An Introduction to Bayesian Data Analysis

Dr. Pablo Emilio Verde
pabloemilio.verde@hhu.de

INNOLEC Lectures 2016
Masaryk University
Brno
Czech Republic

18-19-20 of April 2016

Overview of the course

Day 1

Lecture 1: Introduction to Bayesian Inference

Lecture 2: Bayesian analysis for single parameter models

Lecture 3: Priors distributions single parameters

Day 2

Lecture 4: An introduction to WinBUGS

Lecture 5: Bayesian analysis for multi-parameters models

Lecture 6: Bayesian regression models

Day 3

Lecture 7: Markov Chain Monte Carlo Computations

Lecture 8: Hierarchical statistical modeling

Lecture 9: Longitudinal and multilevel data

Learning objectives and style

- Learning objectives:
  - Understanding of the potential role of Bayesian methods for making inference about real-world problems
  - Learning Bayesian statistical analysis with R and WinBUGS
  - An interest in using Bayesian methods in your own field of work
- Style:
  - Immediately applicable methods rather than latest theory
  - Attention to real problems: case studies
  - Case studies and examples implemented in R and WinBUGS
  - Emphasis to the complementary aspects of Bayesian Statistics to Classical Statistics rather than one vs. the other

Recommended bibliography

- **The BUGS Book: A Practical Introduction to Bayesian Analysis**. David Lunn; Chris Jackson; Nicky Best; Andrew Thomas; David Spiegelhalter. CRC Press, *October 3, 2012*.

- **Bayesian Data Analysis (Third edition)**. Andrew Gelman, John Carlin, Hal Stern and Donald Rubin. 2004 Chapman & Hall/CRC.

- **Bayesian Computation with R (Second edition)**. Jim Albert. 2009. Springer Verlag.

- **An introduction of Bayesian data analysis with R and BUGS: a simple worked example**. Verde, PE. Estadistica (2010), 62, pp. 21-44

Lecture:

Introduction to Bayesian Inference

"I shall not assume the truth of Bayes' axiom (...) theorems which are useless for scientific purposes."

-Ronald A. Fisher (1935) *The Design of Experiments*, page 6.

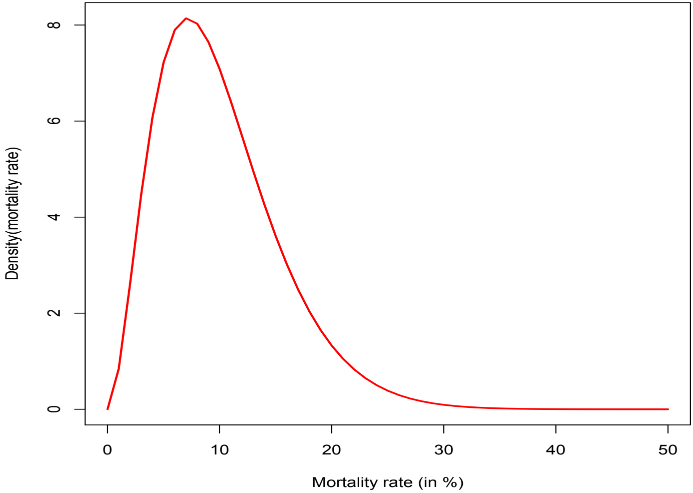
Let's start

- Finals of the football World Cup 2014 in Brazil:
  - Argentina played with Germany
  - The German team won!
- Let be $y \sim \mathit{Binomial}(\theta)$ with $y = 1$ if the German team wins
- After observing this result, the estimated probability that the German team wins against Argentina is ... $\widehat{\theta} = 100\%$ (maximum likelihood estimate: $y/n = 1/1$)
- Now, ... if you think that θ is not 100 % then ... you are doing Bayesian statistical inference
- ...

Probability modeling

Example: surgical procedure

- Suppose a hospital is considering a new high-risk operation
- Experience in other hospitals indicates that the risk θ for each patient is around 10 %
- It would be surprising to be less than 3% or more than 20%
- We can directly express uncertainty about the patient risk θ with a probability distribution



Probability that mortality risk is greater than 15% is $\Pr(\theta > 0.15) = 0.17$

Why a direct probability distribution?

- Tells us what we want: what are plausible values for the parameter of interest?
- No P-values: just calculate relevant tail areas
- No confidence intervals: just report central area that contains 95% of distribution
- Easy to make predictions (see later)
- Fits naturally into decision analysis / risk analysis / cost-effectiveness analysis
- There is a procedure for adapting the distribution in the light of additional evidence: i.e. **Bayes theorem** allows us to *learn from experience*

What about disadvantages?

- Requires the specification of what we thought before new evidence is taken into account: *the prior distribution*
- Explicit allowance for quantitative subjective judgment in the analysis
- Analysis may be more complex than a traditional approach
- Computation may be more difficult
- Currently no established standards for Bayesian reporting

Why does Bayesian statistics is popular today?

- Bayesian methods optimally combine multiple sources of information in a common model
- The computational revolution produced by the rediscovery of Markov chain Monte Carlo (MCMC) techniques in statistics
- Free available software implementation of MCMC (e.g. WinBUGS, JAGS, STAN, large number of packages in R, etc.)
- As a result, we can routinely construct sophisticated statistical models that may reflect the complexity for phenomena of interest

Bayes theorem for observable quantities

Let A and B be events; then it is provable from axioms of probability theory that

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

- $P(A)$ is the marginal probability of A, i.e., **prior** to taking account of the information in B
- $P(A|B)$ is the conditional probability of A given B, i.e., **posterior** to taking account of the value of B
- $P(B|A)$ is the conditional probability of B given A
- $P(B)$ is the marginal probability of B
- Sometimes is useful to work with $P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A})$, which is curiously called "extending the conversation" (Lindley 2006, pag. 68)

Example: Security risk analysis in airports

- A new anti-terrorist alarm system in an airport claimed to have 99.9 % sensitivity (true positive rate) and 99.8% specificity (true negative rate)

- Suppose that a-priory the chance that passenger is suspected to be a terrorist is $1/10{,}000$

- Now, a passenger has triggered the alarm system and the security personal take control over this person

Question: What is the chance that this passenger is actually a terrorist ?

- Let A be the event that passenger is truly a terrorist
- Let B be the event that the passenger triggered the alarm system
- We want $p(A|B)$
- 99.9% sensitivity means that $p(B|A) = 0.999$
- 99.8% specificity means that $p(B|\overline{A}) = 1 - 0.998 = 0.002$

Now Bayes theorem says

$$p(A|B) = \frac{0.999 \times 0.0001}{0.999 \times 0.0001 + 0.002 \times .9999} = 0.048.$$

Answer: With this alarm system 95% of alarms are in fact, false alarms!

Example: Problems with statistical significance

- Suppose that a priory only 10% of clinical trials are truly effective treatments

- Assume each trial is carried out with a design with enough sample size such that $\alpha = 5\%$ and power $1 - \beta = 80\%$

Question: What is the chance that the treatment is true effective given a significant test results?

$$p(H_1|\text{"significant results"})?$$

- Let A be the event that H_1 is true, then $p(H_1) = 0.1$
- Let B be the event that the statistical test is significant
- We want $p(A|B) = p(H_1|\text{"significant results"})$
- We have: $p(B|A) = p(\text{"significant results"}|H_1) = 1 - \beta = 0.8$
- We have: $p(B|\overline{A}) = p(\text{"significant results"}|H_0) = \alpha = 0.05$
- Now, Bayes theorem says

$$p(H_1|\text{"significant results"}) = \frac{(1-\beta)\times 0.1}{(1-\beta)\times 0.1 + \alpha \times 0.9} = 0.64$$

Answer: This says that if truly effective treatments are relatively rare, *then a "statistically significant" results stands a good chance of being a false positive.*

Some comments

- These examples illustrated that our intuition is poor when processing probabilistic evidence
- Bayes theorem applied to *observable quantities* (e.g. diagnostic testing) is uncontroversial and well established
- More controversial is the *Bayesian Inference*, i.e., the use of Bayes theorem in general statistical analysis, where *parameters* are the unknown quantities and their prior distribution needs to be specified.

Bayesian inference for unknown quantities

- **Makes fundamental distinction between:**
  - **Observable quantities** y, i.e., data.
  - **Unknown quantities** θ, that can be statistical parameters, missing data, predicted values, mismeasured data, indicators of variable selected, etc.
  - Technically, in the Bayesian framework **parameters are treated as values of random variables**.
- **Differences with classical statistical inference:**
  - In Bayesian inference, we make probability statements about model parameters
  - In the classical framework, parameters are *fixed* non-random quantities and the probability statements concern the data

Bayesian Inference

- Suppose that we have observed some data y

- We want to make inference about unknown quantities θ: model parameters, missing data, predicted values, mismeasured data, etc.

- The Bayesian analysis starts like a classical statistical analysis by specifying the sampling model:

  $$p(y|\theta)$$

  this is the **likelihood function**.

- From a Bayesian point of view, θ is unknown so should have a *probability distribution* reflecting our *uncertainty*. We specify a **prior distribution**

  $$p(\theta)$$

- Together they define a **full probability model**:

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

Then we use the Bayes theorem to obtain the conditional probability distribution *for* unobserved quantities of interest given the data:

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

This is **the posterior distribution** *for* θ,

$$\text{posterior } \propto \text{likelihood} \times \text{prior.}$$

Inference with binary data

Example: Inference on proportions using a discrete prior

- Suppose I have 3 coins in my pocket:
  1. A biased coin: $p(heads) = 0.25$
  2. A fair coin: $p(heads) = 0.5$
  3. A biased coin: $p(heads) = 0.75$
- I randomly select one coin, I flip it once and it comes a head.
- What is the probability that I have chosen coin number 3 ?

Inference with binary data

1. Let $y = 1$ the event that I observed a head
2. Let θ denote the probability of a head: $\theta \in (0.25, 0.5, 0.75)$
3. Prior: $p(\theta = 0.25) = p(\theta = 0.5) = p(\theta = 0.75) = 1/3$
4. Sampling distribution:

$$y|\theta \sim \text{Binomial}(\theta, 1),$$

with likelihood

$$p(y|\theta) = \theta^{y}(1-\theta)^{(1-y)}.$$

- If we observe a single positive response ($y = 1$), how is our belief revised?

Coin	θ	Prior	Likelihood	Likelihood $\times$ prior	Posterior
		$p(\theta)$	$p(y = 1\|\theta)$	$p(y = 1\|\theta)p(\theta)$	$p(\theta\|y = 1)$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
$\sum$		1.0	1.50	0.495	1.0

- So, observing a head on a single flip of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% that the chance of heads is 0.25.

- Note that if we normalize the likelihood $p(y = 1|\theta)$ we have exactly the same results.

Posterior predictive distributions

The predictive posterior distribution for a new observation y^{new} is given by

$$p(y^{new}|y) = \int p(y^{new}|y, \theta)p(\theta|y)d\theta.$$

Assuming that past and future observations are conditionally independent given θ,
this simplify to

$$p(y^{new}|y) = \int p(y^{new}|\theta)p(\theta|y)d\theta.$$

For the discrete case of θ, integrals are replaced by sums:

$$p(y^{new}|y) = \sum_{\theta_i} p(y^{new}|\theta_i)p(\theta_i|y)$$

where the $p(\theta_i|y)$ can be thought of as "posterior weights".

Example: Three coins continue ...

Suppose we want to predict the probability that in the next toss is a head. We have:

$$\begin{aligned} p(y^{new} = 1|y = 1) &= \sum_{\theta_i} \theta_i \, , p(\theta_i|y = 1) \\ &= (0.25 \times 0.167) + (0.50 \times 0.333) + (0.75 \times 0.5) \\ &= 0.5833 \end{aligned}$$

Sequential learning

Suppose we obtain data y_1 and form the posterior $p(\theta|y_1)$ and then we obtain further data y_2. The posterior based on y_1, y_2 is given by:

$$p(\theta|y_1, y_2) \propto p(y_2|\theta) \times p(\theta|y_1).$$

Today's posterior is tomorrow's prior!

The resultant posterior is the same as if we have obtained the data y_1, y_2 together:

$$p(\theta|y_1, y_2) \propto p(y_1, y_2|\theta) \times p(\theta).$$

Example: Three coins continue ...

- Now suppose that after observing $y_1 = 1$ we observe $y_2 = 1$, how is our belief revised?

Coin	θ	Prior	Likelihood	Likelihood $\times$ prior	Posterior
		$p(\theta)$	$p(y = 1 \mid \theta)$	$p(y = 1 \mid \theta)p(\theta)$	$p(\theta \mid y = 1)$
1	0.25	0.167	0.25	0.042	0.071
2	0.50	0.333	0.50	0.167	0.286
3	0.75	0.500	0.75	0.375	0.644
$\sum$		1.0	1.50	0.583	1.0

- After observing a second head, there is now a 64.4% probability that the chance of heads is 0.75 and only a 7.1% that the chance of heads is 0.25.

A bit of philosophy: Probability "for" and probability "of"

- The prior distribution $p(\theta)$, expresses our uncertainty about θ *before* seeing the data, that could be objective or subjective
- The **Bayesian inference** *allows* the combination of different types of probabilities
- **Subjective probability** implied a *mental construct* where probabilities are used to express **our** uncertainty. This is why we use a pedantic: **"probability for an event..."**
- In the classical setting probabilities are defined in terms of long run frequencies and are interpreted as physical properties of systems. In this way we use: **"probability of ..."**
- In general we follow **Bayesian Statistical Modeling**, which is dynamic view of data analysis, which includes **model building** and **model checking** as well as **statistical inference**

Summary

Bayesian statistics:

- Formal combination of external information with data model by the Bayesian rule
- Uses of subjective probability to make inductive inference
- Straightforward inference by manipulating probabilities
- No need of repeated sampling ideas
- Specification of prior or external evidence may be difficult or controversial
- Prediction and sequential learning is straightforward.

Practical with R

- Exercise 1: Repeat the calculations of this lecture with R

- Exercise 2: Investigate the shape of the Beta distribution in R using the function curve() and dbeta(). Take a look in the R help of these functions.

Solution Exercise 1:

```
theta <- c(0.25, 0.5, 0.75)
prior <- rep(1/3, 3)
plot(theta, prior, type = "h", ylim=c(0,.6),
          ylab="Probability",
          xlab = expression(theta))
# Likelihood
lik <- function(theta, x)
    {theta^x*(1-theta)^(1-x)}
curve( lik(x, 1)/sum(lik(theta, 1)),
       from=0, to =1, add=TRUE,
       col="red", lwd=2, lty=2)
product <- prior * lik(theta, 1)
posterior <- product / sum(product)
points(theta, posterior, col="blue", cex=2, pch=19)
legend(0.3, 0.5, col=c("black", "red", "blue"),
 legend=c("prior", "likelihood", "posterior"), lty=c(1,2,1))
```

Solution Exercise 2:

```
# Beta flexibility
par(mfrow=c(2,2))
curve(dbeta(x, 1, 1), from =0, to=1, lty = 1, lwd=2,
               main="Beta(1, 1)", xlab=expression(theta))

    curve(dbeta(x, 1/2, 1/2), from = 0, to = 1, lty = 2,
              lwd=2, col="red", main="Beta(1/2, 1/2)",
              xlab=expression(theta))

    curve(dbeta(x, 2, 5), from = 0, to = 1, lty = 2, lwd=2,
             col="blue", main="Beta(2, 5)",
             xlab=expression(theta))

    curve(dbeta(x, 2, 2), from = 0, to = 1, lty = 2,
      lwd=2, col="green", main="Beta(2, 2)",
      xlab=expression(theta))
par(mfrow=c(1,1))
```

Lecture :

Bayesian Inference for Single Parameter Models

Summary

1. Conjugate Analysis for:
    - Binomial model
    - Normal model known variance
    - Normal model known mean
    - Poisson model
2. Using R for:
    - Graphical visualization of posteriors
    - Direct Monte Carlo Simulation Methods
    - Calculations of posteriors for functional parameters

Inference of proportions using a continuous prior

Suppose we observe r positive responses out of n patients. Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood

$$p(r|n,\theta) = \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1-\theta)^{n-r} \propto \theta^r (1-\theta)^{n-r}$$

We consider the response rate θ to be **a continuous parameter**, i.e., we need to give a continuous prior distribution.

Suppose that before the data is observed we believe all values for θ are equally likely (very unrealistic!), then we give $\theta \sim \text{Unif}(0,1)$, i.e., $p(\theta) = 1$.

Posterior is then

$$p(\theta|r,n) \propto \theta^r (1-\theta)^{n-r} \times 1$$

this has a form of the *kernel* of a $\text{Beta}(r+1, n-r+1)$.

To represent external evidence that some response rates are more plausible than others, it is mathematical convenient to use a Beta(a, b) prior distribution for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta|r,n) &\propto p(r|\theta,n)p(\theta) \\ &\propto \theta^{r}(1-\theta)^{n-r} \\ &= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \\ &\propto \mathtt{Beta}(r+a, n-r+b). \end{aligned}$$

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood

- A Beta(a, b) distribution has

  $$\begin{aligned} E(\theta) &= a/(a+b) \\ var(\theta) &= ab/[(a+b)^2(a+b+1)] \end{aligned}$$

  Hence the posterior mean is $E(\theta|r,n) = (r+a)/(n+a+b)$

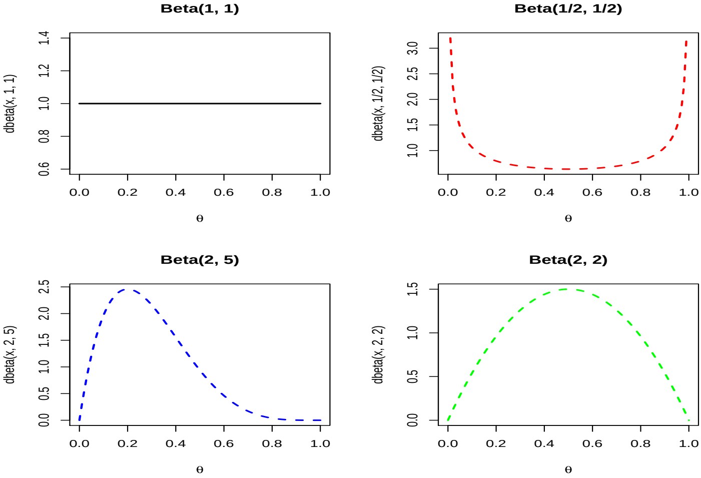
- a and b are equivalent to observing a prior $a-1$ successes in $a+b-2$ trials, then it can be elicited.

- With fixed a and b, as r and n increase, $E(\theta|r,n) \to r/n$ (the MLE).

- A Beta(1,1) is equivalent to Uniform(0,1).

Shape of the Beta density function

The Beta(a, b) prior is a flexible distribution.



Beta(1, 1)

Beta(1/2, 1/2)

Beta(2, 5)

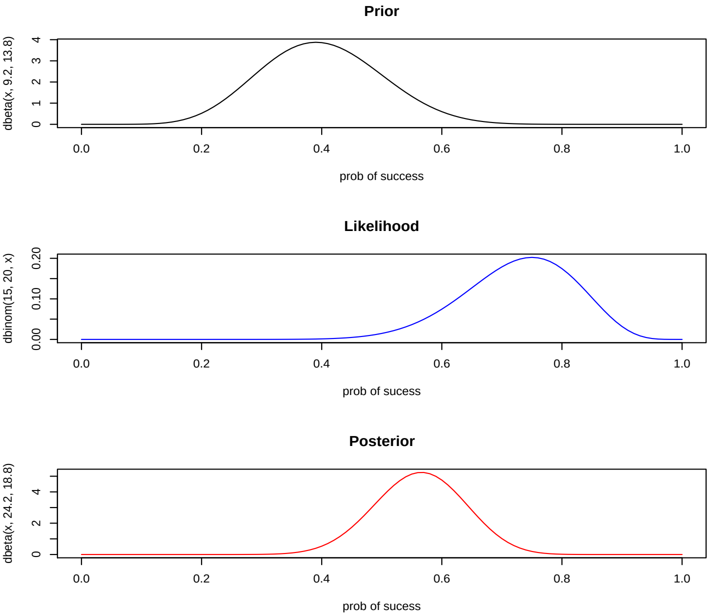
Beta(2, 2)

Example: drug investigation

- Consider an early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- Interpret this as a distribution with mean=0.4 and standard deviation 0.1
- A Beta(9.2, 13.8) distribution has these properties
- Suppose we treat $n = 20$ volunteers with the compound and observe $r = 15$ positive responses.
- Then we update the prior and the posterior is Beta$(15 + 9.2, 5 + 13.8)$

R script to perform the analysis:

```
> par(mfrow=c(3,1)) # graphical output: 3 by 1 panels
# draw a curve for a beta density
>  curve(dbeta(x,9.2, 13.8),from=0, to =1,
               xlab= "prob of success",main="Prior")
# draw a curve for a binomial density
>  curve(dbinom(15, 20,x),from=0, to =1,
               col="blue", xlab="prob of sucess",
               main="Likelihood")
# draw the posterior
>  curve(dbeta(x, 24.2, 18.8),from=0, to =1,
               col="red", xlab="prob of sucess",
               main="Posterior")
>  par(mfrow=c(1,1))
```


Prior
dbeta(x, 9.2, 13.8)
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
prob of success
Likelihood
dbinom(15, 20, x)
0.00
0.10
0.20
0.0
0.2
0.4
0.6
0.8
1.0
prob of sucess
Posterior
dbeta(x, 24.2, 18.8)
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
prob of sucess

Monte Carlo Simulation

- If we are able to sample values θ^* from the posterior $p(\theta|r, n)$, then we can extend the inferential scope.

- For example, one important application of this simulation process is when we need to estimate the posterior of a functional parameter, e.g., the odds ratio:

$$\phi = f(\theta) = \frac{\theta}{1-\theta}$$

- For simple models we can directly simulate from the posterior density and use this values to empirically approximate the posterior quantiles.

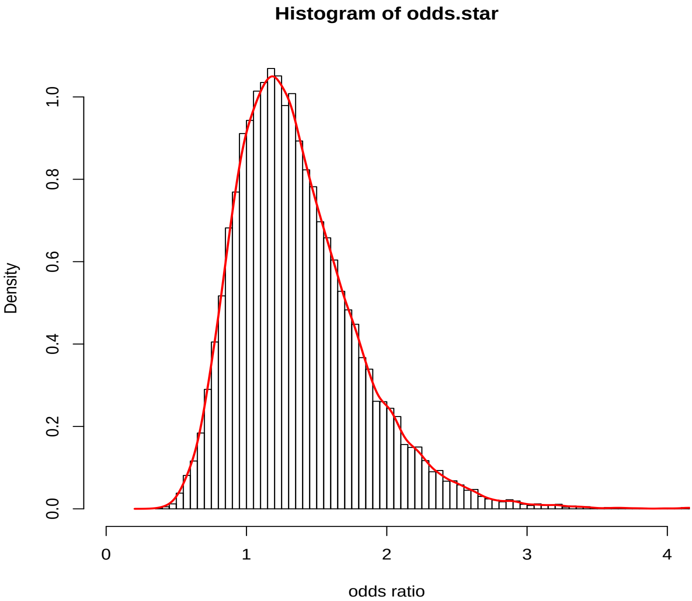
Example: Posterior for the odds ratio

The posterior of

$$\phi = \frac{\theta}{1-\theta}$$

is calculated in R as follows:

```
> theta.star <- rbeta(20000, 24.2, 18.8)
> odds.star <- theta.star/(1-theta.star)
> quantile(odds.star, prob =  c(0.05, 0.5, 0.75, 0.95))
       5%       50%       75%       95%
0.7730276 1.2852558 1.5784318 2.1592678

> hist(odds.star, breaks=100, xlab="odds ratio",
freq=FALSE, xlim=c(0, 4))
> lines(density(odds.star), lwd =2, lty = 1, col ="red")
```


Histogram of odds.star
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
odds ratio

Making predictions for binary data

- Suppose that we want to make predictions from the model
- The predictive posterior distribution for r^{new} (the number of successes in m trials) follows a **Beta-Binomial** distribution with density:

$$p(r^{new}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\begin{array}{c} m \\ r^{new} \end{array} \right) \frac{\Gamma(a+r^{new})\Gamma(b+m-r^{new})}{\Gamma(a+b+m)}$$

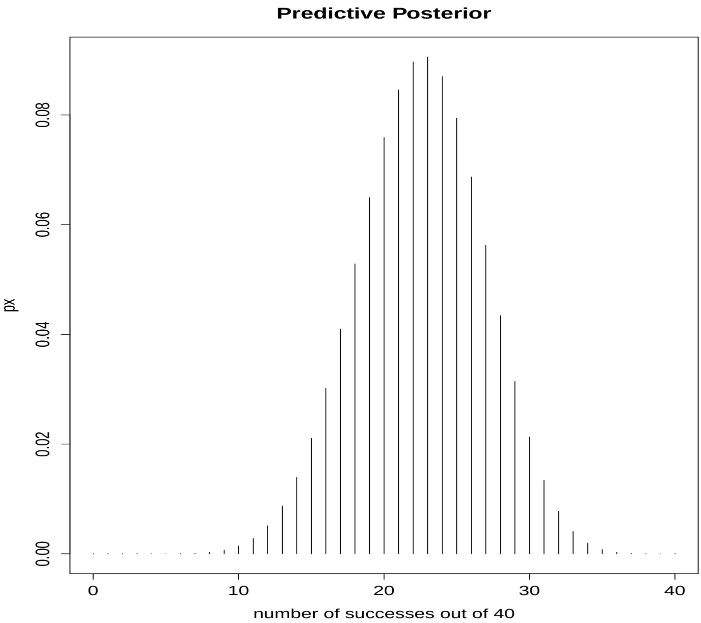
- In R:

```
# Beta-binomial density
betabin <- function(r,a,b,m)
  {
    gamma(a+b)/(gamma(a)*gamma(b)) * choose(m,r) *
             gamma(a+r)*gamma(b+m-r)/gamma(a+b+m)
  }
```

Example: drug investigation continue ...

