# Contents

1 Basic concepts in Bayesian methods ........................................... 1
   0.1 Preface ................................................................ 2
   0.2 Time schedule of the course .................................................... 7
   0.3 Aims of the course .......................................................... 9
1 Modes of statistical inference ................................................................. 10
   1.1 The frequentist approach: a critical reflection............................................ 11
      1.1.1 The classical statistical approach .............................................. 12
      1.1.2 The P-value as a measure of evidence........................................... 15
      1.1.3 The confidence interval as a measure of evidence 26
   1.2 Statistical inference based on the likelihood function................................. 28
      1.2.1 The likelihood function ................................................... 29
      1.2.2 The likelihood principles ................................................... 34
   1.3 The Bayesian approach: some basic ideas.............................................. 43
      1.3.1 Introduction ........................................................ 44
      1.3.2 Bayes theorem – Discrete version for simple events - 1 ................................... 48
   1.4 Outlook ............................................................... 54
2 Bayes theorem: computing the posterior distribution ............................................................. 59
   2.1 Introduction ........................................................... 60
   2.2 Bayes theorem – The binary version.................................................. 61
   2.3 Probability in a Bayesian context ....................................................... 63
   2.4 Bayes theorem – The categorical version .............................................. 66
   2.5 Bayes theorem – The continuous version .............................................. 67
   2.6 The binomial case .......................................................... 69
   2.7 The Gaussian case ........................................................... 77
   2.8 The Poisson case ........................................................... 102
2.9 The prior and posterior of derived parameter............................................ 113
2.10 Bayesian versus likelihood approach ................................................ 116
2.11 Bayesian versus frequentist approach ............................................... 117
2.12 The different modes of the Bayesian approach........................................... 119
2.13 An historical note on the Bayesian approach............................................ 120
3 Introduction to Bayesian inference ............................................................... 127
   3.1 Introduction ............................................................. 128
   3.2 Summarizing the posterior with probabilities ............................................ 129
   3.3 Posterior summary measures ......................................................... 130
      3.3.1 Posterior mode, mean, median, variance and SD 131
      3.3.2 Credible/credibility interval ................................................. 135
   3.4 Predictive distributions ........................................................... 140
      3.4.1 Introduction ........................................................ 141
      3.4.2 Posterior predictive distribution: General case 148
   3.5 Exchangeability .............................................................. 157

Bayesian Biostatistics - Piracicaba 2014
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6</td>
<td>A normal approximation to the posterior</td>
<td>158</td>
</tr>
<tr>
<td>3.7</td>
<td>Numerical techniques to determine the posterior</td>
<td>159</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Numerical integration</td>
<td>160</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Sampling from the posterior distribution</td>
<td>162</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Choice of posterior summary measures</td>
<td>175</td>
</tr>
<tr>
<td>3.8</td>
<td>Bayesian hypothesis testing</td>
<td>176</td>
</tr>
<tr>
<td>3.8.1</td>
<td>The Bayes factor</td>
<td>183</td>
</tr>
<tr>
<td>3.9</td>
<td>Medical literature on Bayesian methods in RCTs</td>
<td>185</td>
</tr>
<tr>
<td>4</td>
<td>More than one parameter</td>
<td>196</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>197</td>
</tr>
<tr>
<td>4.2</td>
<td>Joint versus marginal posterior inference</td>
<td>199</td>
</tr>
<tr>
<td>4.3</td>
<td>The normal distribution with ( \mu ) and ( \sigma^2 ) unknown</td>
<td>201</td>
</tr>
<tr>
<td>4.3.1</td>
<td>No prior knowledge on ( \mu ) and ( \sigma^2 ) is available</td>
<td>202</td>
</tr>
<tr>
<td>4.3.2</td>
<td>An historical study is available</td>
<td>214</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Expert knowledge is available</td>
<td>218</td>
</tr>
<tr>
<td>4.4</td>
<td>Multivariate distributions</td>
<td>220</td>
</tr>
<tr>
<td>4.5</td>
<td>Frequentist properties of Bayesian inference</td>
<td>227</td>
</tr>
<tr>
<td>4.6</td>
<td>The Method of Composition</td>
<td>228</td>
</tr>
<tr>
<td>4.7</td>
<td>Bayesian linear regression models</td>
<td>231</td>
</tr>
<tr>
<td>4.7.1</td>
<td>The frequentist approach to linear regression</td>
<td>232</td>
</tr>
<tr>
<td>4.7.2</td>
<td>A noninformative Bayesian linear regression model</td>
<td>235</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Posterior summary measures for the linear regression model</td>
<td>236</td>
</tr>
<tr>
<td>4.7.4</td>
<td>Sampling from the posterior distribution</td>
<td>239</td>
</tr>
<tr>
<td>4.8</td>
<td>Bayesian generalized linear models</td>
<td>243</td>
</tr>
<tr>
<td>4.8.1</td>
<td>More complex regression models</td>
<td>244</td>
</tr>
<tr>
<td>5</td>
<td>Choosing the prior distribution</td>
<td>246</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>247</td>
</tr>
<tr>
<td>5.2</td>
<td>The sequential use of Bayes theorem</td>
<td>248</td>
</tr>
<tr>
<td>5.3</td>
<td>Conjugate prior distributions</td>
<td>249</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Conjugate priors for univariate data distributions</td>
<td>250</td>
</tr>
<tr>
<td>5.4</td>
<td>Noninformative prior distributions</td>
<td>266</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Introduction</td>
<td>267</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Expressing ignorance</td>
<td>268</td>
</tr>
<tr>
<td>5.4.3</td>
<td>General principles to choose noninformative priors</td>
<td>271</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Improper prior distributions</td>
<td>273</td>
</tr>
<tr>
<td>5.4.5</td>
<td>Weak/vague priors</td>
<td>274</td>
</tr>
<tr>
<td>5.5</td>
<td>Informative prior distributions</td>
<td>279</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Introduction</td>
<td>280</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Data-based prior distributions</td>
<td>285</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Elicitation of prior knowledge</td>
<td>286</td>
</tr>
<tr>
<td>5.5.4</td>
<td>Archetypal prior distributions</td>
<td>294</td>
</tr>
<tr>
<td>5.6</td>
<td>Prior distributions for regression models</td>
<td>302</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Normal linear regression</td>
<td>303</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Generalized linear models</td>
<td>305</td>
</tr>
<tr>
<td>5.7</td>
<td>Modeling priors</td>
<td>306</td>
</tr>
<tr>
<td>5.8</td>
<td>Other regression models</td>
<td>311</td>
</tr>
<tr>
<td>6</td>
<td>Markov chain Monte Carlo sampling</td>
<td>314</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>315</td>
</tr>
<tr>
<td>6.2</td>
<td>The Gibbs sampler</td>
<td>327</td>
</tr>
<tr>
<td>6.2.1</td>
<td>The bivariate Gibbs sampler</td>
<td>328</td>
</tr>
<tr>
<td>6.2.2</td>
<td>The general Gibbs sampler</td>
<td>340</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Remarks</td>
<td>353</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Review of Gibbs sampling approaches</td>
<td>354</td>
</tr>
<tr>
<td>6.3</td>
<td>The Metropolis-Hastings algorithm</td>
<td>356</td>
</tr>
<tr>
<td>6.3.1</td>
<td>The Metropolis algorithm</td>
<td>357</td>
</tr>
<tr>
<td>6.3.2</td>
<td>The Metropolis-Hastings algorithm</td>
<td>367</td>
</tr>
</tbody>
</table>
Bayesian Biostatistics - Piracicaba 2014

II Bayesian tools for statistical modeling

9 Hierarchical models ............................................................................................ 477
9.1 The Poisson-gamma hierarchical model ....................................................... 480
9.1.1 Introduction .......................................................................................... 481
9.1.2 Model specification ............................................................................. 485
9.1.3 Posterior distributions ....................................................................... 491
9.1.4 Estimating the parameters .................................................................. 495
9.1.5 Posterior predictive distributions ....................................................... 502
9.2 Full versus Empirical Bayesian approach ................................................... 505
9.3 Gaussian hierarchical models ..................................................................... 509
9.3.1 Introduction ........................................................................................ 510
9.3.2 The Gaussian hierarchical model ....................................................... 511
9.3.3 Estimating the parameters .................................................................. 512
9.4 Mixed models .............................................................................................. 521
9.4.1 Introduction ........................................................................................ 522

Bayesian Biostatistics - Piracicaba 2014

III More advanced Bayesian modeling

11 Advanced modeling with Bayesian methods .................................................. 600
11.1 Example 1: Longitudinal profiles as covariates ....................................... 601
11.2 Example 2: relating caries on deciduous teeth with caries on permanent teeth ............................................................ 612

Bayesian Biostatistics - Piracicaba 2014
Are you a Bayesian?

Part I

Basic concepts in Bayesian methods
0.1 Preface

Mouthwash trial

▷ Trial that tests whether daily use of a new mouthwash before tooth brushing reduces plaque when compared to using tap water only
▷ Result: New mouthwash reduces 25% of the plaque with a 95% CI = [10%, 40%]
▷ From previous trials on similar products: overall reduction in plaque lies between 5% to 15%
▷ Experts: plaque reduction from a mouthwash does not exceed 30%
▷ What to conclude?
  ▷ Classical frequentist analysis: 25% reduction + 95% CI
  ▷ Conclusion ignores what is known from the past on similar products
  ▷ Likely conclusion in practice: truth must lie somewhere in-between 5% and 25%

A significant result on a small trial

▷ Small sized study with an unexpectedly positive result about a new medication to treat patients with oral cancer
▷ First reaction (certainly of the drug company) = “great!”
▷ Past: none of the medications had such a large effect and new medication is not much different from the standard treatment
▷ Second reaction (if one is honest) = be cautious
▷ Then, You are a Bayesian (statistician)

Incomplete information

▷ Some studies have not all required data, to tackle research question
▷ Example: Determining prevalence of a disease from fallible diagnostic test
▷ Expert knowledge can fill in the gaps
▷ Bayesian approach is the only way to tackle the problem!

Bayesian approach mimics our natural life where learning is done by combining past and present experience
Most of the material is obtained from

Bayesian Biostatistics

0.2 Time schedule of the course

- **Day 1**
  - Morning: Chapters 1 and 2
  - Afternoon:
    - Chapter 3, until Section 3.6
    - Brief practical introduction to R & Rstudio

- **Day 2**
  - Morning: refreshing day 1 + Chapter 3, from Section 3.7
  - Afternoon:
    - Discussion clinical papers
    - Computer exercises in R

- **Day 3**
  - Morning: Chapters 4 + 6 + computer exercises in R
  - Afternoon: Chapter 8 via interactive WinBUGS computer session

- **Day 4**
  - Morning: selection of topics in Chapters 5 and 7 + WinBUGS exercises
  - Afternoon: Chapter 9 + WinBUGS exercises

- **Day 5**
  - Morning: Chapter 10 + (R2)WinBUGS computer exercises
  - Afternoon: exercises and advanced Bayesian modeling + wrap up

0.3 Aims of the course

- Understand the Bayesian paradigm
- Understand the use of Bayesian methodology in medical/biological papers
- Be able to build up a (not too complex) WinBUGS program
- Be able to build up a (not too complex) R2WinBUGS program
Chapter 1
Modes of statistical inference

Aims:
 orthogonal reflects on the ‘classical approach’ for statistical inference
 orthogonal look at a precursor of Bayesian inference: the likelihood approach
 orthogonal A first encounter of the Bayesian approach

1.1 The frequentist approach: a critical reflection

- Review of the ‘classical’ approach on statistical inference

1.1.1 The classical statistical approach

Classical approach:

- Mix of two approaches (Fisher & Neyman and Pearson)
- Here: based on $P$-value, significance level, power and confidence interval
- Example: RCT

Example 1.1: Toenail RCT

- Randomized, double blind, parallel group, multi-center study (Debacker et al., 1996)
- Two treatments ($A$: Lamisil and $B$: Itraconazol) on $2 \times 189$ patients
- 12 weeks of treatment and 48 weeks of follow up (FU)
- Significance level $\alpha = 0.05$
- Sample size to ensure that $\beta \leq 0.20$
- Primary endpoint = negative mycology (negative microscopy & negative culture)
- Here unaffected nail length at week 48 on big toenail
- 163 patients treated with $A$ and 171 treated with $B$
A: \( \mu_1 \) & B: \( \mu_2 \)

\( H_0: \Delta = \mu_1 - \mu_2 = 0 \)

Completion of study: \( \hat{\Delta} = 1.38 \) with \( t_{obs} = 2.19 \) in 0.05 rejection region

Neyman-Pearson: reject that \( A \) and \( B \) are equally effective

Fisher: 2-sided \( P = 0.030 \) \( \Rightarrow \) strong evidence against \( H_0 \)

Wrong statement: Result is significant at 2-sided \( \alpha \) of 0.030. This gives \( P \)-value an a 'priori status'.

### 1.1.2 The \( P \)-value as a measure of evidence

Use and misuse of \( P \)-value:

- The \( P \)-value is not the probability that \( H_0 \) is (not) true
- The \( P \)-value depends on fictive data (Example I.2)
- The \( P \)-value depends on the sample space (Examples I.3 and I.4)
- The \( P \)-value is not an absolute measure
- The \( P \)-value does not take all evidence into account (Example I.5)

---

The \( P \)-value is not the probability that \( H_0 \) is (not) true

Often \( P \)-value is interpreted in a wrong manner

- \( P \)-value = probability that observed or a more extreme result occurs under \( H_0 \)

\( \Rightarrow P \)-value = surprise index

- \( P \)-value \( \neq \) \( p(H_0 \mid y) \)
- \( p(H_0 \mid y) \) = Bayesian probability

The \( P \)-value depends on fictive data

- \( P \)-value = probability that observed or a more extreme result occurs under \( H_0 \)

\( \Rightarrow P \)-value is based not only on the observed result but also on fictive (never observed) data

- Probability has a long-run frequency definition
- Example I.2
Example I.2: Graphical representation of \( P \)-value

\( P \)-value of RCT (Example I.1)

\[ \text{Predicted } y \text{-values} \]

\[ \text{Observed } t \text{-value} \]

\[ \text{Density} \]

\[ 0.05 \]

\[ 0.025 \]

\[ 0.0025 \]

\[ 0.0005 \]

\[ 0.000025 \]

\[ 0.00000025 \]

\[ \text{Example I.3: Accounting for interim analyses in a RCT} \]

2 identical RCTs except for the number of analyses:

- RCT 1: 4 interim analyses + final analysis
  - Correction for multiple testing
  - Group sequential trial: Pocock’s rule
  - Global \( \alpha = 0.05 \), nominal significance level=0.016

- RCT 2: 1 final analysis
  - Global \( \alpha = 0.05 \), nominal significance level=0.05

- If both trials run until the end and \( P = 0.02 \) for both trials, then for RCT 1: NO significance, for RCT 2: Significance

\[ \begin{array}{c|c|c}
\text{Treatment} & \text{Controls} & \text{Cases} \\
\hline
\text{No Chemo} & 160 & 11 \\
\text{Chemo} & 251 & 138 \\
\hline
\text{Total} & 411 & 149 \\
\end{array} \]

\[ \ \]

\[ \text{Example I.4: Kaldor et al’s case-control study} \]

- Case-control study (Kaldor et al., 1990) to examine the impact of chemotherapy on leukemia in Hodgkin’s survivors

- 149 cases (leukemia) and 411 controls

- Question: Does chemotherapy induce excess risk of developing solid tumors, leukemia and/or lymphomas?
• Pearson $\chi^2(1)$-test: $P = 7.8959 \times 10^{-13}$
• Fisher’s Exact test: $P = 1.487 \times 10^{-14}$
• Odds ratio = 7.9971 with a 95% confidence interval = [4.19, 15.25]
• Reason for difference: 2 sample spaces are different
  • Pearson $\chi^2(1)$-test: condition on $n$
  • Fisher’s Exact test: condition on marginal totals

The $P$-value is not an absolute measure

• Small $P$-value does not necessarily indicate large difference between treatments, strong association, etc.
• Interpretation of a small $P$-value in a small/large study

The $P$-value does not take all evidence into account

• Studies are analyzed in isolation, no reference to historical data
• Why not incorporating past information in current study?

Example 1.5: Merseyside registry results

• Subsequent registry study in UK
• Preliminary results of the Merseyside registry: $P = 0.67$
• Conclusion: no excess effect of chemotherapy (?)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Controls</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Chemo</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Chemo</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
1.1.3 The confidence interval as a measure of evidence

- 95% confidence interval: expression of uncertainty on parameter of interest
- Technical definition: in 95 out of 100 studies true parameter is enclosed
- In each study confidence interval includes/does not include true value
- Practical interpretation has a Bayesian nature

Example 1.6: 95% confidence interval toenail RCT

- 95% confidence interval for $\Delta = [0.14, 2.62]$
- Interpretation: most likely (with 0.95 probability) $\Delta$ lies between 0.14 and 2.62
  = a Bayesian interpretation

1.2 Statistical inference based on the likelihood function

- Inference purely on likelihood function has not been developed to a full-blown statistical approach
- Considered here as a pre-cursor to Bayesian approach
- In the likelihood approach, one conditions on the observed data

1.2.1 The likelihood function

- Likelihood was introduced by Fisher in 1922
- Likelihood function = plausibility of the observed data as a function of the parameters of the stochastic model
- Inference based on likelihood function is QUITE different from inference based on $P$-value
Example I.7: A surgery experiment

- New but rather complicated surgical technique
- Surgeon operates \( n = 12 \) patients with \( s = 9 \) successes

- Notation:
  - Result on \( i \)th operation: success \( y_i = 1 \), failure \( y_i = 0 \)
  - Total experiment: \( n \) operations with \( s \) successes
  - Sample \( \{y_1, \ldots, y_n\} \equiv y \)
  - Probability of success \( p(y_i) = \theta \)

⇒ Binomial distribution:
Expresses probability of \( s \) successes out of \( n \) experiments.

Binomial distribution

\[
f_\theta(s) = \binom{n}{s} \theta^s (1-\theta)^{n-s} \quad \text{with} \quad s = \sum_{i=1}^{n} y_i
\]

- \( \theta \) fixed & function of \( s \):
  - \( f_\theta(s) \) is discrete distribution with \( \sum_{s=0}^{n} f_\theta(s) = 1 \)

- \( s \) fixed & function of \( \theta \):
  ⇒ binomial likelihood function \( L(\theta|s) \)

Example I.7 – Determining MLE

To determine MLE first derivative of likelihood function is needed:

- \( \ell(\theta|s) = c + [s \ln \theta + (n - s) \ln(1-\theta)] \)

- \( \frac{d}{d\theta} \ell(\theta|s) = \frac{s}{\theta} - \frac{(n-s)}{(1-\theta)} = 0 \Rightarrow \hat{\theta} = s/n \)

- For \( s = 9 \) and \( n = 12 \) ⇒ \( \hat{\theta} = 0.75 \)
1.2.2 The likelihood principles

Two likelihood principles (LP):

- **LP 1**: All evidence, which is obtained from an experiment, about an unknown quantity \( \theta \), is contained in the likelihood function of \( \theta \) for the given data.
  - Standardized likelihood
  - Interval of evidence

- **LP 2**: Two likelihood functions for \( \theta \) contain the same information about \( \theta \) if they are proportional to each other.

---

**Example 1.7 (continued)**

- Maximal evidence for \( \theta = 0.75 \)

- **Likelihood ratio** \( L(0.5|s)/L(0.75|s) = \) relative evidence for 2 hypotheses \( \theta = 0.5 \) & \( \theta = 0.75 \) \( (0.21 \Rightarrow ?) \)

- Standardized likelihood: \( L_S(\theta|s) \equiv L(\theta|s)/L(\hat{\theta}|s) \)

- \( L_S(0.5|s) = 0.21 \) = test for hypothesis \( H_0 \) without involving fictive data

- **Interval of (\( \geq 1/2 \) maximal) evidence**
**Likelihood principle 2**

LP 2: Two likelihood functions for $\theta$ contain the same information about $\theta$ if they are proportional to each other.

- **LP 2 = Relative likelihood principle**

  $\Rightarrow$ When likelihood is proportional under two experimental conditions, then information about $\theta$ must be the same!

---

**Example I.8: Another surgery experiment**

- **Surgeon 1:** (Example I.7) Operates $n = 12$ patients, observes $s = 9$ successes (and 3 failures).
- **Surgeon 2:** Operates $n$ patients until $k = 3$ failures are observed ($n = s + k$). And, suppose $s = 9$

  - Surgeon 1: $s = \sum_{i=1}^{n} y_i$ has a binomial distribution
    $\Rightarrow$ binomial likelihood $L_1(\theta|s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$

  - Surgeon 2: $s = \sum_{i=1}^{n} y_i$ has a negative binomial (Pascal) distribution
    $\Rightarrow$ negative binomial likelihood $L_2(\theta|s) = \binom{s+k-1}{s}\theta^s(1-\theta)^{k}$

---

**Frequentist inference:**

- $H_0: \theta = 0.5$ & $H_A: \theta > 0.5$

  - **Surgeon 1:** Calculation $P$-value = 0.0730
    $$p[s \geq 9|\theta = 0.5] = \sum_{s=9}^{12} \binom{12}{s} 0.5^s (1-0.5)^{12-s}$$

  - **Surgeon 2:** Calculation $P$-value = 0.0337
    $$p[s \geq 9|\theta = 0.5] = \sum_{s=9}^{\infty} \binom{2+s}{s} 0.5^s (1-0.5)^{s}$$

**Frequentist conclusion $\neq$ Likelihood conclusion**

---

**Inference on the likelihood function:**

LP 2: 2 experiments give us the same information about $\theta$.
Conclusion:

Design aspects (stopping rule) are important in frequentist context

When likelihoods are proportional, it is not important in the likelihood approach how the data were obtained

1.3 The Bayesian approach: some basic ideas

- Bayesian methodology = topic of the course
- Statistical inference through different type of “glasses”

1.3.1 Introduction

- Examples I.7 and I.8: combination of information from a similar historical surgical technique could be used in the evaluation of current technique = Bayesian exercise

- Medical device trials:
  - The effect of a medical device is better understood than that of a drug
  - It is very difficult to motivate surgeons to use concurrent controls to compare the new device with the control device
  - Can we capitalize on the information of the past to evaluate the performance of the new device? Using only a single arm trial??

- Planning phase III study:
  - Comparison new ⇔ old treatment for treating breast cancer
  - Background information is incorporated when writing the protocol
  - Background information is not incorporated in the statistical analysis
  - Suppose small-scaled study with unexpectedly positive result ($P < 0.01$)
  - Reaction???
Central idea of Bayesian approach:

Combine likelihood (data) with Your prior knowledge (prior probability) to update information on the parameter to result in a revised probability associated with the parameter (posterior probability).

Example I.9: Examples of Bayesian reasoning in daily life

- **Tourist example**: Prior view on Belgians + visit to Belgium (data) ⇒ posterior view on Belgians
- **Marketing example**: Launch of new energy drink on the market
- **Medical example**: Patients treated for CVA with thrombolytic agent suffer from SBAs. Historical studies (20% - prior), pilot study (10% - data) ⇒ posterior

1.3.2 Bayes theorem – Discrete version for simple events - 1

- $A$ (diseased) & $B$ (positive diagnostic test)
  - $A^C$ (not diseased) & $B^C$ (negative diagnostic test)
  - $p(A, B) = p(A) \cdot p(B | A) = p(B) \cdot p(A | B)$

Bayes theorem = Theorem on Inverse Probability

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

Bayes theorem - version II:

$$p(B | A) = \frac{p(A | B) \cdot p(B)}{p(A | B) \cdot p(B) + p(A | B^C) \cdot p(B^C)}$$
Example I.10: Sensitivity, specificity, prevalence and predictive values

- \( B = \) "diseased", \( A = \) "positive diagnostic test"

- Characteristics of diagnostic test:
  - Sensitivity (\( Se \)) = \( p(A \mid B) \)
  - Specificity (\( Sp \)) = \( p(A^C \mid B^C) \)
  - Positive predictive value (\( pred+ \)) = \( p(B \mid A) \)
  - Negative predictive value (\( pred- \)) = \( p(B^C \mid A^C) \)
  - Prevalence (\( prev \)) = \( p(B) \)

- \( pred+ \) calculated from \( Se, Sp \) and \( prev \) using Bayes theorem

- Bayes theorem:
  \[
p(D^+ \mid T^+) = \frac{p(T^+ \mid D^+) \cdot p(D^+)}{p(T^+ \mid D^+) \cdot p(D^+) + p(T^+ \mid D^-) \cdot p(D^-)}
  \]

- In terms of \( Se, Sp \) and \( prev \):
  \[
  pred+ = \frac{Se \cdot prev}{Se \cdot prev + (1 - Sp) \cdot (1 - prev)}
  \]

- Obvious, but important: it is not possible to find the probability of having a disease based on test results without specifying the disease’s prevalence.

- Folin-Wu blood test: screening test for diabetes (Boston City Hospital)

<table>
<thead>
<tr>
<th>Test</th>
<th>Diabetic</th>
<th>Non-diabetic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>56</td>
<td>49</td>
<td>105</td>
</tr>
<tr>
<td>-</td>
<td>14</td>
<td>461</td>
<td>475</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>510</td>
<td>580</td>
</tr>
</tbody>
</table>

- \( Se = 56/70 = 0.80 \)
- \( Sp = 461/510 = 0.90 \)
- \( prev = 70/580 = 0.12 \)

- For \( p(B) = 0.03 \) ⇒ \( pred+ = 0.20 \) & \( pred- = 0.99 \)
- For \( p(B) = 0.30 \) ⇒ \( pred+ = ?? \) & \( pred- = ?? \)

- Individual prediction: combine prior knowledge (prevalence of diabetes in population) with result of Folin-Wu blood test on patient to arrive at revised opinion for patient

- Folin-Wu blood test: prior (prevalence) = 0.10 & positive test ⇒ posterior = 0.47
1.4 Outlook

- Bayes theorem will be further developed in the next chapter ⇒ such that it becomes useful in statistical practice
- Reanalyze examples such as those seen in this chapter
- Valid question: What can a Bayesian analysis do more than a classical frequentist analysis?
- Six additional chapters are needed to develop useful Bayesian tools
- But it is worth the effort!

Example 1.12: Toenail RCT – A Bayesian analysis

Re-analysis of toenail data using WinBUGS (most popular Bayesian software)

- Figure 1.4 (a): AUC on pos x-axis represents our posterior belief that $\Delta$ is positive ($= 0.98$)
- Figure 1.4 (b): AUC for the interval $[1, \infty) = our posterior belief that \mu_1/\mu_2 > 1$
- Figure 1.4 (c): incorporation of skeptical prior that $\Delta$ is positive (a priori around -0.5) ($= 0.95$)
- Figure 1.4 (d): incorporation of information on the variance parameters ($\sigma_2^2/\sigma_1^2$ varies around 2)

True stories ...

A Bayesian and a frequentist were sentenced to death. When the judge asked what their final wishes were, the Bayesian replied that he wished to teach the frequentist the ultimate lesson. The judge granted his request and then repeated the question to the frequentist. He replied that he wished to get the lesson again and again and again …

A Bayesian is walking in the fields expecting to meet horses. However, suddenly he runs into a donkey. Looks at the animal and continues his path concluding that he saw a mule.
Take home messages

- The $P$-value does not bring the message as it is perceived by many, i.e. it is not the probability that $H_0$ is true or false
- The confidence interval is generally accepted as a better tool for inference, but we interpret it in a Bayesian way
- There are other ways of statistical inference
- Pure likelihood inference = Bayesian inference without a prior
- Bayesian inference builds on likelihood inference by adding what is known of the problem

Chapter 2
Bayes theorem: computing the posterior distribution

Aims:
- Derive the general expression of Bayes theorem
- Exemplify the computations

2.1 Introduction

In this chapter:
- Bayes theorem for binary outcomes, counts and continuous outcomes case
- Derivation of posterior distribution: for binomial, normal and Poisson
- A variety of examples

2.2 Bayes theorem – The binary version

$D^+ \equiv \theta = 1$ and $D^- \equiv \theta = 0$ (diabetes)
$T^+ \equiv y = 1$ and $T^- \equiv y = 0$ (Folin-Wu test)

$p(\theta = 1 | y = 1) = \frac{p(y = 1 | \theta = 1) \cdot p(\theta = 1)}{p(y = 1 | \theta = 1) \cdot p(\theta = 1) + p(y = 1 | \theta = 0) \cdot p(\theta = 0)}$

- $p(\theta = 1), p(\theta = 0)$ prior probabilities
- $p(y = 1 | \theta = 1)$ likelihood
- $p(\theta = 1 | y = 1)$ posterior probability

$\Rightarrow$ Now parameter has also a probability
Baye’s theorem

**Shorthand notation**

\[
p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}
\]

where \( \theta \) can stand for \( \theta = 0 \) or \( \theta = 1 \).

