example for a Bernouilli Trial

Say we have 50 Bernouilli observations of an event, where $y \sim Binomial(n, \theta)$. We use the Beta distribution, since it is **conjugate**. Therefore,

Example for a Bernouilli Trial

Say we have 50 Bernouilli observations of an event, where $y \sim Binomial(n, \theta)$. We use the Beta distribution, since it is **conjugate**. Therefore,

$$
\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&= Binomial(n, \theta) \times Beta(\alpha, \beta) \\
&= \left[\binom{n}{y}\theta^{y}(1-\theta)^{(n-y)}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}\right] \\
&\propto \theta^{y}(1-\theta)^{(n-y)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)} \\
p(\theta|y) &\propto \theta^{(y+\alpha-1)}(1-\theta)^{(n-y+\beta-1)}
\end{aligned}
$$

Example for a Bernouilli Trial

Say we have 50 Bernouilli observations of an event, where $y \sim Binomial(n, \theta)$. We use the Beta distribution, since it is **conjugate**. Therefore,

$$
\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&= Binomial(n, \theta) \times Beta(\alpha, \beta) \\
&= \left[\binom{n}{y}\theta^{y}(1-\theta)^{(n-y)}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}\right] \\
&\propto \theta^{y}(1-\theta)^{(n-y)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)} \\
p(\theta|y) &\propto \theta^{(y+\alpha-1)}(1-\theta)^{(n-y+\beta-1)}
\end{aligned}
$$

So, the **posterior** is a Beta$(y+\alpha, n-y+\beta)$. But what if we don't have conjugacy? How do we compute if Prior X Likelihood isn't a known distribution?

Markov Chain Monte Carlo (MCMC)

The Monte Carlo Principle:

"Anything we want to learn about a random variable θ can be learned by sampling many times from $f(\theta)$, the density of θ".

- MCMC produces a chain of simulated draws from a distribution, where each draw is **dependant** only on the previous draw.
- The chain should eventually converge to a **stationary distribution** (the posterior probability distribution), given that sampling conditions are met.

Gibbs Sampling

The basic idea: we resample from one variable, based on all other variables from the previous draw, and then do it again, one variable at a time. More formally:

$$
\begin{aligned}
\theta_1^{t+1} &\sim p(\theta_1 | \theta_2^t, \ldots, \theta_k^t, y) \\
\theta_2^{t+1} &\sim p(\theta_2 | \theta^{t+1}, \ldots, \theta_k^t, y) \\
&\vdots \\
\theta_k^{t+1} &\sim p(\theta_k | \theta^{t+1}, \ldots, \theta_k^t - 1, y) \\
\theta^{t+1} &\leftarrow (\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_k^{t+1})'
\end{aligned}
$$

Each draw from the parameters is from the posterior. The result gives a caterpillar traceplot that helps with identifying convergence.

Model Example

Example code (in JAGS):

```
model.jags <- function()  {

  for(i in 1:N){
    moral[i]~dnorm(mu[i], tau)
    mu[i]<-alpha + beta1*hetero[i] + beta2*mobility[i]
  }

  alpha~dnorm(0, .01)
  beta1~dunif(-100,100)
  beta2~dunif(-100,100)
  tau~dgamma(.01,.01)
```

Posterior Distribution Examples

Figure : Trace Plots

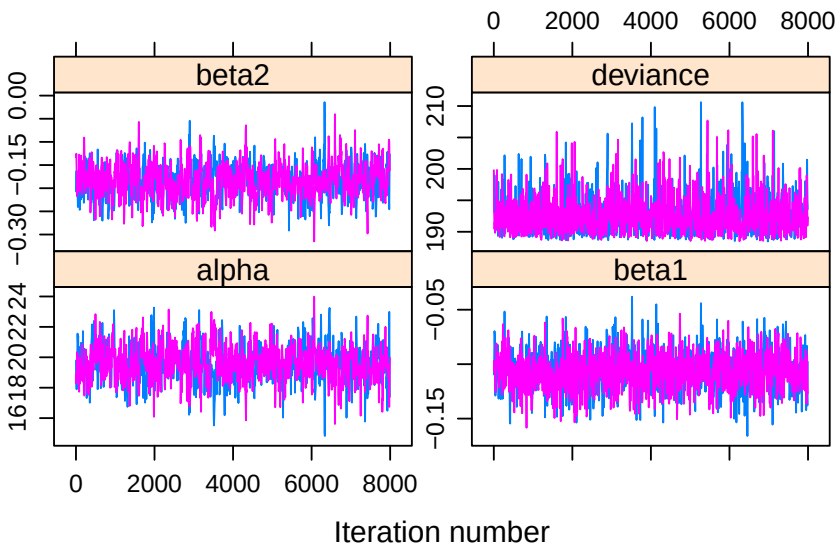
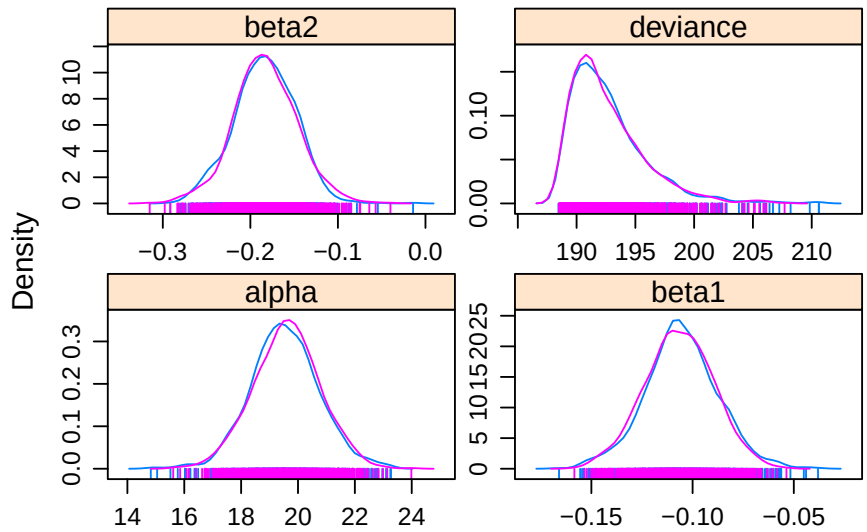


Figure : Density Plots



Softwares

Here are some software normally used:

- WinBUGS
  - Standalone program
  - Can be called from R (and somewhat from Stata)
  - Windows only
- JAGS
  - Needs to be called from R
  - Windows or Mac
- MCMC package
  - R package
  - Somewhat slower and limited compared to BUGS/JAGS
- STAN
  - Written in C++. Can be interfaced through R, python or command prompt
  - Software by Andrew Gelman. Ideal for complex hierarchical models

Further Readings