- Suppose that we are interested in the predictive posterior of the number of success in the next 40 trials:
- Using the betabin function in R we have:

```
# Beta-binomial distribution of the number of
# successes x in the next
# 40 trials with mean 22.5 and standard deviation 4.3

x <- 0:40; px <- betabin(0:40, a=24.2,b=18.8,m=40)

plot(x, px, type="h", xlab="number of successes out
of 40", main="Predictive Posterior")
```


Predictive Posterior
px
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
number of successes out of 40

Example: drug investigation continue ...

- Suppose that we would consider continuing a development program if the drug managed to achieve at least a further 25 successes out of these 40 future trials

- In R we can calculate:

```
# probability of at least 25 successes out of
# 40 further trials:
> sum(betabin(25:40,24.2,18.8,m=40))
[1] 0.3290134
>
```

Simulations for predictive data

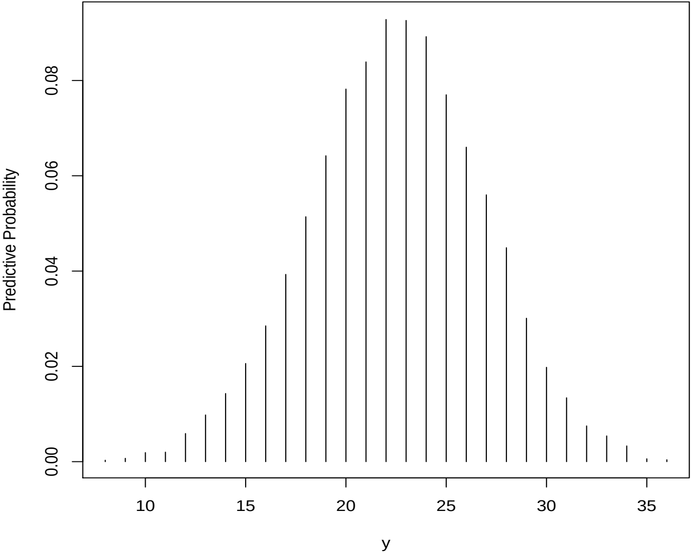
Instead of using the analytical approximation based on the Beta-Binomial distribution, we can simulate predicted observations in the following way:

- We simulate $\theta_1^*, \dots, \theta_B^*$ from the posterior Beta(24.2, 18.8)
- Then we simulate y^* from a binomial distribution with rate θ^* and $n = 40$.
- We tabulate and normalize predictive frequencies.

Simulations for predictive data

In R notation

```
> # Simulation of predictive data with a Beta-Binomial mode
> theta.star <- rbeta(10000, 24.2, 18.8)
> y <- rbinom(10000, 40, theta.star)
> freq.y <- table(y)
> ys <- as.integer(names(freq.y))
> predprob <- freq.y/sum(freq.y)
> plot(ys, predprob, type = "h", xlab = "y",
+      ylab = "Predictive Probability")
>
> # Probability of at least 25 out of future 40 trials.
> sum(predprob[ys>24])
[1] 0.3244
```


Predictive Probability
0.00
0.02
0.04
0.06
0.08
10
15
20
25
30
35
y

Bayesian analysis for Normal data

Known variance, unknown mean

Suppose we have a sample of Normal data

$$y_i \sim \mathtt{N}(\mu, \sigma^2), \quad i = 1, \ldots, n$$

where σ^2 is *known* and μ is *unknown*. The conjugate prior of μ is

$$\mu \sim \mathtt{N}(\mu_0, \tau^2).$$

It is convenient to write τ^2 as σ^2/n_0, where n_0 represents the "*effective number of observations*" in the prior distribution.

Then the posterior distribution for μ is given by

$$p(\mu|y) = \mathtt{N}\left(\frac{n_0\mu_0 + n\bar{y}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

- Prior variance is based on an "implicit" sample size n_0
- As n_0 tends to 0, the distribution becomes "flatter"
- Posterior mean is a weighted average of the prior mean μ_0 and parameter estimate $\bar{x}$, weighted by their precisions, i.e., relative sample sizes.
- Posterior variance is based on an implicit sample size n_0 and the data sample size n.

Alternative expressions for the posterior mean μ_n are :

$$\begin{aligned} \mu_n &= w\mu_0 + (1-w)\bar{y} \quad \text{where} \quad w = \frac{n_0}{n+n_0}, \\ \mu_n &= \mu_0 + (\bar{y} - \mu_0)\frac{n}{n+n_0}, \\ \mu_n &= \bar{y} - (\bar{y} - \mu_0)\frac{n_0}{n+n_0}. \end{aligned}$$

That shows ”shrinkage” towards prior mean.

Prediction

Denoting the posterior mean and variance as μ_n and $\sigma_n^2 = \sigma^2/(n_0 + n)$, the *predictive posterior distribution* for a new observation y^* is

$$p(y^*|y) = \int p(y^*|y,\mu)p(\mu|y)d\mu$$

which is generally equal to

$$p(y^*|y) = \int p(y^*|\mu)p(\mu|y)d\mu$$

which can be shown to give

$$p(y^*|y) \sim \mathtt{N}(\mu_n, \sigma_n^2 + \sigma^2)$$

The predictive posterior distribution is centered around the posterior mean μ_n with variance equal to sum of the posterior variance of μ plus the data variance.

Bayesian inference for Normal data

Unknown variance, know mean

Suppose we have sample of Normal data

$$y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n,$$

where μ is *known* and σ^2 is *unknown*.
It is convenient to change parameterization to the precision $w = 1/\sigma^2$. The conjugate prior for w is then

$$w \sim \text{Gamma}(\alpha, \beta),$$

where

$$p(w) \propto w^{\alpha-1} \exp(-\beta w).$$

Then σ^2 is then said to have an inverse-gamma distribution.

The posterior distribution for w takes the form

$$p(w|\mu, y) \propto w^{\alpha-1} \exp(-\beta w) \times w^{\frac{n}{2}} \exp\left[-\frac{w}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right].$$

Collecting terms gives

$$p(w|\mu, y) = \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

- Clearly we can think of $\alpha = n_0/2$, where n_0 is the "effective prior sample size".

- Since $\sum_{i=1}^{n}(y_i - \mu)^2/n$ estimate $\sigma^2 = 1/w$, then we interpret 2β as representing $n_0 \times$ prior estimate ofσ_0^2.

- Alternative, we can write our conjugate prior as

$$w \sim \text{Gamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right),$$

which can be seen as a scale $\chi^2_{n_0}$ distribution. This is useful when assessing prior distributions for sample variances.

Bayesian inference with count data

Suppose we have an independent sample of counts $y_1 \ldots, y_n$ which can be assumed to follow a Poisson distribution with unknown mean μt_i, where μ is the rate per unit t :

$$p(y|\mu) = \prod_i \frac{(\mu t_i)^{y_i} \exp(-\mu t_i)}{y_i!}$$

The kernel of the Poisson likelihood as a function of μ has the same form as a Gamma(a, b) prior for μ:

$$p(\mu) \propto \mu^{a-1} \exp(-b\mu).$$

This implies the following posterior

$$\begin{aligned}
p(\mu|y) &\propto p(\mu)p(y|\mu) \\
&\propto \mu^{(a-1)}e^{-b\mu}\prod_{i=1}^{n} e^{-\mu t_i}\mu_i^{y} \\
&\propto \mu^{a+Y_n-1}e^{-(b+T_n)\mu} \\
&= \texttt{Gamma}(a+Y_n, b+T_n).
\end{aligned}$$

where $Y_n = \sum_{i=1}^{n} y_i$ and $T_n = \sum_{i=1}^{n} t_i$.

The posterior mean is:

$$E(\mu|y) = \frac{a + Y_n}{b + T_n} = \frac{Y_n}{T_n}\left(\frac{n}{n+b}\right) + \frac{a}{b}\left(1 - \frac{n}{n+b}\right).$$

The posterior mean is a compromise between the prior mean a/b and the MLE $\frac{Y_n}{T_n}$.

Thus b can be interpreted as an effective exposure and a/b as a prior estimate of the Poisson mean.

Example: Annual number of cases of haemolytic uraemic syndrome Henderson and Matthews, (1993)

Annual numbers of cases were available from two a specialist center at Birmingham, from 1970 to 1988. For analysis we consider observed values from the first decade.

year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
cases, x_i	1	5	3	2	1	0	0	2	1	1

Because we are interested in counts of disease over time, a simple

model is a Poisson process.

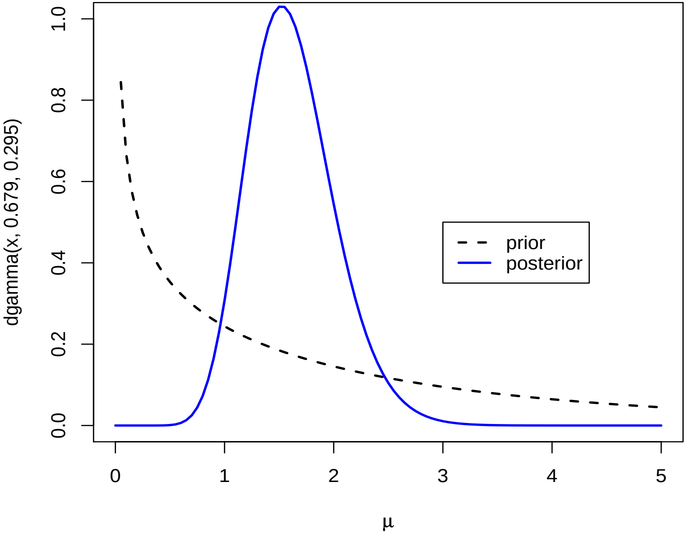
$$y_i|\mu \sim \texttt{Poisson}(\mu) \quad i = 1, \ldots, 10.$$

Similar data is collected by another centre, given a mean rate of 2.3 with s.d. 2.79.

- with prior mean $a/b = 2.3$
- prior sd $\sqrt{a}/b = 2.79$
- Solving for a and b, this information can be translated to a prior Gamma(0.679, 0.295) distribution

Then

$$\begin{aligned} p(\mu|y) &= \mathtt{Gamma}(\sum_i x_i + 0.679, 10 + 0.295) = \mathtt{Gamma}(16.679, 10.295) \\ E(\mu|y) &= \frac{16.679}{10.295} = 1.620; \quad sd(\mu|y) = \frac{\sqrt{16.679}}{10.295} = 0.396 \end{aligned}$$


1.0
0.8
0.6
0.4
0.2
0.0
dgamma(x, 0.679, 0.295)
0
1
2
3
4
5
μ
prior
posterior

The predictive posterior distribution for a new count y^* is

$$y^*|y \sim \text{Negative-Binomial}(a + n\bar{y}, b + n)$$

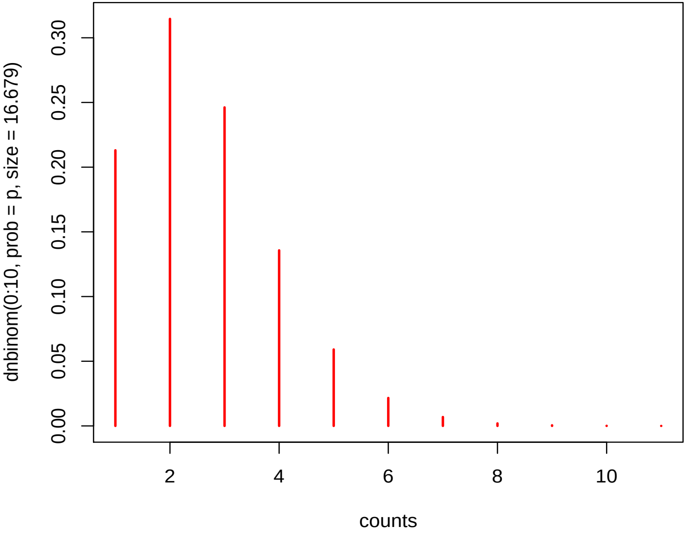
If we ask what is the probability to observe 7 or more cases in the next year we have

```
> sum(dnbinom(6:17, prob=10.295/(1+10.295), size=16.679))
[1] 0.0096
>
```

Note: the observed values for the following years were 7,11,4,7,10,... Indicating a possible structural break.

We can visualize the density by

```
plot(dnbinom(0:10, prob=p, size=16.679, type="h",
                lwd=2, col="red", xlab= "counts")
```


dnbinom(0:10, prob = p, size = 16.679)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
2
4
6
8
10
counts

Some comments

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less that each of the prior s.d. and the s.e. (MLE)
- As $n \to \infty$,
  - the posterior mean → the MLE
  - the posterior s.d. → the s.e. (MLE)
  - the posterior does not depend on the prior.

These observations are generally true, when the MLE exists and is unique.

Priors

- When the posterior is in the same family as the prior then we have what is known as *conjugacy*.

- Conjugate models have the advantage that:
  - Prior parameters can usually be interpreted as *implied prior sample size*: n_0
  - They involve simple mathematical models, examples include:

Distribution of y	Parameter	conjugate prior
Binomial	Prob. of success	Beta
Poisson	Mean	Gamma
Exponential	Reciprocal of mean	Gamma
Normal	Mean (variance known)	Normal
Normal	Variance (mean known)	Inverse Gamma

- Unfortunately conjugate priors only exists for small catalog of likelihoods.

Practical

Exercise: Conjugate inference for a binomial experiment

Drug investigation example from this Lecture

We treat $n = 20$ volunteers with a new compound and observe $r = 15$ positive responses. We use as prior $\theta \sim Beta(9.2, 13.8)$:

1. What is the posterior mean and median for the response rate?
2. What are the 2.5th and 97.5th percentiles of the posterior?
3. What is the probability that the true response rate is greater than 0.6?
4. How is this value affected if a uniform prior is adopted?
5. Using the original Beta(9.2, 13.8) prior, suppose 40 more patients were entered into the study. What is the chance that at least 25 of them respond positively?

Solution

```
# 1) Posteriors mean and median
# The posterior  mean is:  (r+a) / (n + a+b)
> (15 + 9.2) /( 20 + 9.2 + 13.8)
[1] 0.5627907

# the posterior median is
> qbeta(0.5, 24.2, 18.8 )
[1] 0.5637731
>
# 2) Posterior percentiles:
> qbeta(0.025, 24.2, 18.8 )  # 2.5%
[1] 0.4142266
> qbeta(0.975, 24.2, 18.8 )  # 97.5%
[1] 0.7058181
>
```

```
# 3) Posterior prob that the rate is greter than 0.6
> 1 - pbeta(0.6, 24.2, 18.8 )
[1] 0.3156323

# 4) Uniform prior is beta(1,1)  then the posterior
# is beta( r+1, n-r+1)

> 1 - pbeta(0.6, 15+1 ,20-15+1)
[1] 0.9042598

# 5) Posterior probability of at least 25 successes out of
# 40 further trials
> sum(betabin(25:40,24.2,18.8,m=40))
[1] 0.3290134
```

Lecture :

Priors Distributions for Single Parameters

Summary

- Misunderstandings about prior distributions
- Non-informative Priors and Jeffreys invariance priors
- Sensitivity analysis and making priors predictions
- Adjustment of priors based on historical data and judgement
- Mixture of priors

Misunderstandings about prior distributions

It is worth pointing out some misunderstanding regarding prior distributions:

- **The name prior suggests a temporal relationship, however, this is misleading.** The prior distribution models the uncertainty given by the *external evidence*. Cox (1999)
- **The prior is not necessarily unique!** In a recent article Lambert et. al. (2005) analyze the use of 13 different priors for the between study variance parameter in random-effects meta-analysis.
- **There is no such thing as the'correct' prior**. Bayesian analysis is regarded as transforming prior into posterior opinion, rather than producing *'the' posterior distribution*.
- **The prior may not be completely specified**. In Empirical Bayes inference priors have unknown parameters that are estimated from the data.

- **Priors can be overparametrized.** Sometimes we intentionally overparametrized the priors in order to accelerate convergence of simulation methods, see Gelman, Carlin, Stern and Rubin (2004) Chapter 6.

- **Inference may rely only on priors.** There are situations where no further data are available to combine with our priors or there is no intention to update the priors. This is the typical case of *risk analysis*, *sample size determination in experiments*, *simulation of complex process* , etc. In these analytical scenarios priors are usually used to simulate hypothetical data and we refer to that *prior predictive analysis*.

- **Finally, priors are not necessarily important!** In many scientific applications, as the amount of data increases, the prior is overwhelmed by the likelihood and the influence of the prior disappears, see Box and Tiao (1973) (pag. 20-25).

Non-informative Priors: Classical Bayesian Perspective

We may be interested to introduce an initial state of "ignorance" in our Bayesian analysis.

But representing ignorance raises formidable difficulties!

There has been a long and complex search for various **"non-informative"**, **"reference"** or **"objective"** priors during the last century. A sort of **"off-the-shelf objective prior"** that can be applied in all circumstances.

The synthesis of this search is that those *magic priors do not exists*, although useful guidance exists (Berger, 2006).

Problems with uniform priors for continuous parameters

Example: uniform prior on proportions

Let θ be the chance that a bias coin comes down heads, we assume

$$\theta \sim \mathsf{Uniform}(0,1).$$

Let $\phi = \theta^2$ the chance that it coming down heads in both of the next 2 throws.

Now, the density of ϕ is

$$p(\phi) = \frac{1}{2\sqrt{\phi}},$$

which corresponds to a Beta(0.5,1) distribution and is certainly not uniform!

Jeffreys' invariance priors

Consider a 1-to-1 transformation of $\theta : \phi = g(\theta)$

Transformation of variables: prior $p(\theta)$ is equivalent to prior on ϕ of $p(\phi) = p(\theta) \mid \frac{d\theta}{d\phi} \mid$

Jeffreys proposed defining a non-informative prior for θ as

$$p(\theta) \propto I(\theta)^{1/2}$$

where $I(\theta)$ is Fisher information for θ

$$I(\theta) = -E_{x|\theta}\left[\frac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right] = E_{x|\theta}\left[\left(\frac{\partial \log p(X|\theta)}{\partial\theta}\right)^2\right].$$

Non-informative priors for proportions

Data model is Binomial: we consider r successes from n trials

$$r|\theta \sim \text{Binomial}(\theta, n)$$

we have

$$\log p(x|\theta) = r\log(\theta) + (n-r)\log(1-\theta) + C$$

then

$$I(\theta) = \frac{n}{\theta(1-\theta)}.$$

So the Jeffreys' prior is

$$p(\theta) \propto (\theta(1-\theta))^{-1/2},$$

which is a Beta(1/2, 1/2).

Non-informative priors for location parameters

A location parameters θ is such that $p(\theta|y)$ is a function of $(y-\theta)$ and so the distribution of $(y-\theta)$ is independent of θ.

Example: data model is Normal with unknown mean θ and known variance v

$$x_1, x_2, \ldots, x_n|\theta \sim \text{Normal}(\theta, v)$$

then we have

$$\log p(x|\theta) = -\sum \frac{(x_i-\theta)^2}{2v} + C$$

with

$$I(\theta) = \frac{n}{v}.$$

So the Jeffreys' prior is

$$p(\theta) \propto 1$$

which is the Uniform distribution for θ.

Non-informative priors for scale parameters

A scale parameters θ is such that $p(y|\theta)$ is a function of $1/\theta f(y/\theta)$ and so the distribution of (y/θ) is independent of θ.

Example: data model is Normal with known mean m and unknown variance θ

$$x_1, x_2, \ldots, x_n|\theta \sim \text{Normal}(m, \theta)$$

then we have

$$\log p(x|\theta) = -n/2 \log(\theta) - \frac{s}{2\theta},$$

where $s = \sum(x_i - m)^2$, then

$$I(\theta) = \frac{n}{2\theta^2}.$$

So the Jeffreys' prior on variance is

$$p(\theta) \propto \frac{1}{\theta}$$

This Jeffreys' improper prior is approximated by a Gamma(ϵ, ϵ) with $\epsilon \to 0$.

Priors for counts and rates

Data model is Poisson: $x|\theta \sim \text{Poisson}(\theta)$, then we have

$$\log p(x|\theta) = -\theta + x \log \theta + C$$

with

$$I(\theta) = 1/\theta.$$

So the Jeffreys' prior is

$$p(\theta) \propto \theta^{-1/2}.$$

This improper prior is approximated by a Gamma distribution with $\alpha = 1/2$ and $\beta \to 0$.

Comments Jeffreys' rule

- They are invariant, whatever the scale we choose to measure the unknown parameter, the same prior results when the scale is transformed to any particular scale

- Some inconsistencies associated with Jeffreys' priors have been discussed:
  - Applying this rule to the normal case with both mean and variance parameters unknown does not lead to the same prior as applying separately the rule for the mean and the variance and assuming a priori independence between these parameters.

- Although Jeffreys' rule is suggestive, it cannot be applied blindly. It should be thought of as guideline to consider particularly if there is no other obvious way of finding a prior distribution.

Predictive Prior Analysis

In practice it is not necessary to adopt a full Bayesian approach. Sometimes is very useful to use Bayesian methods for some analyzes and classical approaches for others.

Example: predictive distribution of the power in sample size calculations

- a randomize trial is planned with n patients in each of two arms.
- the response within each treatment arm is assumed to have between-patient standard deviation σ.
- the treatment estimate $\hat{\theta} = \bar{y}_1 - \bar{y}_2$ is approximately distributed as Normal$(\theta, 2\sigma^2/n)$.

Predictive Prior Analysis

- a trial designed to have two-sided Type I error α and Type II error β in detecting a true difference of θ in mean response between the groups will require a sample size per group of

$$n = \frac{2\sigma^2}{\theta^2}(z_{1-\beta} - z_{1-\alpha/2})^2,$$

- Alternatively, for fixed n, the power of this experiment is

$$\text{Power} = \Phi\left(\sqrt{\frac{n\theta^2}{2\sigma^2}} - z_{1-\alpha/2}\right).$$

- If we assume $\theta/\sigma = 0.5$, $\alpha = 0.05$ and $\beta = 0.2$, we have $z_{1-\beta} = 0.84$, $z_{1-\alpha/2} = 1.96$, then the power of the trial is 80% and $n = 63$ in each trial arm.

- However, we accept uncertainty about θ and σ and we wish to include this feature into the sample size and power calculations.

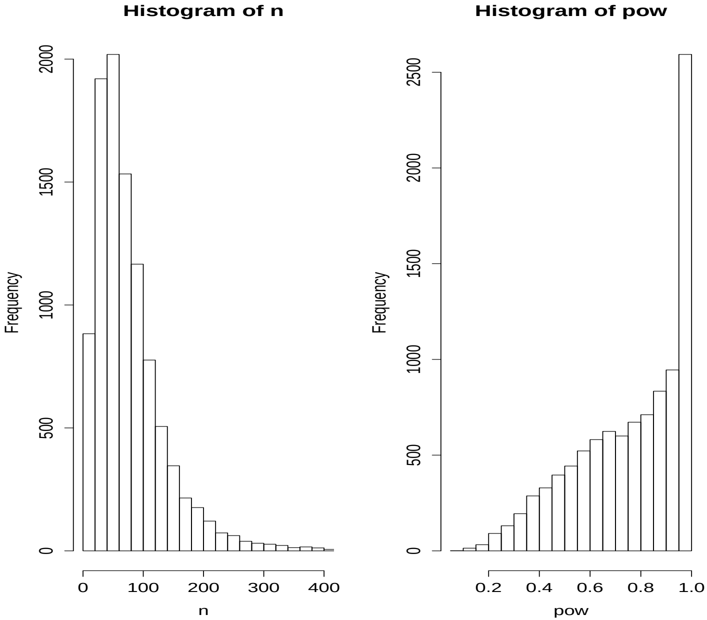
- We assume from previous studies that it is reasonable that

$$\theta \sim N(0.5, 1) \quad \text{and} \quad \sigma \sim N(1, 0.3^2)I(0, \infty).$$

Then

1. Simulate values $\theta^* \sim N(0.5, 1)$ and $\sigma^* \sim N(1, 0.3^2)$ (subject to the constrain of σ being positive).
2. Substitute them in the formulae and generate n^* and Power*.
3. Use the histogram of n^* and Power* as their corresponding predictive distribution.
4. In R we have

```
## predictive prior distribution for sample size and po
 set.seed(123)
 theta <- rnorm(10000, 0.5, 0.1)
 sigma <- rnorm(10000, 1, 0.3)
 sigma <- ifelse(sigma <0, -1*sigma, sigma)
 n <- 2*sigma^2 /(theta^2)*(0.84 + 1.96)^2
 pow <- pnorm( sqrt( 63/2) * theta /sigma - 1.96)

 par(mfrow=c(1,2))
 hist(n,  xlim = c(0, 400), breaks=50)
 hist(pow)
 par(mfrow=c(1,1))
```


Histogram of n
Frequency
0
500
1000
1500
2000
0
100
200
300
400
n
Histogram of pow
Frequency
0
500
1000
1500
2000
2500
0.2
0.4
0.6
0.8
1.0
pow

```
> round(quantile(n),2)
     0%     25%     50%     75%    100%
   0.00   37.09   62.03   99.45 1334.07
> round(quantile(pow),2)
  0%  25%  50%  75% 100%
0.09 0.61 0.81 0.95 1.00
> sum(pow<0.7)/10000
[1] 0.3645
>
```

It is clear that there is huge uncertainty to the appropriate sample size.

For n=63 the median power is 81% and a trial of 63 patients per group could be seriously underpowered. There is a 37% chance that the power is less than 70%.

Mixture of priors for proportions

- We may want to express a more complex prior opinion that can not be encapsulated by a beta distribution

- A prior which is a mixture of beta distributions

  $$p(\theta) = qp_1(\theta) + (1-q)p_2(\theta)$$

  where $p_i = \texttt{Beta}(a_i, b_i)$.