### 2.3 Probability in a Bayesian context

**Bayesian probability** = expression of **Our/Your uncertainty** of the parameter value

- Coin tossing: truth is there, but unknown to us
- Diabetes: from population to individual patient

Probability can have two meanings: limiting proportion (objective) or personal belief (subjective)

#### Other examples of Bayesian probabilities

Subjective probability varies with individual, in time, etc.

- Tour de France
- FIFA World Cup 2014
- Global warming
- ...

#### Subjective probability rules

Let \( A_1, A_2, \ldots, A_K \) mutually exclusive events with total event \( S \)

Subjective probability \( p \) should be **coherent**:

- \( A_k: p(A_k) \geq 0 \ (k=1, \ldots, K) \)
- \( p(S) = 1 \)
- \( p(A^c) = 1 - p(A) \)
- With \( B_1, B_2, \ldots, B_L \) another set of mutually exclusive events:

\[
p(A_i \mid B_j) = \frac{p(A_i, B_j)}{p(B_j)}
\]
### 2.4 Bayes theorem – The categorical version

- Subject can belong to \( K > 2 \) classes: \( \theta_1, \theta_2, \ldots, \theta_K \)
- \( y \) takes \( L \) different values: \( y_1, \ldots, y_L \) or continuous

\[ p(\theta_k | y) = \frac{p(y | \theta_k)p(\theta_k)}{\sum_{k=1}^{K} p(y | \theta_k)p(\theta_k)} \]

### 2.5 Bayes theorem – The continuous version

- 1-dimensional continuous parameter \( \theta \)
- i.i.d. sample \( y = y_1, \ldots, y_n \)
- Joint distribution of sample = \( p(y | \theta) = \prod_{i=1}^{n} p(y_i | \theta) = \) likelihood \( L(\theta | y) \)
- Prior density function \( p(\theta) \)
- Split up: \( p(y, \theta) = p(y | \theta)p(\theta) = p(\theta | y)p(y) \)

\[ p(\theta | y) = \frac{L(\theta | y)p(\theta)}{p(y)} = \frac{L(\theta | y)p(\theta)}{\int L(\theta | y)p(\theta)d\theta} \]

### 2.6 The binomial case

- Shorter: \( p(\theta | y) \propto L(\theta | y)p(\theta) \)
- \( \int L(\theta | y)p(\theta)d\theta = \) averaged likelihood
- Averaged likelihood \( \Rightarrow \) posterior distribution involves integration
- \( \theta \) is now a random variable and is described by a probability distribution. All because we express our uncertainty on the parameter
- In the Bayesian approach (as in likelihood approach), one only looks at the observed data. No fictive data are involved
- One says: In the Bayesian approach, one conditions on the observed data. In other words, the data are fixed and \( p(y) \) is a constant!

### Example II.1: Stroke study – Monitoring safety

- \( \Delta \) Rt-PA: thrombolytic for ischemic stroke
- \( \Delta \) Historical studies ECASS 1 and ECASS 2: complication SICH
- \( \Delta \) ECASS 3 study: patients with ischemic stroke (Tx between 3 & 4.5 hours)
- \( \Delta \) DSMB: monitor SICH in ECASS 3
- Fictive situation:
  - First interim analysis ECASS 3: 50 rt-PA patients with 10 SICHs
  - Historical data ECASS 2: 100 rt-PA patients with 8 SICHs
- Estimate risk for SICH in ECASS 3 \( \Rightarrow \) construct Bayesian stopping rule
Comparison of 3 approaches

- Frequentist
- Likelihood
- Bayesian - different prior distributions
  - Prior information is available: from ECASS 2 study
  - Experts express their opinion: subjective prior
  - No prior information is available: non-informative prior
- Exemplify mechanics of calculating the posterior distribution using Bayes theorem

Notation

- SICH incidence: \( \theta \)
- i.i.d. Bernoulli random variables \( y_1, \ldots, y_n \)
- SICH: \( y_i = 1 \), otherwise \( y_i = 0 \)
- \( y = \sum_i y_i \) has \( \text{Bin}(n, \theta) \): \( p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \)

Frequentist approach

- MLE \( \hat{\theta} = y/n = 10/50 = 0.20 \)
- Test hypothesis \( \theta = 0.08 \) with binomial test (8% = value of ECASS 2 study)
- Classical 95% confidence interval = [0.089, 0.31]

Likelihood inference

- MLE \( \hat{\theta} = 0.20 \)
- No hypothesis test is performed
- 0.95 interval of evidence = [0.09, 0.36]
Bayesian approach: prior obtained from ECASS 2 study

1. Specifying the (ECASS 2) prior distribution
2. Constructing the posterior distribution
3. Characteristics of the posterior distribution
4. Equivalence of prior information and extra data

1. Specifying the (ECASS 2) prior distribution

- ECASS 2 likelihood: \( L(\theta | y_0) = \binom{n_0}{y_0} \theta^{y_0}(1 - \theta)^{n_0 - y_0} \) \((y_0 = 8 \& n_0 = 100)\)

- ECASS 2 likelihood expresses prior belief on \( \theta \) but is not (yet) prior distribution

- As a function of \( \theta \) \( L(\theta | y_0) \neq \) density (AUC \( \neq 1 \))

- How to standardize?
  Numerically or analytically?

Some beta densities

- Kernel of binomial likelihood \( \theta^{y_0}(1 - \theta)^{n_0 - y_0} \propto \) beta density Beta\((\alpha, \beta)\):

  \[
p(\theta) = \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0 - 1}(1 - \theta)^{\beta_0 - 1}
\]

with \( B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \) gamma function

- \( \alpha_0(= 9) \equiv y_0 + 1 \)
- \( \beta_0(= 100 - 8 + 1) \equiv n_0 - y_0 + 1 \)
2. Constructing the posterior distribution

- Bayes theorem needs:
  - Prior $p(\theta)$ (ECASS 2 study)
  - Likelihood $L(\theta|y)$ (ECASS 3 interim analysis), $y=10$ & $n=50$
  - Averaged likelihood $\int L(\theta|y)p(\theta)d\theta$

- Numerator of Bayes theorem
  \[ L(\theta|y)p(\theta) = \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \theta^{y-1}(1-\theta)^{\beta_0+n-y-1} \]

- Denominator of Bayes theorem = averaged likelihood
  \[ p(y) = \int L(\theta|y)p(\theta)d\theta = \binom{n}{y} \frac{B(\alpha_0+y, \beta_0+n-y)}{B(\alpha_0, \beta_0)} \]

\[ \Rightarrow \text{Posterior distribution} = \text{Beta}(\overline{\alpha}, \overline{\beta}) \]

\[ p(\theta|y) = \frac{1}{B(\overline{\alpha}, \overline{\beta})} \theta^{\overline{\alpha}-1}(1-\theta)^{\overline{\beta}-1} \]

with
\[ \overline{\alpha} = \alpha_0 + y \]
\[ \overline{\beta} = \beta_0 + n - y \]

3. Characteristics of the posterior distribution

- Posterior = compromise between prior & likelihood
- Posterior mode: $\hat{\theta}_M = \frac{n_0}{n_0+y} \theta_0 + \frac{n}{n_0+y} \hat{\theta}$
- Shrinkage: $\theta_0 \leq \hat{\theta}_M \leq \hat{\theta}$ when $(y_0/n_0 \leq y/n)$
- Posterior more peaked than prior & likelihood, but not in general
- Posterior = beta distribution = prior (conjugacy)
- Likelihood dominates the prior for large sample sizes
- **NOTE:** Posterior estimate $\hat{\theta} = \text{MLE of combined ECASS 2 data & interim data ECASS 3 (!!!)}$, i.e.
  \[ \hat{\theta}_M = \frac{y_0+y}{n_0+n} \]
4. Equivalence of prior information and extra data

- **Beta**($\alpha_0$, $\beta_0$) prior
  \[ \equiv \text{binomial experiment} \] with ($\alpha_0 - 1$) successes in ($\alpha_0 + \beta_0 - 2$) experiments

  \[ \Rightarrow \text{Prior} \approx \text{extra data to observed data set: ($\alpha_0 - 1$) successes and ($\beta_0 - 1$) failures} \]

Bayesian approach: using a subjective prior

- Suppose DSMB neurologists ‘believe’ that SICH incidence is probably more than 5% but most likely not more than 20%

- If prior belief = ECASS 2 prior density \( \Rightarrow \) posterior inference is the same

- The neurologists could also combine their qualitative prior belief with ECASS 2 data to construct a prior distribution \( \Rightarrow \) adjust ECASS 2 prior

Example subjective prior
Bayesian approach: no prior information is available

- Suppose no prior information is available
- Need: a prior distribution that expresses ignorance = noninformative (NI) prior
- For stroke study: NI prior = \( p(\theta) = 1_{[0,1]} = \text{flat prior on } [0,1] \)
- Uniform prior on [0,1] = Beta(1,1)

2.7 The Gaussian case

Example II.2: Dietary study – Monitoring dietary behavior in Belgium

- IBBENS study: dietary survey in Belgium
- Of interest: intake of cholesterol
- Monitoring dietary behavior in Belgium: IBBENS-2 study

Assume \( \sigma \) is known

1. Specifying the (IBBENS) prior distribution

Histogram of the dietary cholesterol of 563 bank employees \( \approx \) normal

\( y \sim N(\mu, \sigma^2) \) when

\[
\begin{align*}
    f(y) &= \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{(y - \mu)^2}{2\sigma^2}\right] \\
    L(\mu|y) &\propto \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right] \\
    &\propto \exp \left[-\frac{1}{2} \left(\frac{\mu - \bar{y}}{\sigma/\sqrt{n}}\right)^2\right] \equiv L(\mu|\bar{y})
\end{align*}
\]

Bayesian approach: prior obtained from the IBBENS study

1. Specifying the (IBBENS) prior distribution
2. Constructing the posterior distribution
3. Characteristics of the posterior distribution
4. Equivalence of prior information and extra data
Histogram and likelihood IBBENS study:

- Denote sample $n_0$ IBBENS data: $y_0 \equiv \{y_{0,1}, \ldots, y_{0,n_0}\}$ with mean $\overline{y}_0$
- Likelihood $\propto N(\mu_0, \sigma^2_0)$
  \[ \mu_0 \equiv \overline{y}_0 = 328 \]
  \[ \sigma_0 = \sigma/\sqrt{n_0} = 120.3/\sqrt{563} = 5.072 \]
- IBBENS prior distribution ($\sigma$ is known)
  \[ p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp \left\{ -\frac{1}{2} \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \right\} \]
  with $\mu_0 \equiv \overline{y}_0$

2. Constructing the posterior distribution

- **IBBENS-2 study:**
  - sample $y$ with $n=50$
  - $\overline{y} = 318$ mg/day & $s = 119.5$ mg/day
  - 95% confidence interval = $[284.3, 351.9]$ mg/day ⇒ wide
- Combine IBBENS prior distribution IBBENS-2 normal likelihood:
  - IBBENS-2 likelihood: $L(\mu | \overline{y})$
  - IBBENS prior density: $p(\mu) = N(\mu | \mu_0, \sigma^2_0)$
- Posterior distribution $\propto p(\mu)L(\mu|\overline{y})$:
  \[ p(\mu|y) \equiv p(\mu|\overline{y}) \propto \exp \left\{ -\frac{1}{2} \left[ \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 + \left( \frac{\mu - \overline{y}}{\sigma/\sqrt{n}} \right)^2 \right] \right\} \]

- **Integration constant** to obtain density?
- **Recognize** standard distribution: exponent (quadratic function of $\mu$)
- **Posterior distribution:**
  \[ p(\mu|y) = N(\overline{\mu}, \overline{\sigma}^2), \]
  with
  \[ \overline{\mu} = \frac{1}{\sigma^2_0 + \frac{n}{\sigma^2}} \mu_0 + \frac{n}{\sigma^2} \overline{y} \quad \text{and} \quad \overline{\sigma}^2 = \frac{1}{\sigma^2_0 + \frac{n}{\sigma^2}} \]
- Here: $\overline{\mu} = 327.2$ and $\overline{\sigma} = 4.79$. 
3. Characteristics of the posterior distribution

- Posterior distribution: compromise between prior and likelihood
- Posterior mean: weighted average of prior and the sample mean
  \[ \bar{\mu} = \frac{w_0}{w_0 + w_1} \mu_0 + \frac{w_1}{w_0 + w_1} \bar{y} \]
  with
  \[ w_0 = \frac{1}{\sigma^2_0} \quad \text{and} \quad w_1 = \frac{1}{\sigma^2/n} \]
- The posterior precision = 1/posterior variance:
  \[ \frac{1}{\sigma^2} = w_0 + w_1 \]
  with \( w_0 = 1/\sigma^2_0 = \text{prior precision} \) and \( w_1 = 1/(\sigma^2/n) = \text{sample precision} \)

4. Equivalence of prior information and extra data

- Posterior is always more peaked than prior and likelihood
- When \( n \to \infty \) or \( \sigma_0 \to \infty \): \( p(\mu|y) = \mathcal{N}(\bar{y}, \sigma^2/n) \)
  ⇒ When sample size increases the likelihood dominates the prior
- Posterior = normal = prior ⇒ conjugacy

\[ \sigma^2_0 = \sigma^2 \quad \text{(unit information prior)} \implies \sigma^2 = \sigma^2 / (n + 1) \]
  ⇒ Prior information = adding one extra observation to the sample
- General: \( \sigma^2_0 = \sigma^2/n_0 \), with \( n_0 \) general
  \[ \bar{\mu} = \frac{n_0}{n_0 + n} \mu_0 + \frac{n}{n_0 + n} \bar{y} \]
  and
  \[ \sigma^2 = \frac{\sigma^2}{n_0 + n} \]
Bayesian approach: using a subjective prior

- Discounted IBBENS prior: increase IBBENS prior variance from 25 to 100
- Discounted IBBENS prior + shift: increase from $\mu_0 = 328$ to $\mu_0 = 340$

Discounted priors:

\[
\begin{align*}
\text{(a) Discounted prior} & \\
\text{(b) Discounted prior + shift}
\end{align*}
\]

Bayesian approach: no prior information is available

- Non-informative prior: $\sigma_0^2 \to \infty$
  \[\Rightarrow \text{Posterior: } N(\bar{y}, \sigma^2/n)\]
2.8 The Poisson case

- Take $y \equiv \{y_1, \ldots, y_n\}$ independent counts $\Rightarrow$ Poisson distribution

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

- Mean and variance $= \theta$

- Poisson likelihood:
  
  \[ L(\theta|y) \equiv \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \left( \frac{\theta^{y_i}}{y_i!} \right) e^{-\theta} \]

Example II.6: Describing caries experience in Flanders

The Signal-Tandmobiel\textsuperscript{®} (STM) study:

- Longitudinal oral health study in Flanders
- Annual examinations from 1996 to 2001
- 4468 children (7% of children born in 1989)
- Caries experience measured by dmft-index ($\min=0$, $\max=20$)

Frequentist and likelihood calculations

- MLE of $\theta$: $\hat{\theta} = \bar{y} = 2.24$
- Likelihood-based 95% confidence interval for $\theta$: [2.1984, 2.2875]

Bayesian approach: prior distribution based on historical data

1. Specifying the prior distribution
2. Constructing the posterior distribution
3. Characteristics of the posterior distribution
4. Equivalence of prior information and extra data
1. Specifying the prior distribution

- Information from literature:
  - Average dmft-index 4.1 (Liège, 1983) & 1.39 (Gent, 1994)
  - Oral hygiene has improved considerably in Flanders
  - Average dmft-index bounded above by 10

- Candidate for prior: \( \text{Gamma}(\alpha_0, \beta_0) \)
  \[
p(\theta) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\beta_0 \theta}
\]
  - \( \alpha_0 \) = shape parameter & \( \beta_0 \) = inverse of scale parameter
  - \( E(\theta) = \frac{\alpha_0}{\beta_0} \) & \( \text{var}(\theta) = \frac{\alpha_0}{\beta_0^2} \)

- STM study: \( \alpha_0 = 3 \) & \( \beta_0 = 1 \)

2. Constructing the posterior distribution

- Posterior
  \[
p(\theta | y) \propto e^{-n \theta} \prod_{i=1}^{n} (\theta y_i)^{\alpha_0} \beta_0^{\alpha_0-1} e^{-\beta_0 y_i} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\beta_0 \theta}
  \]
  \[
  \propto \theta^{\sum y_i + \alpha_0 - 1} e^{-(n+\beta_0)\theta}
  \]

- Recognize kernel of a \( \text{Gamma}(\sum y_i + \alpha_0, n + \beta_0) \) distribution
  \[
  \Rightarrow p(\theta | y) \equiv p(\theta | \bar{y}) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\beta_0 \theta}
  \]
  with \( \alpha = \sum y_i + \alpha_0 = 9758 + 3 = 9761 \) and \( \beta = n + \beta_0 = 4351 + 1 = 4352 \)
  \( \Rightarrow \) STM study: effect of prior is minimal
3. Characteristics of the posterior distribution

- Posterior is a compromise between prior and likelihood
- Posterior mode demonstrates shrinkage
- For STM study posterior more peaked than prior likelihood, but not in general
- Prior is dominated by likelihood for a large sample size
- Posterior = gamma = prior $\Rightarrow$ conjugacy

4. Equivalence of prior information and extra data

- Prior = equivalent to experiment of size $\beta_0$ with counts summing up to $\alpha_0 - 1$
- STM study: prior corresponds to an experiment of size 1 with count equal to 2

Bayesian approach: no prior information is available

- Gamma with $\alpha_0 \approx 1$ and $\beta_0 \approx 0$ = non-informative prior

2.9 The prior and posterior of derived parameter

- If $p(\theta)$ is prior of $\theta$, what is then corresponding prior for $h(\theta) = \psi$?
  - Same question for posterior density
  - Example: $\theta = \text{odds ratio}$, $\psi = \log(\text{odds ratio})$

- Why do we wish to know this?
  - Prior: prior information on $\theta$ and $\psi$ should be the same
  - Posterior: allows to reformulate conclusion on a different scale

- Solution: apply transformation rule $p(h^{-1}(\psi)) \left| \frac{dh^{-1}(\psi)}{d\psi} \right|$

- Note: parameter is a random variable!
Example II.4: Stroke study – Posterior distribution of \( \log(\theta) \)

- Probability of ‘success' is often modeled on the log-scale (or logit scale)
- Posterior distribution of \( \psi = \log(\theta) \)

\[
p(\psi|y) = \frac{1}{B(\bar{\alpha}, \bar{\beta})} \exp \psi (1 - \exp \psi)^{\bar{\beta} - 1}
\]

with \( \bar{\alpha} = 19 \) and \( \bar{\beta} = 133 \).

2.10 Bayesian versus likelihood approach

- Bayesian approach satisfies 1st likelihood principle in that inference does not depend on never observed results
- Bayesian approach satisfies 2nd likelihood principle:

\[
p_2(\theta|y) = \frac{L_2(\theta|y)p(\theta) \int L_2(\theta|y)p(\theta)d\theta}{\int cL_1(\theta|y)p(\theta)d\theta}
\]

\[
= \frac{cL_1(\theta|y)p(\theta) \int cL_1(\theta|y)p(\theta)d\theta}{\int cL_1(\theta|y)p(\theta)d\theta}
\]

\[
= p_1(\theta|y)
\]

- In Bayesian approach parameter is stochastic
  \( \Rightarrow \) different effect of transformation \( h(\theta) \) in Bayesian and likelihood approach

2.11 Bayesian versus frequentist approach

- **Frequentist approach:**
  - \( \theta \) fixed and data are stochastic
  - Many tests are based on asymptotic arguments
  - Maximization is key tool
  - Does depend on stopping rules

- **Bayesian approach:**
  - Condition on observed data (\textbf{data fixed}), uncertainty about \( \theta \) (\( \theta \) \textbf{stochastic})
  - No asymptotic arguments are needed, all inference depends on posterior
  - Integration is key tool
  - Does not depend on stopping rules
2.12 The different modes of the Bayesian approach

- Subjectivity ↔ objectivity
- Subjective (proper) Bayesian ↔ objective (reference) Bayesian
- Empirical Bayesian
- Decision-theoretic (full) Bayesian: use of utility function
- 46656 varieties of Bayesians (De Groot)
- Pragmatic Bayesian = Bayesian ??

2.13 An historical note on the Bayesian approach

- **Thomas Bayes** was probably born in 1701 and died in 1761
- He was a Presbyterian minister, studied logic and theology at Edinburgh University, and had strong mathematical interests
- Bayes theorem was submitted posthumously by his friend Richard Price in 1763 and was entitled An Essay toward a Problem in the Doctrine of Chances
- Up to 1950 Bayes theorem was called Theorem of Inverse Probability
- Fundament of Bayesian theory was developed by **Pierre-Simon Laplace** (1749-827)
- Laplace first assumed indifference prior, later he relaxed this assumption
- Much opposition: e.g. Poisson, Fisher, Neyman and Pearson, etc
Fisher strong opponent to Bayesian theory
Because of his influence
⇒ dramatic negative effect
Opposed to use of flat prior and
that conclusions change when putting flat prior
on $h(\theta)$ rather than on $\theta$
Some connection between
Fisher and (inductive) Bayesian approach,
but much difference with N&P approach

Proponents of the Bayesian approach:
- de Finetti: exchangeability
- Jeffreys: noninformative prior, Bayes factor
- Savage: theory of subjective and personal probability and statistics
- Lindley: Gaussian hierarchical models
- Geman & Geman: Gibbs sampling
- Gelfand & Smith: introduction of Gibbs sampling into statistics
- Spiegelhalter: (Win)BUGS

Recommendation

The theory that would not die. How Bayes rule cracked the enigma
code, hunted down Russian submarines & emerged triumphant from
two centuries of controversy

Mc Grayne (2011)

Take home messages

- Bayes theorem = model for learning
- Probability in a Bayesian context:
  - data: classical, parameters: expressing what we belief/know
- Bayesian approach = likelihood approach + prior
- Inference:
  - Bayesian:
    - based on parameter space (posterior distribution)
    - conditions on observed data, parameter stochastic
  - Classical:
    - based on sample space (set of possible outcomes)
    - looks at all possible outcomes, parameter fixed
• Prior can come from: historical data or subjective belief
• Prior is equivalent to extra data
• Noninformative prior can mimic classical results
• Posterior = compromise between prior and likelihood
• For large sample, likelihood dominates prior
• Bayesian approach was obstructed by many throughout history
• ... but survived because of a computational trick ... (MCMC)

Chapter 3
Introduction to Bayesian inference

Aims:
▷ Introduction to basic concepts in Bayesian inference
▷ Introduction to simple sampling algorithms
▷ Illustrating that sampling can be useful alternative to analytical/other numerical techniques to determine the posterior
▷ Illustrating that Bayesian testing can be quite different from frequentist testing

3.1 Introduction

More specifically we look at:
• Exploration of the posterior distribution:
  ▷ Summary statistics for location and variability
  ▷ Interval estimation
  ▷ Predictive distribution
• Normal approximation of posterior
• Simple sampling procedures
• Bayesian hypothesis tests

3.2 Summarizing the posterior with probabilities

Direct exploration of the posterior: \( P(a < \theta < b|y) \) for different \( a \) and \( b \)

Example III.1: Stroke study – SICH incidence
• \( \theta \) = probability of SICH due to rt-PA at first ECASS-3 interim analysis
  \( p(\theta|y) = \text{Beta}(19, 133)-distribution \)
• \( P(a < \theta < b|y) \):
  ▷ \( a = 0.2, b = 1.0 \): \( P(0.2 < \theta < 1|y) = 0.0062 \)
  ▷ \( a = 0.0, b = 0.08 \): \( P(0 < \theta < 0.08|y) = 0.033 \)
3.3 Posterior summary measures

- We now summarize the posterior distribution with some simple measures, similar to what is done when summarizing collected data.
- The measures are computed on a (population) distribution.

3.3.1 Posterior mode, mean, median, variance and SD

- Posterior mode: \( \hat{\theta}_M \) where posterior distribution is maximum
- Posterior mean: mean of posterior distribution, i.e. \( \bar{\theta} = \int \theta p(\theta|y) d\theta \)
- Posterior median: median of posterior distribution, i.e. \( 0.5 = \int_{\hat{\theta}_M} p(\theta|y) d\theta \)
- Posterior variance: variance of posterior distribution, i.e. \( \sigma^2 = \int (\theta - \bar{\theta})^2 p(\theta|y) d\theta \)
- Posterior standard deviation: sqrt of posterior variance, i.e. \( \sigma \)

Note:
- Only posterior median of \( h(\theta) \) can be obtained from posterior median of \( \theta \)
- Only posterior mode does not require integration
- Posterior mode = MLE with flat prior

Example III.2: Stroke study – Posterior summary measures

- Posterior at 1st interim ECASS 3 analysis: \( \text{Beta}(\alpha, \beta) \) \( (\alpha = 19 \& \beta = 152) \)
- Posterior mode: maximize \( (\alpha - 1) \ln(\theta) + (\beta - 1) \ln(1 - \theta) \) wrt \( \theta \)
  \( \Rightarrow \hat{\theta}_M = (\alpha - 1)/((\alpha + \beta - 2)) = 18/150 = 0.12 \)
- Posterior mean: integrate \( \frac{1}{B(\alpha, \beta)} \int_0^1 \theta \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta \)
  \( \Rightarrow \bar{\theta} = B(\alpha + 1, \beta)/B(\alpha, \beta) = \alpha/((\alpha + \beta)) = 19/152 = 0.125 \)
- Posterior median: solve \( 0.5 = \frac{1}{B(\alpha, \beta)} \int_{\hat{\theta}_M}^1 \theta \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta \) for \( \hat{\theta}_M = 0.122 \) (R-function qbeta)
- Posterior variance: calculate also \( \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^2 \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta \)
  \( \Rightarrow \sigma^2 = \frac{\alpha \beta}{((\alpha + \beta)(\alpha + \beta + 1))} = 0.0267^2 \)

Graphical representation measures:
Posterior summary measures:

- Posterior for $\mu$ (based on IBBENS prior):
  
  Gaussian with $\mu_M \equiv \mu \equiv \mu_M$

- Posterior mode=mean=median: $\hat{\mu}_M = 327.2$ mg/dl

- Posterior variance & SD: $\sigma^2 = 22.99$ & $\sigma = 4.79$ mg/dl

3.3.2 Credible/credibility interval

- $[a, b] = 95\%$ credible interval (CI) for $\theta$ if $\Pr(a \leq \theta \leq b | y) = 0.95$

- Two types of 95% credible interval:
  
  - $95\%$ equal tail CI $[a, b]$: $AUC = 0.025$ is left to $a$ & $AUC = 0.025$ is right to $b$
  
  - $95\%$ highest posterior density (HPD) CI $[a, b]$: 
    $[a, b]$ contains most plausible values of $\theta$

- Properties:
  
  - $100(1-\alpha)\%$ HPD CI = shortest interval with size $(1-\alpha)$ (Press, 2003)
  
  - $h$(HPD CI) $\neq$ HPD CI, but $h$(equal-tail CI) = equal-tail CI
  
  - Symmetric posterior: equal tail = HPD CI

Example III.4: Dietary study – Interval estimation of dietary intake

- Posterior $= N(\mu, \sigma^2)$

- Obvious choice for a 95% CI is $[\mu - 1.96\sigma, \mu + 1.96\sigma]$

- Equal 95% tail CI = 95% HPD interval

- Results IBBENS-2 study:
  
  - IBBENS prior distribution $\Rightarrow$ 95% CI $= [317.8, 336.6]$ mg/dl
  
  - $N(328; 10,000)$ prior $\Rightarrow$ 95% CI $= [285.6, 351.0]$ mg/dl
  
  - Classical (frequentist) 95% confidence interval $= [284.9, 351.1]$ mg/dl

Example III.5: Stroke study – Interval estimation of probability of SICH

- Posterior $= Beta(19, 133)$-distribution

- 95% equal tail CI ($R$ function $qbeta$) = $[0.077, 0.18]$ (see figure)

- 95% HPD interval $= [0.075, 0.18]$ (see figure)

- Computations HPD interval: use $R$-function optimize
3.4 Predictive distributions

3.4.1 Introduction

- **Predictive distribution** = distribution of a future observation $\tilde{y}$ after having observed the sample $\{y_1, \ldots, y_n\}$

- Two assumptions:
  - Future observations are independent of current observations given $\theta$
  - Future observations have the same distribution ($p(y | \theta)$) as the current observations

- We look at three cases: (a) binomial, (b) Gaussian and (c) Poisson

- We start with a binomial example
Example III.7: Stroke study – Predicting SICH incidence in interim analysis

- Before 1st interim analysis but given the pilot data:
  Obtain an idea of the number of (future) rt-PA treated patients who will suffer from SICH in sample of size \( m = 50 \)
- Distribution of \( \tilde{y} \) (given the pilot data)?