- Now if we observe r successes out of n trials,the posterior is

  $$p(\theta|r,n) = q^* p_1(\theta|r,n) + (1-q^*)p_2(\theta|r,n)$$

  where

  $$\begin{aligned} p_i(\theta|r,n) &\propto p_i(\theta)p(r|\theta,n) \\ q^* &= \frac{qp_1(r|n)}{qp_1(r|n) + (1-q)p_2(r|n)} \end{aligned}$$

- $p_i(r|n)$ is a beta-binomial predictive probability or r successes in n trials assuming θ has distribution $\texttt{Beta}(a_i, b_i)$.

- The posterior is a mixture of beta posteriors, with mixture weights adapted to support prior that provides best prediction for the observed data.

- In R:

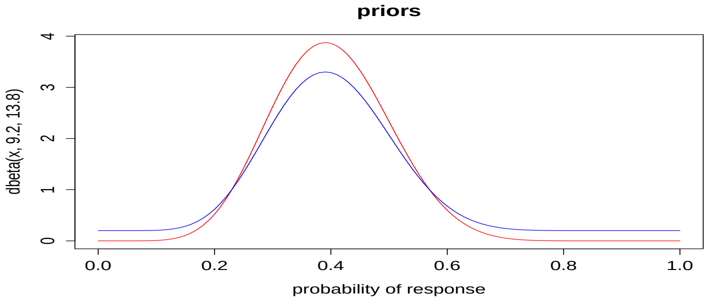
```
# mixture of betas
 mixbeta <- function(x,r,a1,b1,a2,b2,q,n)
  {
    qstar <- q*betabin(r,a1,b1,n)/
    (q*betabin(r,a1,b1,n)+(1-q)*betabin(r,a2,b2,n))

    p1 <- dbeta(x,a1+r,n-r+b1)
    p2 <- dbeta(x,a2+r,n-r+b2)
    posterior <- qstar*p1 + (1-qstar)*p2
  }
```

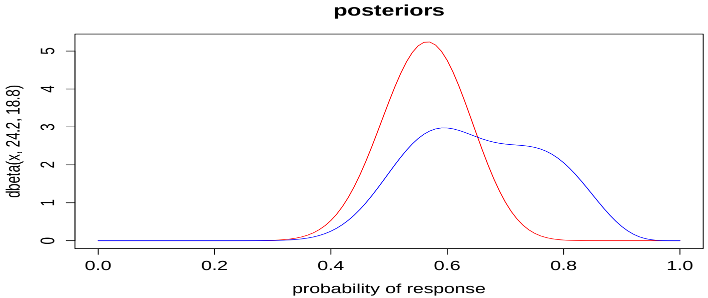
Example: drug investigation continue ...

- We want to combine:
  - an informative Beta(9.2, 13.8)
  - a non-informative Beta(1, 1)
  - we give 80 % of prior weight to the informative prior
- Suppose we treat $n = 20$ volunteers with the compound and observe $r = 15$ positive responses.

We can visualize this analysis in R as follows:

```
 par(mfrow=c(2,1))
# informative beta prior
curve(dbeta(x,9.2,13.8),from=0, to=1,col="red",
     xlab="probability of response",main="priors")
# mixture beta prior with 80% informative and 20% flat
  curve(0.8*dbeta(x, 9.2, 13.8)+0.2*dbeta(x, 1, 1),from=0,
     to=1,col="blue",add=T)
# beta posterior
 curve(dbeta(x,24.2,18.8),from=0,to=1,col="red",
      xlab="probability of response", main="posteriors")
# posterior from a mixture prior
  curve(mixbeta(x, r=15, a1=9.2, b1=13.8, a2=1, b2=1, q=.8
     from=0,to=1, col="blue",add=T)
par(mfrow=c(1,1))
```


priors
dbeta(x, 9.2, 13.8)
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
probability of response


posteriors
dbeta(x, 24.2, 18.8)
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
probability of response

Further work on priors:

- Adjustment of priors based on historical data and judgement
    1. Power priors (Ibrahim and Shen, 2000)
    2. Bias modelling
- Hirearchical priors for large dimentional problems

Summary:

- The need for priors distributions should not be an embarrassment
- It is reasonamble that the prior should influence the analysis, as long as the influence is recognized and justified
- Importance of transparency and sensitivity analysis

Lecture :

Introduction to WinBUGS

Summary

- Introduction to BUGS
- The BUGS language
- Some simple examples
- Making predictions
- Connecting WinBUGS with R

Introduction to BUGS

The BUGS project began at the Medical Research Council Biostatistics Unit in Cambridge in 1989, before the classic Gibbs sampling paper by Gelfand and Smith in 1990. An excellent review and future directions of the BUGS project is given in Lunn et al. (2009).

BUGS stands for **B**ayesian inference **u**sing **G**ibbs **s**ampling, reflecting the basic computational techniques originally adopted.

BUGS has been just one part of the tremendous growth in the application of Bayesian ideas over the last 20 years.

At this time BUGS has approximately over 30,000 registered users worldwide, and an active on-line community comprising over 8,000 members.

Think different ...

The modelling philosophy of BUGS was strongly influenced by developments in artificial intelligence in the 1980's. In particular the focus was on *Expert systems*, where a basic principle was to separate:

- Knowledge base:
  - assumed model for the world
  - makes use of a declarative form of programming
  - structures described using a series of local relationships
- Inference engine:
  - used to draw conclusions in specific circumstances

This approach forces to think first about the model use to describe the problem at hand. The declarative programming approach, sometimes, confuses practitioners of procedural statistical languages (SAS, SPSS, R, etc).

The BUGS language:

- Language for specifying complex Bayesian models.
- Constructs object-oriented internal representation of the model graph by identifying parents and children. This is done with a DAG (Directed Acyclic Graph).
- Builds up an arbitrary complex model through specification of local structure.
- Simulation from full conditionals using Gibbs sampling.
- Current version is WinBUGS 1.4.3, it runs in Windows, and incorporates the DoodleBUGS graphical model editor and a script language for running in batch mode.

WinBUGS is freely available from
`http://www.mrc-bsu.cam.ac.uk/bugs`

OpenBUGS is freely available from
`http://www.openbugs.net/w/FrontPage`

Example: Drug

In $n = 20$ patients we observed $r = 15$ positive responses.

$$y \sim \texttt{Bin}(\theta, n)$$

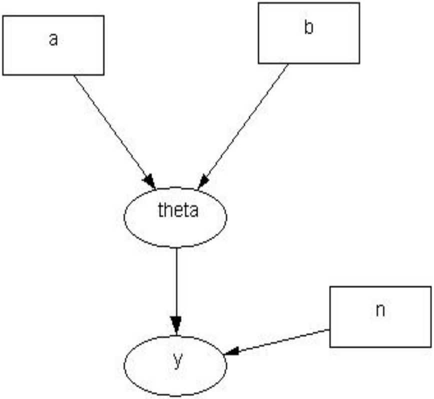
and we assume a conjugate prior for θ:

$$\theta \sim \texttt{Beta}(a, b)$$

Of course, we know the posterior distribution is

$$\theta | y \sim \texttt{Beta}(a + y, n - r + b)$$

and no simulation is necessary. But just to illustrate WinBUGS ...


a
b
theta
n
y

- Directed Acyclic Graph (DAG) representation:
  - *Ovals nodes* represent random variables
  - *Rectangular nodes* represent constants
  - *Arrows* parent child relationships

1. Write BUGS code to specify the model ...

   ```
   model {
       y ~ dbin(theta, n)
       theta ~ dbeta(a, b)
   }
   ```

2. Check the model syntax ...

3. Load the data ...

   ```
       list(n=20, y = 15, a = 3, b =2)
   ```

   ... and compile.

4. Then load initial values ...

   ```
       list(theta = 0.5)
   ```

5. Select the nodes (variables) to monitor (just one in this case)...

6. Set the trace for all selected nodes (*) to monitor the MCMC simulations ...

7. Using the Update tools, select 10,000 simulations ...

8. Results

Example: Drug continue ...

Now we assume a different prior. A logit transform of

$$\phi = \log\left(\frac{\theta}{1-\theta}\right),$$

so that $-\infty < \phi < \infty$

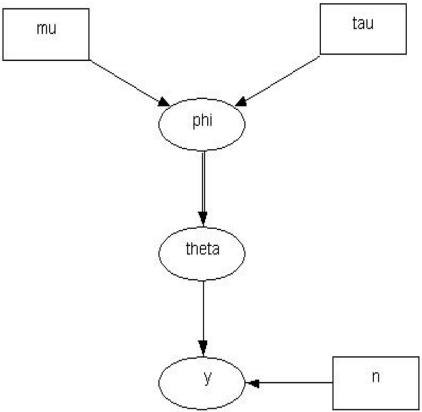
We assume a normal prior for ϕ:

$$\phi \sim \texttt{Normal}(\mu, \tau)$$

for suitable mean μ and precision $\tau = 1/\sigma^2$.

This is a non-conjugate prior with no simple form for the posterior.

Straightforward in WinBUGS!



Double arrows represent a logical node or a mathematical functional relationship.

The BUGS code for the model is ...

```
model
{
    y ~ dbin(theta, n)
    logit(theta) <- phi
    phi ~ dnorm(0, 0.001)
}
```

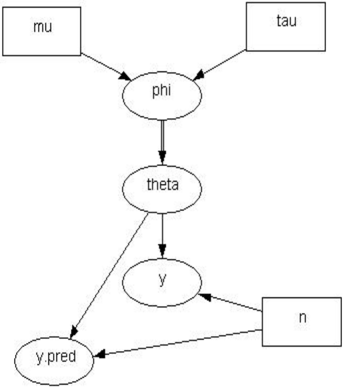
Making predictions

- Important to be able to predict unobserved quantities for
  - 'filling-in" missing or censored data
  - model checking - are predictions 'similar' to observed data?
  - making predictions!
- Easy in MCMC/WinBUGS, just specify a stochastic node without a data value - it automatically predicted
- Provides automatic imputation of missing data
- Easiest case is where there is no data at all!! Just 'forwards sampling' from prior to make a Monte Carlo analysis.

Example: making predictions

The BUGS code for the model is ...

```
model
{
    y ~ dbin(theta, n)
    logit(theta) <- phi
    phi ~ dnorm(mu, tau)
    y.pred ~ dbin(theta, n) # defines a predictions nod
}
```



The prediction node `y.pred` is conditionally independent of the data node `y` given rate θ.

Summary: Running WinBUGS

1. Open *Specification tool* and *Update* from *Model menu*, and *Samples* from *Inference* menu.
2. Highlight `model` by double-click. Click on *Check model*.
3. Highlight start of data. Click on *Load data*.
4. Click on *Compile*.
5. Highlight start of initial values. Click on *Load inits*.
6. Click on *Gen Inits* if model initials values are needed.
7. Click on *Update* to burn in.
8. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
9. Perform more updates.
10. Type * into *Sample Monitor*, and click *stats*, etc. to see results on all monitored nodes.

The R2WinBUGS package in R

The R package R2WinBUGS offers a versatile approach for making MCMC computations within WinBUGS and returning them to R.

Let see a step by step example of linking R and WinBUGS with R2WinBUGS.

Example: non-conjugate priors inference for a binomial experiment

First save the following file with WinBUGS code in your working directory ...

```
# WinBUGS code: binary problem non-conjugate analysis
  model
    { y ~ dbin(theta, n)
      logit(theta) <- phi
      phi ~ dnorm(0,0.001)
      y.pred ~ dbin(theta,n) # making prediction
    }
```

Then in R console ...

```
# load R2WinBUGS package
library(R2WinBUGS)

# setup WinBUGS directory and your working directory
bugsdir <-  "C:/Programme/WinBUGS14"
workdir <- getwd()

# define you data nodes
n <- 20
y <- 15
data1 <- list ("n", "y")
```

```
# define parameters of interest
par1 <- c("theta", "y.pred")

# run the bugs() function:
m1 <- bugs(data1, inits=NULL, par1, "model1.txt",
n.chains = 1, n.iter = 2000,  n.thin=1,
bugs.directory = bugsdir,
working.directory = getwd(),
debug=TRUE)
```

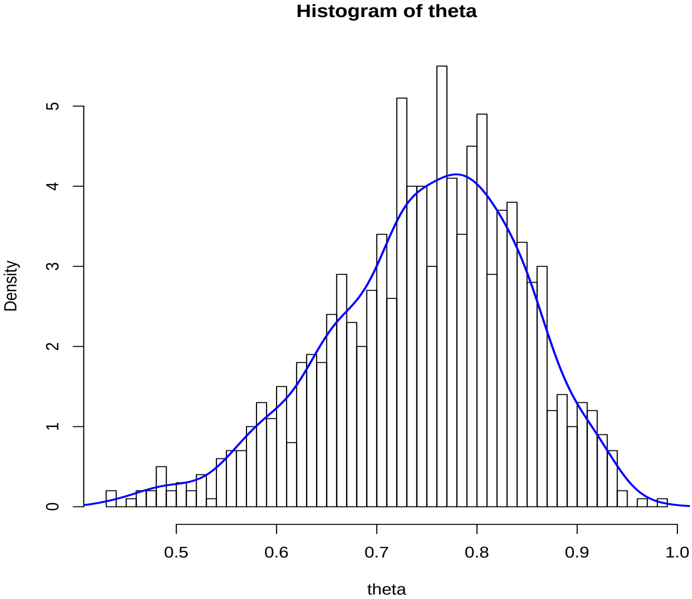
```
> print(m1, digits.summary = 3)
Inference for Bugs model at "model1.txt", fit using WinBUGS
 1 chains, each with 2000 iterations (first 1000 discarded)
 n.sims = 1000 iterations saved
           mean    sd  2.5%    25%    50%    75%  97.5%
theta     0.749 0.096 0.543  0.690  0.760  0.819  0.914
y.pred   14.962 2.733 9.000 13.000 15.000 17.000 19.000
deviance  4.277 1.584 3.198  3.286  3.667  4.551  9.014

DIC info (using the rule, pD = Dbar-Dhat)
pD = 1.1 and DIC = 5.3
DIC is an estimate of expected predictive error ...
```

The R object `m1` generated in this analysis belong to the class `bug` and can be further manipulated, transformed, etc.

```
> class(m1)
[1] "bugs"
> names(m1)
 [1] "n.chains"        "n.iter"          "n.burnin"
 [4] "n.thin"          "n.keep"          "n.sims"
 [7] "sims.array"      "sims.list"       "sims.matrix"
[10] "summary"         "mean"            "sd"
[13] "median"          "root.short"      "long.short"
[16] "dimension.short" "indexes.short"   "last.values"
[19] "isDIC"           "DICbyR"          "pD"
[22] "DIC"             "model.file"      "program"
```

The object m1 is a list in R, so we can extract elements of the list by using the "$" operator, for example:

```
> m1$pD
[1] 1.065
> m1$n.chains
[1] 1
>  m1$sims.array[1:10, ,"theta"]
 [1] 0.8389 0.9124 0.7971 0.8682 0.7025 0.7696 0.8417 0.678
[10] 0.8647
>  m1$sims.array[1:10, ,"y.pred"]
 [1] 15 19 19 20 15 20 18 17 10 17
> theta <- m1$sims.array[1:1000, ,"theta"]
> hist(theta, breaks = 50, prob = TRUE)
> lines(density(theta), col = "blue", lwd =2)
```


Histogram of theta
Density
0
1
2
3
4
5
0.5
0.6
0.7
0.8
0.9
1.0
theta

Some aspects of the BUGS language

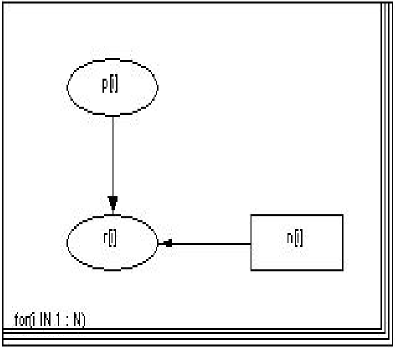
- `<-` represents logical dependence, e.g. `m <- a + b*x`

- `~` represents stochastic dependence, e.g. `r ~ dunif(a,b)`

- Can use arrays and loops

```
 model{
 ....
     for (i in 1:N){
         r[i] ~ dbin(p[i], n[i])
     }
...
}
```

A for loop is represented as a "plate" in the DAG's model



Some aspects of the BUGS language

- Some functions can appear on the left-hand-side of an expression, e.g.

```
logit(p[i]) <- a + b*x[i]
log(m[i]) < - c + d*y[i]
```

- mean(p[]) to take mean of whole array, mean(p[m : n]) to take mean of elements m to n. Also for sum(p[])

- dnorm(0, 1)I(0,) means the random variable will be restricted to the range $(0, \infty)$.

Functions in the BUGS language

- `p <- step(0.05 - x)` $= 1$ if $x \leq 0.05$, 0 otherwise. Hence monitoring `p` and recording its mean will give the probability that $x \leq 0.05$. This is useful to calculate Bayesian p-values.

- `p <- equals(x, 0.7) = 1` if $x = 0.7$, 0 otherwise.

- `tau <- 1/pow(s,2)` sets $\tau = 1/s^2$.

- `s <- 1/sqrt(tau)` sets $s = 1/\sqrt(\tau)$

- `p[i,k] <- inprod(p[], Lambda[i,k])` sets $p_{ik} = \sum_j \pi_j \Lambda_{ij}$.

- See 'Model Specification/Logical nodes' in the manual for full syntax.

Data transformations

Although transformations of data can always be carried out before using WinBUGS, it is convenient to be able to try various transformations of dependent variables within a model description.

For example, we may wish to try both `y` and `sqrt(y)` as dependent variables without creating a separate variable `z = sqrt(y)` in the data file.

The BUGS language therefore permits the following type of structure to occur:

```
for (i in 1:N) {
z[i] <- sqrt(y[i])
z[i] ~ dnorm(mu, tau)
}
```

Strictly speaking, this goes against the declarative structure of the model specification.

Some common distributions

- Binomial: $\texttt{r} \sim \texttt{dbin}(\texttt{p}, \texttt{n})$
- Normal: $\texttt{x} \sim \texttt{dnorm}(\texttt{mu}, \texttt{tau})$
- Poisson: $\texttt{r} \sim \texttt{dpois}(\texttt{lambda})$
- Uniform: $\texttt{x} \sim \texttt{dunif}(\texttt{a}, \texttt{b})$
- Gamma: $\texttt{x} \sim \texttt{dgamma}(\texttt{a}, \texttt{b})$

Note: The normal distribution is parameterized in terms of its mean and $\texttt{precision} = 1/\texttt{variance} = 1/\sigma^2$.

Functions cannot be used as arguments in distributions. You need to create new nodes.

The WinBUGS data formats

WinBUGS accepts data files in:

1. Rectangular formant

   ```
   n[] r[]
   50 2
   ....
   20 4
   END
   ```

2. R list format:

   ```
   list(N =12, n = c(50,12,...), r = c(2, 1,...))
   ```

Double indexing: Specifying categorical explanatory variables

```
y[] x[]
12  1   #subject with x=level 1
34  2   #subject with x=level 2
 ...
 for( i in 1:N) {
    y[i] ~dnorm(mu[i], tau)
    mu[i] <- alpha + beta[x[i]]
    }
 alpha ~ dunif(-100,100)
 beta[1] <- 0   # alias first level of beta
 beta[2] ~ dunif(-100, 100)
 beta[3] ~ dunif(-100, 100)
 tau ~ dgamma(0.1,0.1)
```

Practical: Statistical modeling with WinBUGS

Exercise: Inference on the sex ratio

- A particular maternal condition during pregnancy was thought to influence the sex of the child. The proportion of female birth in the population was $p = 0.485$. A sample of 98 births to women with the condition resulted in 43 females. Is this evidence that the condition reduces the proportion of female births?
- If θ denotes the proportion of female births to women with the condition, we can assume that the observed number, y of female births is $y \sim \text{Binomial}(\theta, n)$ where $n = 98$.

Different models based on three different priors:

1. *Non-informative prior:* $\theta \sim \text{Beta}(1, 1)$.
2. *Informative prior:* Beta with mean 0.485 and take $a + b = 100$, so that $\theta \sim \text{Beta}(48.5, 51.5)$
3. *Non-conjugate prior:* it could be argued that priors 1) and 2) are unrealistic because they allow values of θ close to 0 or 1. So the triangular distribution on (0.4, 0.6) has been proposed. This has zero probability outside the range (0.4, 0.6). In order to sample from this distribution use the results that the sum of two Uniforms variables has a triangular distribution, e.g., $U_1 \sim U(0.2, 0.3)$ and $U_2 \sim U(0.2, 0.3)$ then $Y = U_1 + U_2$ follows a triangular with parameters $(0.4, 0.6)$.

Once you have implemented these three models in WinBUGS perform the following analyzes:

1. Draw a DAG for the model with Beta prior for θ.
2. Draw a DAG for the model with Triangular prior for θ.
3. Estimate the posterior for θ under the three different prior models.
4. Estimate the posterior for the odds $(1-\theta)/\theta$ under the three different prior models.
5. Estimate the posterior probability that $\theta < 0.485$ for the three prior models.
6. Which are your conclusions after comparing theses results.
7. Now, suppose that had been 437 female births out of 980 instead of 43 out of 98. Repeat the analysis with the same three priors for θ. How much difference does the choice of prior make in this case?

Practical: Statistical modeling with WinBUGS

Exercise: Survival data

Aitchison & Dunsmore (1975) give data on survival times (in weeks) of 20 patients after treatment for a particular carcinoma. The data set is

```
list(survtime=c(25, 45, 238, 194, 16, 23, 30, 16,
                       22, 123, 51, 412, 45, 162,
                         14, 72, 5, 43, 45, 91),  N=20)
```

1. Write a WinBUGS script for the following model:

   $$\begin{aligned} y_i &= \log(\text{survival[i]}) \\ y_i|\mu,\tau &\sim N(\mu,\tau), \quad \tau = 1/\sigma^2 \end{aligned}$$

   with non-informative priors for $\mu \sim \mathrm{N}(0, 10^{-3})$ and $\tau \sim \mathrm{Gamma}(10^{-3}, 10^{-3})$.

2. Use as initial values

   ```
           list( mu = 0, tau = 1)
   ```

3. Draw a DAG for this model.

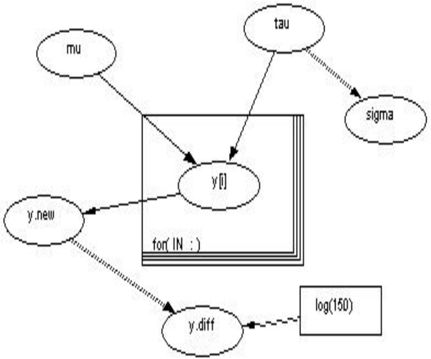
4. Run the model with 10,000 iterations and discard the first 5,000. Analyze visually convergence.