Example III.7: Known incidence rate

- MLE of \( \theta \) (incidence SICH) = \( \frac{8}{100} = 0.08 \) for (fictive) ECASS 2 study
  - Assume now that 8\% is the true incidence
  - Predictive distribution: \( \text{Bin}(50, 0.08) \)
  - 95\% predictive set: \( \{0, 1, \ldots, 7\} \approx 94\% \) of the future counts
    \[ \Rightarrow \text{observed result of 10 SICH patients out of 50 is extreme} \]
- But, 8\% is likely not the true value

Example III.7: Unknown incidence rate

- Uncertainty of \( \theta \) expressed by posterior (ECASS 2 prior) = Beta(\( \alpha, \beta \))-distribution with \( \alpha = 9 \) and \( \beta = 93 \)
- Distribution (likelihood) of \( m = 50 \) future SICH events is given by \( \text{Bin}(m, \theta) \)
- Distribution of \( m = 50 \) future SICH events without knowing \( \theta \) is weighted by posterior distribution = posterior predictive distribution (PPD)
- PPD = Beta-binomial distribution \( \text{BB}(m, \alpha, \beta) \)
  \[
p(y|y) = \frac{1}{B(\tilde{y} + \alpha, m - \tilde{y} + \beta)} \int_0^1 \left( \frac{m}{\tilde{y}} \right) \theta^{\tilde{y}} (1 - \theta)^{m - \tilde{y}} B(\tilde{y} + \alpha, m - \tilde{y} + \beta) \, d\theta \\
  = \left( \frac{m}{\tilde{y}} \right) B(\tilde{y} + \alpha, m - \tilde{y} + \beta) B(\alpha, \beta)
  \]
  Ignores uncertainty in \( \theta \)
Result:

- $BB(m, \pi, \beta)$ shows more variability than $Bin(m, \theta)$
- 94.4% PPS is $\{0, 1, \ldots, 9\}$ ⇒ 10 SICH patients out of 50 less extreme

3.4.2 Posterior predictive distribution: General case

- Central idea: Take the posterior uncertainty into account
- Three cases:
  - All mass (AUC $\approx 1$) of $p(\theta | y)$ at $\hat{\theta}_M$ ⇒ distribution of $\tilde{y}$: $p(\tilde{y} | \hat{\theta}_M)$
  - All mass at $\theta_1, \ldots, \theta^K$ ⇒ distribution of $\tilde{y}$: $\sum_{k=1}^{K} p(\tilde{y} | \theta^k)p(\theta^k | y)$
  - General case: posterior predictive distribution (PPD)

  $\Rightarrow$ distribution of $\tilde{y}$

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

- PPD expresses what we know about the distribution of the (future) $y$s

Example III.7: Second interim analysis

- Posterior after 1st interim analysis can be used:
  - To compute PPD of the number SICH events in subsequent $m$ treated rt-PA patients, to be evaluated in 2nd interim analysis
  - As prior to be combined with the results in the 2nd interim analysis
  - This is an example of the sequential use of Bayes theorem
Example III.6: SAP study – 95% normal range

- Serum alkaline phosphatase (alp) was measured on a prospective set of 250 'healthy' patients by Topal et al (2003)
- Purpose: determine 95% normal range
- Recall: $\sigma^2$ is known

Example III.6: Frequentist approach

- $y_i = 100/\sqrt{\text{alp}_i}$ $(i = 1, \ldots, 250) \approx$ normal distribution $N(\mu, \sigma^2)$
- $\mu = \text{known}$
  - 95% normal range for $y$: $[\mu - 1.96 \sigma, \mu + 1.96 \sigma]$, with $\mu = \bar{y} = 7.11$ ($\sigma = 1.4$)
  - $\Rightarrow$ 95% normal range for alp = [104.45, 508.95]
- $\mu = \text{unknown}$
  - 95% normal range for $y$: $[\bar{y} - 1.96 \sigma \sqrt{1 + 1/n}, \bar{y} + 1.96 \sigma \sqrt{1 + 1/n}]$
  - $\Rightarrow$ 95% normal range for alp = [104.33, 510.18]

Example III.6: Bayesian approach

- PPD: $\tilde{y} | y \sim N(\bar{y}, \sigma^2 + \sigma^2)$
- 95% normal range for $y$: $[\bar{y} - 1.96 \sqrt{\sigma^2 + \sigma^2}, \bar{y} + 1.96 \sqrt{\sigma^2 + \sigma^2}]$
- Prior variance $\sigma_0^2$ large $\Rightarrow$ $\sigma^2 = \sigma^2/n$
  - $\Rightarrow$ Bayesian 95% normal range = frequentist 95% normal range
  - $\Rightarrow$ 95% normal range for alp = [104.33, 510.18]
- Same numerical results BUT the way to deal with uncertainty of the true value of the parameter is different
Example III.8: Caries study – PPD for caries experience

• Poisson likelihood + gamma prior ⇒ gamma posterior

• PPD = negative binomial distribution $NB(\alpha, \beta)$

$$p(y|\alpha, \beta) = \frac{\Gamma(\alpha + \tilde{y})}{\Gamma(\alpha) \tilde{y}!} \left( \frac{\beta}{\beta + 1} \right)^\alpha \left( \frac{1}{\beta + 1} \right)^{\tilde{y}}$$

More applications of PPDs

See later:

• Determining when to stop a trial (see medical literature further)

• Given the past responses, predict the future responses of a patient

• Imputing missing data

• Model checking (Chapter 10)

3.5 Exchangeability

• Exchangeable:
  - Order in \{y_1, y_2, \ldots, y_n\} is not important
  - $y_i$ can be exchanged for $y_j$ without changing the problem/solution
  - Extension of independence

• Exchangeability = key idea in (Bayesian) statistical inference

• Exchangeability of patients, trials, \ldots
3.6 A normal approximation to the posterior

- When the sample is large and the statistical model is relatively simple (not many parameters), then often the posterior looks like a normal distribution.
- This property is typically used in the analysis of clinical trials.
- In the book, we show that the normal approximation can work well for small sample sizes (Example III.10).
- This is a nice property, but not crucial in Bayesian world, since all we need is the posterior.

3.7 Numerical techniques to determine the posterior

- To compute the posterior, the product of likelihood with the prior needs to be divided with the integral of that product.
- In addition, to compute posterior mean, median, SD we also need to compute integrals.
- Here we see 2 possible ways to obtain in general the posterior and posterior summary measures:
  - Numerical integration
  - Sampling from the posterior

3.7.1 Numerical integration

Example III.11: Caries study – Posterior distribution for a lognormal prior

- Take first 10 children from STM study.
- Assume dmft-index has Poisson(θ) with θ = mean dmft-index.
- With prior: Gamma(3, 1) ⇒ Posterior: Gamma(29, 11).
- Replace gamma prior by lognormal prior, then posterior
  \[ \propto \theta^{\sum_{i=1}^{n} y_i - 1} e^{-n\theta - \left(\frac{\ln(\theta) - \mu_0}{\sigma_0}\right)^2}, (\theta > 0) \]
- Posterior moments cannot be evaluated & AUC not known.
- Mid-point approach provides AUC.

Calculation AUC using mid-point approach:
3.7.2 Sampling from the posterior distribution

- Monte Carlo integration: usefulness of sampling idea
- General purpose sampling algorithms

Monte-Carlo integration

- Monte Carlo integration: replace integral by a Monte Carlo sample \( \{\tilde{\theta}_1, \ldots, \tilde{\theta}_K\} \)
- Approximate \( p(\theta|y) \) by sample histogram
- Approximate posterior mean by:

\[
\int \theta p(\theta|y) \, d\theta \approx \frac{1}{K} \sum_{k=1}^{K} \tilde{\theta}_k, \quad \text{for } K \text{ large}
\]

- Classical 95% confidence interval to indicate precision of posterior mean

\[
\left[ \hat{\theta} - 1.96 \frac{s_\tilde{\theta}}{\sqrt{K}}, \hat{\theta} + 1.96 \frac{s_\tilde{\theta}}{\sqrt{K}} \right]
\]

with \( s_\tilde{\theta}/\sqrt{K} = \) Monte Carlo error

\( \triangleright \) Also 95% credible intervals can be computed (approximated)

Example III.12: Stroke study – Sampling from the posterior distribution

- Posterior for \( \theta = \) probability of SICH with rt-PA = Beta(19, 133) (Example II.1)
- 5,000 sampled values of \( \theta \) from Beta(19, 133)-distribution
- Posterior of \( \log(\theta) \): one extra line in \( R \)-program
- Sample summary measures \( \approx \) true summary measures
- 95% equal tail CI for \( \theta \): [0.0782, 0.182]
- 95% equal tail CI for \( \log(\theta) \): [-2.56, -1.70]
- Approximate 95% HPD interval for \( \theta \): [0.0741, 0.179]

Sampling approximation:
Monte Carlo error ($K=50$)

- General purpose sampling algorithms
  - Many algorithms are available to sample from standard distributions
  - Dedicated procedures/general purpose algorithms are needed for non-standard distributions
  - An important example: Accept-reject (AR) algorithm used by e.g. WinBUGS
    - Generate samples from an instrumental distribution
    - Then reject certain generated values to obtain sample from posterior

- Black box = true posterior mean & red box = sampled posterior mean

Accept-reject algorithm – 1

- Sampling in two steps:
  - Stage 1: sample from $q(\theta)$ (proposal distribution) ⇒ $\tilde{\theta}$
  - Stage 2: reject $\tilde{\theta}$ ⇒ sample from $p(\theta | y)$ (target)
- Assumption: $p(\theta | y) < A q(\theta)$ for all $\theta$
  - $q =$ envelope distribution
  - $A =$ envelope constant

Accept-reject algorithm – 2

- Stage 1: $\tilde{\theta}$ & $u$ are drawn independently from $q(\theta)$ & $U(0, 1)$
- Stage 2:
  - Accept: when $u \leq p(\tilde{\theta} | y)/A q(\tilde{\theta})$
  - Reject: when $u > p(\tilde{\theta} | y)/A q(\tilde{\theta})$
- Properties AR algorithm:
  - Produces a sample from the posterior
  - Only needs $p(y | \theta) p(\theta)$
  - Probability of acceptance = $1/A$
Example III.13: Caries study – Sampling from posterior with lognormal prior

Accept-reject algorithm

- Lognormal prior is maximized for \( \log(\theta) = \mu_0 \)

\[ A_q(\theta) = \theta^{\sum_i h_i} e^{-n\theta} \propto \text{Gamma}(\sum_i y_i, n) \text{-distribution} \]

- Data from Example III.11

- Prior: lognormal distribution with \( \mu_0 = \log(2) \) & \( \sigma_0 = 0.5 \)

- 1000 \( \theta \)-values sampled: 840 accepted
  - Sampled posterior mean (median) = 2.50 (2.48)
  - Posterior variance = 0.21
  - 95% equal tail CI = [1.66, 3.44]

Adaptive Rejection Sampling algorithms – 1

Adaptive Rejection Sampling (ARS) algorithm:

- Builds envelope function/density in an adaptive manner

- Builds squeezing function/density in an adaptive manner

- Envelope and squeezing density are log of piecewise linear functions with knots at sampled grid points

- Two special cases:
  - Tangent method of ARS
  - Derivative-free method of ARS

Adaptive Rejection Sampling algorithms – 2
### Adaptive Rejection Sampling algorithms – 3

Properties ARS algorithms:
- Envelope function can be made arbitrarily close to target.
- 5 to 10 grid points determine envelope density.
- Squeezing density avoids (many) function evaluations.
- Derivative-free ARS is implemented in WinBUGS.

**ARMS algorithm**: combination with Metropolis algorithm for non log-concave distributions, implemented in Bayesian SAS procedures.

### 3.7.3 Choice of posterior summary measures

- **In practice**: posterior summary measures are computed with sampling techniques.
- **Choice driven by available software**:
  - Mean, median and SD because provided by WinBUGS.
  - Mode almost never reported, but useful to compare with frequentist solution.
  - Equal tail CI (WinBUGS) & HPD (CODA and BOA).

### 3.8 Bayesian hypothesis testing

Two Bayesian tools for hypothesis testing $H_0: \theta = \theta_0$.

- **Based on credible interval**:
  - Is $\theta_0$ contained in $100(1-\alpha)$% CI? If not, then reject $H_0$ in a Bayesian way!
  - Popular when many parameters need to be evaluated, e.g., in regression models.
  - **Contour probability $p_B$**: $P[p(\theta \mid y) > p(\theta_0 \mid y)] \equiv (1 - p_B)$.

- **Bayes factor**: change of prior odds for $H_0$ due to data (below).

### Example III.14: Cross-over study – Use of CIs in Bayesian hypothesis testing

- **30** patients with systolic hypertension in cross-over study:
  - **Period 1**: randomization to $A$ or $B$.
  - **Washout period**.
  - **Period 2**: switch medication ($A$ to $B$ and $B$ to $A$).

  - $\theta = \mathbb{P}(A$ better than $B) \& H_0 : \theta = \theta_0(= 0.5)$.

- **Result**: **21** patients better of with $A$ than with $B$.

- **Testing**:
  - **Frequentist**: 2-sided binomial test ($P = 0.043$).
  - **Bayesian**: $U(0,1)$ prior + $L_{Bin}(21 \mid 30, \theta) = \text{Beta}(22, 10)$ posterior ($p_B = 0.023$).
Predictive distribution of power and necessary sample size

- Classical power and sample size calculations make use of formulas to compute power for a given sample size, and necessary sample size for a chosen power
- However, there is often much uncertainty involved in the calculations because:
  - Comparison of means: only a rough idea of the common SD is available to us, and there is no agreement on the aimed difference in efficacy
  - Comparison of proportions: only a rough idea of the control rate is available to us, and there is no agreement on the aimed difference in efficacy
- Typically one tries a number of possible values of the required parameter values and inspects the variation in the calculation power/sample size
- Alternatively: use Bayesian approach with priors on parameters
- Here 2 examples: comparison of means & comparison of proportions

Comparison of means

- Explicit formulas for power given sample size & sample size given power are available:
  - Power given sample size: \( \text{power} = \Phi \left( \frac{\sqrt{\frac{\Delta^2}{\sigma^2}} + z_{\alpha/2}}{\text{sample size}} \right) \)
  - Sample size given power: \( n = \frac{\sigma^2}{\Delta^2} \left( z_{\beta} - z_{\alpha/2} \right)^2 \)
- Formulas correspond to \( Z \)-test to compare 2 means with:
  - \( \Delta \) = assumed difference in means
  - \( \sigma \) = assumed common SD in the two populations
  - \( z_{\alpha/2} \) = (negative) threshold corresponding to 2-sided \( \alpha \)-level
  - \( z_{\beta} \) = (positive) threshold corresponding to power = 1 - \( \beta \)

Example 6.5 (Spiegelhalter et al., 2004)

- \( \sigma = 1 \), \( \Delta = 1 \), \( \alpha = 0.05 \Rightarrow z_{\alpha/2} = -1.96 \)
- Power given sample size: \( n = 63 \) (sample size in each treatment group)
- Sample size given power: power = 0.80 \( \Rightarrow z_{\beta} = 0.84 \)
- Priors: \( \Delta \sim N(0.5, 0.1) \) & \( \sigma \sim N(1, 0.3) \mid [0, \infty) \)
- Predictive distributions of power and \( n \) with WinBUGS

Graphical representation of \( p_B \):
Pure Bayesian ways to determine necessary sample size

- The above approach is a combination of a Bayesian technique to take into account the prior uncertainty of parameter values and the classical approach of determining the sample size.

- Alternative, Bayesian approaches for sample size calculation have been suggested based on the size of the posterior CI.

- Note that the calculations require a design prior, i.e. a prior used to express the uncertainty of the parameters at the design stage. This design prior does not need to be the same as the prior used in the analysis.

3.8.1 The Bayes factor

- Posterior probability for $H_0$

$$p(H_0 \mid y) = \frac{p(y \mid H_0) p(H_0)}{p(y \mid H_0) p(H_0) + p(y \mid H_a) p(H_a)}$$

- Bayes factor: $BF(y)$

$$\frac{p(H_0 \mid y)}{1 - p(H_0 \mid y)} = \frac{p(y \mid H_0)}{p(y \mid H_a)} \times \frac{p(H_0)}{1 - p(H_0)}$$

- Bayes factor = factor that transforms prior odds for $H_0$ into posterior odds after observed the data.

- Central in Bayesian inference, but not to all Bayesians.

3.9 Medical literature on Bayesian methods in RCTs

Next three papers use above introduced summary measures + techniques.
RCT example 1

Example 1: P.L. Shah et al. (Lancet, 378, 2011)
Bronchoscopic lung-volume reduction with Eckardt airway stents for emphysema (EASE trial): randomised, sham-controlled, multicentre trial

Statistical analysis

We used a Bayesian adaptive approach to sample size, with interim looks scheduled at 225, 270, 315, 360, 405, and 450 participants. Bayesian statistics is an axiomatic

Based on posterior probabilities

Combined with classical statistical tests for other endpoints

We did primary endpoint analyses 6 months after the last patient completed the index procedure. We defined success for the primary intent-to-treat efficacy analysis when the posterior probability of responding to treatment in the airway bypass arm (P') was superior to the posterior probability of responding to treatment for the sham control arm (P), with probability (P') of 96.5%. Trial success requires that n=0.965, where n=P[P'>P].

RCT example 2


- Phase II neoadjuvant HER2/neu-positive breast cancer trial: planned accrual was 164 patients randomized to chemotherapy with and without trastuzumumab
- Primary endpoint: pathological complete response (pCR) of the tumor
- Accrual was slower than expected
- DSMB assessed results when data on 34 patients was available for assessing pCR with question:

What is likely result at the end?

Based on posterior predictive probabilities

Example 2: What is likely result at the end?

- 1st DSMB interim analysis: 16 control patients with 4 complete response ⇔ 18 patients under experimental treatment with 12 complete response
- Bayesian predictive calculation: probability to obtain a classical statistical significant result at the end ≈ 95%
- DSMB advised to stop the trial
Example 2: Generation of fictive future studies given results obtained so far

Example 3: Cannon et al. (Am Heart J, 2009, 158: 513-9.e3)

Design of the DEFINE trial: Determining the Efficacy and Tolerability of CETP Inhibition with Anacetrapib

Statistical design and analysis

The statistical analysis uses the prior data on the CETP inhibitor torcetrapib to assist with decision making from the limited number of CV events expected in the DEFINE trial. To accomplish this, use of commonly used frequentist-based approaches was deemed to be inadequate due to the limited power associated with the low number of events. Instead, a Bayesian approach was constructed for which predefined assumptions underlying the CV event rate are needed. These assumptions include definition of the underlying event rate distribution and an estimate of the event rate incidence. In addition, a pre-stated level of confidence is required to define how much the true incidence rate in CV events may vary versus

Example 3: Decision table
Take home messages

• Posterior distribution contains all information for statistical inference
• To characterize posterior, we can use:
  ▶ posterior mode, mean, median, variance, SD
  ▶ credible intervals: equal-tail, HPD
• Integration is key in Bayesian inference
• PPD = Bayesian tool to predict future
• No need to rely on large samples for Bayesian inference

Bayesian hypothesis testing can be based on:
  ▶ credible intervals
  ▶ contour probability
  ▶ Bayes factor

 Sampling is a useful tool to replace integration

Chapter 4
More than one parameter

Aims:
  ▶ Moving towards practical applications
  ▶ Illustrating that computations become quickly involved
  ▶ Illustrating that frequentist results can be obtained with Bayesian procedures
  ▶ Illustrating a multivariate (independent) sampling algorithm

4.1 Introduction

• Most statistical models involve more than one parameter to estimate
• Examples:
  ▶ Normal distribution: mean $\mu$ and variance $\sigma^2$
  ▶ Linear regression: regression coefficients $\beta_0, \beta_1, \ldots, \beta_d$ and residual variance $\sigma^2$
  ▶ Logistic regression: regression coefficients $\beta_0, \beta_1, \ldots, \beta_d$
  ▶ Multinomial distribution: class probabilities $\theta_1, \theta_2, \ldots, \theta_d$ with $\sum_{j=1}^{d} \theta_j = 1$

• This requires a prior for all parameters (together): expresses our beliefs about the model parameters
• Aim: derive posterior for all parameters and their summary measures
4.2 Joint versus marginal posterior inference

- Bayes theorem:

\[ p(\theta \mid y) = \frac{L(\theta \mid y)p(\theta)}{\int L(\theta \mid y)p(\theta) \, d\theta} \]

- Hence, the same expression as before but now \( \theta = (\theta_1, \theta_2, \ldots, \theta_d)^T \)
- Now, the prior \( p(\theta) \) is multivariate. But often a prior is given for each parameter separately
- Posterior \( p(\theta \mid y) \) is also multivariate. But we usually look only at the (marginal) posteriors \( p(\theta_j \mid y) \) (\( j = 1, \ldots, d \))

- We also need for each parameter: posterior mean, median (and sometimes mode), and credible intervals

4.3 The normal distribution with \( \mu \) and \( \sigma^2 \) unknown

Acknowledging that \( \mu \) and \( \sigma^2 \) are unknown

- Sample \( y_1, \ldots, y_n \) of independent observations from \( N(\mu, \sigma^2) \)
- Joint likelihood of \( (\mu, \sigma^2) \) given \( y \):

\[ L(\mu, \sigma^2 \mid y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \]

- The posterior is again product of likelihood with prior divided by the denominator which involves an integral
- In this case analytical calculations are possible in 2 of the 3 cases
4.3.1 No prior knowledge on $\mu$ and $\sigma^2$ is available

- Noninformative joint prior $p(\mu, \sigma^2) \propto \sigma^{-2}$ ($\mu$ and $\sigma^2$ a priori independent)
- Posterior distribution $p(\mu, \sigma^2 \mid y) \propto \frac{1}{\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (n-1)s^2 + n(\bar{y} - \mu)^2 \right] \right\}$

Justification prior distribution

- Most often prior information on several parameters arrives to us for each of the parameters separately and independently $\Rightarrow p(\mu, \sigma^2) = p(\mu) \times p(\sigma^2)$
- And, we do not have prior information on $\mu$ nor on $\sigma^2$ $\Rightarrow$ choice of prior distributions:
- The chosen priors are called flat priors

Motivation:
- If one is totally ignorant of a location parameter, then it could take any value on the real line with equal prior probability.
- If totally ignorant about the scale of a parameter, then it is as likely to lie in the interval 1-10 as it is to lie in the interval 10-100. This implies a flat prior on the log scale.
- The flat prior $p(\log(\sigma)) = c$ is equivalent to chosen prior $p(\sigma^2) \propto \sigma^{-2}$

Marginal posterior distributions

- Marginal posterior distributions are needed in practice
- $p(\mu \mid y)$
- $p(\sigma^2 \mid y)$
- Calculation of marginal posterior distributions involve integration:

$$p(\mu \mid y) = \int p(\mu, \sigma^2 \mid y) d\sigma^2 = \int p(\mu \mid \sigma^2, y)p(\sigma^2 \mid y)d\sigma^2$$

- Marginal posterior is weighted sum of conditional posteriors with weights = uncertainty on other parameter(s)
Conditional & marginal posterior distributions for the normal case

• Conditional posterior for $\mu$:  
  \[ p(\mu \mid \sigma^2, y) = N(\bar{y}, \sigma^2/n) \]

• Marginal posterior for $\mu$:  
  \[ p(\mu \mid y) = t_{n-1}(\bar{y}, s^2/n) \]
  \[ \Rightarrow \frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1} \ (\mu \text{ is the random variable}) \]

• Marginal posterior for $\sigma^2$:  
  \[ p(\sigma^2 \mid y) \equiv \text{Inv-}\chi^2(n-1, s^2) \]
  \[ \text{(scaled inverse chi-squared distribution)} \]
  \[ \Rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \ (\sigma^2 \text{ is the random variable}) \]

= special case of $\text{IG}(\alpha, \beta) \ (\alpha = (n-1)/2, \beta = 1/2)$

Some inverse-gamma densities

• Joint posterior distribution  
  
  \[ p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y) = N(\bar{y}, \sigma^2/n) \text{Inv-}\chi^2(n-1, s^2) \]

• Normal-scaled-inverse chi-square distribution  
  \[ = \text{N-Inv-}\chi^2(\bar{y}, n, (n-1)s^2) \]

\[ \Rightarrow \text{A posteriori } \mu \text{ and } \sigma^2 \text{ are dependent} \]
Posterior summary measures and PPD

For $\mu$:
- Posterior mean = mode = median = $\bar{y}$
- Posterior variance = $\frac{(n-1)s^2}{n(n-2)}$
- 95% equal tail credible and HPD interval:
  \[ [\bar{y} - t(0.025; n - 1) s/\sqrt{n}, \bar{y} + t(0.025; n - 1) s/\sqrt{n}] \]

For $\sigma^2$:
- Posterior mean, mode, median, variance, 95% equal tail CI all analytically available
- 95% HPD interval is computed iteratively

PPD:
- $t_{n-1} [\bar{y}, s^2 (1 + \frac{1}{n})]$-distribution

Implications of previous results

Frequentist versus Bayesian inference:
- Numerical results are the same
- Inference is based on different principles

Example IV.1: SAP study – Noninformative prior

- Example III.6: normal range for $a/p$ is too narrow
- Joint posterior distribution = N-Inv-$\chi^2$ (NI prior + likelihood, see before)
- Marginal posterior distributions (red curves) for $y = 100/\sqrt{alp}$

Normal range for alp:
- PPD for $y = t_{249}(7.11, 1.37)$-distribution
- 95% normal range for alp = [104.1, 513.2], slightly wider than before
4.3.2 An historical study is available

- Posterior of historical data can be used as prior to the likelihood of current data

Prior = \( \text{N-Inv-}\chi^2(\mu_0, \kappa_0, \nu_0, \sigma^2_0) \) -distribution (from historical data)

Posterior = \( \text{N-Inv-}\chi^2(\mu, \kappa, \nu, \sigma^2) \) -distribution (combining data and \( \text{N-Inv-}\chi^2 \) prior)

- \( \text{N-Inv-}\chi^2 \) is conjugate prior

- Again shrinkage of posterior mean towards prior mean

- Posterior variance = weighted average of prior-, sample variance and distance between prior and sample mean

\( \Rightarrow \) posterior variance is not necessarily smaller than prior variance!