5. Estimate the probability that a new patient has survival time more than 150 weeks.

Solution

```
# WinBUGS script:
model{
 for ( i in 1:N)
{
    y[i] <- log(survtime[i])    #data transform
    y[i] ~ dnorm(mu, tau)       #sampling model
}
sigma <- 1/sqrt(tau)            #standard deviation
mu ~ dnorm(0, 1.0E-3)           #prior for mu
tau ~ dgamma(1.0E-3, 1.0E-3)    #prior for tau
y.new ~ dnorm(mu, tau)          #predicted data
y.dif <- step(y.new - log(150)) #pr(survival>150)
}
```

Solution



Lecture :

Introduction to Multiparameter Models

Summary

- Introduction to Multiparameter Inference
- Normal model with unknown mean and variance: standard non-informative priors
- Using R for predictive model checking
- Multinomial Model conjugate analysis and its application in R
- Comparing classical and Bayesian multiparameter models
- Multivariate Normal Models
- Complex Contingency Tables

Introduction

- The basic ideas is similar to one parameter models:
  - We have a model for the observed data which defines a likelihood $p(y|\theta)$ on the vector parameter θ
  - We specify a joint prior distribution $p(\theta)$ for the possible values of θ
  - We use Bayes' rule to obtain the posterior of θ,

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

- Problems:
  - In many applications is difficut to specify priors on multivariate θ
  - Computations could be very difficult, they involve multivariate integration

Example: Normal with unknown mean μ and variance σ^2

- In this example we use `marathontimes` in the R package `LearnBayes`

- The obervations are the mean times (in minutes) for men running Marathon with age classes between 20 to 29 years old:

```
> library(LearnBayes)
>  data(marathontimes)
>  marathontimes$time
 [1] 182 201 221 234 237 251 261 266 267 273 286 291 29
>
```

- We assume a Normal model for the data given the mean μ and the variance σ^2:

$$y_1, \dots, y_{20} | \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

- We use a non-informative Jeffreys' prior assuming independence between location and scale:

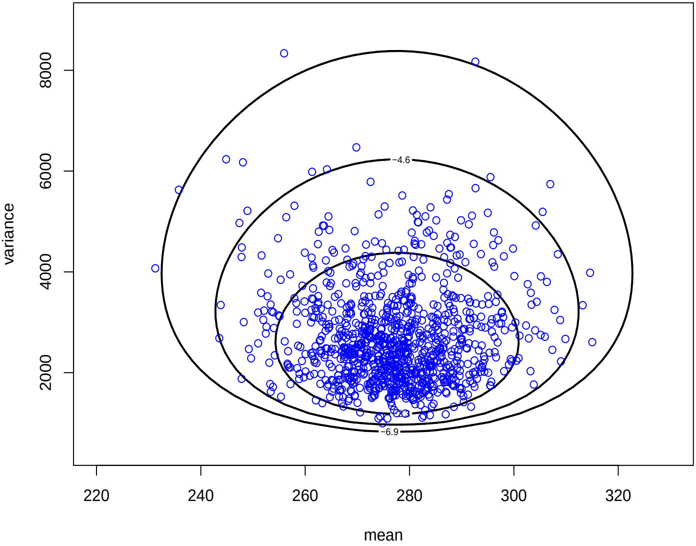
$$p(\mu, \sigma) \propto \frac{1}{\sigma},$$

- Then posterior density for the mean and the variance $p(\mu, \sigma | y)$ is called the Normal-χ^2 distribution
  - This posterior delivers same results as classical analysis: $E(\mu|y) = \bar{y}$, $E(\sigma^2|y) = S^2$
  - The posterior $p(\mu|y)$ is proportional to a t-distribution with $n-1$ df, and so on
  - Direct simulation from these posteriors is straghtforward

- The Normal-χ^2 posterior is implemented in the function `normchis2post` in LearnBayes, which computes the logarithm of the joint posterior density of (μ, σ^2)
- We can visualize the α% levels contours with the function `mycontour`. The arguments include the log-density to plot, the rectangular area $(x_{lo}, x_{hi}, y_{lo}, y_{hi})$ and the data:

  ```
  > mycontour(normchi2post,c(220,330,500,9000),time,
  xlab="mean",ylab="var")
  ```
- We can simulate directly form the marginal posteriors of μ and τ:

  ```
  > SS <- sum((time - mean(time))^2)
  > n <- length(time)
  > sigma2 <- SS/rchisq(1000, n - 1)
  > mu <- rnorm(1000, mean = mean(time),
    sd = sqrt(sigma2)/sqrt(n))
  > points(mu, sigma2, col="blue")
  > quantile(mu, c(0.025, 0.975))  # 95% posterior interval fo
      2.5%    97.5%
  253.1121 301.1133
  ```


−4.6
−2.3
−6.9
8000
6000
4000
2000
variance
220
240
260
280
300
320
mean

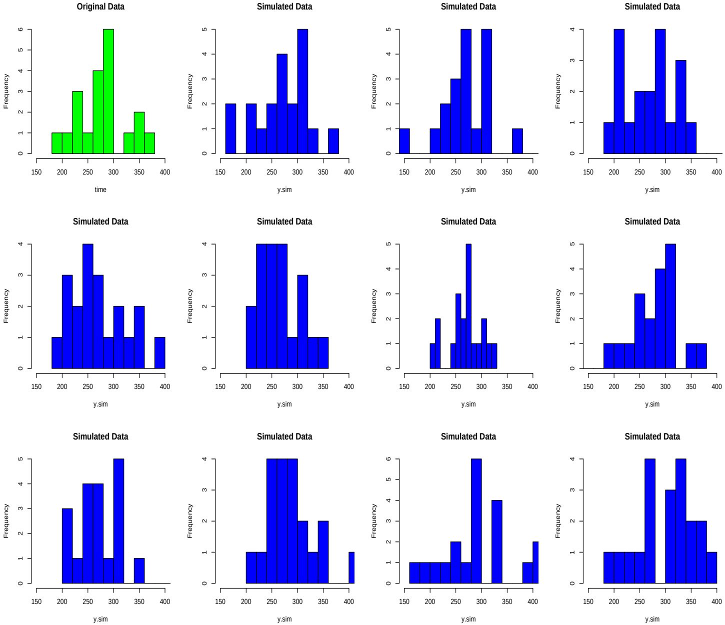
Predictive model checking

Is the statistical model consistent with the data ?

We answer this question by simulating **predictive** y^* values from the model and comparing some **data features** with the predictive data.

We start by looking at some histograms between the **original data** and the **simulated data** with the **same sample size.**

```
#predictive data
y.star <- rnorm(1000, mean = mu, sd =sqrt(sigma2))
par(mfrow=c(3,4))
hist(time, breaks=10, xlim =c(150, 400),
     main="Original Data", col="green")
for(i in 1:11){
   y.sim <- sample(y.star, 20)
   hist(y.sim,breaks=10,xlim=c(150, 400),
   main="Simulated Data",col="blue")
 }
```


Original Data
Frequency
0
1
2
3
4
5
6
150
200
250
300
350
400
time
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim
Simulated Data
Frequency
y.sim

Predictive model checking

We define the following features between the observed data y and the predictive data y^*:

$$T_1^* = min(y^*), \quad T_2^* = max(y^*), \quad T3^* = q_{75}(y^*) - q_{25}(y^*)$$

and for asymmetry

$$T_4^* = |y^*_{(18)} - \mu^*| - |y^*_{(2)} - \mu^*|$$

where the 18th and 2sd order statistics approximate the 90% and 10% respectively.

These measures are compared with the corresponding values based on the observed data:

$$T_1 = min(y), \quad T_2 = max(y), \quad T3 = q_{75}(y) - q_{25}(y)$$

and

$$T_4 = |y_{(18)} - \mu^*| - |y_{(2)} - \mu^*|.$$

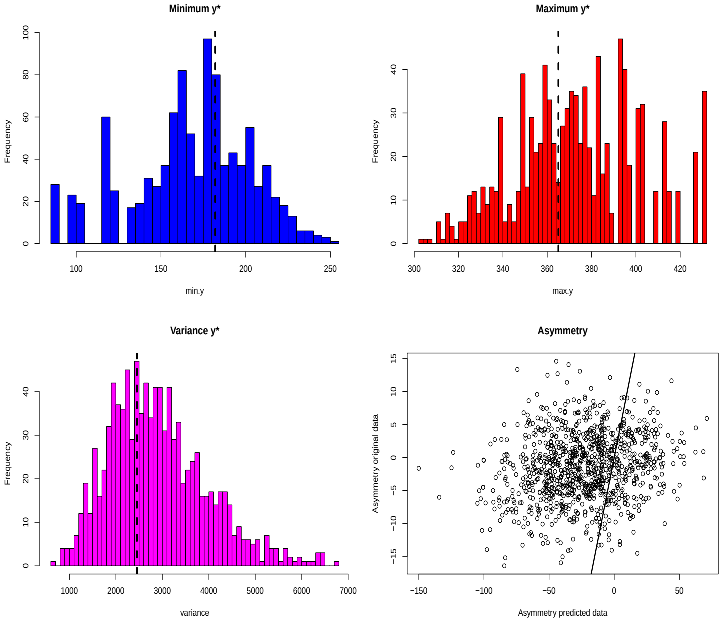
In R we have:

```
# Analysis of the minimum, maximum, variability and asymmetry

min.y <- max.y <- asy1 <- asy2 <- inter.q <-  rep(0,1000)
time <- sort(time)
for (b in 1:1000){
 y.sim <- sample(y.star, 20)
 mu.star <- sample(mu, 20)
 min.y[b] <- min(y.sim)
 max.y[b] <- max(y.sim)
 y.sim <- sort(y.sim)
 asy1[b] <- abs(y.sim[18]-mean(mu.star))-abs(y.sim[2]-mean(mu.st
 asy2[b] <- abs(time[18]-mean(mu.star))-abs(time[2]-mean(mu.star
 inter.q[b] <- quantile(y.sim, prob=0.75)-quantile(y.sim, prob=0
}
```

To display this quantities

```
par(mfrow =c(2,2))
   hist(min.y, breaks = 50, col="blue", main = "Minimum y*")
   abline(v=min(time), lwd=3, lty=2)
   hist(max.y, breaks = 50, col="red", main = "Maximum y*")
   abline(v=max(time), lwd=3, lty=2)
   hist(inter.q, breaks = 50, col="magenta",
        main = "Variability y*")
   abline(v=quantile(time,prob=0.75)-quantile(time,prob=0.25),
        lwd=3,lty=2)
   plot(asy1, asy2, main ="Asymmetry",
       xlab="Asymmetry predicted data",
       ylab ="Asymmetry original data")
   abline(a=0, b=1, lwd=2)
par(mfrow=c(1,1))
```

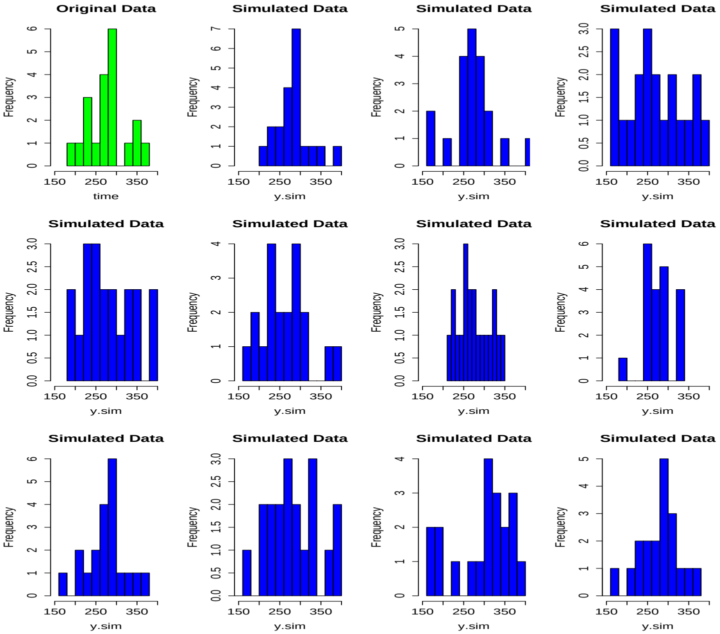
This analysis shows that the data and the model are compatibles and deviations can be easily explained by sampling variation.

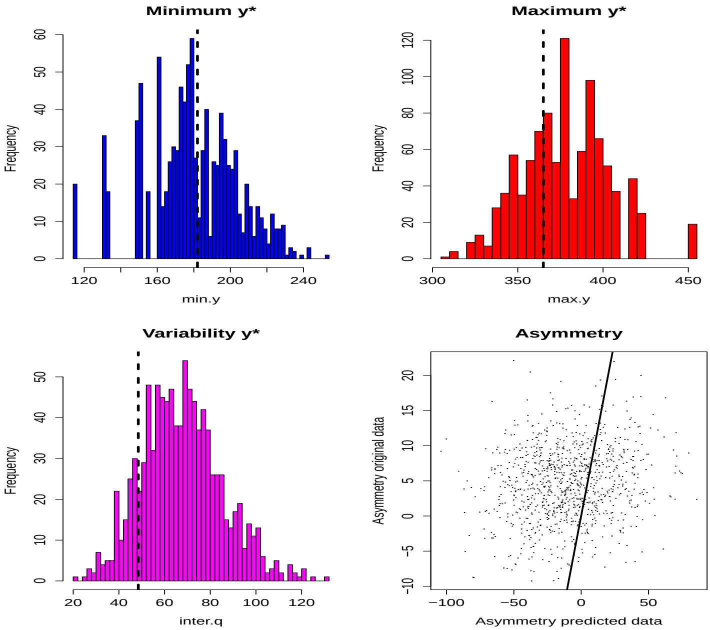

Minimum y*
Frequency
0
20
40
60
80
100
100
150
200
250
min.y
Maximum y*
Frequency
0
10
20
30
40
300
320
340
360
380
400
420
max.y
Variance y*
Frequency
0
10
20
30
40
1000
2000
3000
4000
5000
6000
7000
variance
Asymmetry
Asymmetry original data
−15
−10
−5
0
5
10
15
−150
−100
−50
0
50
Asymmetry predicted data

Marathon times with WinBUGS

Different types of likelihood functions to handle implausible values

```
model
   {
   # Non informative Priors
   mu ~ dnorm(0, 0.0001)
   tau ~ dgamma(0.0001, 0.0001)
   sigma <- pow(tau,-0.5)
   # Data model
   for( i in 1: n)
   {
   #y[i] ~ dnorm(mu, tau)         # Normal Likelihood
   y[i] ~ dnorm(mu, tau) I(115, ) # Truncated Normal Likelihood
   #y[i] ~ dt(mu, tau, nu)        # t Likelihood with nu df.
   }
   y.pred ~ dnorm(mu, tau)I(115, )# Predictions
   }
```


Original Data
Frequency
0
1
2
3
4
5
6
150
250
350
time
Simulated Data
Frequency
0
1
2
3
4
5
6
7
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
5
150
250
350
y.sim
Simulated Data
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
150
250
350
y.sim
Simulated Data
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
150
250
350
y.sim
Simulated Data
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
5
6
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
5
6
150
250
350
y.sim
Simulated Data
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
150
250
350
y.sim
Simulated Data
Frequency
0
1
2
3
4
5
150
250
350
y.sim


Minimum y*
Frequency
0
10
20
30
40
50
60
120
160
200
240
min.y
Maximum y*
Frequency
0
20
40
60
80
100
120
300
350
400
450
max.y
Variability y*
Frequency
0
10
20
30
40
50
20
40
60
80
100
120
inter.q
Asymmetry
Asymmetry original data
−10
−5
0
5
10
15
20
−100
−50
0
50
Asymmetry predicted data

Multinomial distribution with unknown cell probabilities

In this model we observe n independent trials each of which has p possible outcomes with associated probabilities

$$\theta = (\theta_1, \ldots, \theta_p).$$

Letting $y = (y_1, \ldots, y_p)$ be the number of times each outcome is observe we have

$$y|\theta \sim \text{Multinomial}(n, \theta),$$

with likelihood

$$p(y|\theta) = \frac{(\sum y_i)!}{\prod y_i!} \prod \theta_i^{y_i}, \quad \sum_i y_i = n, \sum \theta_i = 1.$$

The kernel of this likelihood is proportional to a Dirichlet distribution. If a-prior we assume

$$\theta \sim \text{Dirichlet}(a_1, \ldots, a_p), \quad a_i > 0,$$

then

$$p(\theta) \propto \prod_i \theta^{a_i - 1}$$

and the posterior for θ is

$$p(\theta|y) \propto \prod_i \theta^{y_i + a_i - 1}$$

so taking $\alpha_i = a_i + y_i$, then the posterior for θ is

$$p(\theta|y) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_p).$$

Some comments

- The prior distribution is equivalent to a likelihood resulting from $\sum a_i$ observations, with a_i observations in the ith category.
- A uniform density is obtained by setting $a_i = 1$ for all i. This distribution assigns equal probability to each θ_i.
- In general it is very difficult to work with this distribution in practice. Proportions are in general not independent and modeling prior information in this context it is very difficult.
- We are going to use for contingency tables a surrogate Poisson model for the multinomial distribution.
- Moments and properties for the Dirichlet distribution are in Appendix A of Gelman et.al.

Example: A model for a two by two contingency table

	Intervention		
	New	Control	
Death	$\theta_{1,1}$	$\theta_{1,2}$	
No death	$\theta_{2,1}$	$\theta_{2,2}$	
			N

Data model:

$$p(y|\theta) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{y_{i,j}}$$

Prior model:

$$p(\theta) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{a_{i,j}-1}$$

Posterior model:

$$p(\theta|y) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{y_{i,j}+a_{i,j}-1}$$

We are interested in make inference for the Odds ratio:

$$\Psi = \frac{\theta_{1,1}\theta_{2,2}}{\theta_{1,2}\theta_{2,1}}.$$

We use direct simulation methods:

- Simulate a large number of values for the vector θ from its posterior.
- For each simulated value calculate Ψ^*.
- Inference for Ψ is based on the histogram of Ψ^*.

Example: GREAT trial, Spiegelhalter et.al. pag 69.

Intervention: Thrombolytic therapy after myocardial infarction, given at home by general practitioners.

Aim of study: to compare a new drug treatment to be given at home as soon as possible after a myocardial infarction and placebo.

Outcome measure: Thirty-day mortality rate under each treatment, with the benefit of the new treatment measured by the odds ratio, i.e., the ratio of the odds of death following the new treatment to the odds of death on the conventional: $OR < 1$ therefore favors the new treatment.

Prospective Bayesian analysis: NO. It was carried out after the trial reported its results.

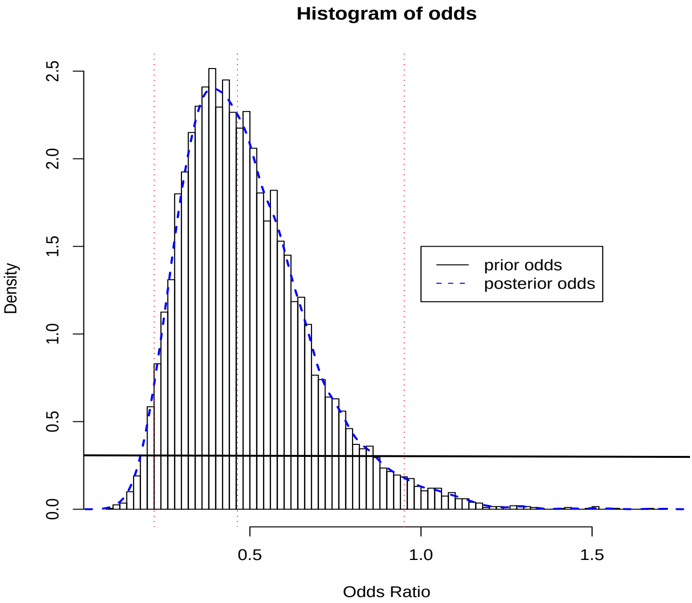
Example: GREAT trial continue

	Intervention		
	New	Control	
Death	13	23	36
No death	150	125	275
	163	148	311

We use a Dirichlet prior with parameters $a_{1,1} = a_{1,2} = a_{2,1} = a_{2,2} = 1$, which corresponds to a uniform distribution for $(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2})$.

We can simulate from a Dirichlet distributions with the function `rdirichlet()` from the package LearnBayes:

```
> library(LearnBayes)
>  draws <- rdirichlet(10000, c(13,23,150,125) )
>  odds <-draws[,1]*draws[,4]/(draws[,2]*draws[,3])
>  hist(odds, breaks = 100, xlab="Odds Ratio", freq = FALSE
>  lines(density(odds), lty =2, lwd=2, col ="blue")
>  abline(v=quantile(odds, prob=c(0.025, 0.5, 0.975)),
    lty=3, col="red")
>
>  quantile(odds, prob=c(0.025, 0.5, 0.975))
     2.5%       50%     97.5%
0.2157422 0.4649921 0.9698272
```


Histogram of odds
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.5
1.0
1.5
Odds Ratio
prior odds
posterior odds

Now suppose that we want to calculate the ”p-value” of

$$H_0 : \Psi \geq 1 \quad \text{vs.} \quad H_1 : \Psi < 1$$

then,

```
>  sum(odds > 1)/10000
[1] 0.0187
```

It is interesting to compare this results with the exact Fisher test:

```
>  fisher.test(matrix(c(13, 23, 150, 125),
   nrow=2, byrow=TRUE), alternative="less")
        Fisher’s Exact Test for Count Data
p-value = 0.02817
```

Thus, the uniform prior on $(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2})$ does not corresponds to a standard statistical analysis.

Now, if we work with a prior with parameters $a_{1,1} = 2$, $a_{1,2} = 1$, $a_{2,1} = 1$ and $a_{2,2} = 2$, with density

$$p(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}) \propto \theta_{1,1}\theta_{2,2}$$

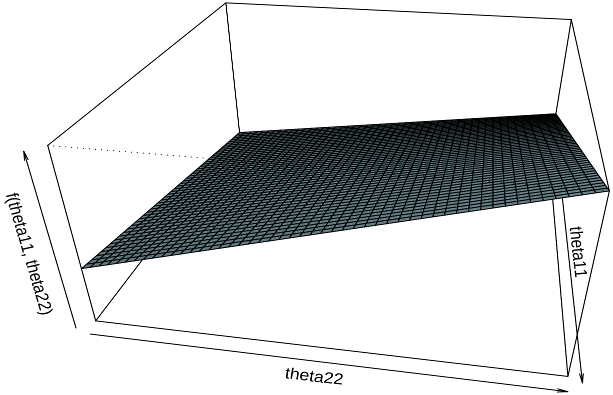
we get the following results

```
> # Fisher exact test
>  param <- c(2,1,1,2)  # parameter of the dirichlet
>  draws <- rdirichlet(10000, c(13,23,150,125)+param )
>  odds <-draws[,1]*draws[,4]/(draws[,2]*draws[,3])
>  sum(odds>1)/10000
[1] 0.0277
```

This shows (empirically) that the Fisher test is NOT based in a non-informative prior. Some weakly information of association is implied in the test.

However, there is a difference in interpretation between the Bayesian and sampling theory results.

Dirichlet with a11=2, a12=1, a21=1, a22=2



Multivariate Normal Distribution

Multivariate Normal for p-dimensional vector y

$$y \sim \text{Normal}_p(\mu, \Sigma)$$

The conjugate prior for μ is also a MVN. WinBUGS follows parametrization using the precision matrix $\Omega = \Sigma^{-1}$.

In BUGS notation we have

```
y[1:p] ~ dmnorm(mu[], Omega[,])
mu[1:p] ~ dmnorm(mu.prior[], Omega.prior[])
Sigma[1:p, 1:p] <- inverse(Omega[,])
```

Priors on precision matrix of multivariate normals

Conjugate prior is the Wishart distribution, which is analogous to Gamma or χ^2.

Arises in classical statistics as the distribution of the sum-of-squares and products matrix in multivariate normal sampling.

The Wishart distribution $W_p(k, R)$ for a symmetric positive definite $p \times p$ matrix Ω has density

$$p(\Omega) \propto |R|^{k/2} |\Omega|^{(k-p-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(R\Omega)\right),$$

defined for a real scalar $k > p - 1$ and a symmetric positive definite matrix R.

Some characteristics

- When $p = 1$
$$\mathrm{W}_1(k, R) \equiv \mathrm{Gamma}(k/2, R/2) \equiv \chi^2_k/R.$$
- The expectation of the $\mathrm{W}_p(k, R)$ is
$$E[\Omega] = kR^{-1}.$$
- Jeffreys prior is
$$p(\Omega) \propto |\Sigma|^{-(p+1)/2},$$
equivalent to $k \to 0$. This is not currently implemented in WinBUGS.
- For "weakly informative" we can set R/k to be a rough prior guess at the unknown true covariance matrix and taking $k = p$ indicates minimal "effective prior sample size".

Known problems

- Every element must have same precision
- Incomprehensible! It is a good idea to make some prior predictions in practice to understand what we are doing.
- Gelman at al (2004) page 483 outline alternative strategies
- If do not use Wishart priors for precision matrices in BUGS, you need to make sure that the covariance matrix at each iteration is positive-definite, otherwise may crash.

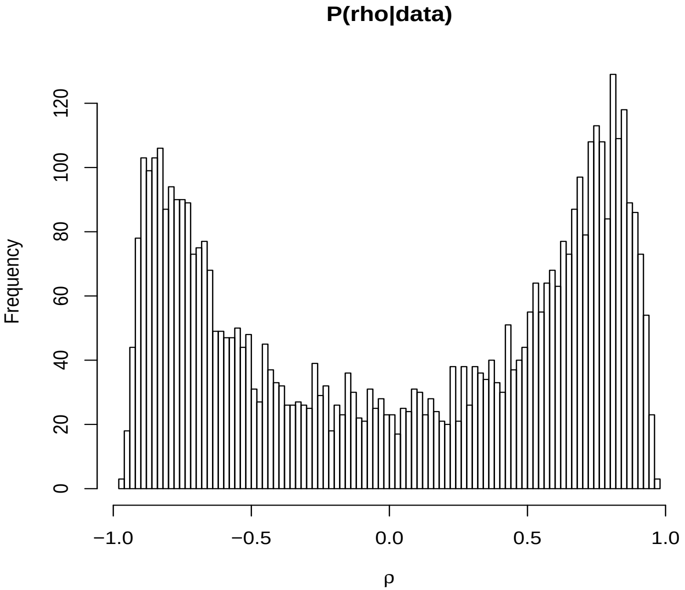
Exercise 1: **Correlation Analysis with missing data**

The following is an artificial data from the book *Tools for Statistical Inference* (Tanner pag 63.(1994)):

```
> Y
      [,1] [,2]
 [1,]    1    1
 [2,]    1   -1
 [3,]   -1    1
 [4,]   -1   -1
 [5,]    2   NA
 [6,]    2   NA
 [7,]   -2   NA
 [8,]   -2   NA
...
```

The following script in WinBUGS implement a Bayesian analysis for this problem.

```
model
    {
        for (i in 1 : 12){
            Y[i, 1 : 2] ~ dmnorm(mu[], tau[ , ])
        }
        mu[1] <- 0
        mu[2] <- 0
        tau[1 : 2,1 : 2] ~ dwish(R[ , ], 2)
        R[1, 1] <- 0.001
        R[1, 2] <- 0
        R[2, 1] <- 0
        R[2, 2] <- 0.001
        Sigma2[1 : 2,1 : 2] <- inverse(tau[ , ])
        rho <- Sigma2[1, 2] / sqrt(Sigma2[1, 1] * Sigma2[2,
```



Using predictive posterior values in each iteration WinBUGS

Multinomial model: non-conjugate analysis

In some applications the cell probabilities of contingency tables can be made function of more basic parameters

Example: Population genetics example

	Offspring genotype		
Maternal genotype	AA	AB	BB
AA	427	95	-
AB	108	161	71
BB	-	64	74

The model equations are given by the following table:

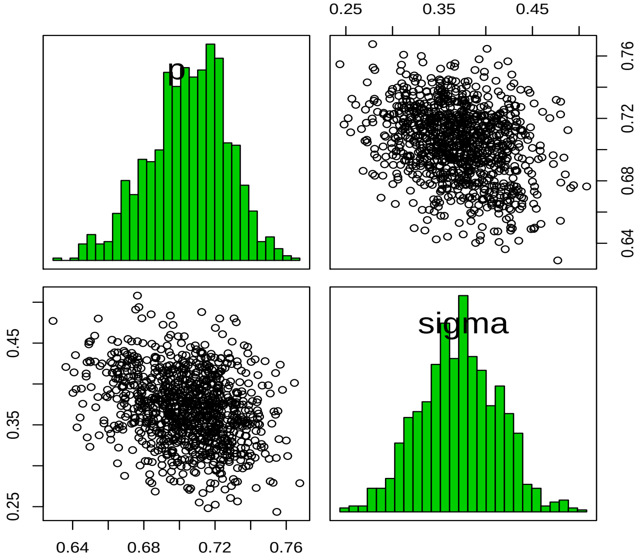
	Offspring genotype		
Maternal genotype	AA	AB	BB
AA	$(1-\sigma)p+\sigma$	$(1-\sigma)q$	-
AB	$(1-\sigma)p/2+\sigma/4$	$1/2$	$(1-\sigma)q/2+\sigma/4$
BB	-	$(1-\sigma)q$	$(1-\sigma)q+\sigma$

where p is the frequency of A in outcross pollen, σ is the rate of self-fertilization and $q = 1 - p$

To implement this model in WinBUGS, we equate cell probabilites with requred function:

```
model{
XAA[1]<- (1-sigma)*p + sigma;  XAA[2]<- (1- sigma)*q;
XAA[3]<- 0
XAB[1]<-(1-sigma)*p/2 +sigma/4; XAB[2]<-0.5;
XAB[3]<-(1-sigma)*q/2 +sigma/4
XBB[1]<- 0; XBB[2]<- (1-sigma)*p;
XBB[3]<- (1- sigma)*q + sigma
KAA <- sum(NAA[]); KAB <- sum(NAB[]); KBB <- sum(NBB[])

NAA[1:3]   ~ dmulti(XAA[], KAA)
NAB[1:3]   ~ dmulti(XAB[], KAB)
NBB[1:3]   ~ dmulti(XBB[], KBB)
p ~ dunif(0, 1)                     # uniform prior for p
sigma ~ dunif(0, 1)                 # uniform prior for si
q <- 1 -p
}
list(NAA = c(427, 95, 0), NAB=c(108, 161, 71),
 NBB = c(0, 64, 74))
```

```
> print(m.popgen, digits=3)
Inference for Bugs model at "popgen.bug", fit using WinBUG
 2 chains, each with 10000 iterations (first 5000 discarde
 n.sims = 10000 iterations saved
           mean    sd   2.5%    25%    50%    75%  97.5%  
p         0.705 0.024  0.655  0.690  0.706  0.721  0.749 1.
sigma     0.371 0.042  0.288  0.343  0.371  0.400  0.451 1.
deviance 27.151 1.972 25.210 25.730 26.540 27.960 32.410 1.

For each parameter, n.eff is a crude measure of effective 
and Rhat is the potential scale reduction factor (at conve

DIC info (using the rule, pD = Dbar-Dhat)
pD = 2.0 and DIC = 29.1
```


p
sigma
0.25
0.35
0.45
0.76
0.72
0.68
0.64
0.45
0.35
0.25
0.64
0.68
0.72
0.76

Practical: Multiple Parameters with R

Exercise 1:

- Repeat the Marathon times example of the lecture
- Make a change in the Marathon times as following:

```
marathontimes$time[1:5] <- rnorm(5, mean = 130, sd = 3
```

- Repeat the posterior distribution predictive checks

Exercise 2:

- Repeat the Example of the GREAT trial
- Calculate the posterior distribution of the difference between treatments' the probability of death. Hint:

```
diff <- draws[, 1] - draws[,2]
```

Lecture :

Bayesian Regression Models

Modeling Examples

- Multiple linear regression: model diagnostics and variable selection
- ANOVA models: an alternative to multiple testing
- Logistic regression: combining multiple information in Risk analysis

Introduction

Standard (and non standard) regression models can be easily formulated within a Bayesian framework.