- Similar results for posterior measures and PPD as in first case

Example IV.2: SAP study – Conjugate prior

- Prior based on retrospective study (Topal et al., 2003) of 65 'healthy' subjects:

  - Mean (SD) for \( y = 100/\sqrt{\alpha}p = 5.25 (1.66) \)
  - Conjugate prior = \( \text{N-Inv-}\chi^2(5.25, 65, 64, 2.76) \)
  - Note: mean (SD) prospective data: 7.11 (1.4), quite different

- Posterior = \( \text{N-Inv-}\chi^2(6.72, 315, 314, 2.61) \):

  - Posterior mean \textit{in-between} between prior mean & sample mean, but:
    - Posterior precision \( \neq \) prior + sample precision
    - Posterior variance \( < \) prior variance and \( > \) sample variance
    - Posterior \textit{informative} variance \( > \) \textit{NI} variance
    - Prior information did not lower posterior uncertainty, reason: conflict of likelihood with prior

Marginal posteriors:

Red curves = marginal posteriors from informative prior (historical data)

Histograms retro- and prospective data:
4.3.3 Expert knowledge is available

- Expert knowledge available on each parameter separately

  \[ \text{Joint prior } N(\mu_0, \sigma_0^2) \times \text{Inv-Chi}^2(\nu_0, \tau_0^2) \neq \text{conjugate} \]

- Posterior cannot be derived analytically, but numerical/sampling techniques are available

4.4 Multivariate distributions

- Distributions with a multivariate response:
  - Multivariate normal distribution: generalization of normal distribution
  - Multivariate Student’s t-distribution: generalization of location-scale t-distribution
  - Multinomial distribution: generalization of binomial distribution

- Multivariate prior distributions:
  - N-inv-Chi^2-distribution: prior for \( N(\mu, \sigma^2) \)
  - Dirichlet distribution: generalization of beta distribution
  - (Inverse-)Wishart distribution: generalization of (inverse-) gamma (prior) for covariance matrices (see mixed model chapter)

What now?

Computational problem:
- ‘Simplest problem’ in classical statistics is already complicated
- Ad hoc solution is still possible, but not satisfactory
- There is the need for another approach

Example IV.3: Young adult study – Smoking and alcohol drinking

- Study examining life style among young adults

<table>
<thead>
<tr>
<th></th>
<th>Smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>180</td>
</tr>
<tr>
<td>Yes</td>
<td>216</td>
</tr>
<tr>
<td>Total</td>
<td>396</td>
</tr>
</tbody>
</table>

- Of interest: association between smoking & alcohol-consumption
Likelihood part:

2×2 contingency table = multinomial model \( \text{Mult}(n, \theta) \)

- \( \theta = \{\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22} = 1 - \theta_{11} - \theta_{12} - \theta_{21}\} \) and \( 1 = \sum_{i,j} \theta_{ij} \)
- \( y = \{y_{11}, y_{12}, y_{21}, y_{22}\} \) and \( n = \sum_{i,j} y_{ij} \)

\[ \text{Mult}(n, \theta) = \frac{n!}{y_{11}! y_{12}! y_{21}! y_{22}!} \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}} \]

Dirichlet prior:

Conjugate prior to multinomial distribution = Dirichlet prior \( \text{Dir}(\alpha) \)

\[ \theta \sim \frac{1}{B(\alpha)} \prod_{i,j} \theta_{ij}^{\alpha_{ij}-1} \]

- \( \alpha = \{\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}\} \)
- \( B(\alpha) = \prod_{i,j} \Gamma(\alpha_{ij}) / \Gamma \left( \sum_{i,j} \alpha_{ij} \right) \)

\[ \Rightarrow \text{Posterior distribution} = \text{Dir}(\alpha + y) \]

- Note:
  - Dirichlet distribution = extension of beta distribution to higher dimensions
  - Marginal distributions of a Dirichlet distribution = beta distribution

Measuring association:

- Association between smoking and alcohol consumption:
  \[ \psi = \frac{\theta_{11} \theta_{22}}{\theta_{12} \theta_{21}} \]

- Needed \( p(\psi | y) \), but difficult to derive

- Alternatively replace analytical calculations by sampling procedure

Analysis of contingency table:

- Prior distribution: \( \text{Dir}(1, 1, 1, 1) \)
- Posterior distribution: \( \text{Dir}(180+1, 41+1, 216+1, 64+1) \)
- Sample of 10,000 generated values for \( \theta \) parameters
- 95% equal tail CI for \( \psi \): \([0.839, 2.014]\)
- Equal to classically obtained estimate
Posterior distributions:

4.5 Frequentist properties of Bayesian inference

- Not of prime interest for a Bayesian to know the sampling properties of estimators
- However, it is important that Bayesian approach gives most often the right answer
- What is known?
  - Theory: posterior is normal for a large sample (BCLT)
  - Simulations: Bayesian approach may offer alternative interval estimators with better coverage than classical frequentist approaches

4.6 The Method of Composition

A method to yield a random sample from a multivariate distribution

- Stagewise approach
- Based on factorization of joint distribution into a marginal & several conditionals
  \[ p(\theta_1, \ldots, \theta_d | y) = p(\theta_d | y) p(\theta_{d-1} | \theta_d, y) \ldots p(\theta_1 | \theta_2, y) \]

- Sampling approach:
  - Sample \( \hat{\theta}_d \) from \( p(\hat{\theta}_d | y) \)
  - Sample \( \hat{\theta}_{(d-1)} \) from \( p(\theta_{(d-1)} | \hat{\theta}_d, y) \)
  - ... 
  - Sample \( \hat{\theta}_1 \) from \( p(\hat{\theta}_1 | \hat{\theta}_{d-1}, \ldots, \hat{\theta}_2, y) \)

Sampling from posterior when \( y \sim N(\mu, \sigma^2) \), both parameters unknown

- Sample first \( \sigma^2 \), then given a sampled value of \( \sigma^2 (\tilde{\sigma}^2) \) sample \( \mu \) from \( p(\mu | \tilde{\sigma}^2, y) \)
- Output case 1: No prior knowledge on \( \mu \) and \( \sigma^2 \) on next page
4.7 Bayesian linear regression models

- Example of a classical multiple linear regression analysis

- Non-informative Bayesian multiple linear regression analysis:
  - Non-informative prior for all parameters + classical linear regression model
  - Analytical results are available + method of composition can be applied

4.7.1 The frequentist approach to linear regression

Classical regression model: $y = X\beta + \varepsilon$
- $y$ = a $n \times 1$ vector of independent responses
- $X = n \times (d + 1)$ design matrix
- $\beta = (d + 1) \times 1$ vector of regression parameters
- $\varepsilon = n \times 1$ vector of random errors $\sim N(0, \sigma^2 I)$

Likelihood:
$$L(\beta, \sigma^2 | y, X) = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right]$$
- MLE = LSE of $\beta$: $\hat{\beta} = (X^T X)^{-1} X^T y$
- Residual sum of squares: $S = (y - X\beta)^T(y - X\beta)$
- Mean residual sum of squares: $s^2 = S/(n - d - 1)$

Example IV.7: Osteoporosis study: a frequentist linear regression analysis
- Cross-sectional study (Boonen et al., 1996)
- 245 healthy elderly women in a geriatric hospital
- Aim: Find determinants for osteoporosis
- Average age women = 75 yrs with a range of 70-90 yrs
- Marker for osteoporosis = $tbbmc$ (in kg) measured for 234 women
- Simple linear regression model: regressing $tbbmc$ on bmi
- Classical frequentist regression analysis:
  - $\hat{\beta}_0 = 0.813 (0.12)$
  - $\hat{\beta}_1 = 0.0404 (0.0043)$
  - $s^2 = 0.29$, with $n - d - 1 = 232$
  - $\text{corr}(\hat{\beta}_0, \hat{\beta}_1) = -0.99$
4.7.2 A noninformative Bayesian linear regression model

Bayesian linear regression model = prior information on regression parameters & residual variance + normal regression likelihood

- Noninformative prior for $(\beta, \sigma^2)$: $p(\beta, \sigma^2) \propto \sigma^{-2}$
- Notation: omit design matrix $X$
- Posterior distributions:

$$p(\beta, \sigma^2 | y) = N_{(d+1)} \left[ \beta | \hat{\beta}, \sigma^2(X^T X)^{-1} \right] \times \text{Inv} - \chi^2(\sigma^2 | n - d - 1, s^2)$$

$$p(\beta | \sigma^2, y) = N_{(d+1)} \left[ \beta | \hat{\beta}, \sigma^2(X^T X)^{-1} \right]$$

$$p(\sigma^2 | y) = \text{Inv} - \chi^2(\sigma^2 | n - d - 1, s^2)$$

$$p(\beta | y) = T_{n-d-1} \left[ \beta | \hat{\beta}, s^2(X^T X)^{-1} \right]$$

4.7.3 Posterior summary measures for the linear regression model

- Posterior summary measures of
  - (a) regression parameters $\beta$
  - (b) parameter of residual variability $\sigma^2$
- Univariate posterior summary measures
  - The marginal posterior mean (mode, median) of $\beta_j = \text{MLE (LSE)} \hat{\beta}_j$
  - 95% HPD interval for $\beta_j$
  - Marginal posterior mode and mean of $\sigma^2$
  - 95% HPD-interval for $\sigma^2$

Multivariate posterior summary measures

- Multivariate posterior summary measures for $\beta$
  - Posterior mean (mode) of $\beta = \hat{\beta}$ (MLE=LSE)
  - 100(1-\alpha)%-HPD region
  - Contour probability for $H_0: \beta = \beta_0$
4.7.4 Sampling from the posterior distribution

- Most posteriors can be sampled via standard sampling algorithms
- What about \( p(\beta \mid y) = \text{multivariate } t\)-distribution? How to sample from this distribution? (R function \texttt{rmvt} in \texttt{mvtnorm})
- Easy with Method of Composition: Sample in two steps
  - Sample from \( p(\sigma^2 \mid y) = \text{scaled inverse chi-squared distribution} \Rightarrow \tilde{\sigma}^2\)
  - Sample from \( p(\beta \mid \tilde{\sigma}^2, y) = \text{multivariate normal distribution} \)

Example IV.8: Osteoporosis study – Sampling with Method of Composition

- Sample \(\tilde{\sigma}^2\) from \( p(\sigma^2 \mid y) = \text{Inv-} \chi^2(\sigma^2 \mid n-d-1, s^2)\)
- Sample from \(\tilde{\beta}\) from \( p(\beta \mid \tilde{\sigma}^2, y) = N_{(d+1)} \left[ \beta \mid \tilde{\beta}, \tilde{\sigma}^2(\mathbf{X}^T \mathbf{X})^{-1} \right] \)
- Sampled mean regression vector = \((0.816, 0.0403)\)
- 95% equal tail CIs = \(\beta_0: [0.594, 1.040]\) & \(\beta_1: [0.0317, 0.0486]\)
- Contour probability for \( H_0: \beta = 0 < 0.001\)
- Marginal posterior of \((\beta_0, \beta_1)\) has a ridge \( r(\beta_0, \beta_1) = -0.99\)

PPD:

- Distribution of a future observation at \(bmi=30\)
- Sample future observation \(\tilde{y}\) from \( N(\tilde{\mu}_{30}, \tilde{\sigma}_{30}^2)\):
  - \(\tilde{\mu}_{30} = \tilde{\beta}^T (1, 30)\)
  - \(\tilde{\sigma}_{30}^2 = \sigma^2 \left[ 1 + (1, 30)(\mathbf{X}^T \mathbf{X})^{-1}(1, 30)^T \right]\)
- Sampled mean and standard deviation = \(2.033\) and \(0.282\)
4.8 Bayesian generalized linear models

Generalized Linear Model (GLIM): extension of the linear regression model to a wide class of regression models

- Examples:
  - Normal linear regression model with normal distribution for continuous response and \( \sigma^2 \) assumed known
  - Poisson regression model with Poisson distribution for count response, and \( \log(\text{mean}) = \) linear function of covariates
  - Logistic regression model with Bernoulli distribution for binary response and logit of probability = linear function of covariates

4.8.1 More complex regression models

- Considered multiparameter models are limited
  - Weibull distribution for \( \alpha \rho \)?
  - Censored/truncated data?
  - Cox regression?

- Postpone to MCMC techniques

Take home messages

- Any practical application involves more than one parameter, hence immediately Bayesian inference is multivariate even with univariate data.

- A multivariate prior is needed and a multivariate posterior is obtained, but the marginal posterior is the basis for practical inference

- Nuisance parameters:
  - Bayesian inference: average out nuisance parameter
  - Classical inference: profile out (maximize out nuisance parameter)

- Multivariate independent sampling can be done, if marginals can be computed

- Frequentist properties of Bayesian estimators (with NI priors) often good
Chapter 5
Choosing the prior distribution

Aims:
- Review the different principles that lead to a prior distribution
- Critically review the impact of the subjectivity of prior information

5.1 Introduction

Incorporating prior knowledge

- Unique feature for Bayesian approach
- But might introduce subjectivity
- Useful in clinical trials to reduce sample size

In this chapter we review different kinds of priors:

- Conjugate
- Noninformative
- Informative

5.2 The sequential use of Bayes theorem

- Posterior of the $k$th experiment = prior for the $(k+1)$th experiment (sequential surgeries)
- In this way, the Bayesian approach can mimic our human learning process
- Meaning of ‘prior’ in prior distribution:
  - Prior: prior knowledge should be specified independent of the collected data
  - In RCTs: fix the prior distribution in advance

5.3 Conjugate prior distributions

In this section:

- Conjugate priors for univariate & multivariate data distributions
- Conditional conjugate and semi-conjugate distributions
- Hyperpriors
5.3.1 Conjugate priors for univariate data distributions

- In previous chapters, examples were given whereby combination of prior with likelihood, gives posterior of the same type as the prior.
- This property is called conjugacy.
- For an important class of distributions (those that belong to exponential family) there is a recipe to produce the conjugate prior.

### Table conjugate priors for univariate discrete data distributions

<table>
<thead>
<tr>
<th>Exponential family member</th>
<th>Parameter</th>
<th>Conjugate prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>( \theta )</td>
<td>Beta((\alpha_0, \beta_0))</td>
</tr>
<tr>
<td>Binomial</td>
<td>( \theta )</td>
<td>Beta((\alpha_0, \beta_0))</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>( \theta )</td>
<td>Beta((\alpha_0, \beta_0))</td>
</tr>
<tr>
<td>Poisson</td>
<td>( \lambda )</td>
<td>Gamma((\alpha_0, \beta_0))</td>
</tr>
</tbody>
</table>

### Table conjugate priors for univariate continuous data distributions

<table>
<thead>
<tr>
<th>Exponential family member</th>
<th>Parameter</th>
<th>Conjugate prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal-variance fixed</td>
<td>( \mu, \sigma^2 )</td>
<td>( N(\mu_0, \sigma_0^2) )</td>
</tr>
<tr>
<td>Normal-mean fixed</td>
<td>( \mu, \sigma^2 )</td>
<td>( IG(\alpha_0, \beta_0) )</td>
</tr>
<tr>
<td>Normal</td>
<td>( \mu, \sigma^2 )</td>
<td>( NIG(\mu_0, \kappa_0, \alpha_0, \beta_0) )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( \lambda )</td>
<td>Gamma((\alpha_0, \beta_0))</td>
</tr>
</tbody>
</table>

### Recipe to choose conjugate priors

For a random sample \( y = \{y_1, \ldots, y_n\} \) of i.i.d. elements:

\[
p(y \mid \theta) = b(y) \exp \left[ c(\theta)^T t(y) + d(\theta) \right]
\]

- \( d(\theta), b(y) \) = scalar functions, \( c(\theta) = (c_1(\theta), \ldots, c_d(\theta))^T \)
- \( t(y) = d \)-dimensional sufficient statistic for \( \theta \) (canonical parameter)
- Examples: Binomial distribution, Poisson distribution, normal distribution, etc.
Recipe to choose conjugate priors

For the exponential family, the class of prior distributions $ℑ$ is closed under sampling =

$$p(\theta | \alpha, \beta) = k(\alpha, \beta) \exp \left[ c(\theta)^T \alpha + \beta d(\theta) \right]$$

- $\alpha = (\alpha_1, \ldots, \alpha_d)^T$ and $\beta$ hyperparameters
- Normalizing constant: $k(\alpha, \beta) = 1/\int \exp \left[ c(\theta)^T \alpha + \beta d(\theta) \right] d\theta$

Proof of closure:

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

$$= \exp \left[ c(\theta)^T t(y) + n d(\theta) \right] \exp \left[ c(\theta)^T \alpha + \beta d(\theta) \right]$$

$$= \exp \left[ c(\theta)^T \alpha^* + \beta^* d(\theta) \right],$$

with $\alpha^* = \alpha + t(y), \beta^* = \beta + n$

Practical advantages when using conjugate priors

A (natural) conjugate prior distribution for the exponential family is convenient from several viewpoints:

- mathematical
- numerical
- interpretational (convenience prior):
  - The likelihood of historical data can be easily turned into a conjugate prior. The natural conjugate distribution = equivalent to a fictitious experiment
  - For a natural conjugate prior, the posterior mean = weighted combination of the prior mean and sample estimate

Example V.2: Dietary study – Normal versus $t$-prior

- Example II.2: IBBENS-2 normal likelihood was combined with $N(328, 100)$ (conjugate) prior distribution
- Replace the normal prior by a $t_{30}(328, 100)$-prior
  - posterior practically unchanged, but 3 elegant features of normal prior are lost:
    - Posterior cannot be determined analytically
    - Posterior is not of the same class as the prior
    - Posterior summary measures are not obvious functions of the prior and the sample summary measures
5.3.2 Conjugate prior for normal distribution – mean and variance unknown

\( N(\mu, \sigma^2) \) with \( \mu \) and \( \sigma^2 \) unknown ∈ two-parameter exponential family

- Conjugate = product of a normal prior with inverse gamma prior
- Notation: \( \text{NIG}(\mu_0, \kappa_0, a_0, b_0) \)

Mean known and variance unknown

- For \( \sigma^2 \) unknown and \( \mu \) known
- Natural conjugate is inverse gamma (IG)
- Equivalently: scaled inverse-\( \chi^2 \) distribution (Inv-\( \chi^2 \))

5.3.3 Multivariate data distributions

Priors for two popular multivariate models:

- Multinomial model
- Multivariate normal model

Table conjugate priors for multivariate data distributions

<table>
<thead>
<tr>
<th>Exponential family member</th>
<th>Parameter</th>
<th>Conjugate prior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MULTIVARIATE CASE</strong></td>
<td><strong>Discrete distributions</strong></td>
<td></td>
</tr>
<tr>
<td>Multinomial</td>
<td>Mult(n,( \theta ))</td>
<td>( \theta ) Dirichlet(( \alpha_0 ))</td>
</tr>
<tr>
<td><strong>Continuous distributions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal-covariance fixed</td>
<td>( N(\mu, \Sigma) )-( \Sigma ) fixed</td>
<td>( \mu ) ( N(\mu_0, \Sigma_0) )</td>
</tr>
<tr>
<td>Normal-mean fixed</td>
<td>( N(\mu, \Sigma) )-( \mu ) fixed</td>
<td>( \Sigma ) IW(( \Lambda_0, \nu_0 ))</td>
</tr>
<tr>
<td>Normal*</td>
<td>( N(\mu, \Sigma) )</td>
<td>( \mu, \Sigma ) NIW(( \mu_0, \kappa_0, a_0, b_0 ))</td>
</tr>
</tbody>
</table>
Multinomial model

\( \text{Mult}(n, \theta) \): \[ p(y | \theta) = \frac{n!}{y_1! \cdots y_k!} \prod_{j=1}^{k} \theta_j^{y_j} \in \text{exponential family} \]

Natural conjugate: Dirichlet(\( \alpha_0 \)) distribution

\[ p(\theta | \alpha_0) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_{0j})}{\sum_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} \theta_j^{\alpha_{0j}-1} \]

Properties:

\( \triangleright \) Posterior distribution = Dirichlet(\( \alpha_0 + y \))
\( \triangleright \) Beta distribution = special case of a Dirichlet distribution with \( k = 2 \)
\( \triangleright \) Marginal distributions of the Dirichlet distribution = beta distributions
\( \triangleright \) Dirichlet(1, 1, \ldots, 1) = extension of the classical uniform prior Beta(1,1)

Multivariate normal model

The \( p \)-dimensional multivariate normal distribution:

\[ p(y_1, \ldots, y_n | \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right] \]

Conjugates:

\( \triangleright \) \( \Sigma \) known and \( \mu \) unknown: \( \text{N}(\mu_0, \Sigma_0) \) for \( \mu \)
\( \triangleright \) \( \Sigma \) unknown and \( \mu \) known: inverse Wishart distribution \( \text{IW}(\Lambda_0, \nu_0) \) for \( \Sigma \)
\( \triangleright \) \( \Sigma \) unknown and \( \mu \) unknown:

Normal-inverse Wishart distribution \( \text{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0) \) for \( \mu \) and \( \Sigma \)

5.3.4 Conditional conjugate and semi-conjugate priors

Example \( \theta = (\mu, \sigma^2) \) for \( y \sim \text{N}(\mu, \sigma^2) \)

- **Conditional conjugate** for \( \mu \): \( \text{N}(\mu_0, \sigma_0^2) \)
- **Conditional conjugate** for \( \sigma^2 \): \( \text{IG}(\alpha, \beta) \)
- **Semi-conjugate prior** = product of conditional conjugates
- Often conjugate priors cannot be used in WinBUGS, but semi-conjugates are popular

5.3.5 Hyperpriors

Conjugate priors are restrictive to present prior knowledge

\( \Rightarrow \) Give parameters of conjugate prior also a prior

Example:

- Prior: \( \theta \sim \text{Beta}(1, 1) \)
- Instead: \( \theta \sim \text{Beta}(\alpha, \beta) \) and \( \alpha \sim \text{Gamma}(1, 3), \beta \sim \text{Gamma}(2, 4) \)

\( \triangleright \) \( \alpha, \beta = \text{hyperparameters} \)

\( \triangleright \) Gamma(1, 3) \( \times \) Gamma(2, 4) = hyperprior/hierarchical prior

- **Aim**: more flexibility in prior distribution (and useful for Gibbs sampling)
5.4 Noninformative prior distributions

5.4.1 Introduction

Sometimes/often researchers cannot or do not wish to make use of prior knowledge
⇒ prior should reflect this absence of knowledge

• Prior that express no knowledge = (initially) called a noninformative (NI)

• Central question: What prior reflects absence of knowledge?
  ▷ Flat prior?
  ▷ Huge amount of research to find best NI prior
  ▷ Other terms for NI: non-subjective, objective, default, reference, weak, diffuse, flat, conventional and minimally informative, etc

• Challenge: make sure that posterior is a proper distribution!

5.4.2 Expressing ignorance

• Equal prior probabilities = principle of insufficient reason, principle of indifference, Bayes-Laplace postulate

• Unfortunately, but ... flat prior cannot express ignorance

Ignorance at different scales:

Ignorance on $\sigma$-scale is different from ignorance on $\sigma^2$-scale
Ignorance cannot be expressed mathematically

5.4.3 General principles to choose noninformative priors

A lot of research has been spent on the specification of NI priors, most popular are Jeffreys priors:

- Result of a Bayesian analysis depends on choice of scale for flat prior: $p(\theta) \propto c$ or $p(h(\theta)) \equiv p(\psi) \propto c$

- To preserve conclusions when changing scale: Jeffreys suggested a rule to construct priors based on the invariance principle/rule (conclusions do not change when changing scale)

- Jeffreys rule suggests a way to choose a scale to take the flat prior on

Jeffreys rule also exists for more than one parameter (Jeffreys multi-parameter rule)

5.4.4 Improper prior distributions

- Many NI priors are improper (= AUC is infinite)

- Improper prior is technically no problem when posterior is proper

- Example: Normal likelihood ($\mu$ unknown + $\sigma^2$ known) + flat prior on $\mu$

$$p(\mu \mid y) = \frac{p(y \mid \mu) p(\mu)}{\int p(y \mid \mu) p(\mu) \, d\mu} = \frac{p(y \mid \mu) c}{\int p(y \mid \mu) c \, d\mu} = \frac{1}{\sqrt{2\pi} \sigma / \sqrt{n}} \exp \left[ -\frac{n}{2} \left( \frac{\mu - y}{\sigma} \right)^2 \right]$$

- Complex models: difficult to know when improper prior yields a proper posterior (variance of the level-2 obs in Gaussian hierarchical model)

- Interpretation of improper priors?