- Specify probability distribution for the data
- Specify form of relationship between response and explanatory variables
- Specify prior distribution for regression coefficients and any other unknown (nuisance) parameters

Some advantages of a Bayesian formulation in regression modeling include:

- Easy to include parameter restrictions and other relevant prior knowledge
- It is simple to extended to non-linear regression
- We can easily robustified a model
- Easy to make inference about functions of regression parameters and/or predictions
- We can handle missing data and covariate measurement error
- Transparent variable selection by prior modeling regression coefficients

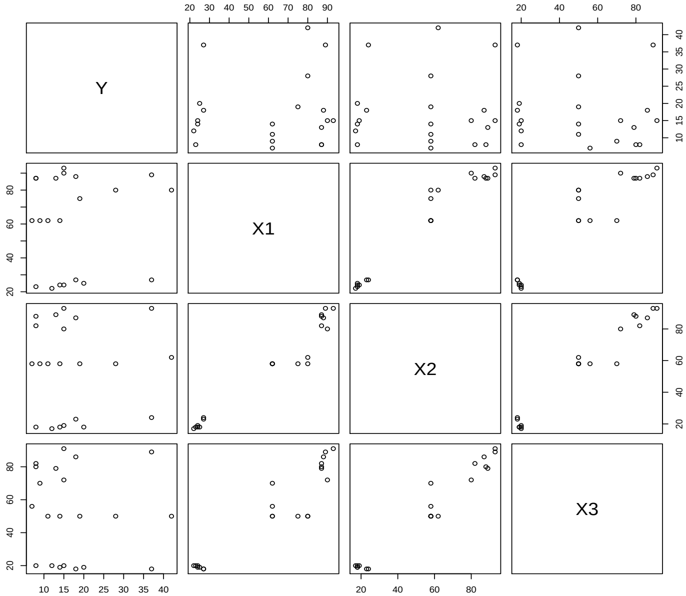
Linear regression

Example: Stack loss data, WinBUGS Volume 1

In this example, we illustrate a more complete regression analysis including outlier checking, model adequacy and variable selection methods.

This is a very often analyzed data of Brownlee (1965, p. 454).

- 21 daily responses of stack loss, y_i the amount of ammonia escaping from industrial chimneys
- Covariates: air flow x_1, temperature x_2 and acid concentration x_3
- Transformed covariates: $z_{k,i} = (x_{k,i} - \bar{x_k})/sd(x_k)$ for $k = 1, 2, 3$.


Y
X1
X2
X3
20 30 40 50 60 70 80 90
20 40 60 80
10 15 20 25 30 35 40
20 40 60 80
20 40 60 80
10 15 20 25 30 35 40
20 40 60 80
20 40 60 80

Model 1 specification:

$$
\begin{array}{rcl}
y_i & \sim & \text{Normal}(\mu_i, \tau) \quad i = 1, \ldots, 21 \\
\tau & = & 1/\sigma^2 \\
\mu_i & = & \beta_0 + \beta_1 z_{1,i} + \beta_2 z_{2,i} + \beta_3 z_{3,i} \\
\sigma & \sim & \text{Uniform}(0.01, 100) \\
\beta_k & \sim & \text{Normal}(0, 0.001), \quad k = 0, \ldots, 3.
\end{array}
$$

In these model we also calculate some diagnostic quantities:

By analogy to the classical regression analysis we calculate *the standardized residuals* as

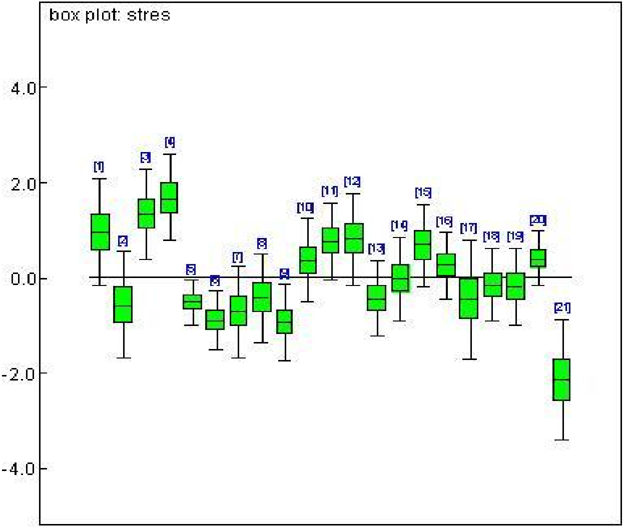
$$d_i = \frac{y_i - E(y_i|z)}{\sqrt{Var(y_i|z)}}.$$

In a Bayesian setting these residuals have posterior distributions as well. They can be used similarly to residual analysis in classical statistics.

We specify the model in WinBUGS as:

```
model{
# Model 1
   for (i in 1 : N) {
      y[i] ~ dnorm(mu[i], tau)
     mu[i] <- beta0 + beta[1] * z[i, 1] +
              beta[2] * z[i, 2] + beta[3] * z[i, 3]
     stres[i] <- (y[i] - mu[i]) / sigma
 # Priors
beta0 ~  dnorm(0, 0.001)
for (j in 1 : p) {
beta[j] ~ dnorm(0, 0.001) }  # coefficients independent
tau <- 1/(sigma*sigma)
sigma ~ dunif(0.01, 100)     # Gelman's prior
}
```

We run the model in WinBUGS, we look at the possible outliers.



As an alternative model we replace the Normal likelihood by a t-distribution with $\nu = 4$ degrees of freedom. The idea is to have a more robust model against outliers in the y's.

Model 2:

$$\begin{aligned} y_i &\sim \mathrm{t}(\mu_i, \tau, \nu) \quad i = 1, \ldots, 21 \\ \tau &= 1/\sigma^2 \\ \mu_i &= \beta_0 + \beta_1 z_{1,i} + \beta_2 z_{2,i} + \beta_3 z_{3,i} \\ \sigma &\sim \mathrm{Uniform}(0.01, 100) \\ \beta_k &\sim \mathrm{Normal}(0, 0.001), \quad k = 0, \ldots, 3. \end{aligned}$$

In the WinBUGS code we modify the model section by commenting out the normal model and adding one line with the model based on the t-distribution and we run the MCMC again.

```
#y[i] ~ dnorm(mu[i], tau)
  y[i] ~ dt(mu[i], tau, 4)
```

```
#DIC Normal
Dbar Dhat pD DIC
Y 110.537 105.800 4.736 115.273
total 110.537 105.800 4.736 115.273
#DIC t
Dbar Dhat pD DIC
Y 109.043 104.103 4.940 113.983
total 109.043 104.103 4.940 113.983
```

A very modest difference between the two models.

Variable Selection

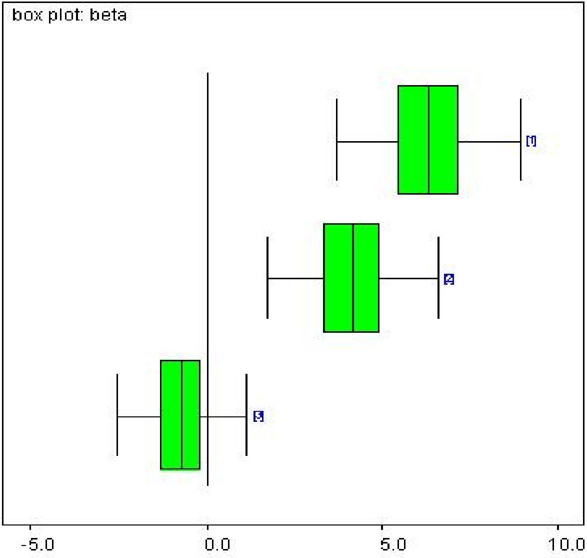
We modify the structure of the distribution of the regression coefficients by adding a **common distribution with unknown variance**. This is called a ridge regression model, where the βs are assumed exchangeable.

Model 3 specification:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \tau), \quad \tau = 1/\sigma^2, \quad i = 1, \ldots, 21 \\
\mu_i &= \beta_0 + \beta_1 z_{1,i} + \beta_2 z_{2,i} + \beta_3 z_{3,i} \\
\sigma &\sim \text{Uniform}(0.01, 100) \\
\beta_k &\sim \text{Normal}(0, \phi), \quad \phi = 1/\sigma_\beta^2 \quad k = 0, \ldots, 3 \\
\sigma_\beta &\sim \text{Uniform}(0.01, 100).
\end{aligned}
$$

In WinBUGS

```
for (j in 1 : p) {
# beta[j] ~ dnorm(0, 0.001)    # coefs independent
beta[j] ~ dnorm(0, phi)  # coefs exchangeable (ridge regres
}

phi <- 1/(sigmaB*sigmaB)
sigmaB ~ dunif(0.01, 100) #Gelman's prior
```


box plot: beta
[1]
[2]
[3]
-5.0
0.0
5.0
10.0

We see that β_3 is not a relevant variable in the model. Now, we try another variable selection procedure, by modifying the distribution of the regression coefficients.

Model 5 specification:

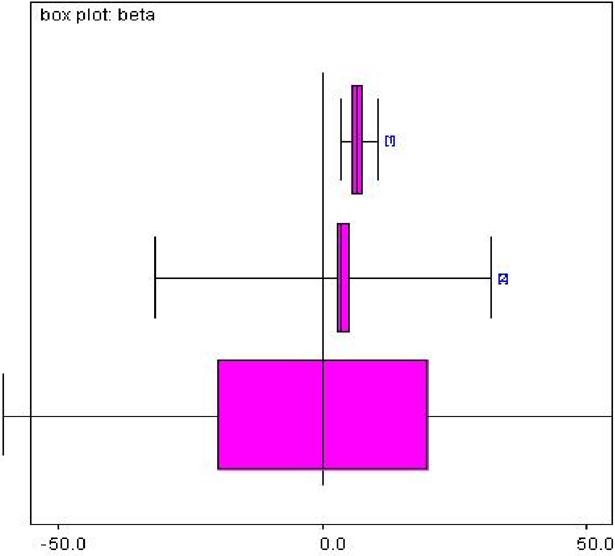
$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \quad i = 1, \ldots, 21 \\
\mu_i &= \beta_0 + \beta_1\,\pi_1\,z_{1,i} + \beta_2\,\pi_2\,z_{2,i} + \beta_3\,\pi_3\,z_{3,i} \\
\sigma &\sim \text{Uniform}(0.01, 100) \\
\beta_k &\sim \text{Normal}(0, 100), \quad k = 0, \ldots, 3, \\
\pi_1 &\sim \text{Bernoulli}(0.5) \\
\pi_2 &\sim \text{Bernoulli}(0.5) \\
\pi_3 &\sim \text{Bernoulli}(0.5).
\end{aligned}
$$

In WinBUGS we change the equation of μ_i:

```
#mean equation model
mu[i] <- beta0 + beta[1] * pi[1]* z[i, 1] +
    beta[2]* pi[2] * z[i, 2] + beta[3] * pi[3]* z[i, 3]

#priors
for (j in 1 : p) {
  beta[j] ~ dnorm(0, 0.001)
   pi[j] ~ dbern(0.5)
}
```

We run the model for 40,000 iterations and we discard the first 20,000. Results are:

```
pi[1] 0.9987 0.0367 <- X1 very important
pi[2] 0.8361 0.3702 <- X2 non meaningful
pi[3] 0.0425 0.2018 <- X1 definitively no important
```


box plot: beta
[1]
[2]
-50.0
0.0
50.0

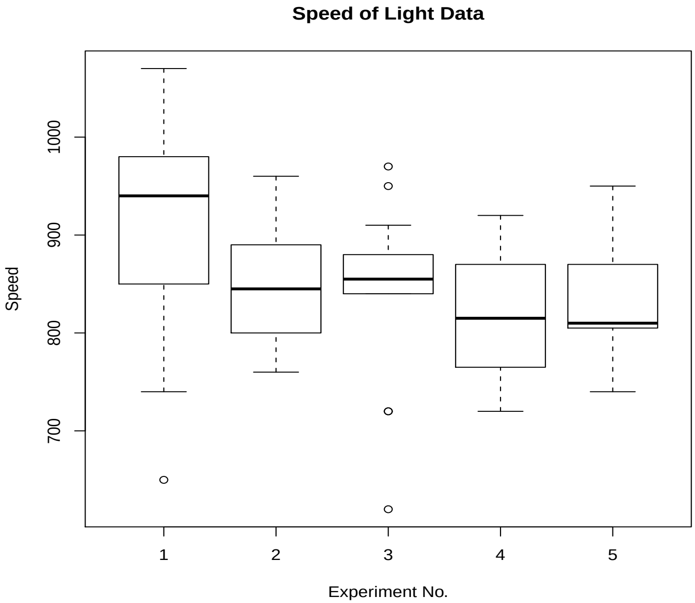
ANOVA and Experimental Design

Example: Speed of light measurements by Michelson

This is a classical data set available in R and corresponds to Measurements of the speed of light in air, made between 5th June and 2nd July, 1879. The data consists of five experiments, each consisting of 20 consecutive runs. The response is the speed of light in km/s, less 299000. The currently accepted value, on this scale of measurement, is 734.5.

We start by comparing the 5 experiments with boxplots:

```
> library(MASS)
> data(michelson)
> attach(michelson)
> plot(Expt, Speed, main="Speed of Light Data",
       xlab="Experiment No.",ylab="Speed")
```


Speed of Light Data
Speed
1000
900
800
700
1
2
3
4
5
Experiment No.

ANOVA model: Bayesian analysis in WinBUGS

The Bayesian approach to the ANOVA model is similar to the regression model, with mean equation according to the mean groups parametrization. So if we have n_j observations in J groups with $n = \sum_{j=1}^{J} n_j$.

Model 1

$$\begin{aligned}
y_{i,j} &\sim \text{Normal}(\mu_j, \tau), \quad \tau = 1/\sigma^2 \quad i = 1, \ldots, n_j, \\
\mu_j &= \mu + \alpha_j, \quad j = 1, \ldots, J, \\
\alpha_j &\sim \text{Normal}(0, 0.001), \quad j = 2, \ldots, J, \\
\mu &\sim \text{Normal}(0, 0.001), \\
\sigma &\sim \text{Uniform}(0.01, 100).
\end{aligned}$$

The WinBUGS implementation of Model 1 is:

```
model{
for(i in 1:n){
y[i] ~ dnorm( mu[i], tau)
mu[i] <- mu0 + alpha[ a[i] ]
}

# Corner constrains parametrization
# alpha[1] <- 0
# Sum to zero parametrization
alpha[1] <- -sum( alpha[2:J] )

# group means:
for( i in 1:J){
m.g[i] <- mu0 + alpha[i]}
# mean differences:
d[1] <- (m.g[2] - m.g[1] )
d[2]<-  (m.g[3] - m.g[1] )
...
# Priors
mu0 ~ dnorm(0, 0.0001)
for( j in 2:J){
alpha[j] ~ dnorm(0, 0.0001)
}
tau <- 1/(sigma*sigma)
sigma ~ dunif(0.01, 100) #Gelman's prior
}
```

One serious problem of ANOVA modeling is the lack of variance homogeneity between groups. We can extend our Bayesian set up to include this data feature as follows:

Model 2

$$\begin{aligned}
y_{i,j} &\sim \text{Normal}(\mu_j, \tau_j), \quad \tau_j = 1/\sigma_j^2 \quad i = 1, \ldots, n_j, \\
\mu_j &= \mu + \alpha_j, \quad j = 1, \ldots, J, \\
\alpha_j &\sim \text{Normal}(0, 0.001), \quad j = 2, \ldots, J, \\
\mu &\sim \text{Normal}(0, 0.001), \\
\sigma_j &\sim \text{Uniform}(0.01, 100), \quad j = 1, \ldots, J, .
\end{aligned}$$

This model includes a structural dispersion sub-model for each treatment group.

The WinBUGS implementation of Model 2 change to:

```
{
for(i in 1:n){
#y[i] ~ dnorm( mu[i], tau)
y[i] ~ dnorm( mu[i], tau[i])

mu[i] <- mu0 + alpha[ a[i] ]
tau[i] <- gamma[ a[i] ]
}
# Priors for groups dispersion model
for( j in 1:J){
   gamma[j] ~ dgamma(0.001, 0.001)
   sigma2[j] <- 1/ gamma[j]
   sigma[j] <- pow(sigma2[j], 0.5)
}
```

For models with more than 3 groups the estimation based on simple means can be improved by adding exchangeability structure to the α_i's (Stein, 1956, Lindley 1962, Casella and Berger, pag.574). This assumed a sub-model for α_i in the following way:

Model 3

$$\begin{aligned}
y_{i,j} &\sim \text{Normal}(\mu_j, \tau_j), \quad \tau_j = 1/\sigma_j^2 \quad i = 1, \ldots, n_j, \\
\mu_j &= \mu + \alpha_j, \quad j = 1, \ldots, J, \\
\alpha_j &\sim \text{Normal}(0, \phi), \quad \phi = 1/\sigma_\alpha \quad j = 2, \ldots, J, \\
\mu &\sim \text{Normal}(0, 0.001), \\
\phi &\sim \text{Uniform}(0.01, 100), \quad j = 1, \ldots, J, \\
\sigma_j &\sim \text{Uniform}(0.01, 100), \quad j = 1, \ldots, J, .
\end{aligned}$$

The WinBUGS implementation of Model 3 change to:

```
# Priors
mu0 ~ dnorm(0, 0.0001)
for( j in 2:J){
alpha[j] ~ dnorm(0,  xi)
}
xi <- 1/(sigma.alpha*sigma.alpha)
sigma.alpha ~ dunif(0.01, 100) #Gelman's prior
```

One alternative to exchangeability between α_i's is to use a mixture model to investigate the structure of the data, for example:

Model 4

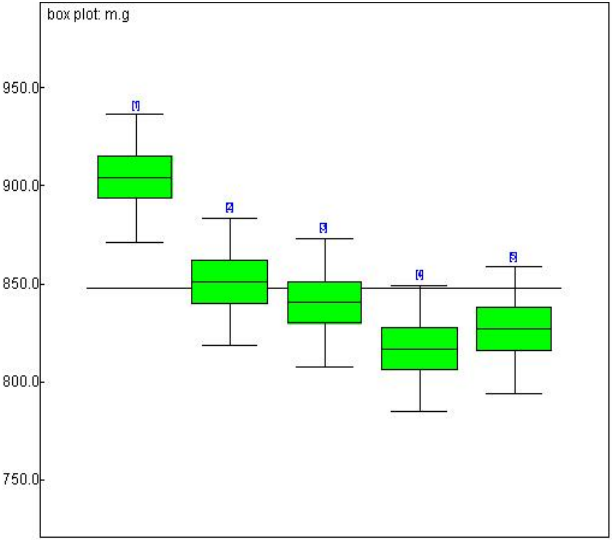
$$\begin{aligned}
y_{i,j} &\sim \text{Normal}(\mu_j, \tau_j), \quad \tau_j = 1/\sigma_j^2 \quad i = 1, \ldots, n_j, \\
\mu_j &= \mu + \alpha_j \pi_j, \quad j = 1, \ldots, J, \\
\alpha_j &\sim \text{Normal}(0, 0.001), \quad j = 2, \ldots, J, \\
\mu &\sim \text{Normal}(0, 0.001), \\
\pi_j &\sim \text{Bernoulli}(0.5), \quad j = 2, \ldots, J, \\
\sigma_j &\sim \text{Uniform}(0.01, 100), \quad j = 1, \ldots, J, .
\end{aligned}$$

The WinBUGS implementation of Model 3 change to:

```
# group means mixture model:
for( i in 1:J){
m.g[i] <- mu0 + alpha[i]*pi[i]
pi[i] ~ dbern(0.5)
}
```

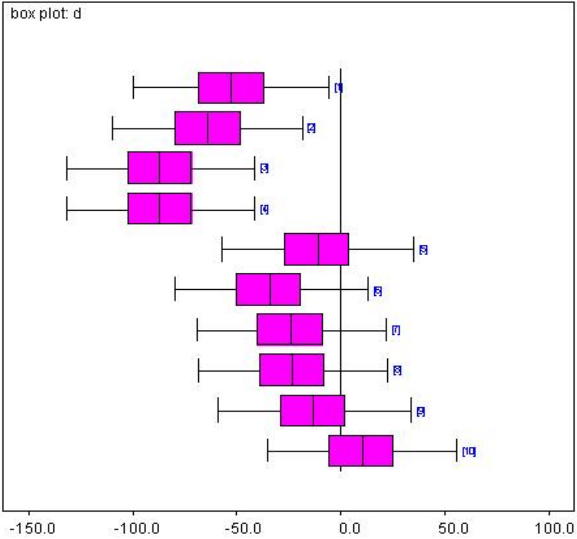
The next slides present some graphical results of the main features of each model. The full WinBUGS script for this analysis is `anova.odc`, this includes some other calculations for residual analysis as well.

ANOVA: classical model



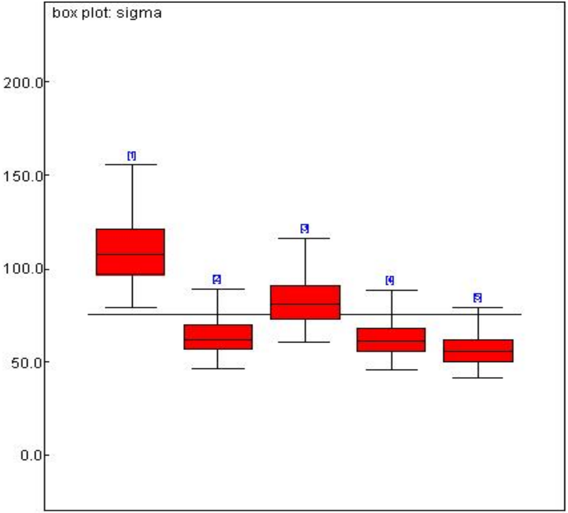
Posterior distributions for mean groups μ_j, Model 1.

ANOVA: classical model



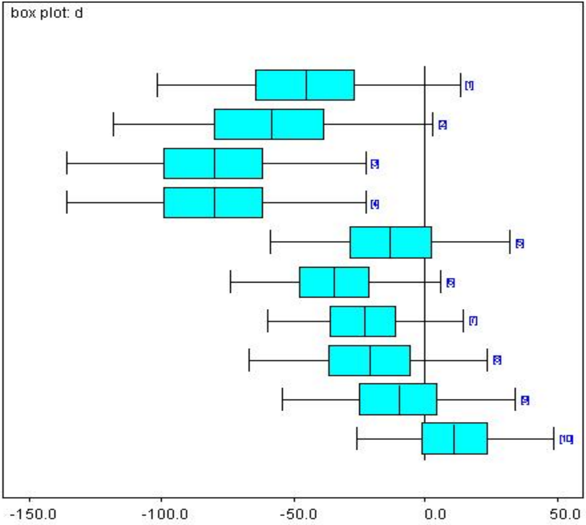
Posterior distributions for mean differences between all groups (d_i's): Model 1

ANOVA: variance heterogeneity model



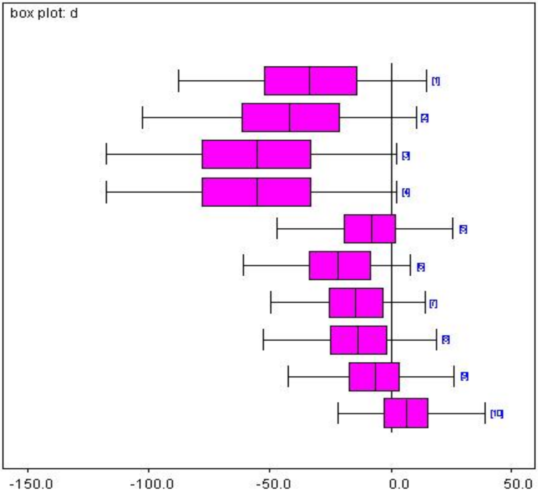
Posterior distributions for σ_j, Model 2.

ANOVA: variance heterogeneity model



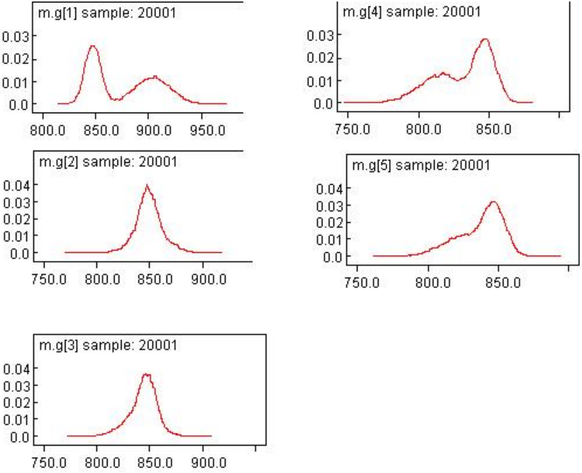
Posterior distributions for mean differences between all groups

ANOVA: exchangeability model



Posterior distributions for mean differences between all groups with exchangeability, Model 3.

ANOVA: Mixture models



Posterior distributions for the group means with mixture distributions: Model 4

Logistic regression: combining multiple information in Risk analysis

Example: The Challenger O-Ring Data: A Bayesian Risk Analysis

Historical background

- The Space Shuttle Challenger's final mission was on January 28th 1986, on an unusually cold morning (31F/-0.5C)
- It disintegrated 73 seconds into its flight after an **O-ring** seal in its right solid rocket booster failed
- The uncontrolled flame caused structural failure of the external tank, and the resulting aerodynamic forces broke up the orbiter
- All seven members of the crew were killed

Technical background

Dalal, Fowlkes, and Hoadley: Risk Analysis of the Space Shuttle

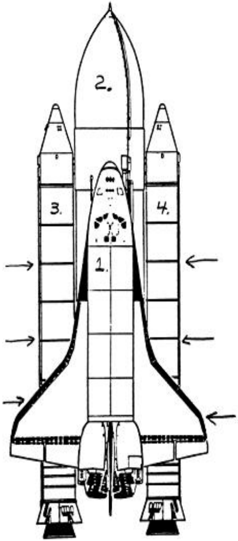


Figure 2. Space Shuttle: Orbiter, External Tank, Solid Rocket Motors.

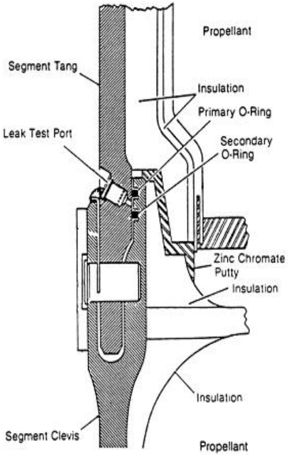


Figure 3. Solid Rocket Motor Cross Section: Tang, Clevis, and O-Rings.

The Challenger's O-Ring Thermal-Distress Data

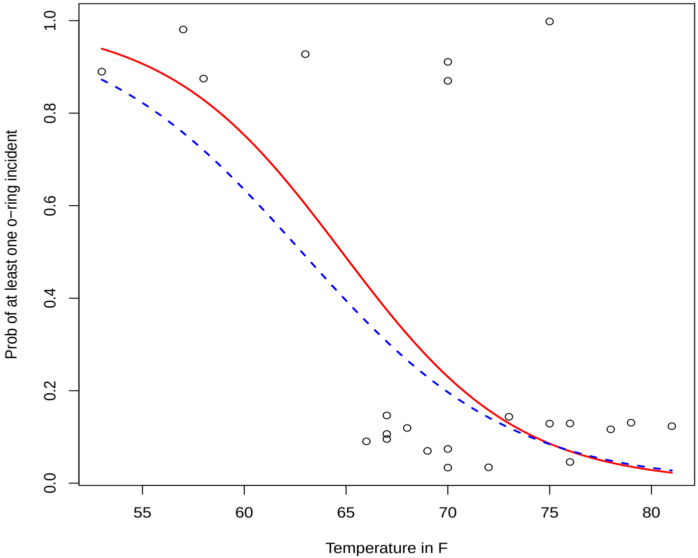
Example: Regression modeling with change point

- On the night of January 27, 1986, the night before the space shuttle Challenger accident, there was a tree-hour teleconference among people at Marton Thiokol (manufacturer of the solid rocket motor), Marshal Space Center and Kennedy Space Center.

- The discussion focused on the forecast of a 31 F temperature for lunch time the next morning, and the effect of low temperature on O-ring performance.

- The available data of previous shuttle flights consisted on occurrence of thermal distress (yes, no) and temperature at launch time. Let's take a look at these data and fit a logistic regression in R:

```
> #distress yes=1 , no =0
> y = c(1,1,1, 1,0,0,0, 0,0,0,0, 0,1,1,0,
 0,0,1,0, 0,0,0,0)
> # temperature at launch time
> x = c(53,57,58, 63,66, 67,67, 67,68, 69,70,
 70,70, 70,72, 73,75, 75,76, 76,78, 79,81)
> # number of previous flights
> N = 23
> summary(f.b1 <- glm(y ~ x, family=binomial))
...
AIC: 24.315

> summary(f.b2 <- update(f.b1, family=binomial(link=cloglog
...
AIC: 23.531
```


Prob of at least one o−ring incident
0.0
0.2
0.4
0.6
0.8
1.0
55
60
65
70
75
80
Temperature in F

- These data have been analyzed several times in the literature. I found that fitting a change point model makes an improvement on previous published analysis. The model is:

  $$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta_0 + \beta_1 z_i\right)$$

  where

  $$z_i = \begin{cases} 1 & \text{for} \qquad\qquad x_i \geq K, \\ 0 & \text{otherwise.} \end{cases}$$

- In this model K **is unknown** and represents the temperature from which the probability of distress is the lowest.

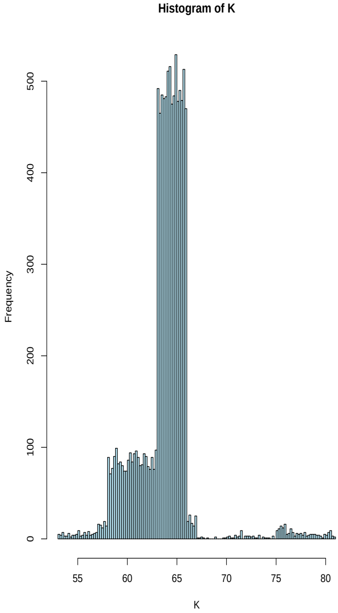
- The parameter β_1 represents the reduction of risk under temperatures greater than K.

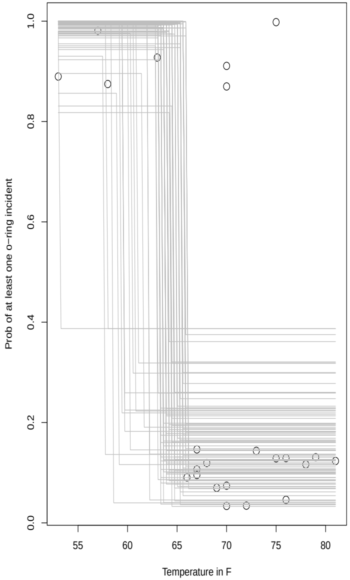
This model is implemented in WinBUGS as following:

```
model
  {
    for(i in 1 : N) {
     y[i] ~ dbern(p[i])
     logit(p[i]) <- beta0 + beta1 * z[i] }

    for(i in 1:N){z[i] <- step(x[i] - K) } # step = 0 u

  #priors
    K ~dunif(53, 81)
    beta0 ~ dnorm(0.0, 0.01)
    beta1 ~ dnorm(0.0, 0.01)
}
```

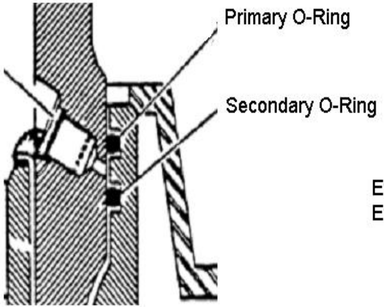
We run the model with R2WinBUGS:

```
>  data.b <- c("x", "y", "N")
>  par.b <- c("beta0", "beta1", "K")
>
>  chall.1 <- bugs(data.b, inits=NULL, par.b,
            "challenger-1.txt", n.chains=1,
            n.iter=20000,  n.thin=1,
            bugs.directory = bugsdir,
            working.directory = getwd(),
            clearWD=TRUE, debug=TRUE)

>  print(chall.1)
         mean  sd  2.5%   25%  50%  75% 97.5%
beta0     6.2 3.9  -0.2   3.3  5.8  8.4  15.3
beta1    -8.0 3.9 -17.2 -10.2 -7.6 -5.1  -2.1
K        63.8 2.9  58.2  63.1 64.1 65.1  67.0
```


Histogram of K
Frequency
0
100
200
300
400
500
55
60
65
70
75
80
K


Prob of at least one o−ring incident
0.0
0.2
0.4
0.6
0.8
1.0
55
60
65
70
75
80
Temperature in F

Example: Probabilistic Risk Assessment



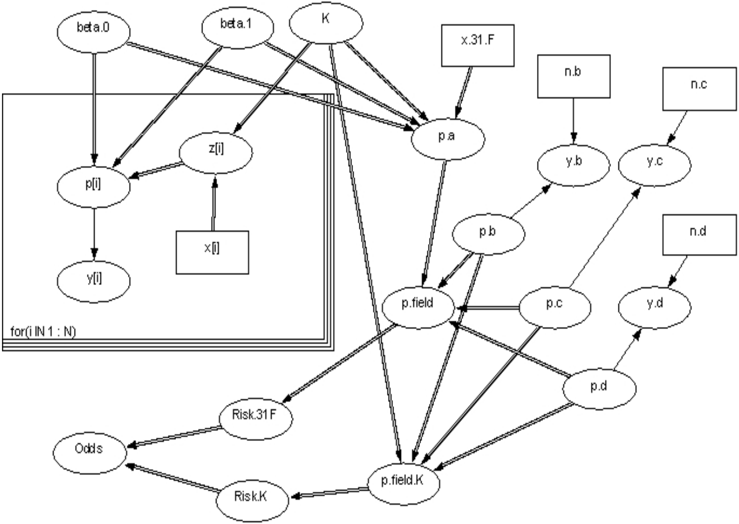
Detail: O-Ring Cross-Section

Event ***a***: It Erodes
Event ***b***: It suffers blow-by | a

Event ***c***: It erodes | a, b
Event ***d***: It suffers blow-by | a,b,c

- **A catastrophic failure of a field joint** is expected when all four events occur during the operation of the solid rocket boosters:
  $$p_{field} = p_a \times p_b \times p_c \times p_d$$
- Since there are 6 field joints per launch, assuming the all 6 joint failures are independent, the probability of at least one failure is
  $$\text{Risk} = 1 - (1 - p_{field})^6$$
- p_a is calculated from the logistic regression model as:
  $$p_a = \Pr(\text{O-Ring Thermal-Distress at 31F})$$
- The available data for the other events was very sparse:
  - Event b: 2 in 7 flights
  - Event c: 1 in 2 flights
  - Event d: 0 in 1 flight

We model the probabilities of events b, c and d with a beta-binomial model with uninformative priors Beta(0.5, 0.5).


beta.0
beta.1
K
x.31.F
n.b
n.c
p.a
z[i]
p[i]
y.b
y.c
p.b
x[i]
y[i]
n.d
p.field
p.c
y.d
for(i IN 1 : N)
p.d
Risk.31F
Odds
p.field.K
Risk.K

In WinBUGS we modify the model file ”challenger-1.txt” by adding these risk calculations. We add, also, the calculation of risk at temperature K or greater.

To run the model in R by adding the date of events b, c, and d:

```
# Risk analysis
y.b <- 2; y.c <- 1; y.d <- 0
data.r <- c("x", "y", "N", "y.b", "y.c", "y.d")
par.r <- c("beta0", "beta1", "K", "p.cat", "p.catK")
chall.2<- bugs(data.r, inits=NULL, par.r,
           "challenger-1.txt", n.chains = 1,
...
> print(chall.2, 3)
          mean    sd   2.5%    25%    50%    75%  97.5%
K       63.726 2.653 58.190 63.090 64.120 65.120 66.010
p.cat    0.173 0.204  0.000  0.019  0.088  0.258  0.736
p.catK   0.038 0.061  0.000  0.003  0.014  0.047  0.213
```

Summary results of our analysis:

- The risk of a catastrophic failure with launching temperature of 31 F is 17%
- The risk of a catastrophic failure with temperatures greater than 63 F is 3.8%
- The odds ratio of the risk between these temperatures is 9.8

One interesting historical note:

The Roger Commission, appointed by president Reagan to investigate the causes of the accident, pointed out:

... a mistake in the analysis of the thermal distress data was that the flights with zero incidents were left off because it was felt that these flights did not contribute any information about the temperature effect. (p. 145)

Practical

Exercise: a nonconjugate nonlinear model

Volume 2 in WinBUGS help: Dugongs

Originally, Carlin and Gelfand (1991) consider data on length y_i and age x_i measurements for 27 dugongs (sea cows) and use the following nonlinear growth curve with no inflection point and an asymptote as x_i tends to infinity:

$$\begin{aligned} y_i &\sim \texttt{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \alpha - \beta\gamma^{x_i}, \end{aligned}$$

where $\alpha, \beta > 0$ and $\gamma \in (0, 1)$.

Practical

- Open the file in WinBUGS and run the model
- Change the last observation 2.57 to 2.17 and run the model. Did you see different results?
- Modify the WinBUGS code as:

```
#    y[i] ~ dnorm(mu[i], tau)
     y[i] ~ dt(mu[i], tau, df)  # fit robust t distribut
...
df <- 5
```

- Run the model with this change and compare results

Lecture:

Bayesian Computations with MCMC Methods

Summary

- Introduction to discrete Markov chains.
- Construction of Markov chain Monte Carlo algorithms.
- Gibbs sampling methods.
- Metropolis and Metropolis-Hastings method.
- Issues in applications of MCMC (convergence checking, proposal distributions, etc.)

Why is computation important ?

- Bayesian inference centers around the posterior distribution

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

  where θ may be a vector of thousand components!

- $p(y|\theta)$ and $p(\theta)$ are usually available in closed form, but $p(\theta|y)$ is usually not analytical tractable. Also, we may want to obtain
  - marginal posteriors $p(\theta_1|y) = \int p(\theta|y) d\theta_{(-1)}$
  - calculate properties of $p(\theta_1|y)$, such as mean $E(\theta_1|y) = \int \theta_1 p(\theta|y) d\theta_{(-1)}$, tail areas, etc.

- We see that numerical integration becomes crucial in Bayesian inference!

General Monte Carlo integration

If we have algorithms for sampling from arbitrary (typically high-dimensional) posterior distributions, we could use 'Monte Carlo' methods for Bayesian inference:

- Suppose we can draw samples from the joint posterior distribution for θ, i.e.

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)} \sim p(\theta|y)$$

- Then Monte Carlo integration
  - $\theta^{(1)}, \theta^{(2)}, \ldots \theta^{(N)} \sim p(\theta|y)$
  - $E(g(\theta)) = \int g(\theta)p(\theta|y)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta^{i})$
  - Theorems exist which prove convergence even if the sample is **dependent**,i.e.

$$\frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) \to E(g(\theta)) \quad \text{as} \quad n \to \infty$$

Markov Chain Monte Carlo (MCMC)

- **Independent sampling** from $p(\theta|y)$ may be very difficult in high dimensions
- Alternative strategy based on **dependent sampling**:
  - We know $p(\theta|y)$ up to a normalizing constant
  - Then, we design a **Markov chain** which has $p(\theta|y)$ as its stationary distribution
  - A sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(3)}, \ldots$ forms a Markov chain if
    $$\theta^{(i+1)} \sim p(\theta|\theta^{(i)})$$
    i.e. conditional on the value of $\theta^{(i)}$, $\theta^{(i+1)}$ is independent of $\theta^{(i-1)}, \ldots, \theta^{(0)}$.

- Run the chain until it appears to have settled down to equilibrium, say

$$\theta^{(k)}, \theta^{(k+1)}, \ldots, \theta^{(K)} \sim p(\theta|y)$$

- Use these **sampling values** to empirically estimate the posterior of θ, say,

$$\hat{p}(\theta|y)$$

Problem: Design a Markov chain with $p(\theta|y)$ as its **unique stationary distribution**?

Answer: This is surprisingly easy and **several standard recipes** are available

- Metropolis *et al.* (1953) showed how to do it
- Hastings (1970) generalized Metropolis algorithm
- Geman and Geman (1984) introduced the Gibbs Sampling algorithm
- Gelfand and Smith (1990) popularized Gibbs sampling in statistics
- See Gilks, Richardson and Spiegelhalter (1996) for a gentle introduction and many worked examples.
- Robert and Casella (2004) for more detailed theoretical reference

The Gibbs sampler

Let our vector of unknowns θ consist of k sub-components $\theta = (\theta_1, \ldots, \theta_k)$

1. Choose starting values $\theta_1^0, \ldots, \theta_k^0$
2. Sample from

$$\begin{array}{lcl} \theta_1^{(1)} & \sim & p(\theta_1 | \theta_2^{(0)}, \theta_3^0, \ldots, \theta_k^{(0)}, y) \\ \theta_2^{(1)} & \sim & p(\theta_2 | \theta_1^{(1)}, \theta_3^0, \ldots, \theta_k^{(0)}, y) \\ \ldots & & \\ \theta_k^{(1)} & \sim & p(\theta_k | \theta_1^{(1)}, \theta_2^1, \ldots, \theta_{k-1}^{(1)}, y) \end{array}$$

3. Repeat step 2 many 1000s of times. Eventually we obtain samples from $p(\theta|y)$

Example: Normal distribution with unknown mean and variance

- Suppose that the observations $y_1, \ldots, y_n \sim \mathtt{N}(\mu, \tau^{-1})$
- We assume that y_i are conditionally independent given θ and precision τ, and θ and τ are themselves independent.
- We put conjugate priors on μ and τ:

$$\mu \sim \mathtt{N}(\theta_0, \phi_0^{-1}), \quad \tau \sim \mathtt{Gamma}(a, b)$$

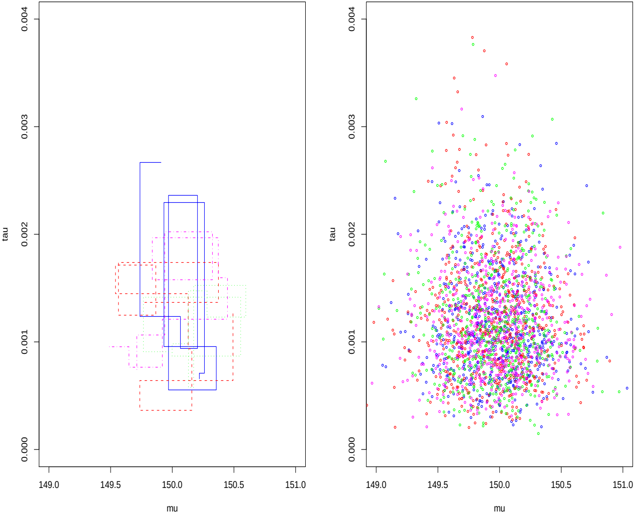
- Then the full conditionals for μ and τ are:

$$\begin{aligned} p(\mu|\tau, y) &= \mathtt{N}\left(\frac{\mu_0\phi_0 + n\bar{y}\tau}{\phi_0 + n\tau}, \frac{1}{\phi_0 + n\tau}\right) \\ p(\tau|\mu, y) &= \mathrm{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum(y_i - \mu)^2\right) \end{aligned}$$

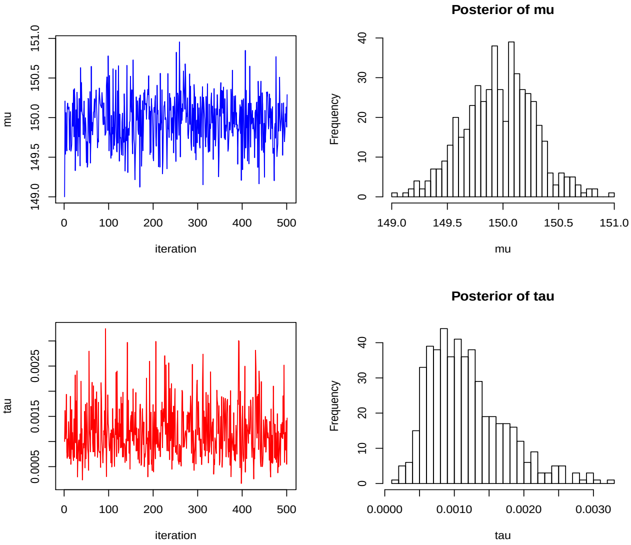
Programming the Gibbs sampler in R

Example: Normal distribution with unknown mean and variance

```
gibbsNormal <- function(y, mu1, tau1, N, mu0, phi0, a, b)
     mu <- numeric(N+1) # N = number of iterations
     mu[1] <- mu1  #  the initial value for mu
     tau <- numeric(N+1)
     tau[1] <- tau1  # the initial value for  tau
     n <- length(y) ;  ybar <- mean(y)
     for(i in 2:(N+1)) {
     # generate samples from full conditional
       mu[i] <- rnorm(1,
       mean = (mu0*phi0 + n*ybar*tau[i-1])/(phi0 + n*tau[i-
       sd =  1/sqrt(phi0 + n*tau[i-1]))
       tau[i] <- rgamma(1, (n+2*a)/2,
                 (sum((y-mu[i])^2) + 2*b)/2)
     }
     output <- cbind(mu, tau) }
```



Four independent sequences of the Gibbs sampler for a normal distribution with unknown mean and variance.
Left panel: first 10 steps. Right panel: last 500 iterations in each



Traces of one sequence of Gibbs simulations. Upper panels: 500 iterations from the posterior of μ. Lower panels: 500 iterations from the posterior of τ.

General properties of a MCMC algorithm

- **Reversibility**:
  - The key property is *reversibility* or *detailed balance*, i.e., a balance in the flow of transition probabilities between the states of the Markov chain
  - A Markov chain which is reversible **has stationary distribution**
  - Thus we build a kernel or transition matrix P such that it ensures reversibility

$$p(x)P(y|x) = p(y)P(x|y)$$

where $P(y|x)$ represents the flow of probability $x \to y$

- **Irreducibility**:
  - An irreducible chain is one in which for any point $\theta^{(k)}$ in the parameter space, it is possible to move from $\theta^{(k)}$ to other point $\theta^{(l)}$ in a finite number of steps.
  - This guarantees the chain can visit all possible values of θ irrespective of the starting value $\theta^{(0)}$.
- **Aperiodicity**:
  - A chain is aperiodic if does not exhibits periodic behavior.
  - If $R_1, R_2, \ldots, R_k$ are disjoint regions in parameter space the chain does not cycle around them.

To sample form $p(\theta|y)$ we construct a Markov chain with transition matrix P which satisfies **reversibility**, **irreducibility** and **aperiodicity**.

Metropolis algorithm

The algorithm proceeds as follows:

- Start with a preliminary guess, say θ^0.
- At iteration t sample a proposal $\theta^t \sim P(\theta^t|\theta^{t-1})$.
- The jumping distribution must be symmetric, satisfying the condition $P(\theta_a|\theta_b) = P(\theta_b|\theta_a)$ for all θ_a and θ_b.
- If $p(\theta^t|y) > p(\theta^{t-1}|y)$ then accept θ^t
- If not, flip a coin with probability $r = \frac{p(\theta^t|y)}{p(\theta^{t-1}|y)}$, if it comes up heads, accept θ^t.
- If the coin toss comes up tails, stay at θ^{t-1}

The algorithm continues, until we sample from $p(\theta|y)$. The coin tosses allow it to go to less plausible θ^ts, and keep it from getting stuck in local maxima.

Some samplers for the Metropolis algorithm

The following are commonly implemented samplers for the proposal distribution $P(\theta^t|\theta^{t-1})$

- Random walk sampler, observations are generated by $\theta_t = \theta_{t-1} + z_t$ with $z_t \sim f$. There are many common choices for f, including the uniform in the unit disc, a multivariate normal or a t - distribution
- The independent sampler, the candidate observation is sample independently from the current state of the chain
- The Gibbs sampler. The Gibbs transition can be regarded as a special case of a Metropolis transition
- The STAN software implements samplers based on Hamiltonian dynamics

Metropolis in R

- The package MCMCpack implements Metropolis for a large number of statistical models.