Examples of Jeffreys priors

- Binomial model: $p(\theta) \propto \theta^{-1/2} (1 - \theta)^{-1/2} \Leftrightarrow$ flat prior on $\psi(\theta) \propto \arcsin \sqrt{\theta}$

- Poisson model: $p(\lambda) \propto \lambda^{-1/2} \Leftrightarrow$ flat prior on $\psi(\lambda) = \sqrt{\lambda}$

- Normal model with $\sigma$ fixed: $p(\mu) \propto c$

- Normal model with $\mu$ fixed: $p(\sigma^2) \propto \sigma^{-2} \Leftrightarrow$ flat prior on $\log(\sigma)$

- Normal model with $\mu$ and $\sigma^2$ unknown: $p(\mu, \sigma^2) \propto \sigma^{-2}$, which reproduces some classical frequentist results !!!
5.4.5 Weak/vague priors

- For practical purposes: sufficient that prior is **locally uniform** also called **vague** or **weak**

- **Locally uniform**: prior \( \approx \) constant on interval outside which likelihood \( \approx \) zero

- **Examples** for \( N(\mu, \sigma^2) \) likelihood:
  - \( \mu \): \( N(0, \sigma_0^2) \) prior with \( \sigma_0 \) large
  - \( \sigma^2 \): \( \text{IG}(\varepsilon, \varepsilon) \) prior with \( \varepsilon \) small \( \approx \) Jeffreys prior

**Vague priors in software:**

- **WinBUGS** allows only (proper) **vague priors** (Jeffreys priors are not allowed)
  - \( \mu \sim \text{dnorm}(0.0, 1.0E-6) \): normal prior with variance = 1000
  - \( \tau^2 \sim \text{dgamma}(0.001, 0.001) \): inverse gamma prior for variance with shape=rate=10^{-3}

- **SAS** allows **improper priors** (allows Jeffreys priors)
5.5 Informative prior distributions

5.5.1 Introduction

- In basically all research some prior knowledge is available
- In this section:
  - Formalize the use of historical data as prior information using the power prior
  - Review the use of clinical priors, which are prior distributions based on either historical data or on expert knowledge
  - Priors that are based on formal rules expressing prior skepticism and optimism
- The set of priors representing prior knowledge = subjective or informative priors
- But, first two success stories how the Bayesian approach helped to find:
  - a crashed plane
  - a lost fisherman on the Atlantic Ocean

Locating a lost plane

- Statisticians helped locate an Air France plane in 2011 which was missing for two years using Bayesian methods
- June 2009: Air France flight 447 went missing flying from Rio de Janeiro in Brazil to Paris, France
- Debris from the Airbus A330 was found floating on the surface of the Atlantic five days later
- After a number of days, the debris would have moved with the ocean current, hence finding the black box is not easy
- Existing software (used by the US Coast Guard) did not help
- Senior analyst at Metron, Colleen Keller, relied on Bayesian methods to locate the black box in 2011
Finding a lost fisherman on the Atlantic Ocean

New York Times (30 September 2014)

“... if not for statisticians, a Long Island fisherman might have died in the Atlantic Ocean after falling off his boat early one morning last summer

The man owes his life to a once obscure field known as Bayesian statistics - a set of mathematical rules for using new data to continuously update beliefs or existing knowledge

It is proving especially useful in approaching complex problems, including searches like the one the Coast Guard used in 2013 to find the missing fisherman, John Aldridge

But the current debate is about how scientists turn data into knowledge, evidence and predictions. Concern has been growing in recent years that some fields are not doing a very good job at this sort of inference. In 2012, for example, a team at the biotech company Amgen announced that they’d analyzed 53 cancer studies and found it could not replicate 47 of them

The Coast Guard has been using Bayesian analysis since the 1970s. The approach lends itself well to problems like searches, which involve a single incident and many different kinds of relevant data, said Lawrence Stone, a statistician for Metron, a scientific consulting firm in Reston, Va., that works with the Coast Guard

5.5.2 Data-based prior distributions

- In previous chapters:
  - Combined historical data with current data assuming identical conditions
  - Discounted importance of prior data by increasing variance

- Generalized by power prior (Ibrahim and Chen):
  - Likelihood historical data: \( L(\theta \mid y_0) \) based on \( y_0 = \{y_{01}, \ldots, y_{0n_0}\} \)
  - Prior of historical data: \( p_0(\theta \mid c_0) \)
  - Power prior distribution:

\[
p(\theta \mid y_0, a_0) \propto L(\theta \mid y_0)^{a_0} p_0(\theta \mid c_0)
\]

with \( 0 \) no accounting \( \leq a_0 \leq 1 \) fully accounting
5.5.3 Elicitation of prior knowledge

- Elicitation of prior knowledge: turn (qualitative) information from ‘experts’ into probabilistic language

- Challenges:
  - Most experts have no statistical background
  - What to ask to construct prior distribution:
    - Prior mode, median, mean and prior 95% CI?
    - Description of the prior: quartiles, mean, SD?
  - Some probability statements are easier to elicit than others

Example V.5: Stroke study – Prior for 1st interim analysis from experts

Prior knowledge on $\theta$ (incidence of SICH), elicitation based on:

- Most likely value for $\theta$ and prior equal-tail 95% CI
- Prior belief $p_k$ on each of the $K$ intervals $I_k \equiv [\theta_{k-1}, \theta_k)$ covering [0, 1]

Elicitation of prior knowledge – some remarks

- Community and consensus prior: obtained from a community of experts
- Difficulty in eliciting prior information on more than 1 parameter jointly
- Lack of Bayesian papers based on genuine prior information

Identifiability issues

- With overspecified model: non-identifiable model
- Unidentified parameter, when given a NI prior also posterior is NI
- Bayesian approach can make parameters estimable, so that it becomes an identifiable model
- In next example, not all parameters can be estimated without extra (prior) information
**Example V.6: Cysticercosis study – Estimate prevalence without gold standard**

**Experiment:**
- 868 pigs tested in Zambia with Ag-ELISA diagnostic test
- 496 pigs showed a positive test
- **Aim**: estimate the prevalence $\pi$ of cysticercosis in Zambia among pigs

If estimate of sensitivity $\alpha$ and specificity $\beta$ available, then:

$$\hat{\pi} = \frac{p^+ + \hat{\beta} - 1}{\hat{\alpha} + \beta - 1}$$

- $p^+ = n^+/n =$ proportion of subjects with a positive test
- $\hat{\alpha}$ and $\hat{\beta} =$ estimated sensitivity and specificity

**Data:**

**Table of results:**

<table>
<thead>
<tr>
<th>Test</th>
<th>Disease (True)</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>$\pi\alpha(1-\pi)(1-\beta)$</td>
<td>$n^+ = 496$</td>
</tr>
<tr>
<td>-</td>
<td>$\pi(1-\alpha)(1-\pi)\beta$</td>
<td>$n^- = 372$</td>
</tr>
<tr>
<td>Total</td>
<td>$\pi (1-\pi)$</td>
<td>$n = 868$</td>
</tr>
</tbody>
</table>

- Only collapsed table is available
- Since $\alpha$ and $\beta$ vary geographically, **expert knowledge is needed**

**Prior and posterior:**

- Prior distribution on $\pi$ ($p(\pi)$), $\alpha$ ($p(\alpha)$) and $\beta$ ($p(\beta)$) is needed
- Posterior distribution:

$$p(\pi, \alpha, \beta \mid n^+, n^-) \propto \binom{n}{n^+} \pi^\alpha (1-\pi)^{(1-\beta)n^+} [\pi(1-\alpha) + (1-\pi)\beta]^{n^-} p(\pi)p(\alpha)p(\beta)$$

- WinBUGS was used

**Posterior of $\pi$:**

(a) Uniform priors for $\pi$, $\alpha$ and $\beta$ (no prior information)
(b) Beta(21,12) prior for $\alpha$ and Beta(32,4) prior for $\beta$ (historical data)
5.5.4 Archetypal prior distributions

- Use of prior information in Phase III RCTs is problematic, except for medical device trials (FDA guidance document)

⇒ Pleas for objective priors in RCTs

- There is a role of subjective priors for interim analyses:
  - Skeptical prior
  - Enthusiastic prior

Example V.7: Skeptical priors in a phase III RCT

Tan et al. (2003):

- Phase III RCT for treating patients with hepatocellular carcinoma
- Standard treatment: surgical resection
- Experimental treatment: surgery + adjuvant radioactive iodine (adjuvant therapy)
- Planning: recruit 120 patients

Frequentist interim analyses for efficacy were planned:

- First interim analysis (30 patients): experimental treatment better ($P = 0.01 < 0.029$) = $P$-value of stopping rule
- But, scientific community was skeptical about adjuvant therapy

⇒ New multicentric trial (300 patients) was set up

Prior to the start of the subsequent trial:

- Pretrial opinions of the 14 clinical investigators were elicited
- The prior distributions of each investigator were constructed by eliciting the prior belief on the treatment effect (adjuvant versus standard) on a grid of intervals
- Average of all priors = community prior
- Average of the priors of the 5 most skeptical investigators = skeptical prior

To exemplify the use of the skeptical prior:

- Combine skeptical prior with interim analysis results of previous trial

⇒ 1-sided contour probability (in 1st interim analysis) = 0.49

⇒ The first trial would not have been stopped for efficacy
A formal skeptical/enthusiastic prior

Formal subjective priors (Spiegelhalter et al., 1994) in normal case:

- Useful in the context of monitoring clinical trials in a Bayesian manner
- $\theta =$ true effect of treatment ($A$ versus $B$)
- **Skeptical normal prior**: choose mean and variance of $p(\theta)$ to reflect skepticism
- **Enthusiastic normal prior**: choose mean and variance of $p(\theta)$ to reflect enthusiasm
- See figure next page & book
5.6 Prior distributions for regression models

5.6.1 Normal linear regression

Normal linear regression model:

\[ y_i = x_i^T \beta + \varepsilon_i, \quad (i = 1, \ldots, n) \]

\[ y = X \beta + \varepsilon \]

Priors

- **Non-informative priors:**
  - Popular NI prior: \( p(\beta, \sigma^2) \propto \sigma^{-2} \) (Jeffreys multi-parameter rule)
  - WinBUGS: product of independent \( N(0, \sigma_0^2) \) (\( \sigma_0^2 \) large) + \( IG(\varepsilon, \varepsilon) \) (\( \varepsilon \) small)

- **Conjugate priors:**
  - Conjugate NIG prior = \( N(\beta_0, \sigma^2 \Sigma_0) \times IG(a_0, b_0) \) (or \( \text{Inv-}\chi^2(\nu_0, \tau_0^2) \))

- **Historical/expert priors:**
  - Prior knowledge on regression coefficients must be given jointly
  - Elicitation process via distributions at covariate values
  - Most popular: express prior based on historical data

5.6.2 Generalized linear models

- In practice choice of NI priors much the same as with linear models
- But, too large prior variance may not be best for sampling, e.g. in logistic regression model
- In SAS: Jeffreys (improper) prior can be chosen
- Conjugate priors are based on fictive historical data
  - Data augmentation priors & conditional mean priors
  - Not implemented in classical software, but fictive data can be explicitly added and then standard software can be used
5.7 Modeling priors

Modeling prior: adapt characteristics of the statistical model

- Multicollinearity: appropriate prior avoids inflation of $\beta$
- Numerical (separation) problems: appropriate prior avoids inflation of $\beta$
- Constraints on parameters: constraint can be put in prior
- Variable selection: prior can direct the variable search

Multicollinearity

Multicollinearity: $|X^TX| \approx 0 \Rightarrow$ regression coefficients and standard errors inflated

Ridge regression:

- Minimize: $(y^* - X\beta)^T(y^* - X\beta) + \lambda\beta^T\beta$ with $\lambda \geq 0$ & $y^* = y - \bar{y}1_n$
- Estimate: $\hat{\beta}_R(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$

= Posterior mode of a Bayesian normal linear regression analysis with:

- Normal ridge prior $N(0, \tau^2I)$ for $\beta$
- $\tau^2 = \sigma^2/\lambda$ with $\sigma$ and $\lambda$ fixed

- Can be easily extended to BGLIM

Numerical (separation) problems

Separation problems in binary regression models: complete separation and quasi-complete separation

Solution: Take weakly informative prior on regression coefficients

Constraints on parameters

Signal-Tandmobiel® study:

- $\theta_k =$ probability of CE among Flemish children in $(k = 1, \ldots, 6)$ school year
- Constraint on parameters: $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_6$
- Solutions:

- Prior on $\theta = (\theta_1, \ldots, \theta_6)^T$ that maps all $\theta$s that violate the constraint to zero
- Neglect the values that are not allowed in the posterior (useful when sampling)
Other modeling priors

- LASSO prior (see Bayesian variable selection)
- ...

5.8 Other regression models

- A great variety of models
- Not considered here: conditional logistic regression model, Cox proportional hazards model, generalized linear mixed effects models
- ...

Take home messages

- Often prior is dominated by the likelihood (data)
- Prior in RCTs: prior to the trial
- Conjugate priors: convenient mathematically, computationally and from an interpretational viewpoint
- Conditional conjugate priors: heavily used in Gibbs sampling
- Hyperpriors: extend the range of conjugate priors, also important in Gibbs sampling

- Noninformative priors:
  - do not exist, strictly speaking
  - in practice vague priors (e.g. locally uniform) are ok
  - important class of NI priors: Jeffreys priors
  - be careful with improper priors, they might imply improper posterior

- Informative priors:
  - can be based on historical data & expert knowledge (but only useful when viewpoint of a community of experts)
  - are useful in clinical trials to reduce sample size
Chapter 6
Markov chain Monte Carlo sampling

Aims:
▷ Introduce the sampling approach(es) that revolutionized Bayesian approach

6.1 Introduction

▷ Solving the posterior distribution analytically is often not feasible due to the difficulty in determining the integration constant
▷ Computing the integral using numerical integration methods is a practical alternative if only a few parameters are involved
⇒ New computational approach is needed
▷ Sampling is the way to go!
▷ With Markov chain Monte Carlo (MCMC) methods:
  1. Gibbs sampler
  2. Metropolis-(Hastings) algorithm

MCMC approaches have revolutionized Bayesian methods!

Intermezzo: Joint, marginal and conditional probability

Two (discrete) random variables $X$ and $Y$

• **Joint** probability of $X$ and $Y$: probability that $X=x$ and $Y=y$ happen together
• **Marginal** probability of $X$: probability that $X=x$ happens
• **Marginal** probability of $Y$: probability that $Y=y$ happens
• **Conditional** probability of $X$ given $Y=y$: probability that $X=x$ happens if $Y=y$
• **Conditional** probability of $Y$ given $X=x$: probability that $Y=y$ happens if $X=x$

Intermezzo: Joint, marginal and conditional probability

IBbens study: 563 (556) bank employees in 8 subsidiaries of Belgian bank participated in a dietary study
Intermezzo: Joint, marginal and conditional probability

**IBBENS study:** 563 (556) bank employees in 8 subsidiaries of Belgian bank participated in a dietary study.

**IBBENS study:** frequency table

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>−50</td>
<td>2</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>50 – 60</td>
<td>1</td>
<td>25</td>
<td>50</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>60 – 70</td>
<td>0</td>
<td>12</td>
<td>54</td>
<td>52</td>
<td>13</td>
<td>1</td>
<td>132</td>
</tr>
<tr>
<td>70 – 80</td>
<td>0</td>
<td>5</td>
<td>42</td>
<td>72</td>
<td>34</td>
<td>0</td>
<td>153</td>
</tr>
<tr>
<td>80 – 90</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>58</td>
<td>32</td>
<td>2</td>
<td>105</td>
</tr>
<tr>
<td>90 – 100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>18</td>
<td>3</td>
<td>41</td>
</tr>
<tr>
<td>100 – 110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>110 – 120</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>120+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>54</td>
<td>163</td>
<td>220</td>
<td>107</td>
<td>8</td>
<td>556</td>
</tr>
</tbody>
</table>

**IBBENS study:** joint probability

<table>
<thead>
<tr>
<th>Weight</th>
<th>−50</th>
<th>50 – 60</th>
<th>60 – 70</th>
<th>70 – 80</th>
<th>80 – 90</th>
<th>90 – 100</th>
<th>100 – 110</th>
<th>110 – 120</th>
<th>120+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>−50</td>
<td>2/556</td>
<td>12/556</td>
<td>4/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>18/556</td>
</tr>
<tr>
<td>50 – 60</td>
<td>1/556</td>
<td>25/556</td>
<td>50/556</td>
<td>14/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>90/556</td>
</tr>
<tr>
<td>60 – 70</td>
<td>0/556</td>
<td>12/556</td>
<td>54/556</td>
<td>52/556</td>
<td>13/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>132/556</td>
</tr>
<tr>
<td>70 – 80</td>
<td>0/556</td>
<td>5/556</td>
<td>42/556</td>
<td>72/556</td>
<td>34/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>153/556</td>
</tr>
<tr>
<td>80 – 90</td>
<td>0/556</td>
<td>0/556</td>
<td>12/556</td>
<td>58/556</td>
<td>32/556</td>
<td>2/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>105/556</td>
</tr>
<tr>
<td>90 – 100</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>20/556</td>
<td>18/556</td>
<td>3/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>41/556</td>
</tr>
<tr>
<td>100 – 110</td>
<td>0/556</td>
<td>0/556</td>
<td>1/556</td>
<td>2/556</td>
<td>7/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>11/556</td>
</tr>
<tr>
<td>110 – 120</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>5/556</td>
</tr>
<tr>
<td>120+</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>1/556</td>
</tr>
<tr>
<td>Total</td>
<td>3/556</td>
<td>54/556</td>
<td>163/556</td>
<td>220/556</td>
<td>107/556</td>
<td>8/556</td>
<td>1/556</td>
<td>1/556</td>
<td>1/556</td>
<td>556</td>
</tr>
</tbody>
</table>

**IBBENS study:** marginal probabilities

<table>
<thead>
<tr>
<th>Weight</th>
<th>−50</th>
<th>50 – 60</th>
<th>60 – 70</th>
<th>70 – 80</th>
<th>80 – 90</th>
<th>90 – 100</th>
<th>100 – 110</th>
<th>110 – 120</th>
<th>120+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>−50</td>
<td>2/556</td>
<td>12/556</td>
<td>4/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>18/556</td>
</tr>
<tr>
<td>50 – 60</td>
<td>1/556</td>
<td>25/556</td>
<td>50/556</td>
<td>14/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>90/556</td>
</tr>
<tr>
<td>60 – 70</td>
<td>0/556</td>
<td>12/556</td>
<td>54/556</td>
<td>52/556</td>
<td>13/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>132/556</td>
</tr>
<tr>
<td>70 – 80</td>
<td>0/556</td>
<td>5/556</td>
<td>42/556</td>
<td>72/556</td>
<td>34/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>153/556</td>
</tr>
<tr>
<td>80 – 90</td>
<td>0/556</td>
<td>0/556</td>
<td>12/556</td>
<td>58/556</td>
<td>32/556</td>
<td>2/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>105/556</td>
</tr>
<tr>
<td>90 – 100</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>20/556</td>
<td>18/556</td>
<td>3/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>41/556</td>
</tr>
<tr>
<td>100 – 110</td>
<td>0/556</td>
<td>0/556</td>
<td>1/556</td>
<td>2/556</td>
<td>7/556</td>
<td>1/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>11/556</td>
</tr>
<tr>
<td>110 – 120</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>5/556</td>
</tr>
<tr>
<td>120+</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>0/556</td>
<td>1/556</td>
</tr>
<tr>
<td>Total</td>
<td>3/556</td>
<td>54/556</td>
<td>163/556</td>
<td>220/556</td>
<td>107/556</td>
<td>8/556</td>
<td>1/556</td>
<td>1/556</td>
<td>1/556</td>
<td>556</td>
</tr>
</tbody>
</table>
Intermezzo: Joint, marginal and conditional probability

IBBENS study: conditional probabilities

<table>
<thead>
<tr>
<th>Length</th>
<th>150</th>
<th>160</th>
<th>170</th>
<th>180</th>
<th>190</th>
<th>200</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>12/54</td>
<td>25/54</td>
<td>50/90</td>
<td>14/90</td>
<td>0/90</td>
<td>0/90</td>
<td>0/90</td>
</tr>
<tr>
<td>50 – 60</td>
<td>1/90</td>
<td>25/90</td>
<td>25/54</td>
<td>50/90</td>
<td>14/90</td>
<td>0/90</td>
<td>0/90</td>
</tr>
<tr>
<td>60 – 70</td>
<td>12/54</td>
<td>25/90</td>
<td>50/90</td>
<td>14/90</td>
<td>0/90</td>
<td>0/90</td>
<td>0/90</td>
</tr>
<tr>
<td>70 – 80</td>
<td>5/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>80 – 90</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>90 – 100</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>100 – 110</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>110 – 120</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>120 –</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
<td>0/54</td>
</tr>
<tr>
<td>Total</td>
<td>54/54</td>
<td>54/54</td>
<td>54/54</td>
<td>54/54</td>
<td>54/54</td>
<td>54/54</td>
<td>54/54</td>
</tr>
</tbody>
</table>

Bayesian Biostatistics - Piracicaba 2014 322

Intermezzo: Joint, marginal and conditional density

Two (continuous) random variables $X$ and $Y$

- **Joint** density of $X$ and $Y$: density $f(x, y)$
- **Marginal** density of $X$: density $f(x)$
- **Marginal** density of $Y$: density $f(y)$
- **Conditional** density of $X$ given $Y=y$: density $f(x|y)$
- **Conditional** density of $Y$ given $X=x$: density $f(y|x)$
Intermezzo: Joint, marginal and conditional density

IBBENS study: conditional densities

6.2 The Gibbs sampler


- Gelfand and Smith (1990) introduced Gibbs sampling to tackle complex estimation problems in a Bayesian manner

6.2.1 The bivariate Gibbs sampler

Method of Composition:

- $p(\theta_1, \theta_2 \mid y)$ is completely determined by:
  - **marginal** $p(\theta_2 \mid y)$
  - **conditional** $p(\theta_1 \mid \theta_2, y)$

- Split-up yields a simple way to sample from joint distribution

Gibbs sampling:

- $p(\theta_1, \theta_2 \mid y)$ is completely determined by:
  - **conditional** $p(\theta_2 \mid \theta_1, y)$
  - **conditional** $p(\theta_1 \mid \theta_2, y)$

- Property yields another simple way to sample from joint distribution:
  - Take starting values $\theta_1^0$ and $\theta_2^0$ (only 1 is needed)
  - Given $\theta_1^k$ and $\theta_2^k$ at iteration $k$, generate the $(k+1)$-th value according to iterative scheme:
    1. Sample $\theta_1^{k+1}$ from $p(\theta_1 \mid \theta_2^k, y)$
    2. Sample $\theta_2^{k+1}$ from $p(\theta_2 \mid \theta_1^{k+1}, y)$
Result of Gibbs sampling:

- Chain of vectors: \( \mathbf{\theta}^k = (\theta_1^k, \theta_2^k)^T, k = 1, 2, \ldots \)
  - Consists of dependent elements
  - Markov property: \( p(\mathbf{\theta}^{(k+1)} | \mathbf{\theta}^k, \mathbf{\theta}^{(k-1)}, \ldots, y) = p(\mathbf{\theta}^{(k+1)} | \mathbf{\theta}^k, y) \)
- Chain depends on starting value + initial portion/burn-in must be discarded
- Under mild conditions: sample from the posterior distribution = target distribution
  \[ \Rightarrow \text{From } k_0 \text{ on: summary measures calculated from the chain consistently estimate the true posterior measures} \]

Gibbs sampling is called a Markov chain Monte Carlo method

Example VI.1: SAP study – Gibbs sampling the posterior with NI priors

- Example IV.5: sampling from posterior distribution of the normal likelihood based on 250 \( \alpha \phi \) measurements of ‘healthy’ patients with NI prior for both parameters
- Now using Gibbs sampler based on \( y = 100/\sqrt{\alpha \phi} \)
- Determine two conditional distributions:
  1. \( p(\mu | \sigma^2, y) = N(\bar{y}, \sigma^2/n) \)
  2. \( p(\sigma^2 | \mu, y) = Inv - \chi^2(n, s_{\mu}^2) \) with \( s_{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 \)
- Iterative procedure: At iteration \((k + 1)\)
  1. Sample \( \mu^{(k+1)} \) from \( N(\bar{y}, (\sigma^2)^{k}/n) \)
  2. Sample \( (\sigma^2)^{(k+1)} \) from \( Inv - \chi^2(n, s_{\sigma^2}^2) \)

Gibbs sampling path and sample from joint posterior:

- Zigzag pattern in the \((\mu, \sigma^2)\)-plane
- 1 complete step = 2 substeps (blue=genuine element)
- Burn-in = 500, total chain = 1,500
Example VI.2: Sampling from a discrete $\times$ continuous distribution

- Joint distribution: $f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}$
  - $x$ a discrete random variable taking values in $\{0, 1, \ldots, n\}$
  - $y$ a random variable on the unit interval
  - $\alpha, \beta > 0$ parameters
- Question: $f(x)$?

Example VI.3: SAP study – Gibbs sampling the posterior with l priors

- Example VI.1: now with independent informative priors (semi-conjugate prior)
  - $\mu \sim N(\mu_0, \sigma^2_0)$
  - $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \tau^2_0)$
- Posterior:
  $$p(\mu, \sigma^2 | y) \propto \frac{1}{\sigma_0} e^{-\frac{1}{2\sigma_0}(\mu-\mu_0)^2} \times (\sigma^2)^{-\frac{\nu_0}{2} - 1} e^{-\sigma_0^2 \nu_0 \tau^2 / 2\sigma^2} \times \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2}$$
  $$\propto \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} e^{-\frac{1}{2\sigma^0_0}(\mu-\mu_0)^2} (\sigma^2)^{-\frac{\nu_0 + n + 1}{2} - 1} e^{-\sigma_0^2 \nu_0 \tau^2 / 2\sigma^2}$$
Conditional distributions:

- Determine two conditional distributions:
  1. \(p(\mu \mid \sigma^2, y)\):
     \[
     p(\mu \mid \sigma^2, y) = \prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \frac{1}{\sqrt{2\pi\sigma^2}} (\mu, \sigma^2) \]
   2. \(p(\sigma^2 \mid \mu, y)\):
     \[
     p(\sigma^2 \mid \mu, y) = \text{Inv-}\chi^2(\nu_0 + n, \sum_{i=1}^n (y_i - \mu)^2 + \nu_0 \tau_0^2 / \nu_0 + n)
     \]

- Iterative procedure: At iteration \((k + 1)\)
  1. Sample \(\mu^{(k+1)}\) from \((\bar{\mu}, \sigma^2)^k\)
  2. Sample \((\sigma^2)^{(k+1)}\) from \(\text{Inv-}\chi^2(\nu_0 + n, \sum_{i=1}^n (y_i - \mu)^2 + \nu_0 \tau_0^2 / \nu_0 + n)\)

### 6.2.2 The general Gibbs sampler

Starting position \(\theta^0 = (\theta_1^0, \ldots, \theta_d^0)^T\)

Multivariate version of the Gibbs sampler:

**Iteration \((k + 1)\):**

1. Sample \(\theta_1^{(k+1)}\) from \(p(\theta_1 \mid \theta_1^{(k+1)}, \ldots, \theta_d^{(k+1)}, y)\)
2. Sample \(\theta_2^{(k+1)}\) from \(p(\theta_2 \mid \theta_1^{(k+1)}, \theta_2^{(k+1)}, \ldots, \theta_d^{(k+1)}, y)\)
3. Sample \(\theta_3^{(k+1)}\) from \(p(\theta_3 \mid \theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3^{(k+1)}, y)\)
4. Sample \(\theta_d^{(k+1)}\) from \(p(\theta_d \mid \theta_1^{(k+1)}, \ldots, \theta_{(d-1)}^{(k+1)}, y)\)

- Full conditional distributions: \(p(\theta_j \mid \theta_1^k, \ldots, \theta_{(j-1)}^k, \theta_{(j+1)}^k, \ldots, \theta_{(d-1)}^k, \theta_d^k, y)\)
- Also called: full conditionals
- Under mild regularity conditions:
  \(\theta^k, \theta^{(k+1)}, \ldots\) ultimately are observations from the posterior distribution

With the help of advanced sampling algorithms (AR, ARS, ARMS, etc), sampling the full conditionals is done based on the prior \(\times\) likelihood
Example VI.4: British coal mining disasters data

- British coal mining disasters data set: # severe accidents in British coal mines from 1851 to 1962
- Decrease in frequency of disasters from year 40 (+ 1850) onwards?

Statistical model:

- Likelihood: Poisson process with a change point at $k$
  
  \[ y_i \sim \text{Poisson}(\theta) \text{ for } i = 1, \ldots, k \]
  
  \[ y_i \sim \text{Poisson}(\lambda) \text{ for } i = k + 1, \ldots, n \quad (n = 112) \]

- Priors
  
  \[ \theta: \text{Gamma}(a_1, b_1), \quad (a_1 \text{ constant}, \ b_1 \text{ parameter}) \]
  
  \[ \lambda: \text{Gamma}(a_2, b_2), \quad (a_2 \text{ constant}, \ b_2 \text{ parameter}) \]
  
  \[ k: \quad p(k) = 1/n \]
  
  \[ b_1: \text{Gamma}(c_1, d_1), \quad (c_1, d_1 \text{ constants}) \]
  
  \[ b_2: \text{Gamma}(c_2, d_2), \quad (c_2, d_2 \text{ constants}) \]

Full conditionals:

- $p(\theta \mid y, \lambda, b_1, b_2, k) = \text{Gamma}(a_1 + \sum_{i=1}^{k} y_i, k + b_1)$
- $p(\lambda \mid y, \theta, b_1, b_2, k) = \text{Gamma}(a_2 + \sum_{i=k+1}^{n} y_i, n - k + b_2)$
- $p(b_1 \mid y, \theta, b_2, k) = \text{Gamma}(a_1 + c_1, \theta + d_1)$
- $p(b_2 \mid y, \theta, b_1, k) = \text{Gamma}(a_2 + c_2, \lambda + d_2)$
- $p(k \mid y, \theta, b_1, b_2) = \frac{\pi(y \mid k, \theta, \lambda)}{\sum_{j=1}^{n} \pi(y \mid j, \theta, \lambda)}$

Posterior distributions:

- Posterior mode of $k$: 1891
- Posterior mean for $\theta/\lambda = 3.42$ with 95% CI = [2.48, 4.59]

\[ a_1 = a_2 = 0.5, \quad c_1 = c_2 = 0, \quad d_1 = d_2 = 1 \]
Note:

• In most published analyses of this data set $b_1$ and $b_2$ are given inverse gamma priors. The full conditionals are then also inverse gamma.

• The results are almost the same ⇒ our analysis is a sensitivity analysis of the analyses seen in the literature.

• Despite the classical full conditionals, the WinBUGS/OpenBUGS sampler for $\theta$ and $\lambda$ are not standard gamma but rather a slice sampler. See Exercise 8.10.

Example VI.5: Osteoporosis study – Using the Gibbs sampler

Bayesian linear regression model with NI priors:

\[
\text{Regression model: } \text{tbbmc}_i = \beta_0 + \beta_1 \text{bmii}_i + \epsilon_i \quad (i = 1, \ldots, n = 234)
\]

\[
\text{Priors: } \pi(\beta_0, \beta_1, \sigma^2) \propto \sigma^{-2}
\]

\[
\text{Notation: } \mathbf{y} = (\text{tbbmc}_1, \ldots, \text{tbbmc}_{234})^T, \mathbf{x} = (\text{bmi}_1, \ldots, \text{bmi}_{234})^T
\]

Full conditionals:

\[
p(\sigma^2 | \beta_0, \beta_1, \mathbf{y}) = \text{Inv-}\chi^2(n, s^2_\beta)
\]

\[
p(\beta_0 | \sigma^2, \beta_1, \mathbf{y}) = \mathcal{N}(r_{\beta_0}, \sigma^2/n)
\]

\[
p(\beta_1 | \sigma^2, \beta_0, \mathbf{y}) = \mathcal{N}(r_{\beta_1}, \sigma^2/\mathbf{x}^T\mathbf{x})
\]

with

\[
s^2_\beta = \frac{1}{n} \sum (y_i - \beta_0 - \beta_1 x_i)^2
\]

\[
r_{\beta_0} = \frac{1}{n} \sum (y_i - \beta_1 x_i)
\]

\[
r_{\beta_1} = \frac{1}{n} \sum (y_i - \beta_0) x_i/\mathbf{x}^T\mathbf{x}
\]

Comparison with Method of Composition:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method of Composition</th>
<th>Gibbs sampler</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%  25%  50%  75%  97.5% Mean  SD</td>
<td>2.5%  25%  50%  75%  97.5% Mean  SD</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>0.57  0.74  0.81  0.89  1.05  0.81  0.12</td>
<td>0.67  0.77  0.84  0.91  1.10  0.77  0.11</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.032  0.038  0.040  0.043  0.049  0.040  0.004</td>
<td>0.030  0.036  0.040  0.042  0.046  0.039  0.0041</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.069  0.078  0.083  0.088  0.100  0.083  0.008</td>
<td>0.069  0.077  0.083  0.088  0.099  0.083  0.0077</td>
</tr>
</tbody>
</table>

○ Method of Composition = 1,000 independently sampled values

○ Gibbs sampler: burn-in = 500, total chain = 1,500
Trace plot from Gibbs sampler:

![Trace plot](image)

Trace versus index plot:

Comparison of index plot with trace plot shows:

- $\sigma^2$: index plot and trace plot similar $\Rightarrow$ (almost) independent sampling
- $\beta$: trace plot shows slow mixing $\Rightarrow$ quite dependent sampling

$\Rightarrow$ Method of Composition and Gibbs sampling: similar posterior measures of $\sigma^2$

$\Rightarrow$ Method of Composition and Gibbs sampling: less similar posterior measures of $\beta$

Autocorrelation:

- Autocorrelation of lag 1: correlation of $\beta_1^k$ with $\beta_1^{(k-1)}$ ($k=1, \ldots$)
- Autocorrelation of lag 2: correlation of $\beta_1^k$ with $\beta_1^{(k-2)}$ ($k=1, \ldots$)
- Autocorrelation of lag m: correlation of $\beta_1^k$ with $\beta_1^{(k-m)}$ ($k=1, \ldots$)

High autocorrelation:

- burn-in part is larger $\Rightarrow$ takes longer to forget initial positions
- remaining part needs to be longer to obtain stable posterior measures

6.2.3 Remarks

- Full conditionals determine joint distribution
- Generate joint distribution from full conditionals
- Transition kernel
6.2.4 Review of Gibbs sampling approaches

Sampling the full conditionals is done via different algorithms depending on:

▷ Shape of full conditional (classical versus general purpose algorithm)
▷ Preference of software developer:
  ◦ SAS® procedures GENMOD, LIFEREG and PHREG: ARMS algorithm
  ◦ WinBUGS: variety of samplers

Several versions of the basic Gibbs sampler:

▷ Deterministic- or systematic scan Gibbs sampler: \( d \) dims visited in fixed order
▷ Block Gibbs sampler: \( d \) dims split up into \( m \) blocks of parameters and Gibbs sampler applied to blocks

Review of Gibbs sampling approaches – The block Gibbs sampler

Block Gibbs sampler:

▷ Normal linear regression
  \( p(\sigma^2 \mid \beta_0, \beta_1, y) \)
  \( p(\beta_0, \beta_1 \mid \sigma^2, y) \)

• May speed up considerably convergence, at the expense of more computational time needed at each iteration
• WinBUGS: blocking option on
• SAS® procedure MCMC: allows the user to specify the blocks

6.3 The Metropolis(-Hastings) algorithm

Metropolis-Hastings (MH) algorithm = general Markov chain Monte Carlo technique to sample from the posterior distribution but does not require full conditionals

• Special case: Metropolis algorithm proposed by Metropolis in 1953
• General case: Metropolis-Hastings algorithm proposed by Hastings in 1970
• Became popular only after introduction of Gelfand & Smith’s paper (1990)
• Further generalization: Reversible Jump MCMC algorithm by Green (1995)

6.3.1 The Metropolis algorithm

Sketch of algorithm:

• New positions are proposed by a proposal density \( q \)

• Proposed positions will be:
  ▷ Accepted:
    ◦ Proposed location has higher posterior probability: with probability \( 1 \)
    ◦ Otherwise: with probability proportional to ratio of posterior probabilities
  ▷ Rejected:
    ◦ Otherwise

• Algorithm satisfies again Markov property \( \Rightarrow \) MCMC algorithm
• Similarity with AR algorithm
**Metropolis algorithm:**

Chain is at $\theta^k \Rightarrow$ Metropolis algorithm samples value $\theta^{(k+1)}$ as follows:

1. Sample a candidate $\tilde{\theta}$ from the symmetric proposal density $q(\tilde{\theta} | \theta)$, with $\theta = \theta^k$.

2. The next value $\theta^{(k+1)}$ will be equal to:
   - $\tilde{\theta}$ with probability $\alpha(\theta^k, \tilde{\theta})$ (accept proposal),
   - $\theta^k$ otherwise (reject proposal),

   with
   $$
   \alpha(\theta^k, \tilde{\theta}) = \min \left( r = \frac{p(\tilde{\theta} | y)}{p(\theta^k | y)}, 1 \right)
   $$

Function $\alpha(\theta^k, \tilde{\theta}) = \text{probability of a move}$

---

**Example VI.7: SAP study – Metropolis algorithm for NI prior case**

Settings as in Example VI.1, now apply Metropolis algorithm:

- Proposal density: $\mathcal{N}(\theta^k, \Sigma)$ with $\theta^k = (\mu^k, (\sigma^2)^k)^T$ and $\Sigma = \text{diag}(0.03, 0.03)$

- Jumps to any location in the $(\mu, \sigma^2)$-plane
- Burn-in = 500, total chain = 1,500

**MH-sampling:**

The MH algorithm only requires the product of the prior and the likelihood to sample from the posterior
Marginal posterior distributions:

- Acceptance rate = 40%
- Burn-in = 500, total chain = 1,500

Trace plots:

- Accepted moves = blue color, rejected moves = red color

Second choice of proposal density:

- Proposal density: \( N(\theta^k, \Sigma) \) with \( \theta^k = (\mu^k, (\sigma^2)^k)^T \) and \( \Sigma = \text{diag}(0.001, 0.001) \)

- Acceptance rate = 84%
- Poor approximation of true distribution
Problem:

What should be the acceptance rate for a good Metropolis algorithm?

From theoretical work + simulations:

- Acceptance rate: 45% for $d=1$ and $\approx 24\%$ for $d>1$
6.3.3 Remarks

- The Gibbs sampler is a special case of the Metropolis-Hastings algorithm, but Gibbs sampler is still treated differently
- The transition kernel of the MH-algorithm
- The reversibility condition
- Difference with AR algorithm

6.5. Choice of the sampler

Choice of the sampler depends on a variety of considerations

Example VI.9: Caries study – MCMC approaches for logistic regression

Subset of \( n = 500 \) children of the Signal-Tandmobiel\(^\circ\) study at 1st examination:

- Research questions:
  - Have girls a different risk for developing caries experience (CE) than boys (gender) in the first year of primary school?
  - Is there an east-west gradient (x-coordinate) in CE?
- Bayesian model: logistic regression + \( N(0, 100^2) \) priors for regression coefficients
- No standard full conditionals
- Three algorithms:
  - Self-written R program: evaluate full conditionals on a grid + ICDF-method
  - WinBUGS program: multivariate MH algorithm (blocking mode on)
  - SAS\(^\circ\) procedure MCMC: Random-Walk MH algorithm

<table>
<thead>
<tr>
<th>Program</th>
<th>Parameter</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>MCSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>Intercept</td>
<td>-0.5900</td>
<td>0.2800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gender</td>
<td>-0.0379</td>
<td>0.1810</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>x-coord</td>
<td>0.0052</td>
<td>0.0017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Intercept</td>
<td>-0.5880</td>
<td>0.2840</td>
<td>-0.5860</td>
<td>0.0104</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gender</td>
<td>-0.0516</td>
<td>0.1850</td>
<td>-0.0578</td>
<td>0.0071</td>
<td></td>
</tr>
<tr>
<td></td>
<td>x-coord</td>
<td>0.0052</td>
<td>0.0017</td>
<td>0.0052</td>
<td>6.621E-5</td>
<td></td>
</tr>
<tr>
<td>WinBUGS</td>
<td>Intercept</td>
<td>-0.5800</td>
<td>0.2810</td>
<td>-0.5730</td>
<td>0.0094</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gender</td>
<td>-0.0379</td>
<td>0.1770</td>
<td>-0.0324</td>
<td>0.0060</td>
<td></td>
</tr>
<tr>
<td></td>
<td>x-coord</td>
<td>0.0052</td>
<td>0.0018</td>
<td>0.0053</td>
<td>5.901E-5</td>
<td></td>
</tr>
<tr>
<td>SAS(^\circ)</td>
<td>Intercept</td>
<td>-0.6530</td>
<td>0.2600</td>
<td>-0.6450</td>
<td>0.0317</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gender</td>
<td>-0.0319</td>
<td>0.1950</td>
<td>-0.0443</td>
<td>0.0208</td>
<td></td>
</tr>
<tr>
<td></td>
<td>x-coord</td>
<td>0.0055</td>
<td>0.0016</td>
<td>0.0055</td>
<td>0.00016</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions:

- Posterior means/medians of the three samplers are close (to the MLE)
- Precision with which the posterior mean was determined (high precision = low MCSE) differs considerably
- The clinical conclusion was the same

$\Rightarrow$ Samplers may have quite a different efficiency

Take home messages

- The two MCMC approaches allow fitting basically any proposed model
- There is no free lunch: computation time can be MUCH longer than with likelihood approaches
- The choice between Gibbs sampling and the Metropolis-Hastings approach depends on computational and practical considerations

Chapter 7
Assessing and improving convergence of the Markov chain

Questions:

$\triangleright$ Are we getting the right answer?
$\triangleright$ Can we get the answer quicker?

7.1 Introduction

- MCMC sampling is powerful, but comes with a cost: dependent sampling + checking convergence is not easy:
  - Convergence theorems do not tell us when convergence will occur
  - In this chapter: graphical + formal diagnostics to assess convergence
- Acceleration techniques to speed up MCMC sampling procedure
- Data augmentation as Bayesian generalization of EM algorithm
7.2 Assessing convergence of a Markov chain

7.2.1 Definition of convergence for a Markov chain

Loose definition:

With increasing number of iterations \( k \to \infty \), the distribution of \( \theta^k, p_k(\theta) \), converges to the target distribution \( p(\theta | y) \).

In practice, convergence means:

- Histogram of \( \theta^k \) remains the same along the chain
- Summary statistics of \( \theta^k \) remain the same along the chain

Approaches

- Theoretical research
  - Establish conditions that ensure convergence, but in general these theoretical results cannot be used in practice
- Two types of practical procedures to check convergence:
  - Checking stationarity: from which iteration \( k_0 \) is the chain sampling from the posterior distribution (assessing burn-in part of the Markov chain).
  - Checking accuracy: verify that the posterior summary measures are computed with the desired accuracy
- Most practical convergence diagnostics (graphical and formal) appeal to the stationarity property of a converged chain
- Often done: base the summary measures on the converged part of the chain

7.2.2 Checking convergence of the Markov chain

Here we look at:

- Convergence diagnostics that are implemented in
  - WinBUGS
  - CODA, BOA
- Graphical and formal tests
- Illustrations using the
  - Osteoporosis study (chain of size 1,500)
7.2.3 Graphical approaches to assess convergence

Trace plot:

- A simple exploration of the trace plot gives a good impression of the convergence
- Univariate trace plots (each parameter separately) or multivariate trace plots (evaluating parameters jointly) are useful
- Multivariate trace plot: monitor the (log) of likelihood or log of posterior density:
  - WinBUGS: automatically created variable deviance
  - Bayesian SAS® procedures: default variable LogPost
- Thick pen test: trace plot appears as a horizontal strip when stationarity
- Trace plot evaluates the mixing rate of the chain

7.2.4 Graphical approaches to assess convergence

Trace plot:

- A simple exploration of the trace plot gives a good impression of the convergence
- Univariate trace plots (each parameter separately) or multivariate trace plots (evaluating parameters jointly) are useful
- Multivariate trace plot: monitor the (log) of likelihood or log of posterior density:
  - WinBUGS: automatically created variable deviance
  - Bayesian SAS® procedures: default variable LogPost
- Thick pen test: trace plot appears as a horizontal strip when stationarity
- Trace plot evaluates the mixing rate of the chain
Autocorrelation plot:

- Autocorrelation of lag $m$ ($\rho_m$) \(corr(\theta^k, \theta^{k+m})\) (for $k = 1, \ldots$) estimated by
  - Pearson correlation
  - Time series approach

- Autocorrelation function (ACF): function $m \mapsto \hat{\rho}_m$ ($m = 0, 1, \ldots$)

- Autocorrelation plot: graphical representation of ACF and indicates
  - Mixing rate
  - Minimum number of iterations for the chain to 'forget' its starting position

- Autocorrelation plot is a NOT a convergence diagnostic. Except for the burn-in phase, the ACF plot will not change anymore when further sampling

Running-mean plot:

- Running mean (ergodic mean) $\bar{\theta}$: mean of sampled values up to iteration $k$
  - Checks stationarity for $k > k_0$ of $p_k(\theta)$ via the mean
  - Initial variability of the running-mean plot is always relatively high
  - Stabilizes with increasing $k$ in case of stationarity
Q-Q plot:

- **Stationarity**: distribution of the chain is stable irrespective of part
- Q-Q plot: 1st half of chain (X-axis) versus 2nd half of chain on (Y-axis)
- **Non-stationarity**: Q-Q graph deviating from the bisecting line

Cross-correlation plot:

- **Cross-correlation** of $\theta_1$ with $\theta_2$: correlation between $\theta^k_1$ with $\theta^k_2$ ($k = 1, \ldots, n$)
- **Cross-correlation plot**: scatterplot of $\theta^k_1$ versus $\theta^k_2$
- Useful in case of convergence problems to indicate which model parameters are strongly related and thus whether model is **overspecified**
7.2.5 Formal approaches to assess convergence

Here:

- Four formal convergence diagnostics:
  - Geweke diagnostic: single chain
  - Heidelberger-Welch diagnostic: single chain
  - Raftery-Lewis diagnostic: single chain
  - Brooks-Gelman-Rubin diagnostic: multiple chains

- Illustrations based on the
  - Osteoporosis study (chain of size 1,500)

Geweke diagnostic

- Acts on a single chain $\{\theta^k\}_{k=1}^n$
- Checks only stationarity $\Rightarrow$ looks for $k_0$
- Compares means of (10%) early $\leftrightarrow$ (50%) late part of chain using a $t$-like test

- In addition, there is a dynamic version of the test:
  - Cut off repeatedly $x\%$ of the chain, from the beginning
  - Each time compute Geweke test

- Test cannot be done in WinBUGS, but should be done in R accessing the sampled values of the chain produced by WinBUGS

Geweke diagnostic on osteoporosis regression analysis

- Based on chain of 1,500 iterations
- R program geweke.diag (CODA)
- With standard settings of 10% (1st part) & 50% (2nd part): numerical problems
- Early part = 20%, Geweke diagnostic for $\beta_1$: $Z = 0.057$ and for $\sigma^2$: $Z = 2.00$
- Dynamic version of the Geweke diagnostic (with $K = 20$) next page, results:
  - $\beta_1$: majority of $Z$-values are outside [-1.96,1.96] $\Rightarrow$ non-stationarity
  - $\log(\text{posterior})$: non-stationary
  - $\sigma^2$: all values except first inside [-1.96,1.96]
BGR diagnostic:

- Version 1: Gelman and Rubin ANOVA diagnostic
- Version 2: Brooks-Gelman-Rubin interval diagnostic
- General term: Brooks-Gelman-Rubin (BGR) diagnostic
- Global and dynamic diagnostic
- Implementations in WinBUGS, CODA and BOA

For both versions:

- Take \( M \) widely dispersed starting points \( \theta^0_m (m = 1, \ldots, M) \)
- \( M \) parallel chains \( (\theta^k_m)_{m} (m = 1, \ldots, M) \) are run for \( 2n \) iterations
- The first \( n \) iterations are discarded and regarded as burn-in

Version 1:

- Within chains: chain mean \( \bar{\theta}_m = (1/n) \sum_{k=1}^{n} \theta^k_m (m = 1, \ldots, M) \)
- Across chains: overall mean \( \bar{\theta} = (1/M) \sum_m \bar{\theta}_m \)

ANOVA idea:

- \( W = \frac{1}{M} \sum_{m=1}^{M} s_m^2 \), with \( s_m^2 = \frac{1}{n} \sum_{k=1}^{n} (\theta^k_m - \bar{\theta}_m)^2 \)
- \( B = \frac{n}{M-1} \sum_{m=1}^{M} (\bar{\theta}_m - \bar{\theta})^2 \)

- Two estimates for \( \text{var}(\theta^k | y) \):
  - Under stationarity: \( W \) and \( \tilde{W} \equiv \text{var}(\theta^k | y) = \frac{n-1}{n} W + \frac{B}{n} \) good estimates
  - Non-stationarity: \( W \) too small and \( \tilde{W} \) too large

- Convergence diagnostic: \( \hat{R} = \frac{\tilde{W}}{W} \)

Version 1:

- \( \hat{R} \): the estimated potential scale reduction factor (PSRF)
- \( \hat{R}_{c} = (\hat{d} + 3)/(\hat{d} + 1) \hat{R} \), with \( \hat{d} = 2\hat{V}/\text{var}(\hat{V}) \) corrected for sampling variability
- 97.5\% upper confidence bound of \( \hat{R}_{c} \) is added

In practice:

- Continue sampling until \( \hat{R}_{c} < 1.1 \) or 1.2
- Monitor \( \hat{V} \) and \( W \), both must stabilize
- When convergence: take summary measures from 2nd half of the total chain
Version 1:
CODA/BOA: Dynamic/graphical version

- \( M \) chains are divided into batches of length \( b \)
- \( \hat{V}^{1/2}(s), W^{1/2}(s) \) and \( \hat{R}_c(s) \) are calculated based upon 2nd half of the (cumulative) chains of length \( 2sb \), for \( s = 1, \ldots, \lceil n/b \rceil \)
- 97.5% upper pointwise confidence bound is added

BGR diagnostic:

Version 2:

- \( \alpha = 0.20 \)
- Chain divided in cumulative subchains based on iterations 1-100, 1-200, 1-300, etc...
- For each part compute on the 2nd half of the subchain:
  - (a) \( \hat{R}_I \) (red)
  - (b) \( \hat{V}_I / \max(\hat{V}_I) \) (green)
  - (c) \( W_I / \max(W_I) \) (blue)
- Three curves are produced, which should be monitored
- When convergence: take summary measures from 2nd half of the total chain

BGR diagnostic on osteoporosis regression analysis

- Based on chain of 1,500 iterations
- 8 overdispersed starting values from a classical linear regression analysis
- First version (CODA):
  - \( \beta_1: \hat{R}_c = 1.06 \) with 97.5% upper bound equal to 1.12
  - \( \sigma^2: \hat{R}_c = 1.00 \) with 97.5% upper bound is equal to 1
- Second version (WinBUGS):
  - \( \beta_1: \hat{R}_I \) dropped quickly below 1.1 with increasing iterations and varied around 1.03 from iteration 600 onwards
  - \( \sigma^2: \hat{R}_I \) was quickly around 1.003 and the curves stabilized almost immediately
Conclusion about convergence of osteoporosis regression analysis

- A chain of length 1,500 is inadequate for achieving convergence of the regression parameters
- Convergence was attained with 15,000 iterations
- For accurately estimating certain quantiles 15,000 iterations were not enough
- With 8 chains the posterior distribution was rather well estimated after 1,500 iterations each

7.2.6 Computing the Monte Carlo standard error

Aim: estimate the Monte Carlo standard error (MCSE) of the posterior mean \( \bar{\theta} \)

- MCMC algorithms produce dependent random variables
- \( s/\sqrt{n} \) (\( s \) posterior SD and \( n \) the length of the chain) cannot be used
- Two approaches:
  - Time-series approach (SAS MCMC/CODA/BOA)
  - Method of batch means (WinBUGS/CODA/BOA)
- Chain must have converged!
- Related to this: effective sample size
  - Size of chain if it were generated from independent sampling
**MCSE for osteoporosis regression analysis**

- Chain: total size of 40,000 iterations with 20,000 burn-in iterations

- **MCSE for \( \beta_1 \):**
  - WinBUGS: MCSE = 2.163E-4 (batch means)
  - SAS\(^{®}\): MCSE = 6.270E-4 (time series)
  - CODA: MCSE = 1.788E-4 (time series)
  - CODA: MCSE = – (batch means), computational difficulties
  - ESS: 312 (CODA), 22.9 (SAS)

- **Extreme slow convergence with SAS PROC MCMC:** stationarity only after 500,000 burn-in iterations + 500,000 iterations

**7.2.7 Practical experience with the formal diagnostic procedures**

- Geweke diagnostic:
  - Popular, but dynamic version is needed
  - Computational problems on some occasions

- Gelman and Rubin diagnostics:
  - Animated discussion whether to use multiple chains to assess convergence
  - BGR diagnostic should be part of each test for convergence
  - Problem: how to pick overdispersed and realistic starting values

**It is impossible to prove convergence in practice:**

- Example Geyer (1992)
- AZT RCT example
7.3 Accelerating convergence

Aim of acceleration approaches: lower autocorrelations

- Review of approaches:
  - Choosing better starting values
  - Transforming the variables: centering + standardisation
  - Blocking: processing several parameters at the same time
  - Algorithms that suppress the purely random behavior of the MCMC sample: over-relaxation
  - Reparameterization of the parameters: write your model in a different manner

- More sophisticated MCMC algorithms may be required
- Thinning: look at part of chain, is not an acceleration technique

---

Effect of acceleration on osteoporosis regression analysis

(a) Original
(b) Block
(c) Overrelaxation
(d) Centered bmi
7.4 Practical guidelines for assessing and accelerating convergence

Practical guidelines:

- Start with a pilot run to detect quickly some trivial problems. Multiple (3-5) chains often highlight problems faster.
- Start with a more formal assessment of convergence only when the trace plots show good mixing.
- When convergence is slow, use acceleration tricks.
- When convergence fails despite acceleration tricks: let the sampler run longer/apply thinning/change software/... write your own sampler.
- Ensure that MCError at most 5% of the posterior standard deviation.
- Many parameters: choose a random selection of relevant parameters to monitor.

7.5 Data augmentation

Data augmentation (DA): augment the observed data ($y$) with fictive data ($z$).

Aim: DA technique may simplify the estimation procedure considerably.

- Classical DA technique: EM algorithm
- Fictive data are often called missing data.
- DA technique consists of 2 parts:
  - Part 1. Assume that missing data $z$ (auxiliary variables) are available to give complete(d) data $w = \{y, z\}$
  - Part 2. Take into account that part of the (complete) vector $w$ ($z$) is not available.
- Bayesian DA technique (Tanner & Wong, 1987): Bayesian extension of EM algorithm, replacing maximization with sampling.

Bayesian data augmentation technique:

- Sampling from $p(\theta \mid y, z)$ may be easier than from $p(\theta \mid y)$
- Replace E-step by: sampling the missing data $z$ from $p(z \mid \theta, y)$
- Replace M-step by: sampling $\theta$ from $p(\theta \mid y, z)$

$\Rightarrow$ DA approach: repeatedly sampling from $p(\theta \mid y, z)$ and from $p(z \mid \theta, y)$

$\Rightarrow$ Block Gibbs sampler.