- The current implementation uses a random walk Metropolis with proposal density multivariate Normals.

- The proposal density is centered at the current $\theta^{(t)}$ with variance-covariance matrix given by the user or automatically calculated proportional to the Hessian of the posterior evaluated at its mode.

Example: an interesting bivariate distribution

- It is well know that the pair of **marginal distributions doest not uniquely determine a bivariate distribution**. For example, a bivariate distribution with normal marginal distributions need not be jointly normal (Feller 1966, p. 69)

- In contrast, **the conditional distribution functions uniquely determine a joint density** function (Arnold and Press 1989).

- A natural question arises: **Must a bivariate distribution with normal conditionals be jointly normal?**

The answer is NO!

Gelman and Meng (1991) introduced an interesting distribution, where the join distribution is non-normal:

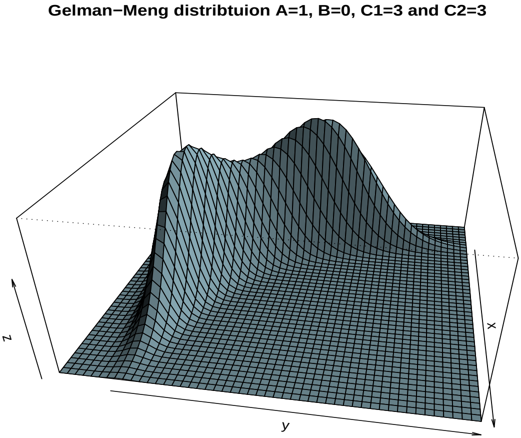
$$f(x,y) \propto \exp\left(-1/2\left[Ax^2y^2 + x^2 + y^2 - 2Bxy - 2C_1 - 2C_2y\right]\right).$$

But it has conditional distributions which are normals:

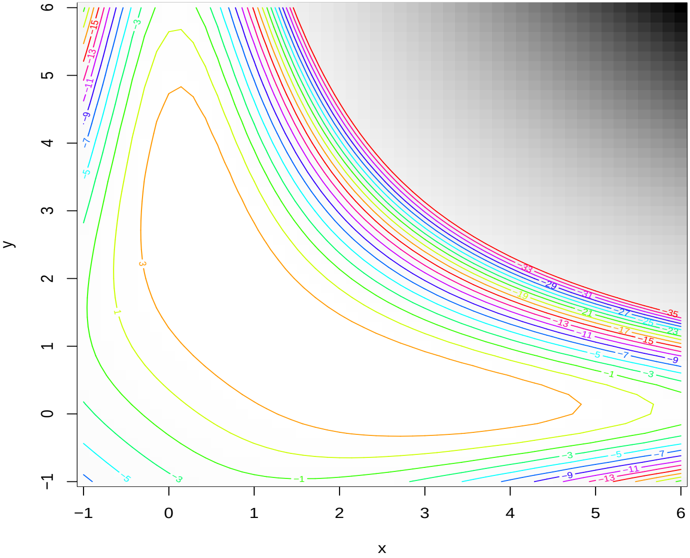
$$x|y \sim N(\frac{By + C_1}{Ay^2+1}, \frac{1}{Ay^2+1}), \quad y|x \sim N(\frac{Bx + C_1}{Ax^2+1}, \frac{1}{Ax^2+1}).$$

The question is how to sample form $f(x,y)$?

Here we show an example where the parameters are $A = 1$, $B = 0$ $C_1 = 3$ and $C_2 = 3$.

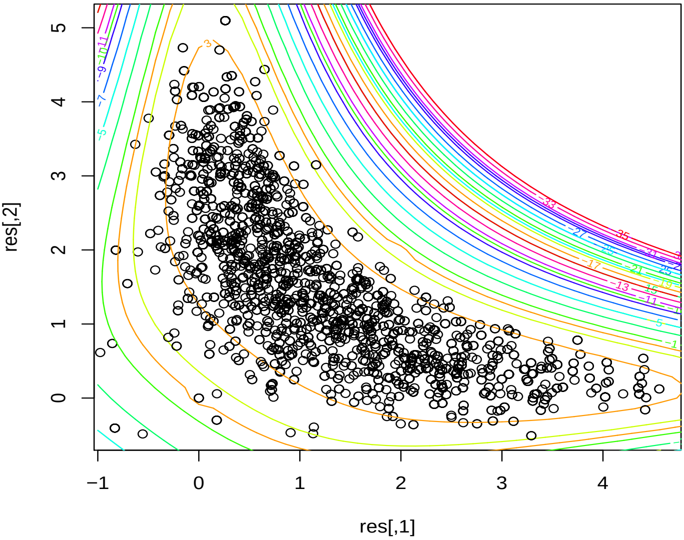


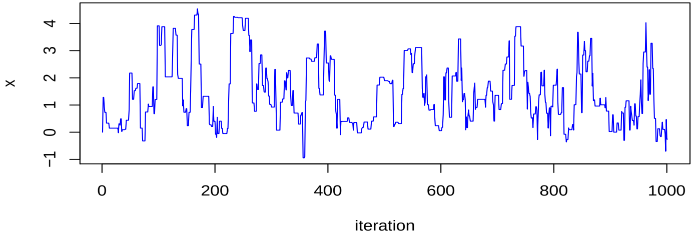
Gelman−Meng distribtuion A=1, B=0, C1=3 and C2=3

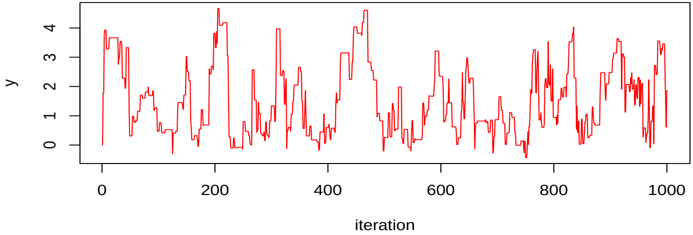


We implement a Metropolis sampling algorithm based on a random walk and bivariate normals as proposals.

```
## Gelman and Meng (1991) kernel function:
  f.g.m <- function(xy , A = 1, B = 0, C1 = 3, C2 = 3)
   { x <- xy[1]; y <- xy[2]
     r <- -.5 * (A * x^2 * y^2 + x^2 + y^2
              - 2 * B * x * y - 2 * C1 * x - 2 * C2 * y)
            as.vector(r) }

## Metropolis sampling with MCMCpack
 res <- MCMCmetrop1R(f.g.m, theta.init=c(0, 1), mcmc=10000,
              burnin=5000, logfun=FALSE)
>
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
The Metropolis acceptance rate was 0.47580
```


res[,2]
res[,1]
−1
0
1
2
3
4
5
−3
−11
−10
−9
−7
−5
−33
−35
−27
−25
−31
−17
−21
−25
−19
−13
−15
−11
−5
−1


x
−1
0
1
2
3
4
0
200
400
600
800
1000
iteration


y
0
1
2
3
4
0
200
400
600
800
1000
iteration

Performance of MCMC methods

There are three main issues to consider

- Convergence: How quickly does the distribution of $\theta^{(t)}$ approach $p(\theta|x)$?
- Efficiency: How well are functionals of $p(\theta|x)$ estimated from $\theta^{(t)}$?
- Simplicity: How convenient is the method to use?

In general computer effort should be measured in seconds, not iterations!

Checking convergence

- Convergence is to target distribution and not to a single value
- Once convergence is reached, samples should look like a random scatter about a stable mean value.
- One approach is to run many long chains with widely differing starting values.
- Plot traces of the simulated chain
- Plot ACF, etc.

Lecture:

Introduction to Hierarchical Models

Learning from the information of the others ...

- During the previous lectures we have seen several examples, where we performed **a single statistical analysis**.
- Now, suppose that **instead of having a single analysis, you have several ones**. Each one giving independent results from each other.

One might ask:

Should we combine these independent results in a single global analysis?

The general answer is:

YES!!

- Hierarchical models give us the statistical framework to combine multiple sources of information in a single analysis.
- Informally, by combining several individual results, each individual analysis is improve by *the information we have from the others*. This information is summarized in an *Empirical Prior distribution*.
- From the classical point of view, Charles Stein demonstrated in the 50s a fundamental theoretical result, which shows the benefit of combining individual results in a single analysis.

Charles Stein (22nd of March 1920) ... and he still working everyday!

Hierarchical Models

It becomes common practice to construct statistical models which reflects the underline complexity of the problem under study. Such us different patterns of heterogeneity, dependence, miss-measurements, missing data, etc.

Statistical models which reflect complexity of the data involve multiple parameters. Examples are:

- ”Study effect” in meta-analysis
- ”Subject effect” in growth curves models
- ”Identification of hidden process” in sequence of observations
- ”Relative risks of disease outcome in different areas/time periods
- ”Frailty effects” in correlated survival data
- ...

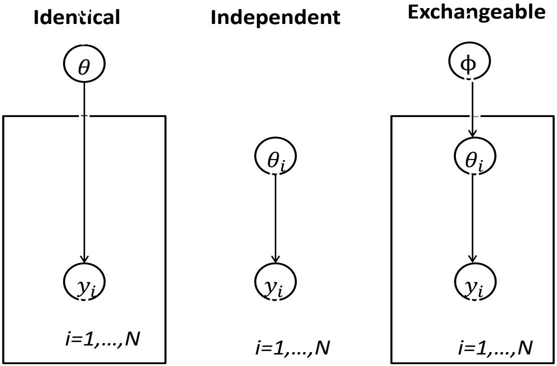
Statistical Inference for Multiple Parameters

How to make inference on multiple parameters $\{\theta_1, \ldots, \theta_N\}$ measured in N units (person, centers, areas, ...) which are related or connected by the structure of the problem?

We can identify three different scenarios

1. **Identical parameters:** All the θ's are identical, i.e. all the data can be pooled and the individual units ignored.
2. **Independent parameters:** All the θ's are entirely unrelated, i.e. the results from each unit can be analyzed independently.
3. **Exchangeable parameters:** The θ's are assumed to be 'similar' in the sense that the "labels" convey no information.

Graphical Models for Multiple Parameters



Exchangeability

- The assumption of exchangeable units is mathematically equivalent to assuming that the θ's are drawn at random from some population distribution.

- "Exchangeability" express formally the idea that we find no systematic reason to distinguish the individual random variables
$$y_1, \ldots, y_n.$$

- We assume that $y_1, \ldots, y_n$ are exchangeable if the probability that we assign to any set of potential outcomes,
$$p(y_1, \ldots, y_n),$$
is unaffected by permutations of the labels attached to the variables.

Example:

Suppose that y_1, y_2, y_3 are the outcomes of the three patients enrolled in a clinical trial. Were $y_i = 1$ indicates positive reaction to a treatment and $y_i = 0$ no reaction.

We may judge

$$p(y_1 = 1, y_2 = 0, y_3 = 1) = p(y_2 = 1, y_1 = 0, y_3 = 1) = p(y_1 = 1, y_3 = 0, y_2 = 1)$$

- i.e. the probability of getting 2 positive outcomes is NOT affected by the particular patient on which the positive outcome comes.
- Note that this is a very strong assumption. In reality we may expect that patients may behave in different way. For example they may fail to comply a treatment.
- Note that ”exchangeability” does not mean we believe that $y_1, y_2, \dots y_n$ are independent!

Exchangeability and Hierarchical Models

Suppose y_{ij} is outcome for individual j, unit i, with unit-specific parameter θ_i

- Assumption of partial exchangeability of individuals within units is represented by

$$\begin{aligned} y_{ij} &\sim p(y_{ij}|\theta_i) \\ \theta_i &\sim p(\theta_i|\phi) \end{aligned}$$

- Assumption of exchangeability of units can be represented by

$$\begin{aligned} \theta_i &\sim p(\theta_i|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

where, $p(\phi)$ can be considered as **a common prior** for all units, but with **unknown parameters**.

- Assuming that $\theta_1, \ldots, \theta_N$ are drawn from some common prior distribution whose parameters are unknown is known as a **hierarchical** or **multilevel** model.

- Bayesian statistical inference is based on:

$$p(\theta_1, \ldots, \theta_N, \phi | \mathbf{y}) \propto p(\phi)\, p(\theta_1, \ldots, \theta_N | \phi)\, p(\mathbf{y} | \theta_1, \ldots, \theta_N, \phi)$$

- The dimension of $(\theta_1, \ldots, \theta_N)$ could be very large in practice.

- Empirical Bayes techniques omit $p(\phi)$. They are useful in practice, but they may biased variability estimates be reusing $\mathbf{y}$ to estimate ϕ.

Exchangeability some further comments

- Note that there does not need to be any actual sampling - perhaps these N units are the only ones that exists - this is very common in meta-analysis.

- The probability structure is a consequence of the belief in exchangeability rather than a physical randomization mechanism.

- We emphasis that an assumption of exchangeability is a judgement based on our knowledge of the context.

Hierarchical models and shrinkage

Suppose in each unit we observe a response y_i assumed to have a Normal likelihood

$$y_i \sim N(\theta_i, s_i^2)$$

Unit means θ_i are assumed to be exchangeable, and to have a Normal distribution

$$\theta_i \sim N(\mu, \tau^2)$$

where μ and τ^2 are "hyper-parameters" for the moment assumed known.

After observing y_i, Bayes theorem gives

$$\theta_i | y_i \sim N(w_i \, \mu + (1 - w_i) \, y_i, (1 - w_i) \, s_i^2)$$

where $w_i = s_i^2/(s_i^2 + \tau^2)$ is the weight given to the prior mean.

Hierarchical models and shrinkage

- An exchangeable model therefore leads to the inferences for each unit having narrowed intervals that if they are assumed independent, but shrunk towards the prior mean response.

- w_i controls the "shrinkage" of the estimate towards μ, and the reduction in the width of the interval for θ_i

- Shrinkage (w_i) depends on precision of the individual unit i relative to the variability between units.

Profile likelihoods and Empirical Bayes

We consider the hierarchical model

$$y_i \sim N(\theta_i, s_i^2), \quad \theta_i \sim N(\mu, \tau^2).$$

The hyperparameters μ and τ are unknown. The marginal distribution of the data is

$$y_i \sim N(\mu, s_i^2 + \tau^2).$$

Let $w_i = 1/(s_i^2 + \tau^2)$ be the weight associated to the ith study. Then the joint log(likelihood) for μ and τ is

$$L(\mu, \tau) = -\frac{1}{2} \sum \left[(y_i - \mu)^2 w_k - \log w_k \right]. \tag{1}$$

By differentiating with respect to μ and setting to 0, the conditional ML estimator of μ is

$$\widehat{\mu(\tau)} = \sum y_i w_i / \sum w_i, \tag{2}$$

with variance $1/\sum w_i$.

- We can substitute (2) in (1) and obtain the profile likelihood for τ:

$$L(\tau) = -\frac{1}{2} \sum \left[(y_i - \widehat{\mu(\tau)})^2 w_k - \log w_k \right]. \qquad (3)$$

- This profile log(likelihood) may be plotted and maximized numerically to obtain the ML estimate $\widehat{\tau}$. This can be then be substituted in (2) to obtain the ML estimate of μ.

- The estimates $\widehat{\mu}$ and $\widehat{\tau}$ can be substituted in the posterior of θ_i and get an Empirical Bayes posterior for θ_i:

$$p(\theta_i | \widehat{\mu}, \widehat{\tau}, data).$$

Boundary estimate problems of the scale parameter

- We illustrate this issue with a simulation example. We assume

$$\begin{aligned} y_i &\sim N(\theta_i, 1), \\ \theta_i &\sim N(0, \tau^2) \end{aligned}$$

  for $i = 1, \ldots, 10$ and we assume a typical $\tau = 0.5$.

- For this model we simulate 1000 data sets $y_1, \ldots, y_{10}$ for each one we determine the marginal likelihood and its ML estimation of τ.

- In this simple model the ML of τ is just $\widehat{\tau^2} = 1/N \sum y_i^2 - 1$ if $1/N \sum y_i^2 > 1$ and $\widehat{\tau^2} = 0$ otherwise.

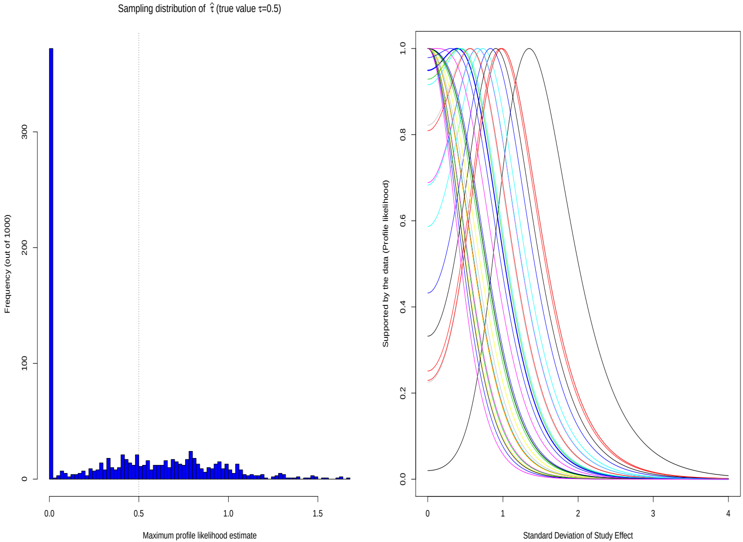


Figure: Boundary issues of the dispersion parameter τ in random-effects meta-analysis. Left panel: Sampling distribution of $\widehat{\tau}$. Right panel: Profile likelihoods of τ simulated data.

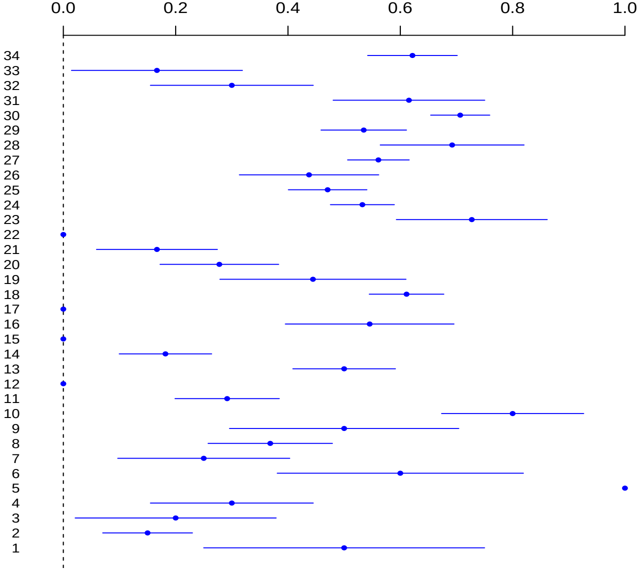
Potential advantages of a full Bayesian Hierarchical Model

- Unified modeling: full probability model with flexible choice of "random-effects" distribution
- Exact likelihoods: not necessary to adopt approximated Normal likelihoods
- Allowing for uncertainty in all parameters: full uncertainty from all the parameters is reflected in the widths of the posterior intervals
- Allowing for other sources of evidence: other sources of evidence can be reflected in the prior distributions for parameters
- Allowing direct probability statements on different scales: Possible to make inference on a variety of scales, such as risk differences, odds ratio, etc.

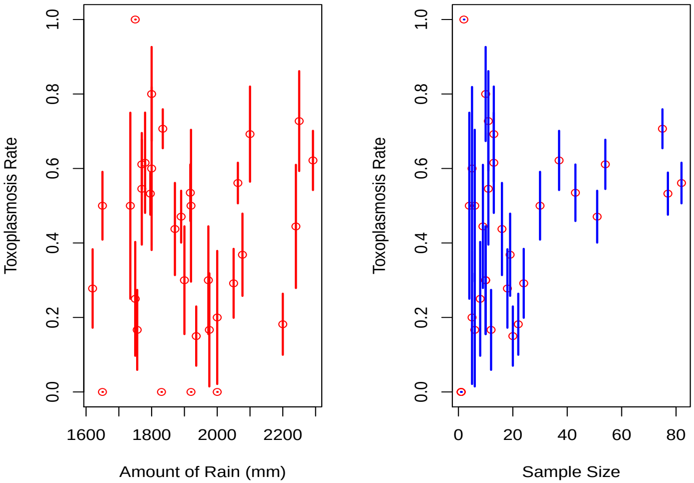
Example: Toxoplasmosis data

- These data present the relation between rainfall and the proportions of people with toxoplasmosis for 34 cities in El Salvador (Efron 1986)
- The question is how to combine results of 34 different cities in a single model?
- The data have been used to illustrate non-linear relationship between the amount of rain and toxoplasmosis prevalence
- We illustrate a series of hierarchical models

...we take some preliminary visualization


Toxoplasmosis Rates: El Salvador 34 cities

Toxoplasmosis Rates: El Salvador 34 cities



Pooled approach:

To start we ignore possible (complex) relation between rainfall and proportion and we fit same beta-binomial model to all the cities

$$\begin{aligned}
\prod_{i=1}^{34} p(r_i|\pi, n_i) &= \prod_{i=1}^{34} \texttt{Binomial}(n_i, \pi) \\
p(\pi) &= \texttt{Beta}(a, b) \\
p(\pi|r, n) &\propto \prod_{i=1}^{34} p(r_i|\pi, n_i)p(\pi) \\
&= \texttt{Beta}\left(a + \sum_i r_i, b + \sum_i (n_i - r_i)\right)
\end{aligned}$$

Some comments:

- Now, is it reasonable to assume **common probability** π of Toxoplasmosis for each city?
- The beta-binomial model assumes that each outcome is independent and identically distributed according to the binomial probability distribution with parameter π
- Does this model adequately describe the random variation in outcomes for each city?
- Are the cities rates more variable that our model assumes ?

Question: How to model the excess of variation in the data?

Modeling the excess of variation

Model 1:

Combining information with a Generalized Linear Mixed Model (GLMM)

$$\begin{aligned} r_i &\sim \texttt{Binomial}(n_i, \pi_i) \\ \texttt{logit}(\pi_i) &\sim \texttt{N}(\mu, \sigma^2) \\ \mu &\sim \texttt{N}(0, 100) \\ \sigma &\sim \texttt{Uniform}(0.01, 10) \end{aligned}$$

This model combines information between cities in a single model.

Model 2:

Modeling each city rate independently

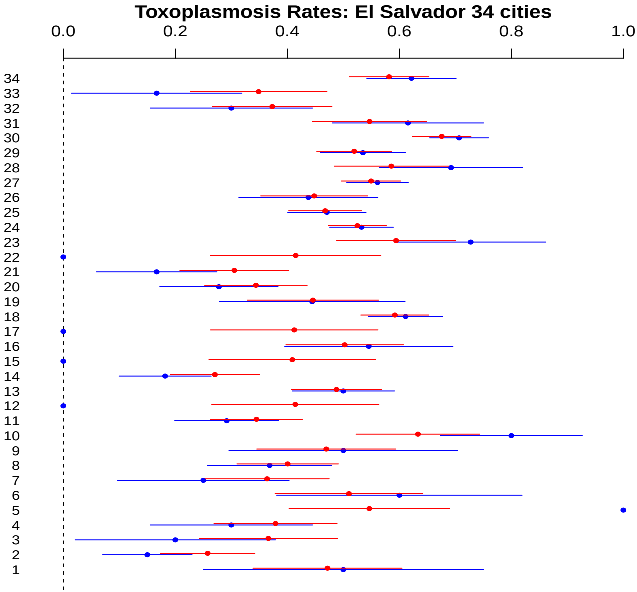
$$\begin{aligned} r_i &\sim \texttt{Binomial}(n_i, \pi_i) \\ \texttt{logit}(\pi_i) &\sim \texttt{N}(0, 100). \end{aligned}$$

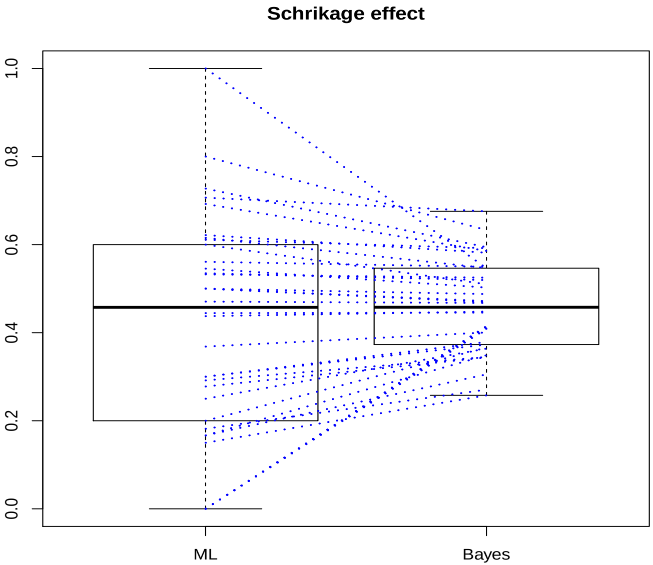
This model does NOT use information between cities. Each rate π_i is estimated independently.