Applications:

- Genuine missing data
- Censored/misclassified observations
- Mixture models
- Hierarchical models
**Example VII.6 – Genetic linkage model**

If two factors are linked with a recombination fraction $\pi$, the intercrosses $Ab/aB \times Ab/aB$ (repulsion) result in the following probabilities ($\theta = \pi^2$):

(a) for $AB$: $0.5 + \theta/4$
(b) for $Ab$: $(1 - \theta)/4$
(c) for $aB$: $(1 - \theta)/4$
(d) for $ab$: $\theta/4$

$n$ subjects classified into $(AB, Ab, aB, ab)$ with $y = (y_1, y_2, y_3, y_4)$ frequencies

$y \sim \text{Mult}(n, (0.5 + \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4))$

Flat prior for $\theta +$ above multinomial likelihood $\Rightarrow$ posterior

$p(\theta \mid y) \propto (2 + \theta)^y_1 (1 - \theta)^y_2 \theta^y_3 \theta^y_4$

Posterior = unimodal function of $\theta$ but has no standard expression

Popular data set introduced by Rao (1973): 197 animals with $y = (125, 18, 20, 34)$

**Likelihood + prior:**

Imagine that: 1st cell was split up into 2 cells with

- Probabilities $1/2$ and $\theta/4$
- Corresponding frequencies $(y_1 - z)$ and $z$
- (Completed) vector of frequencies: $w = (125 - z, z, 18, 20, 34)$

Given $z$: $p(\theta \mid w) = p(\theta \mid y, z) \propto \theta^z (1 - \theta)^{y_2 + y_3} \theta^{y_4}$

Given $\theta$: $p(z \mid \theta, y) = \text{Bin}(y_1, \theta/(2 + \theta))$

These 2 posteriors = full conditionals of posterior obtained based on

- Likelihood: $\text{Mult}(n, (0.5 + \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4))$
- Based on frequencies $(125 - z, z, 18, 20, 34)$
- Priors: flat for $\theta$ and discrete uniform for $z$ on $\{0, 1, \ldots, 125\}$

**Output:**

- 100 iterations, burn-in = 10 initial

**Example VII.7 – Cysticercosis study – Estimating prevalence**

Example V.6: no gold standard available

- Prior information needed to estimate prevalence ($\pi$), SE ($\alpha$) and SP ($\beta$)
- Analysis was done with WinBUGS 1.4.3
- Here we show how data augmentation can simplify the MCMC procedure

$$
\begin{align*}
\text{Test} & \quad \text{Disease (True)} & \quad \text{Observed} \\
+ & \quad - & \quad + & \quad - & \quad \text{Total} \\
+ & \quad \pi \alpha & \quad (1 - \pi)(1 - \beta) & \quad z_1 & \quad y - z_1 & \quad y = 496 \\
- & \quad \pi(1 - \alpha) & \quad (1 - \pi)\beta & \quad z_2 & \quad (n - y) - z_2 & \quad n - y = 372 \\
\text{Total} & \quad \pi & \quad (1 - \pi) & \quad n & \quad n = 868
\end{align*}
$$
Likelihood + Prior:

Priors:
- $\pi$: Beta($\nu_\pi, \eta_\pi$)
- $\alpha$: Beta($\nu_\alpha, \eta_\alpha$)
- $\beta$: Beta($\nu_\beta, \eta_\beta$)

Likelihood: multinomial with probabilities in table

Posterior on completed data:

$$
\begin{align*}
\pi_{\text{z1}} &\sim \text{Bin}\left(y, \frac{\pi \alpha}{\pi \alpha + (1 - \pi)(1 - \beta)}\right) \\
\pi_{\text{z2}} &\sim \text{Bin}\left(n - y, \frac{\pi (1 - \alpha)}{(1 - \alpha)(1 - \pi)\beta}\right) \\
\alpha &\sim \text{Beta}(\nu_\alpha + z_2, \eta_\alpha) \\
\beta &\sim \text{Beta}(n - y - z_2 + \nu_\beta, y - z_1 + \eta_\beta) \\
\pi &\sim \text{Beta}(z_1 + z_2 + \nu_\pi, n - z_1 - z_2 + \eta_\pi)
\end{align*}
$$

Marginal posterior for prevalence: Beta($z_1 + z_2 + \nu_\pi$, $n - z_1 - z_2 + \eta_\pi$)

Full conditionals:

- Observed data: complex
- Completed data: standard

$$
\begin{align*}
z_1 &\mid y, \pi, \alpha, \beta \sim \text{Bin}\left(y, \frac{\pi \alpha}{\pi \alpha + (1 - \pi)(1 - \beta)}\right) \\
z_2 &\mid y, \pi, \alpha, \beta \sim \text{Bin}\left(n - y, \frac{\pi (1 - \alpha)}{(1 - \alpha)(1 - \pi)\beta}\right) \\
\alpha &\sim \text{Beta}(\nu_\alpha + z_2, \eta_\alpha) \\
\beta &\sim \text{Beta}(n - y - z_2 + \nu_\beta, y - z_1 + \eta_\beta) \\
\pi &\sim \text{Beta}(z_1 + z_2 + \nu_\pi, n - z_1 - z_2 + \eta_\pi)
\end{align*}
$$

Identifiability issue:

- Models based on $\pi$, $\alpha$ and $\beta$ and $1 - \pi$, $1 - \alpha$ and $1 - \beta$ are the same
- Only definition of 'diseased' and 'positive test' are interchanged
- If no strong priors are given: Markov chain will swap between values in [0,0.5] and [0.5,1]
- Solution: restrict $\pi$, $\alpha$ and $\beta$ to [0.5,1]
Example VII.8 – Caries study – Analysis of interval-censored data

Signal-Tandmobiel® study: Distribution of emergence of tooth 22 on 500 children

▷ True distribution of the emergence times not known (annual examinations):
  ◦ Left-censored: emergence before 1st examination
  ◦ Right-censored: emergence after last examination
  ◦ Interval-censored: in-between 2 examinations

⇒ Data: \([L_i, R_i]\, (i = 1, \ldots, n)\)

▷ DA approach: assume true emergence times \(y_i \sim N(\mu, \sigma^2)\) \((i = 1, \ldots, n)\) known

▷ Two types of full conditionals:
  ◦ Model parameters \(| y_i \) \((i = 1, \ldots, n)\) (classical distributions)
  ◦ \(y_i | [L_i, R_i]\) (truncated normal distributions)

Output:

Priors:
Vague for \(\mu\) and \(\sigma^2\)

MCMC program:
Self-written R-program
WB-program
Use \(\text{I}(L_i, R_i)\)

10,000 iterations
2,000 burn-in

Results:
• mean prediction
■ [2.5%, 97.5%]

Example VII.9 – MCMC approaches for a probit regression problem

Example VI.9: MCMC approaches for logistic regression model
Here probit regression: \(P(y_i = 1 \mid x_i) = \Phi(x_i^T \beta)\)

▷ Gibbs sampler is logical choice
▷ Making use of DA technique
▷ Discretization idea:
  ◦ \(z_i = x_i^T \beta + \epsilon_i\, (i = 1, \ldots, n)\)
  ◦ \(\epsilon_i \sim N(0, 1)\)
  ◦ \(y_i = 0 (1)\) for \(z_i \leq (> 0)\)
  ◦ \(P(y_i = 1 \mid x_i) = P(z_i > 0) = 1 - \Phi(-x_i^T \beta) = \Phi(x_i^T \beta)\)
Full conditionals:

\[ \beta \text{ given } z: \text{ Conditional on } z = (z_1, z_2, \ldots, z_n) \text{ with design matrix } Z \]

\[ \circ \beta \sim N(\beta_{LS}, (X^T X)^{-1}), \text{ with design matrices of covariates} \]

\[ \Rightarrow \text{ Bayesian regression analysis with } \sigma^2 = 1 \text{ and } \beta_{LS} = (X^T X)^{-1} X^T Z \]

\[ z \text{ given } \beta: \text{ Full conditional of } z_i \text{ on } \beta \text{ and } y_i \]

\[ \circ y_i = 0: \text{ proportional to } \phi_{x_i^T \beta, 1}(z) I(z \leq 0) \]

\[ \circ y_i = 1: \text{ proportional to } \phi_{x_i^T \beta, 1}(z) I(z > 0) \]

\[ \Rightarrow \text{ Sampling from truncated normals} \]

Chapter 8

Software

Aims:

\[ \triangleright \text{ Introduction to the use of WinBUGS (and related software)} \]

\[ \triangleright \text{ Review of other Bayesian software} \]

Take home messages

- MCMC algorithms come with a cost:
  \[ \triangleright \text{ estimation process takes often (much) longer time than classical ML} \]
  \[ \triangleright \text{ more difficult to check if correct estimates are obtained} \]

- Testing convergence:
  \[ \triangleright \text{ strictly speaking, we are never sure} \]
  \[ \triangleright \text{ handful tests (graphical + formal) are used} \]
  \[ \triangleright \text{ when many models need to be fit, then often trace plots are the only source for testing} \]

- WinBUGS + MCMC techniques implied a revolution in use of Bayesian methods

- WinBUGS is the standard software for many Bayesians

- SAS® has developed Bayesian software, but will not be discussed here

- Many R programs for Bayesian methods have been developed, both related to WinBUGS/JAGS/... and stand-alone
In this chapter:

- **WinBUGS, OpenBUGS + related software**

- **Other sources:**
  - WinBUGS manual, WinBUGS Examples manuals I and II
  - SAS®-STAT manual

8.1 WinBUGS and related software

- **WinBUGS**: Windows version of Bayesian inference using Gibbs Sampling (BUGS)
  - Start: 1989 in MRC Biostatistics at Cambridge with BUGS
  - Final version = 1.4.3

- **OpenBUGS** = open source version of WinBUGS

- Both packages **freely available**

- Additional and related software: CODA, BOA, etc.

8.1.1 A first analysis

Input to WinBUGS: `.odc` file

- **Text file**: with program, data and initial values for a MCMC run
- **Compound document** contains various types of information: text, tables, formulae, plots, graphs, etc.

To access `.odc` file:

- Start WinBUGS
- Click on File
- Click on Open...

WinBUGS programming language:

- Similar but not identical to R
- chapter 8 osteoporosis.odc

WinBUGS program on osteoporosis simple linear regression analysis

```plaintext
model
{
  for (i in 1:N)
  {
    t客[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1*bmi[i]
  }

  sigma2 <- 1/tau
  sigma ~ sqrt(sigma2)

  beta0 ~ dnorm(0,1.0E-6)
  beta1 ~ dnorm(0,1.0E-6)
  tau ~ dgamma(1.0E-3,1.0E-3)
}

# data
list(t客=c(1.798, 2.588, 2.325, 2.236, 1.925, 2.304, ....),
     bmi=c(23.61, 30.48, 27.18, 34.68, 26.72, 25.78, 29.24, ....),
     N=234)

# initial values
list(beta0=0.4,beta1=0.025,tau=1.0/0.05)
```
Basic components of a WinBUGS program

Model:

- Components of model
  - Blue part: likelihood
  - Red part: priors
  - Green part: transformations for monitoring
- WinBUGS expresses variability in terms of the precision (= 1/variance)
- BUGS language = declarative: order of the statements does not matter
- WinBUGS does not have if-then-else commands, replaced by step and equals function
- WinBUGS does not allow improper priors for most of its analyses

Data:

- List statement
- Different formats (rectangular, list, etc) for providing data are possible

Initial values:

- List statement: initial values for intercept, slope and residual precision
- No need to specify initial values for all parameters of model
- For derived parameters, initial value must not be specified

Obtain posterior samples

To obtain posterior samples, one requires 3 WinBUGS tools:

(a) Specification Tool
(b) Sample Monitor Tool
(c) Update Tool

Specification Tool

- check model:
  - Checking syntax of program
  - Message data loaded in left bottom corner of screen/error message
- load data:
  - Message data loaded in left bottom corner of screen/error message
- compile:
  - Message data loaded in left bottom corner of screen/error message
- load inits:
  - button switches off/button gen inits
Sample Monitor Tool

▷ Choose Samples...:
  ◦ Fill in node names
  ◦ To end: type in * and click on set
▷ Term node because of connection with Directed Acyclic Graph (DAG)
▷ Dynamic trace plot: click on trace

Update Tool

▷ Choose the number of MCMC iterations:
  ◦ Click on Model of the menu bar and choose Update...
  ◦ Specify the number of iterations
  ◦ To start sampling, click on update

Obtain posterior samples

Extra via Update Tool:

▷ Extra iterations: click again on update and specify extra number of iterations
▷ To apply thinning (10%) at generation of samples: fill in 10 next to thin

Extra via Sample Monitor Tool:

▷ To look at the densities: click on density
▷ To read off posterior summary measures: click on stats
▷ To obtain ACF: click on auto corr
▷ To apply thinning (10%) when showing posterior summary: fill in 10 next to thin

WinBUGS output on osteoporosis simple linear regression analysis
8.1.2 Information on samplers

Information on sampler:

- Blocking: Check Blocking options..., Node Info..., Node Tool: methods
- Standard settings of samplers: Check Update options..., Updater options window
- Example: settings Metropolis sampler
  - 4,000 iterations needed to tune the sampler
  - Acceptance rate between 20% and 40%
  - Effect of reducing tuning phase to 750 iterations: next page

8.1.3 Assessing and accelerating convergence

Checking convergence of Markov chain:

- **Within** WinBUGS: only BGR test is available
  - Choose starting values in an overdispersed manner
  - Number of chains must be specified before compiling
  - Choose bgr diag in the Sample Monitor Tool
Checking convergence using CODA/BOA:

- Export to S+/R: connect with CODA/BOA
  - Click on coda in Sample Monitor Tool
  - 1 index file + x files on sampled values
  - Index file:
    ```
    beta0 1 1500
    beta1 1501 3000
    sigma2 3001 4500
    ```
  - Process files with CODA/BOA
  - See file chapter 8 osteoporosis BGR-CODA.R

(Simple) tricks to improve convergence (hopefully):

- Center + standardize covariates
  - \( \mu[i] \leftarrow \beta_0 + \beta_1 \cdot (\text{bmi}[i] - \text{mean} (\text{bmi}[i])) / \text{sd} (\text{bmi}[i]) \)
  - Vector: indicated by \( [] \)
- Check effect of standardization: cross-correlation plot
- Over-relaxation: switch on over relax in Update Tool

Cross-correlation on osteoporosis multiple linear regression analysis

8.1.4 Vector and matrix manipulations

Example: regress \( \text{tbbmc} \) on \( \text{age, weight, length, bmi, strength} \):

- WinBUGS file: compound document
- See file chapter 8 osteomultipleregression.odc
  - Data collected in matrix \( x \)
  - Regression coefficients in vector \( \beta \)
  - Scalar product: \( \mu[i] \leftarrow \text{inprod} (\beta[], x[i,]) \)
  - Precision matrix: \( \beta[1:6] \sim \text{dmnorm} (\mu.\beta[], \text{prec.}\beta[,,]) \)
  - Data section in two parts:
    - First part: list statement
    - Second part: rectangular part end by END
    - Click on both parts + each time: load data
**WinBUGS program** on osteoporosis multiple linear regression analysis

```r
model
{
  for (i in 1:N)
    fbimu[i] ~ dnorm(0.0,tau)
  x[i,1:4] ~ apd[1,4] x[i,5] <- length[x[i,4]]
  mu[i] <- inprod(xbeta[1],x[i,1])
} for (r in 1:6) { for (s in 1:6) { prec.beta[r,s] <- equals(r,s)*c
} for (r in 1:6) { mu.beta[r] <- 0.0
} for (r in 1:6) { for (s in 1:6) { prec.beta[r,s] <- inprod(x[,r],x[,s])*tau*c
} mu[1:6] ~ dmnorm(mu.beta[,], prec.beta[,])
  tau ~ dgamma(1.0E-3,1.0E-3)
}

data
list(N=186)

app[] <- c(71.00, 157.00, 67.00, 27.18, 96.25, 2.325)

END
```

**Vector and matrix manipulations**

**Ridge regression:**

```r
c <- 5
for (r in 1:6) { for (s in 1:6) { prec.beta[r,s] <- equals(r,s)*c
} }
```

**Zellner’s g-inverse:**

```r
c <- 1/N
for (r in 1:6) { mu.beta[r] <- 0.0
} for (r in 1:6) { for (s in 1:6) { prec.beta[r,s] <- inprod(x[,r],x[,s])*tau*c
} }
```

**Cross-correlation** on osteoporosis multiple linear regression analysis

**8.1.5 Working in batch mode**

Running WinBUGS in **batch mode**:

- **WinBUGS scripts**
- **R2WinBUGS**: interface between R and WinBUGS
- **Example**:

```r
osteo.sim <- bugs(data, inits, parameters, "osteo.model.txt",
n.chains=8, n.iter=1500,
bugs.directory="c:/Program Files/WinBUGS14/",
working.directory=NULL, clearWD=FALSE);
print(osteo.sim)
plot(osteo.sim)
```

- **vague normal prior**
- **ridge normal prior**
Output R2WinBUGS program on osteoporosis multiple linear regression analysis

8.1.6 Troubleshooting

- Errors may occur at:
  - syntax checking
  - data loading
  - compilation
  - providing initial values
  - running the sampler
- TRAP error
- Common error messages: Section Tips and Troubleshooting of WinBUGS manual

8.1.7 Directed acyclic graphs

- A graphical model: pictorial representation of a statistical model
  - Node: random variable
  - Line: dependence between random variables
- Directed Acyclic Graph (DAG) (Bayesian network)
  - Arrow added to line: express dependence child on father
  - No cycles permitted

Local independence assumption:

Directed local Markov property/conditional independence assumption:

- Each node \( v \) is independent of its non-descendants given its parents

Likelihood of DAG:

\[
p(V) = \prod_{v \in V} p(v | \text{parents}(v))
\]

Full conditionals of DAG:

\[
p(v | V \setminus v) = p(v | \text{parents}(v)) \prod_{w \in \text{parents}(v)} p(w | \text{parents}(w))
\]
Doodle option in WinBUGS:

Visualizes a DAG + conditional independencies

- **Three types of nodes:**
  - Constants (fixed by design)
  - Stochastic nodes: variables that are given a distribution (observed = data or unobserved = parameters)
  - Logical nodes: logical functions of other nodes

- **Two types of directed lines:**
  - Single: stochastic relationship
  - Double: logical relationship

- Plates

8.1.8 Add-on modules: GeoBUGS and PKBUGS

- **GeoBUGS**
  - Interface for producing maps of the output from disease mapping and other spatial models
  - Creates and manipulates adjacency matrices

- **PKBUGS**
  - Complex population pharmacokinetic/pharmacodynamic (PK/PD) models

8.1.9 Related software

OpenBUGS

- Started in 2004 in Helsinki
- Based on BUGS language
- Current version: 3.2.1 (http://www.openbugs.info/w/)
- Larger class of sampling algorithms
- Improved blocking algorithms
- New functions and distributions added to OpenBUGS
- Allows censoring $C$(lower, upper) and truncation $T$(lower, upper)
- More details on samplers by default
JAGS

- Platform independent and written in C++
- Current version: 3.1.0 (http://www-fis.iarc.fr/martyn/software/jags/allows)
- New (matrix) functions:
  - mexp(): matrix exponential
  - %*%: matrix multiplication
  - sort(): for sorting elements
- Allows censoring and truncation

Interface with WinBUGS, OpenBUGS and JAGS

- R2WinBUGS
- BRugs and R2OpenBUGS: interface to OpenBUGS
- Interface between WinBUGS and SAS, STATA, MATLAB
- R2jags: interface between JAGS and R

8.2 Bayesian analysis using SAS®

SAS® is a versatile package of programs: provides tools for
- Setting up data bases
- General data handling
- Statistical programming
- Statistical analyses

SAS® has a broad range of statistical procedures
- Most procedures support frequentist analyses
- Version 9.2: Bayesian options in GENMOD, LIFEREG, PHREG, MI, MIXED
- Procedure MCMC version 9.2 experimental, from version 9.3 genuine procedure
- From version 9.3 random statement in PROC MCMC
8.3.1 Additional Bayesian software

Additional programs (based on MCMC): Mainly in R and Matlab
- R: check CRAN website (http://cran.r-project.org/)
- Website of the International Society for Bayesian Analysis (http://bayesian.org/)
- MLwiN (http://www.bristol.ac.uk/cmm/software/mlwin/)

INLA: No sampling involved
- Integrated Nested Laplace Approximations
- For Gaussian models

STAN:
- Based on Hamiltonian sampling
- Faster for complex models, slower for “simple” models

8.3.2 Comparison of Bayesian software

Comparisons are difficult in general
Only comparisons on particular models can be done

Chapter 9
Hierarchical models

Aims:
- Introduction of one of the most important models in (bio)statistics, but now with a Bayesian flavor
- Comparing frequentist with Bayesian solutions
- Comparison Empirical Bayes estimate with Bayesian estimate
• **Bayesian hierarchical models (BHMs):** for hierarchical/clustered data

• **Examples:**
  - Measurements taken repeatedly over time on the same subject
  - Data with a spatial hierarchy: surfaces on teeth and teeth in a mouth
  - Multi-center clinical data with patients within centers
  - Cluster randomized trials where centers are randomized to interventions
  - Meta-analyses

• **BHM:** a classical mixed effects model + prior on all parameters

• As in classical frequentist world: random and fixed effects

In this chapter:

• Introduction to BHM via Poisson-gamma model and Gaussian hierarchical model

• Full Bayesian versus Empirical Bayesian approach

• Bayesian Generalized Linear Mixed Model (BGLMM)

• Choice of non-informative priors

• Examples analyzed with WinBUGS 1.4.3, OpenBUGS 3.2.1 (and Bayesian SAS procedures)

---

### 9.1 The Poisson-gamma hierarchical model

#### 9.1.1 Introduction

• Simple Bayesian hierarchical model: Poisson-gamma model

• Example: Lip cancer study
Example IX.1: Lip cancer study – Description

▷ Mortality from lip cancer among males in the former German Democratic Republic including Berlin (GDR)

▷ 1989: 2,342 deaths from lip cancer among males in 195 regions of GDR

▷ Observed # of deaths: \( y_i \) (\( i = 1, \ldots, n = 195 \))

▷ Expected # of deaths: \( e_i \) (\( i = 1, \ldots, n = 195 \))

▷ \( SMR_i = y_i / e_i \): standardized mortality rate (morbidity rate)

▷ \( SMR_i \): estimate of true relative risk (RR) = \( \theta_i \) in the \( i \)th region
  ◦ \( \theta_i = 1 \): risk equal to the overall risk (GDR) for dying from lip cancer
  ◦ \( \theta_i > 1 \): increased risk
  ◦ \( \theta_i \leq 1 \): decreased risk

To visually pinpoint regions with an increased risk for mortality (or morbidity):

Display SMRs a (geographical) map: disease map

- Problem:
  ▷ \( SMR_i \) is an unreliable estimate of \( \theta_i \) for a relatively sparsely populated region
  ▷ More stable estimate provided by BHM

9.1.2 Model specification

- Lip cancer mortality data are clustered: male subjects clustered in regions of GDR

- What assumptions can we make on the regions?

- What is reasonable to assume about \( y_i \) and \( \theta_i \)?
Example IX.2: Lip cancer study – Basic assumptions

- Assumptions on 1st level:
  - \( y_i \sim \text{Poisson}(\theta_i e_i) + \text{independent} \) \((i = 1, \ldots, n)\)

- Three possible assumptions about the \( \theta_i \)’s:
  - A1: \( n \) regions are unique: MLE of \( \theta_i = \text{SMR} \), with variance \( \theta_i/e_i \)
  - A2: \( \{y_1, y_2, \ldots, y_n\} \) a simple random sample of GDR: \( \theta_1 = \ldots = \theta_n = \theta = 1 \)
  - A3: \( n \) regions (\( \theta_i \)) are related: \( \theta_i \sim p(\theta \mid \cdot) \) \((i = 1, \ldots, n)\) (prior of \( \theta_i \))

- Choice between assumptions A1 and A3: subjective arguments

- Assumption A2: test statistically/simulation exercise (next page)

\( \Rightarrow \) Choose A3

Model specification

Assumption A3: \( p(\theta \mid \psi) \)

- \( \theta = \{\theta_1, \ldots, \theta_n\} \) is exchangeable
- Subjects within regions are exchangeable but not between regions
- Exchangeability of \( \theta_i \) \( \Rightarrow \) borrowing strength (from the other regions)

What common distribution \( p(\theta \mid \psi) \) to take?

- Conjugate for the Poisson distribution: \( \theta_i \sim \text{Gamma}(\alpha, \beta) \) \((i = 1, \ldots, n)\)
- A priori mean: \( \alpha/\beta \), variance: \( \alpha/\beta^2 \)

To establish BHM: \((\alpha, \beta) \sim p(\alpha, \beta)\)
Bayesian Poisson-gamma model:

**Level 1:** $y_i \mid \theta_i \sim \text{Poisson}(\theta_i e_i)$

**Level 2:** $\theta_i \mid \alpha, \beta \sim \text{Gamma}(\alpha, \beta)$

for $(i = 1, \ldots, n)$

**Prior:** $(\alpha, \beta) \sim p(\alpha, \beta)$

\[ p(\alpha, \beta, \theta \mid y) \propto n \prod_{i=1}^{n} p(y_i \mid \theta_i, \alpha, \beta) \prod_{i=1}^{n} p(\theta_i \mid \alpha, \beta) p(\alpha, \beta) \]

Because of hierarchical independence of $y_i$

\[ p(\alpha, \beta, \theta \mid y) \propto n \prod_{i=1}^{n} (\theta_i e_i)^{\theta_i} y_i! \exp(-\theta_i e_i) \prod_{i=1}^{n} \beta^\alpha \theta_i^\alpha - 1 \Gamma(\alpha) \theta_i e_i \beta^\beta e^{-\beta \theta_i} p(\alpha, \beta) \]

Assume: $p(\alpha) = p(\alpha) p(\beta)$

With $p(\alpha) = \lambda_\alpha \exp(-\lambda_\alpha \alpha)$ & $p(\beta) = \lambda_\beta \exp(-\lambda_\beta \beta)$ ($\lambda_\alpha = \lambda_\beta = 0.1$) = Gamma(1,0.1)

Full conditionals:

\[ p(\theta_i \mid \theta_{-i}, \alpha, \beta, y) \propto \theta_i^{\theta_i + \alpha - 1} \exp(-(\beta + \beta)\theta_i) \] (i = 1, \ldots, n)

\[ p(\alpha \mid \theta, \beta, y) \propto \frac{\beta^n \prod_i \theta_i^{\alpha - 1} \exp(-\lambda_\alpha \alpha)}{\Gamma(\alpha)^n} \]

\[ p(\beta \mid \theta, \alpha, y) \propto \beta^{\alpha \sum_i \theta_i} \exp\left[-(\sum_i \theta_i + \lambda_\beta)\beta\right] \]

with $\theta = (\theta_1, \ldots, \theta_n)^T$

$\theta_{-i} = \theta$ without $\theta_i$
9.1.4 Estimating the parameters

Some results:

- \( p(\theta_i | \alpha, \beta, y_i) = \text{Gamma}(\alpha + y_i, \beta + e_i) \)

- Posterior mean of \( \theta_i \) given \( \alpha, \beta \):
  \[
  E(\theta_i | \alpha, \beta, y_i) = \frac{\alpha + y_i}{\beta + e_i}
  \]

- Shrinkage given \( \alpha, \beta \):
  \[
  \frac{\alpha + y_i}{\beta + e_i} = B_i \frac{\alpha}{\beta} + (1 - B_i) \frac{y_i}{e_i}
  \]
  with \( B_i = \beta/(\beta + e_i) \): shrinkage factor

- Also shrinkage for: \( p(\theta_i | y) = \int p(\theta_i | \alpha, \beta, y_i)p(\alpha, \beta | y)\,d\alpha\,d\beta \)

Example IX.3: Lip cancer study – A WinBUGS analysis

- WinBUGS program: see chapter 9 lip cancer PG.odc
- Three chains of size 10,000 iterations, burn-in 5,000 iterations
- Convergence checks: (a) WinBUGS: BGR - OK
  (b) CODA: Dynamic Geweke no convergence, but rest OK
- Posterior summary measures from WinBUGS:

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>5.844</td>
<td>0.8656</td>
<td>0.04197</td>
<td>4.325</td>
<td>5.786</td>
<td>7.712</td>
<td>5001</td>
<td>15000</td>
</tr>
<tr>
<td>beta</td>
<td>4.91</td>
<td>0.7728</td>
<td>0.03763</td>
<td>3.563</td>
<td>4.861</td>
<td>6.573</td>
<td>5001</td>
<td>15000</td>
</tr>
<tr>
<td>mtheta</td>
<td>1.193</td>
<td>0.04556</td>
<td>7.557E-4</td>
<td>1.107</td>
<td>1.191</td>
<td>1.287</td>
<td>5001</td>
<td>15000</td>
</tr>
<tr>
<td>sdttheta</td>
<td>0.4976</td>
<td>0.04384</td>
<td>0.002051</td>
<td>0.4194</td>
<td>0.4951</td>
<td>0.5906</td>
<td>5001</td>
<td>15000</td>
</tr>
<tr>
<td>vartheta</td>
<td>0.2496</td>
<td>0.04443</td>
<td>0.002071</td>
<td>0.1759</td>
<td>0.2451</td>
<td>0.3488</td>
<td>5001</td>
<td>15000</td>
</tr>
</tbody>
</table>
WinBUGS program:

```winbugs
model
for (i in 1 : n )
{ 
# Poisson likelihood for observed counts
observe[i] ~ dpois(lambda[i]); lambda[i] ~ theta[i]*expect[i]
smr[i] ~ observe[i]/expect[i]
theta[i] ~ dgamma(alpha,beta)
predict[i] ~ dpois(lambda[i])
}
# Dist of future observed counts for expected count 100
theta.new ~ dgamma(alpha,beta); lambda.new ~ theta.new*100;
predict.new ~ dpois(lambda.new)
# Prior distributions for "population" parameters
alpha ~ dexp(0.1); beta ~ dexp(0.1)
# Population mean and population variance
mtheta ~ alpha/beta; vartheta ~ alpha/pow(beta,2); sdtheta ~ sqrt(vartheta)
}
```

Random effects output:

- CODA: HPD intervals of parameters

Caterpillar plot

Estimate of random effects:
First 30 $\theta_i$
95% equal tail CIs

Shrinkage effect:

![Shrinkage effect](image)

Original and fitted map of GDR:

![Original and fitted map of GDR](image)

Conclusion?
9.1.5 Posterior predictive distributions

Prediction of future counts from a Poisson-gamma model given observed counts \( y \).