In WinBUGS we run both models simultaneously as follows:

```
model{
   for( i in 1 : I ) {
      r[i] ~ dbin(p[i,1],n[i])
        logit(p[i,1]) <- a[i]
      a[i] ~ dnorm(alpha1, tau1) # Exchangeable
     r2[i] ~ dbin(p[i,2],n[i])   # Copy of r[i]
      logit(p[i,2]) <- b[i]
      b[i] ~ dnorm(0, 0.01)      # Independent
           }
tau1 <- 1/(sigma1*sigma1)        # Priors
sigma1 ~ dunif(0.01, 5)
alpha1 ~ dnorm(0,0.01)
}
```

```
#data
list(I=34,
r=c(2,3,1,3,2,3,2,7,3,8,7,0,15,4 ,0,6 ,0,33,
4,5 ,2 ,0,8,41,24,7, 46,9,23,53,8,3,1,23),

r2=c(2,3,1,3,2,3,2,7,3,8,7,0,15,4 ,0,6 ,0,33,
4,5 ,2 ,0,8, 41,24,7, 46,9,23,53,8,3,1,23),

n=c(4,20,5,10,2,5,8,19,6,10,24,1,30,22,1,11,
1,54,9,18,12,1,11,77,51,16,82,13,43,75,13,10,6,37)
)
```


Toxoplasmosis Rates: El Salvador 34 cities
0.0
0.2
0.4
0.6
0.8
1.0
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1


Schrikage effect
1.0
0.8
0.6
0.4
0.2
0.0
ML
Bayes

Results:

DIC for Model 1 (GLMM) $=$ 145.670

DIC for Model 2 (independent) $=$176.504

- Clearly Model 1 which combines information between cities is the winer.
- The use of Model 2 has the effect of reduce the variability between cities
- The use of Model 2 gives better results at the level of the city.

Clear conclusion: We should take advantage of modeling simultaneously multiple results!

Example: Toxoplasmosis data continue ...

Two important questions are:

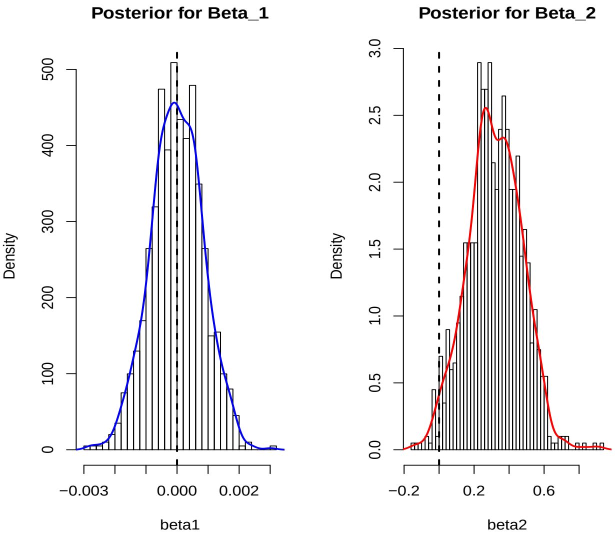
- Which is the influence of the amount of rain in toxoplasmosis prevalence?
- Which is the influence of the number of participants in each town?

Model 3:

$$\begin{aligned}
r_i &\sim \texttt{Binomial}(n_i, \pi_i) \\
\texttt{logit}(\pi_i) &= a_i + \beta_1 \times (x_{i,1} - \bar{x}_1) + \beta_2 \times (x_{i,1} - \bar{x}_1) \\
a_i &\sim \texttt{N}(\alpha, \sigma^2) \\
\sigma &\sim \texttt{Uniform}(0.01, 10) \\
\alpha,\ \beta_1,\ \beta_2 &\sim \texttt{N}(0, 100),
\end{aligned}$$

In WinBUGS:

```
model
{
for(i in 1:N)
{log.n[i] <- log(n[i])}
for( i in 1 : N ) {
      r[i] ~ dbin(p[i], n[i])
  a[i] ~ dnorm(alpha,tau)
    logit(p[i]) <- a[i] +
                   beta1 * (rain[i]-mean(rain[])) +
                   beta2 * (log.n[i] - mean(log.n[]))
   }
# Priors
tau <- 1/(sigma*sigma)
sigma ~ dunif(0.01, 5)
alpha ~ dnorm(0,0.01)
beta1 ~ dnorm(0, 0.01)
beta2 ~ dnorm(0, 0.01)}
```



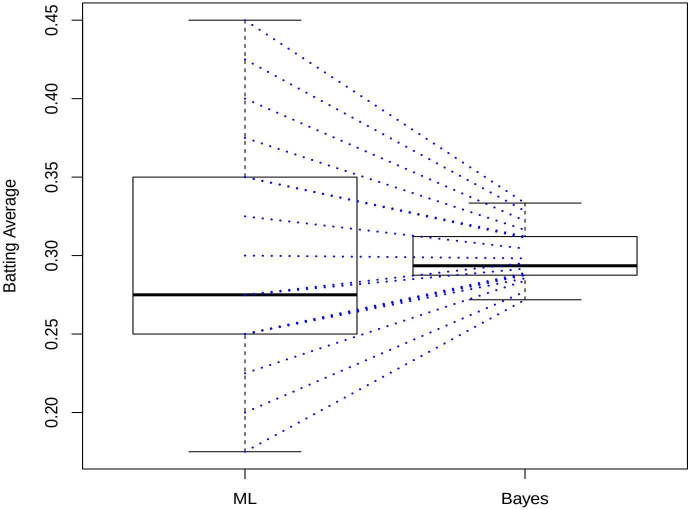
Posterior of β_1 clearly indicates no linear influence of the amount of rain. Posterior of β_2 shows a clear trend.

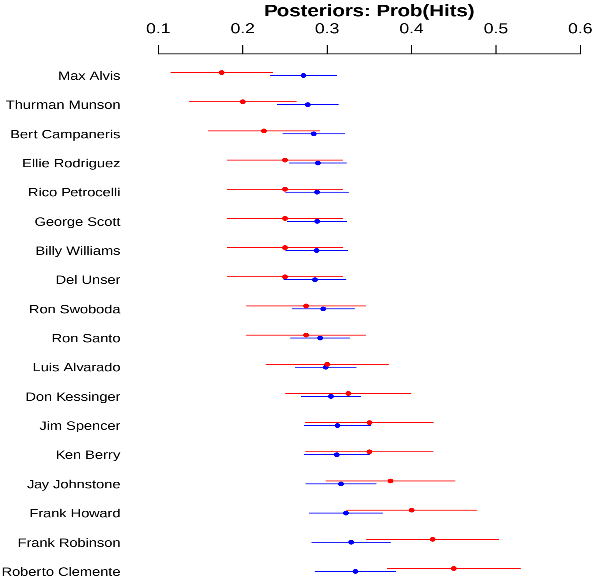
Example: Ranking of the eighteen baseball players (Efron and Morris, 1977)

How can we rank the players ability ? Who was the best player of the season 1970?

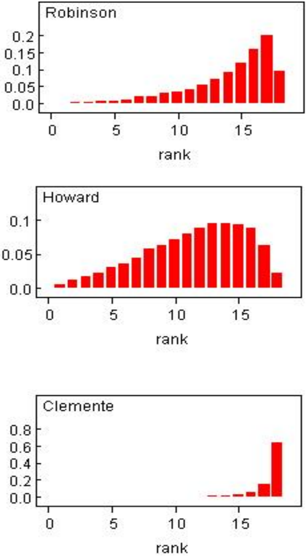
We use the ranking function in WinBUGS to answer this question.

Name	hit/AB	Observed Avg (ML)	**"TRUTH"**	James-Stein
Clemente	18/45	**.400**	.346	.290
Robinson	17/45	**.378**	.298	.286
⋮	⋮	⋮	⋮	⋮
Total Squared error		**.077**		**0.022**

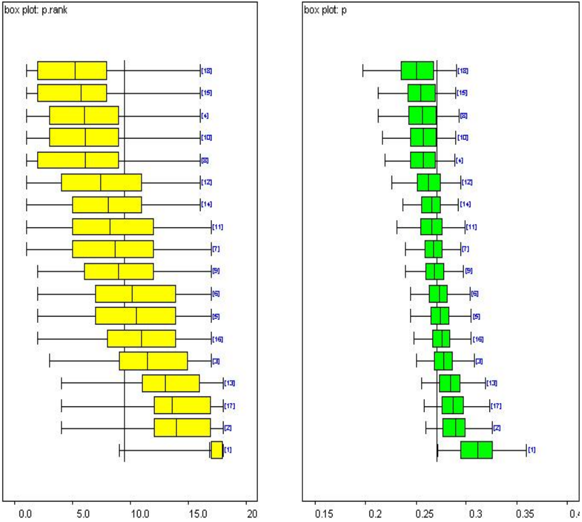

Batting Average
0.20
0.25
0.30
0.35
0.40
0.45
ML
Bayes


Posteriors: Prob(Hits)
0.1
0.2
0.3
0.4
0.5
0.6
Max Alvis
Thurman Munson
Bert Campaneris
Ellie Rodriguez
Rico Petrocelli
George Scott
Billy Williams
Del Unser
Ron Swoboda
Ron Santo
Luis Alvarado
Don Kessinger
Jim Spencer
Ken Berry
Jay Johnstone
Frank Howard
Frank Robinson
Roberto Clemente

```
model{
      for( i in 1 : N ) {
        r[i] ~ dbin(p[i],n[i])
 logit(p[i]) <- b[i]
        b[i] ~ dnorm(mu, tau)
   p.rank[i] <- rank(p[], i)
                            }
# hyper-priors
          mu ~ dnorm(0.0,1.0E-6)
       sigma ~ dunif(0,100)
         tau <- 1/(sigma*sigma)
        }
# Data
list(r = c(145, 144, 160, 76, 128, 140, 167,
       41, 148, 57, 83, 79, 142, 152, 52, 168, 137, 21),
n= c(412, 471, 566, 320, 463, 511, 631, 183,
      555, 245, 322, 315, 480, 583, 231, 603, 453,
      115), N=18 )
```



Rank distributions for Robinson, Howard and Clemente.



Left panel: posterior distributions of ranks. Right panel: posterior distributions of performances.

Summary on hierarchical models

- Hierarchical models allows to "borrow strength" across units
- Posteriors distribution of θ_i for each unit borrows information from the likelihood contributions for all other units. Estimation is more efficient.
- MCMC allows considerable flexibility over choice of random effects distribution (not restricted to normal random errors)
- MCMC allows to make inference on difficult questions, e.g. ranking estimation of random effects
- Easy to extend to more complicated models (e.g. non-linear repeated measurements, etc.)

Lecture:

Longitudinal Data Analysis

Introduction

- There is a great interest in the analysis of hierarchical data resulting form longitudinal studies
- In these problems each experimental unit is measured several times, e.g. patients participating in a clinical study are measured in different periods of the trial.
- The common feature of this type of data is that measurements within units can not be considered statistically independent.
- Therefore, a special modeling technique should be considered. For example, mixed effects modeling

Running Example: HIV children trial

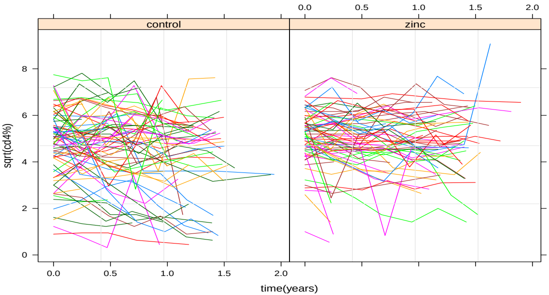
- The data correspond to repeated measurements of HIV positive children during a period of two years.
- The outcome variable is a measurement of the immune system (CD4 percentage of cells).
- There are two treatment groups: The control group corresponds to children without zinc supplement dietary and the treatment group corresponds to children with zinc supplement.
- It is expected that a diet with zinc supplement will improve the response of the immune system.

- Data organized in hierarchical way, with $y = \sqrt{CD4\%}$, time = years, tr = treatment and person = id:

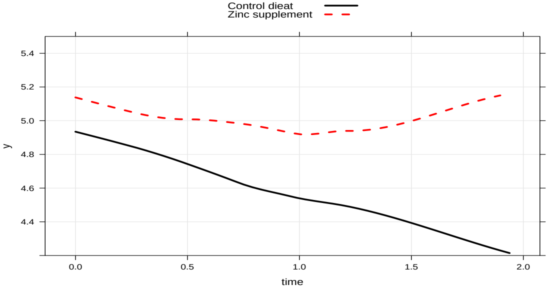
```
       y  time      tr person
1   4.243 0.000 control      1
2   6.083 0.558 control      1
3   3.606 0.788 control      1
4   3.606 1.421 control      1
5   3.464 1.938 control      1
6   1.000 0.000    zinc      2
7   0.548 0.213    zinc      2
8   5.477 0.000 control      3
...
```

- The data are very noisy and it is difficult to observe differences between groups.

HIV positive children (1 to 5 years old)



HIV positive children



Identical parameters model: pool data model

- A model with identical parameters corresponds to a linear model with common intercept parameter α and different slopes per treatment group, β_1 (control group) and β_2 (zinc complement group):

$$y_{i,j} = \alpha + \beta_j \, time_i + \epsilon_{i,j}$$

with

$$\epsilon_{i,j} \sim N(0, \sigma_\epsilon^2)$$

- Independent non-informative prior distributions:

$$\alpha, \beta_1, \beta_2 \sim N(0, 100)$$

and

$$\sigma_\epsilon \sim Uniform(0, 10)$$

- The hypothesis of interest is if $\beta_1 - \beta_2 < 0$

Identical parameters model: WinBUGS

The model in BUGS:

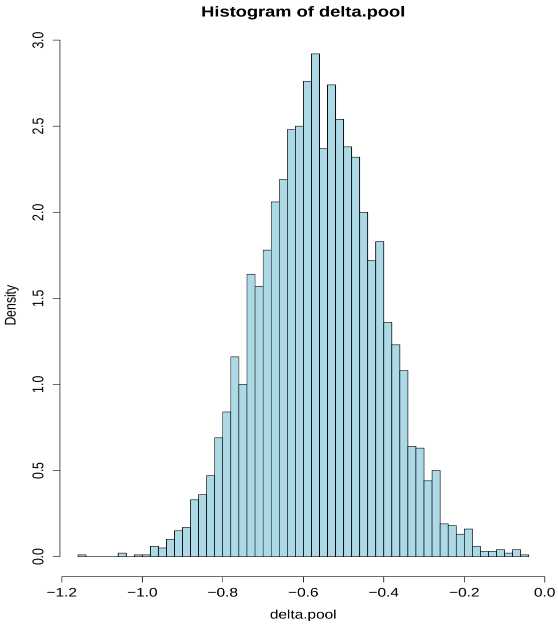
```
model
{
# Priors ...........................................
 alpha ~ dnorm(0, 0.01)
 beta[1] ~ dnorm(0, 0.01)    # Slope control group
 beta[2] ~ dnorm(0, 0.01)    # Slope zinc supplement group
 prec.y <- pow(sigma.y, -2)
 sigma.y ~ dunif(0, 10)
# Data model .......................................
for(i in 1:n)
{
  y[i] ~ dnorm(mu[i], prec.y)
  mu[i] <- alpha + beta[tr[i]] * time[i]
}
}
```

Results identical parameters model

```
Inference for Bugs model at "hiv_pool.bug", fit using WinBUGS,
 2 chains, each with 10000 iterations (first 5000 discarded), n.
 n.sims = 5000 iterations saved
           mean  sd   2.5%   25%   50%   75%  97.5%
alpha       4.9 0.1    4.7   4.8   4.9   4.9    5.0
beta[1]    -0.5 0.1   -0.8  -0.6  -0.5  -0.4   -0.3
beta[2]     0.0 0.1   -0.2  -0.1   0.0   0.1    0.3
sigma.y     1.5 0.0    1.4   1.4   1.5   1.5    1.6

DIC info (using the rule, pD = Dbar-Dhat)
pD = 4.0 and DIC = 2530.8
DIC is an estimate of expected predictive error
```

Posterior for $\delta = \beta_1 - \beta_2$ is:

```
attach.bugs(m.hiv.pool)
delta.pool <- beta[,1] - beta[,2]
hist(delta.pool, breaks = 50, freq = F, col = "lightblue")
```


Histogram of delta.pool
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
delta.pool

Exchangeable parameters model: Mixed-effects model

- In this model children trajectories are modelled with two random components: α_i a random intercept and $\beta_{i,j}$ a random slope:

$$y_{i,j} = \alpha_i + \beta_{i,j}\, time_i + \epsilon_{i,j}$$

with

$$\begin{aligned}
\alpha_i &\sim N(\mu_\alpha, \sigma^2_\alpha) && (4)\\
\beta_{i,1} &\sim N(\mu_{\beta_1}, \sigma^2_\beta) && (5)\\
\beta_{i,2} &\sim N(\mu_{\beta_2}, \sigma^2_\beta) && (6)\\
\epsilon_{i,j} &\sim N(0, \sigma^2_\epsilon) && (7)
\end{aligned}$$

- Independent non-informative prior distributions:

$$\mu_\alpha, \mu_{\beta_1}, \mu_{\beta_2} \sim N(0, 100)$$

and

$$\sigma_\alpha, \sigma_\beta, \sigma_\epsilon \sim Uniform(0, 10)$$

Some comments on the exchangeable parameters model

- In this model each child has his/her own trajectory
- The number of parameters is

$$2 \times \texttt{number of children} + \texttt{hyperparameters}$$

- We can allow α_i and β_i to be correlated
- We can model random effects with non-normal distributions
- We can model outcome variables as non-normal as well
- We can add non-linear terms to the expected response, etc.

Exchangeable parameters model: WinBUGS

```
  "model
  {
  # Priors ...................................
  for(j in 1:J){
   alpha[j] ~ dnorm(mu.alpha, prec.alpha)
   beta[j] ~ dnorm(mu.beta[tr.group[j]], prec.beta)
  # alpha[j] ~ dt(mu.alpha, prec.alpha, 4)
  # beta[j] ~ dt(mu.beta[tr.group[j]], prec.beta, 4)
}
  mu.alpha ~ dnorm(0, 0.01)
  prec.alpha <- pow(sigma.alpha, -2)
  sigma.alpha ~ dunif(0, 10)

  mu.beta[1] ~ dnorm(0, 0.01)    # Slope control group
  mu.beta[2] ~ dnorm(0, 0.01)    # Slope zinc supplement group
  prec.beta <- pow(sigma.beta, -2)
  sigma.beta ~ dunif(0, 10)

  prec.y <- pow(sigma.y, -2)
  sigma.y ~ dunif(0, 10)
```

Exchangeable parameters model: WinBUGS

```
 # Data model .......................................
  for(i in 1:n)
  {
  # Observations at child level
  y[i] ~ dnorm(mu[i], prec.y)

 # y[i] ~ dt(mu[i], prec.y, 4)

 # Random intercept and slope ...

 # mu[i] <- alpha[person[i]] + beta[person[i]] * time[i]
   mu[i] <- alpha[person[i]] + beta[person[i]] * time[i]
 #            + gamma[tr[i]]*pow(time[i], 2)  # quadratic term..
}
  }
```

Results for the exchangeable parameters model

Results for the model based on Normals for random-effects

```
              mean   sd   2.5%   25%   50%   75%  97.5%
mu.alpha       4.9  0.1    4.7   4.8   4.9   5.0    5.1
sigma.alpha    1.3  0.1    1.1   1.2   1.3   1.3    1.4
mu.beta[1]    -0.6  0.1   -0.8  -0.7  -0.6  -0.5   -0.3
mu.beta[2]    -0.4  0.1   -0.7  -0.5  -0.4  -0.4   -0.2
sigma.beta     0.6  0.1    0.4   0.5   0.6   0.6    0.8
gamma[2]       0.3  0.1    0.1   0.2   0.3   0.3    0.5
sigma.y        0.7  0.0    0.7   0.7   0.7   0.7    0.8
```

Comparison between models

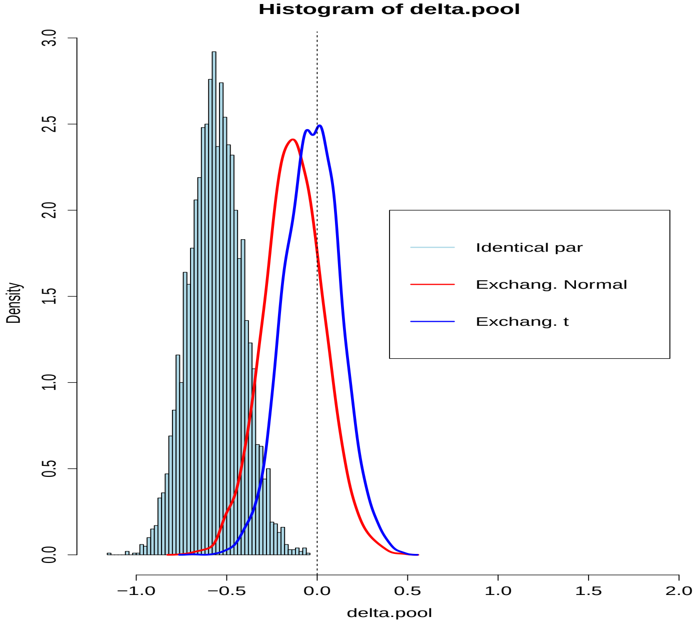
Note the differences between DIC

```
# Identical parameters
pD = 4.0 and DIC = 2530.8
# Exchangeable parameters Normal
pD = 196.8 and DIC = 1683.3
# Exchangeable parameters t-distribution 4 df.
pD = 232.4 and DIC = 1440.0
```

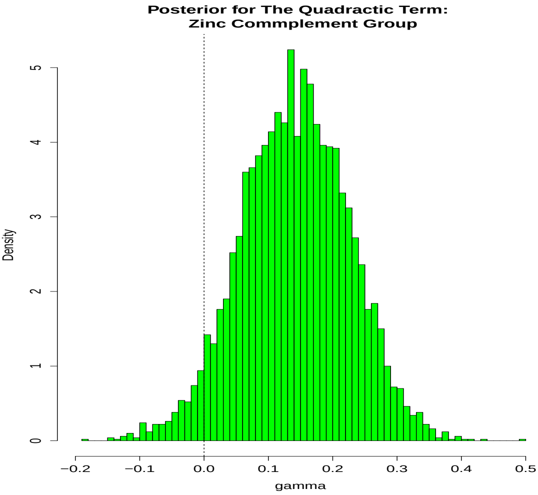
Variability explained by random components:

```
# Variability explained
 tot.var <- sigma.y^2 + sigma.alpha^2 + sigma.beta^2
 mean(sigma.y^2/tot.var) * 100
 13.21
 mean(sigma.alpha^2/tot.var) * 100
 70.69
 mean(sigma.beta^2/tot.var)* 100
 16.08
```

Comparison between models

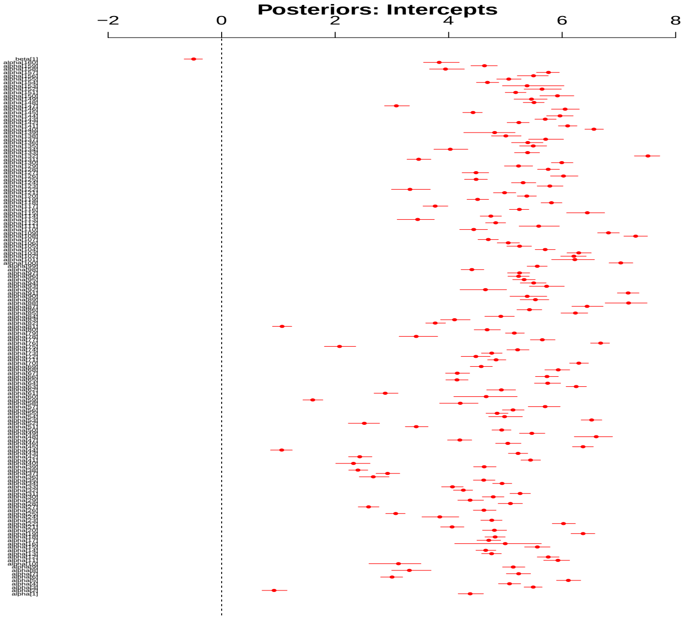


Posterior for the quadratic term

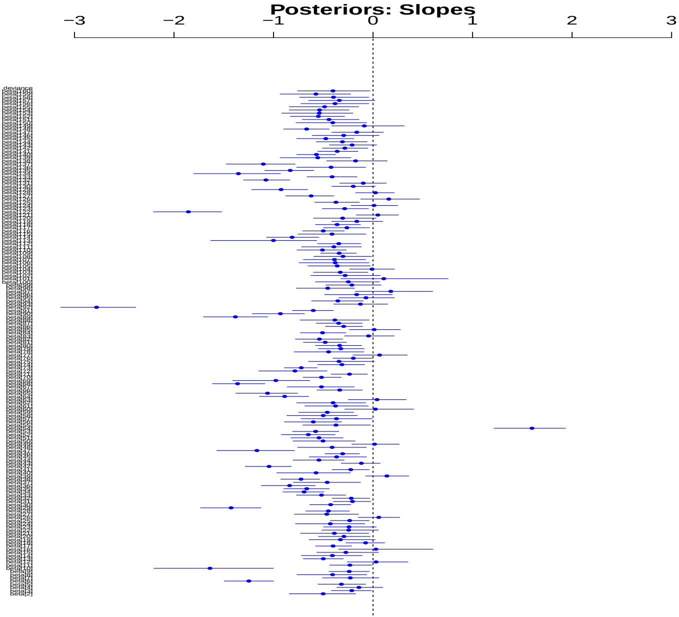


The difference between the groups is explained by $\gamma > 0$ i.e. in the zinc supplement group the children recover after one year!

Posteriors for random intercepts



Posteriors for random slopes



Summary of the course

- A new interpretation of probability as a subjective mental construct
- Subjective probability does not mean a bias analysis
- Relationship with classical statistics and the implied priors
- Uses of predictions for model checking and missing data implutations
- Thinking different about regression models
- Different modeling tools: WinBUGS, DAGs, DIC ...
- Introduction to hierarchical modeling and longitudinal data analysis

Finally

...Is the probability that the German team wins against Argentina 100% ?

Nice to see that most people attended and survived the course!!

Hope that everybody will be now enthusiastic Bayesians ;-)

Take home message...

GOD DOESN'T PLAY DICE WITH
THE UNIVERSE – HE PREFERS
CARD GAMES

MUCHAS GRACIAS !!!