Two cases:

1. Future count of males dying from lip cancer in one of the 195 regions
2. Future count of males dying from lip cancer in new regions not considered before

Example IX.5: Lip cancer study – Posterior predictive distributions

Case 1:

\[
p(\tilde{y}_i \mid y) = \int_{\alpha} \int_{\beta} \int_{\tilde{\theta}_i} p(\tilde{y}_i \mid \tilde{\theta}_i) p(\tilde{\theta}_i \mid \alpha, \beta, y) p(\alpha, \beta \mid y) \, d\tilde{\theta}_i \, d\alpha \, d\beta
\]

with

\[
p(\tilde{y}_i \mid \tilde{\theta}_i) = \text{Poisson}(\tilde{y}_i \mid \tilde{\theta}_i, \tilde{e}_i)
\]

\[
p(\tilde{\theta}_i \mid \alpha, \beta, y) = \text{Gamma}(\tilde{\theta}_i \mid \alpha + y_i, \beta + \tilde{e}_i)
\]

\[
p(\alpha, \beta \mid y) \text{ (above)}
\]

• Sampling with WinBUGS/SAS

• WinBUGS: predict[i] ~ dpois(lambda[i])

• SAS: preddist outpred=lout;

Case 2:

\[
p(\tilde{y} \mid y) = \int_{\alpha} \int_{\beta} \int_{\tilde{\theta}} p(\tilde{y} \mid \tilde{\theta}) p(\tilde{\theta} \mid \alpha, \beta) p(\alpha, \beta \mid y) \, d\tilde{\theta} \, d\alpha \, d\beta
\]

with

\[
p(\tilde{y} \mid \tilde{\theta}) = \text{Poisson}(\tilde{y} \mid \tilde{\theta}, \tilde{e})
\]

\[
p(\tilde{\theta} \mid \alpha, \beta, y) = \text{Gamma}(\tilde{\theta} \mid \alpha, \beta)
\]

\[
p(\alpha, \beta \mid y) \text{ (above)}
\]

• Sampling with WinBUGS

• WinBUGS:

  \[\text{theta.new} ~ \text{dgamma(alpha,beta)}\]
  \[\text{lambda.new} ~ \text{theta.new*100}\]
  \[\text{predict.new} ~ \text{dpois(lambda.new)}\]

9.2 Full versus Empirical Bayesian approach

• Full Bayesian (FB) analysis: ordinary Bayesian approach

• Empirical Bayesian (EB) analysis: classical frequentist approach to estimate (functions of) the ‘random effects’ (level-2 observations) in a hierarchical context

• EB approach on lip cancer mortality data:

  ○ MMLE of \( \alpha \) and \( \beta \): maximum of \( p(\alpha, \beta \mid y) \) when \( p(\alpha) \propto c, p(\beta) \propto c \)

  ○ MMLE of \( \alpha \): \( \overline{\alpha}_{EB} \), MMLE of \( \beta \): \( \overline{\beta}_{EB} \)

• Summary measures for \( \theta_i \):

  \[p(\tilde{\theta}_i \mid \overline{\alpha}_{EB}, \overline{\beta}_{EB}, y_i) = \text{Gamma}(\tilde{\theta}_i \mid \overline{\alpha}_{EB} + y_i, \overline{\beta}_{EB} + \tilde{e}_i)\]

• Posterior summary measures = Empirical Bayes-estimates because hyperparameters are estimated from the empirical result
Bayesians’ criticism on EB: hyperparameters are fixed

- Classical approach: Sampling variability of $\alpha_{\text{EB}}$ and $\beta_{\text{EB}}$ involves bootstrapping
- But is complex

Example IX.6 – Lip cancer study – Empirical Bayes analysis

- All regions:
  - EB: $\alpha_{\text{EB}} = 5.66, \beta_{\text{EB}} = 4.81 \Rightarrow$ mean of $\theta_i$s = 1.18
  - FB: $\alpha_{\text{FB}} = 5.84, \beta_{\text{FB}} = 4.91 \Rightarrow$ mean of $\theta_i$s = 1.19

- Restricted to 30 regions: comparison of estimates + 95% CIs of $\theta_i$s

Comparison 95% EB and FB:

9.3 Gaussian hierarchical models
9.3.1 Introduction

- Gaussian hierarchical model: hierarchical model whereby the distribution at each level is Gaussian
- Here variance component model/random effects model: no covariates involved
- Bayesian (hierarchical) linear model:
  - All parameters are given a prior distribution
  - Fundamentals by Lindley and Smith
- Illustrations: dietary IBBENS study

9.3.2 The Gaussian hierarchical model

- Two-level Bayesian Gaussian hierarchical model:
  - Level 1: \( y_{ij} \mid \theta_i, \sigma^2 \sim N(\theta_i, \sigma^2) \) \( (j = 1, \ldots, m_i; i = 1, \ldots, n) \)
  - Level 2: \( \theta_i \mid \mu, \sigma^2_\theta \sim N(\mu, \sigma^2_\theta) \) \( (i = 1, \ldots, n) \)
- Priors: \( \sigma^2 \sim p(\sigma^2) \) and \( (\mu, \sigma^2_\theta) \sim p(\mu, \sigma^2_\theta) \)
- Hierarchical independence
- Hyperparameters: often \( p(\mu, \sigma^2_\theta) = p(\mu) p(\sigma^2_\theta) \)
- Alternative model formulation is \( \theta_i = \mu + \alpha_i \), with \( \alpha_i \sim N(0, \sigma^2_\theta) \)
- Joint posterior:

\[
p(\theta, \sigma^2, \mu, \sigma^2_\theta \mid y) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} N(y_{ij} \mid \theta_i, \sigma^2) \prod_{i=1}^{n} N(\theta_i \mid \mu, \sigma^2_\theta) p(\sigma^2) p(\mu, \sigma^2_\theta)
\]

9.3.3 Estimating the parameters

Some analytical results are available (when \( \sigma^2 \) is fixed):

- Similar shrinkage as with the simple normal model: \( N(\mu, \sigma^2) \) (with fixed \( \sigma^2 \))
- Much shrinkage with low intra-class correlation

\[
ICC = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2}
\]

Example IX.7 – Dietary study – Comparison between subsidiaries

Aim: compare average cholesterol intake between subsidiaries with WinBUGS

- Priors: \( \mu \sim N(0, 10^6), \sigma^2 \sim IG(10^{-3}, 10^{-3}) \) and \( \sigma_\theta \sim U(0, 100) \)
  - 3 chains each of size 10,000, 3 \times 5,000 burn-in
  - BGR diagnostic in WinBUGS: almost immediate convergence
  - Export to CODA: OK except for Geweke test (numerical difficulties)
- Posterior summary measures on the last 5,000 iterations
  - \( \bar{\mu} = 328.3, \sigma_\mu = 9.44 \)
  - \( \bar{\sigma}_M = 119.5 \)
  - \( \bar{\sigma}_{\theta M} = 18.26 \)
  - \( \bar{T}_i \): see table (not much variation)
  - \( \bar{B}_i \in [0.33, 0.45] \) (\( \approx \) uniform shrinkage)
**Clinical trial applications**

- Hierarchical models have been used in a variety of clinical trial/medical applications:
  - Meta-analyses
  - Inclusion of historical controls to reduce the sample size
  - Dealing with multiplicity

**Meta-analyses**

- A Bayesian meta-analysis differs from a classical meta-analysis in that all parameters are given (a priori) uncertainty
- There are advantages in using a Bayesian meta-analysis. One is that no asymptotics needs to be considered. This is illustrated in the next meta-analysis.
Beta-blocker meta-analysis of 22 clinical trials

- A meta-analysis on 22 RCTs on beta-blockers for reducing mortality after acute myocardial infarction was performed by Yusuf et al. (1985).

- Originally, a classical meta-analysis was performed, below is a WinBUGS program for a Bayesian meta-analysis.

- The WinBUGS program:

```wbrf
model{
  for( i in 1 : Num ) {
    dc[i] ~ dbin(pc[i], nc[i]); dt[i] ~ dbin(pt[i], nt[i])
    logit(pc[i]) <- nu[i]; logit(pt[i]) <- nu[i] + delta[i]
    nu[i] ~ dnorm(0.0,1.0E-5); delta[i] ~ dnorm(mu, itau)
    mu ~ dnorm(0.0,1.0E-6); itau ~ dgamma(0.001,0.001); delta.new ~ dnorm(mu, itau)
    tau <- 1 / sqrt(itau)
  }
  mu ~ dnorm(0.0,1.0E-6); itau ~ dgamma(0.001,0.001); delta.new ~ dnorm(mu, itau)
  tau <- 1 / sqrt(itau)
}
```

Multiplicity issues

- Hierarchical models have been suggested to correct for multiplicity in a Bayesian manner.

- For example:
  - Testing treatment effect in many subgroups
  - Idea: treatment effects in the subgroups have a common distribution with mean to be estimated
  - The type I error is not directly controlled, but because of shrinkage effect some protection against overoptimistic interpretations is obtained

- In other cases, protection is sought via changing settings guided by simulation studies.

9.4 Mixed models
9.4.1 Introduction

- Bayesian linear mixed model (BLMM): extension of Bayesian Gaussian hierarchical model
- Bayesian generalized linear mixed models (BGLMM): extension of BLMM
- Bayesian nonlinear mixed model: extension of BGLMM

9.4.2 The linear mixed model

**Linear mixed model (LMM)**: $y_{ij}$ response for $j$th observation on $i$th subject

\[
y_{ij} = x_{ij}^T \beta + z_{ij}^T b_i + \varepsilon_{ij}
\]

\[
y_i = X_i \beta + Z_i b_i + \varepsilon_i
\]

Distributional assumptions:

- $b_i \sim N_{q}(0, G)$, $G$: $(q \times q)$ covariance matrix
- $G$ with $(j, k)$th element: $\sigma_{b_j b_k}$ ($j \neq k$), $\sigma_{b_j}^2$ ($j = k$)
- $\varepsilon_i \sim N_{m_i}(0, R_i)$, $R_i$: $(m_i \times m_i)$ covariance matrix often $R_i = \sigma^2 I_{m_i}$
- $b_i$ statistically independent of $\varepsilon_i$ ($i = 1, \ldots, n$)

Implications:

\[
y_i \mid b_i \sim N_{m_i}(X_i \beta + Z_i b_i, R_i)
\]

\[
y_i \sim N_{m_i}(X_i \beta, Z_i G Z_i^T + R_i)
\]

- LMM popular for analyzing longitudinal studies with irregular time points + Gaussian response
- Covariates: time-independent or time-dependent
- Random intercept model: $b_{0i}$
- Random intercept + slope model: $b_{0i} + b_{1i} t_{ij}$
- Bayesian linear mixed model (BLMM): all parameters prior distribution
Bayesian linear mixed model:

Level 1: \( y_{ij} \mid \beta, b_i, \sigma^2 \sim N(x_{ij}^{\text{T}}\beta + z_{ij}^{\text{T}}b_i, \sigma^2) \) (\( j = 1, \ldots, m_i; \ i = 1, \ldots, n \))

Level 2: \( b_i \mid G \sim N(0, G) \) (\( i = 1, \ldots, n \))

Priors: \( \sigma^2 \sim p(\sigma^2), \beta \sim p(\beta) \) and \( G \sim p(G) \)

Joint posterior:

\[
p(\beta, G, \sigma^2, b_1, \ldots, b_n \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} p(y_{ij} \mid b_i, \sigma^2, \beta) \prod_{i=1}^{n} p(b_i \mid G)p(\beta)p(G)p(\sigma^2)
\]

Example IX.11 – Dietary study – Comparison between subsidiaries

Aim: compare average cholesterol intake between subsidiaries correcting for age and gender

Model:

\[
y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{gender}_{ij} + b_{0i} + \epsilon_{ij}
\]

Priors: vague for all parameters
Results:

- WinBUGS program: chapter 9 dietary study chol age gender.odc
  - Three chains of each 10,000 iterations, 3 × 5,000 burn-in
  - Rapid convergence

- Posterior summary measures on the last 3 × 5,000 iterations
  - \( \beta_1(SD) = -0.69(0.57) \), \( \beta_2(SD) = -62.67(10.67) \)
  - With covariates: \( \sigma_M = 116.3, \sigma_{b0M} = 14.16 \)
  - Without covariates: \( \sigma_M = 119.5, \sigma_{b0M} = 18.30 \)
  - With covariates: \( ICC = 0.015 \)

Example IX.12 – Toenail RCT – Fitting a BLMM

Aim: compare itraconazol with lamisil on unaffected nail length

- Double-blinded multi-centric RCT (36 centers) sportsmen and elderly people treated for toenail dermatophyte onychomycosis
- Two oral medications: itraconazol (treat=0) or lamisil (treat=1)
- Twelve weeks of treatment
- Evaluations at 0, 1, 2, 3, 6, 9 and 12 months
- Response: unaffected nail length big toenail
- Patients: subgroup of 298 patients

Individual profiles:

Model specification:

- Model:
  \[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij} \times \text{treat}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} \]

- \( b_i \sim N(0, G) \)
- \( \varepsilon_{ij} \sim N(0, \sigma^2) \)

- Priors
  - Vague normal for fixed effects
  - \( G \sim IW(D, 2) \) with \( D = \text{diag}(0.1, 0.1) \)
Results:

- **WinBUGS program: chapter 9 toenail LMM.odc**
  - Three chains of each 10,000 iterations, $3 \times 5,000$ burn-in
  - Rapid convergence

- **Posterior summary measures on the last $3 \times 5,000$ iterations**
  - $\hat{\beta}_1(\text{SD}) = 0.58 (0.043)$, $\hat{\beta}_2(\text{SD}) = -0.057 (0.058)$
  - $\sigma_M = 1.78$
  - $\sigma_{b0M} = 2.71$
  - $\sigma_{b1M} = 0.48$
  - $\text{corr}(b_0, b_1)_M = -0.39$

- **Frequentist analysis with SAS® MIXED:** MLEs close to Bayesian estimates

---

9.4.3 The Bayesian generalized linear mixed model (BGLMM)

Generalized linear model (GLIM):

- $n$ subjects with responses $y_i$, having an exponential distribution
- Examples: binary responses, Poisson counts, gamma responses, Gaussian responses, etc.
- When covariates involved: logistic/probit regression, Poisson regression, etc.
- But clustering in data (if present) is ignored

Regression coefficients have a **population averaged** interpretation

Generalized linear mixed model (GLMM):

- Two-level hierarchy, i.e. $m_i$ subjects in $n$ clusters with responses $y_{ij}$ having an exponential distribution
- Examples: binary responses, Poisson counts, gamma responses, Gaussian responses, etc.
- Correlation of measurements in same cluster often modelled using random effects
- When covariates are involved: logistic/probit **mixed effects** regression, Poisson **mixed effects** regression, etc.

Regression coefficients have a **subject-specific** interpretation

**Bayesian generalized linear mixed model (BGLMM):** GLMM with all parameters given a prior distribution
Example IX.13 – Lip cancer study – Poisson-lognormal model

aff: percentage of the population engaged in agriculture, forestry and fisheries

Aim: verify how much of the heterogeneity of SMR, can be explained by aff

Model:

\[ y_i | \mu_i \sim \text{Poisson}(\mu_i), \text{ with } \mu_i = \theta_i e_i \]
\[ \log(\mu_i/e_i) = \beta_0 + \beta_1 \text{aff}_i + b_{0i} \]

\[ b_{0i} \sim N(0, \sigma^2_{b0}) \]

\[ \text{aff}_i \text{ centered} \]

Prior:

- Independent vague normal for regression coefficients
- \( \sigma_{b0} \sim U(0, 100) \)

Results:

- WinBUGS program: chapter 9 lip cancer PLNT with AFF.odc
  - Three chains of each 20,000 iterations, 3 \times 10,000 burn-in
  - Rapid convergence

- Posterior summary measures on the last 3 \times 5,000 iterations
  - \( \overline{\beta_1 (SD)} = 2.23 (0.33), 95\% \text{ equal tail CIL1.55, 2.93} \)
  - With aff: \( \overline{\sigma_{b0M}} = 0.38 \)
  - Without aff: \( \overline{\sigma_{b0M}} = 0.44 \)
  - aff explains 25\% of variability of the random intercept

- Analysis also possible with SAS® procedure MCMC

Example IX.14 – Toenail RCT – A Bayesian logistic RI model

Aim: compare evolution of onycholysis over time between two treatments

Response: \( y_{ij} \): 0 = none or mild, 1 = moderate or severe

Model:

\[ \text{logit}(\pi_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij} \times \text{treat}_i + b_{0i} \]

\[ b_{0i} \sim N(0, \sigma^2_{b0}) \]

Prior:

- Independent vague normal for regression coefficients
- \( \sigma_{b0} \sim U(0, 100) \)

Results:

Mean profiles:

- Itraconazol
- Lamisil

Bayesian Biostatistics - Piracicaba 2014 538

Bayesian Biostatistics - Piracicaba 2014 539

Bayesian Biostatistics - Piracicaba 2014 540

Bayesian Biostatistics - Piracicaba 2014 541
Results:

- WinBUGS program: chapter 9 toenail RI BGLMM.odc
  - Three chains of each 10,000 iterations, 3 × 5,000 burn-in
  - Rapid convergence

- Posterior summary measures on the last 3 × 5,000 iterations
  - \( \beta_0 (SD) = -1.74(0.34) \)
  - \( \beta_1 (SD) = -0.41(0.045) \)
  - \( \beta_2 (SD) = -0.17(0.069) \)
  - \( \sigma^2_{b0} = 17.44 \)
  - ICC = \( \sigma^2_{b0} / (\sigma^2_{b0} + \pi^2 / 3) \) = 0.84

- Frequentist analysis with SAS® GLIMMIX: similar results, see program chapter 9 toenail binary GLIMMIX and MCMC.sas

Example IX.15 – Toenail RCT – A Bayesian logistic RI+RS model

Aim: compare evolution of onycholysis over time between two treatments

Response: \( y_{ij} \): 0 = none or mild, 1 = moderate or severe

- Model:

\[
\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij} \times \text{treat}_i + b_{0i} + b_{1i} t_{ij}
\]

- Similar distributional assumptions and priors as before

WinBUGS program: chapter 9 toenail RI+RS BGLMM.odc

Numerical difficulties due to overflow: use of min and max functions

Other BGLMMs

Extension of BGLMM with residual term:

\[
g(\mu_{ij}) = x_{ij}^T \beta + x_{ij}^T b_i + \varepsilon_{ij}, \quad (j = 1, \ldots, m_i; i = 1, \ldots, n)
\]

Further extensions:

- WinBUGS program: chapter 9 lip cancer PLNT with AFF.odc
  - Replace Gaussian assumption by a \( t \)-distribution
  - \( t \)-distribution with degrees of freedom as parameter

Other extensions: write down full conditionals

Some further extensions

In practice more models are needed:

- Non-linear mixed effects models (Section 9.4.4)
- Ordinal logistic random effects model
- Longitudinal models with multivariate outcome
- Frailty models
  - ...

See WinBUGS Manuals I and II, and literature
Advantage of Bayesian modelling:

- Much easier to deviate from normality assumptions
- Much easier to fit more complex models

9.4.6 Estimation of the REs and PPDs

Prediction/estimation:

- Bayesian approach: estimating random effects = estimating fixed effects
- Prediction/estimation of individual curves in longitudinal models:
  \[ \lambda_{\beta}^T \hat{\beta} + \lambda_{b_i}^T \hat{b}_i \]
  
  with \( \hat{\beta} \) and \( \hat{b}_i \): posterior means, medians or modes

PPD:

- See Poisson-gamma model, but replace \( \theta_i \) by \( b_i \)
- Estimation via MCMC

Software:

- WinBUGS (CODA command) + past-processing with R
- R2WinBUGS

Example IX.19 – Toenail RCT – Exploring the random effects

Histograms of estimated RI and RS:
Predicting individual evolutions with WinBUGS:

Predicted profiles:

- Predicted evolution for the \( i \)th subject: \( x_{Tij} \hat{\beta} + z_{Tij} \hat{b}_i \)
- Extra WinBUGS commands:
  - For existing subject (id[iobs]):
    newresp[iobs] ~ dnorm(mean[iobs], tau.eps)
  - For new subject: sample random effect from its distribution
- Computation: Stats table WinBUGS + R, CODA + R, R2WinBUGS
- Prediction missing response \( \text{NA} \): automatically done

9.4.7 Choice of the level-2 variance prior

- Bayesian linear regression: NI prior for \( \sigma^2 = \text{Jeffreys prior} \quad p(\sigma^2) \propto 1/\sigma^2 \)
- Bayesian Gaussian hierarchical model: NI prior for \( \sigma_\theta^2 = p(\sigma_\theta^2) \propto 1/\sigma_\theta^2 \) ??
  - Theoretical + intuitive results: improper posterior
  - Note: Jeffreys prior from another model, not Jeffreys prior of hierarchical model
- Possible solution: \( \text{IG}(\varepsilon, \varepsilon) \) with \( \varepsilon \) small (WinBUGS)?
  - Not: see application
- Solutions:
  - \( \text{U}(0,c) \) prior on \( \sigma_\theta \)
  - Parameter expanded model suggestion by Gelman (2006)

Example IX.20 – Dietary study* – NI prior for level-2 variance

- Modified dietary study: chapter 9 dietary study chol2.odc
  - Prior 1: \( \sigma_\theta^2 \sim \text{IG}(10^{-3}, 10^{-3}) \)
  - Prior 2: \( \sigma_\theta \sim \text{U}(0,100) \)
  - Prior 3: suggestion of Gelman (2006) gives similar results as with \( \text{U}(0,100) \)
- Results:
  - Posterior distribution of \( \sigma_\theta \): clear impact of IG-prior
  - Trace plot of \( \mu \): regularly stuck with IG-prior
Choice of the level-2 variance prior

Prior for level-2 covariance matrix?

- Classical choice: Inverse Wishart
- Solution:
  - In general: not clear
  - 2 dimensions: uniform prior on both $\sigma$'s and on correlation
    - $\text{IW}(D, 2)$ with small diagonal elements
    - $\text{IW}(D, 2)$ with small diagonal elements $\approx$ uniform priors, but slower convergence with uniform priors

9.4.8 Propriety of the posterior

- Impact of improper (NI) priors on the posterior: theoretical results available
  - Important for SAS since it allows improper priors
- Proper NI priors can also have too much impact
  - Important for WinBUGS
- Improper priors can yield proper full conditionals
9.4.9 Assessing and accelerating convergence

Checking convergence Markov chain of a Bayesian hierarchical model:

- Similar to checking convergence in any other model
- But, large number of parameters ⇒ make a selection

Accelerating convergence Markov chain of a Bayesian hierarchical model:

- Use tricks seen before (centering, standardizing, overrelaxation, etc.)
- Specific tricks for a hierarchical model:
  - hierarchical centering
  - (reparameterization by sweeping)
  - (parameter expansion)

Hierarchical centering:

**Uncentered** Gaussian hierarchical model:

\[y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\]
\[\alpha_i \sim N(0, \sigma^2_\alpha)\]

**Centered** Gaussian hierarchical model:

\[y_{ij} = \theta_i + \varepsilon_{ij}\]
\[\theta_i \sim N(\mu, \sigma^2_\alpha)\]

- For \(m_i = m\): hierarchical centering implies faster mixing when \(\sigma^2_\alpha > \sigma^2/m\)
- Similar results for multilevel BLMM and BGLMM

Take home messages

- Frequentist solutions are often close to Bayesian solutions, when the frequentist approach can be applied
- There are specific problems that one has to take care of when fitting Bayesian mixed models
- Hierarchical models assume exchangeability of level-2 observations, which might a strong assumption. When satisfied it allows for better prediction making use of the ‘borrowing-strength’ principle

Chapter 10
Model building and assessment

**Aims:**
- Look at Bayesian criteria for model choice
- Look at Bayesian techniques that evaluate chosen model
10.1 Introduction

• Bayesian procedures for model building and model criticism ⇒ select an appropriate model

• In this chapter: model selection from a few good candidate models

• Chapter 11: model selection from a large set of candidate models

• Bayesian model building and criticism = similar to frequentist model and criticism, except for
  ◦ Bayesian model: combination of likelihood and prior ⇒ 2 choices
  ◦ Bayesian inference: based on MCMC techniques while frequentist approaches on asymptotic inference

10.2 Measures for model selection

Use of Bayesian criteria for choosing the most appropriate model:

• Briefly: Bayes factor

• Extensively: DIC

10.2.1 The Bayes factor

• In this chapter: explorative tools that check and improve the fitted model:
  - Model selection: Deviance Information Criterion (DIC)
  - Model criticism: Posterior Predictive Check (PPC)

• For other criteria, see book
10.2.2 Information theoretic measures for model selection

Most popular frequentist information criteria: AIC and BIC

- Both adjust for model complexity: effective degrees of freedom ($\leq$ number of model parameters)
- Akaike’s information criterion (AIC) developed by Akaike (1974)
  \[ \text{AIC} = -2 \log L(\hat{\theta} | y) + 2p \]
- Bayesian Information Criterion (BIC) suggested by Schwartz (1978)
  \[ \text{BIC} = -2 \log L(\hat{\theta} | y) + p \log(n) \]
- Deviance Information Criterion (DIC) suggested by Spiegelhalter et al. (2002)
  \[ \text{DIC} = \text{generalization of AIC to Bayesian models} \]

Practical aspects of AIC, BIC, DIC

- AIC, BIC and DIC adhere Occam’s Razor principle
- Practical rule for choosing model:
  - Choose model with smallest AIC/BIC/DIC
  - Difference of $\geq 5$: substantive evidence
  - Difference of $\geq 10$: strong evidence

Number of parameters $\leftrightarrow$ effective degrees of freedom

What is the effective degrees of freedom $\rho$ of a statistical model?

- For mixed model, number of parameters:
  - Let $p = \#$ fixed effects, $v = \#$ variance parameters, $q = \#$ random effects
  - Marginal model (marginalized over random effects): $p + v$
  - Conditional model (random effects in model): $p + v + q$

- # Free parameters (effective degrees of freedom): $p + v$
  - Marginal model: $\rho = p + v$
  - Conditional model: $p + v \leq \rho < p + v + 1$
    - with
      - $p + v = \#$ degrees of freedom when random effects are integrated out
      - $p + v + 1 = \#$ degrees of freedom when clusters are fixed effects

Example X.5+6: Theoretical example – Conditional/marginal likelihood of LMM

Statistical model of Example IX.7:

- Two-level Bayesian Gaussian hierarchical model:
  - Level 1: $y_{ij} | \theta_i, \sigma^2 \sim N(\theta_i, \sigma^2)$ ($j = 1, \ldots, m_i; i = 1, \ldots, 8$)
  - Level 2: $\theta_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$ ($i = 1, \ldots, 8$)
  - Priors: $\sigma \sim IG(10^{-3}, 10^{-3}), \sigma_\theta \sim U(0,100), \mu \sim N(0, 10^6)$

- Conditional likelihood ($\theta_i$ and $\theta_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$): $1 + 1 \leq \rho \leq 8 + 1$
  \[ L_C(\theta, \sigma | y) = \prod_{i=1}^{8} \prod_{j=1}^{m_i} N(y_{ij} | \theta_i, \sigma^2) \]

- Marginal likelihood (averaged over the random effects): $\rho = 3$
  \[ L_M(\mu, \sigma, \sigma | y) = \int \cdots \int \prod_{i=1}^{8} \prod_{j=1}^{m_i} N(y_{ij} | \theta_i, \sigma^2) N(\theta_i | \mu, \sigma^2) d\theta_1 \ldots d\theta_8 \]
Bayesian definition of effective degrees of freedom

- Bayesian effective degrees of freedom: $p_D = D(\bar{\theta}) - D(\bar{\theta})$
  - Bayesian deviance $= D(\theta) = -2 \log p(y \mid \theta) + 2 \log f(y)$
  - $D(\bar{\theta}) = \text{posterior mean of } D(\theta)$
  - $D(\bar{\theta}) = D(\theta)$ evaluated in posterior mean
  - $f(y)$ typically saturated density but not used in WinBUGS

- Motivation:
  - The more freedom of the model, the more $D(\bar{\theta})$ will deviate from $D(\bar{\theta})$
  - $p_D = \rho$ (classical degrees of freedom) for normal likelihood with flat prior for $\mu$ and fixed variance

Deviance information criterion

- DIC as a Bayesian model selection criterion:
  $$DIC = D(\bar{\theta}) + 2p_D = D(\bar{\theta}) + p_D$$

- Both DIC and $p_D$ can be calculated from an MCMC run:
  - $\theta^1, \ldots, \theta^K = \text{converged Markov chain}$
  - $D(\bar{\theta}) \approx \frac{1}{K} \sum_{k=1}^{K} D(\theta^k)$
  - $D(\bar{\theta}) \approx D(\bar{\theta})$

- Practical rule for choosing model: as with AIC/BIC
- DIC & $p_D$ are subject to sampling variability

Focus of a statistical analysis

- **Focus** of research determines degrees of freedom: population $\leftrightarrow$ cluster

- Prediction in **conditional** and **marginal** likelihood:
  - Predictive ability of **conditional** model: focus on current clusters (current random effects)
  - Predictive ability of **marginal** model: focus on future clusters (random effects distribution)

- Example:
  - Multi-center RCT comparing A with B
  - Population focus: interest in overall treatment effect $= \beta$
  - Cluster focus: interest in treatment effect in center $k = \beta + b_k$

DIC and $p_D$ – Use and pitfalls

Examples to illustrate use and pitfalls when using $p_D$ and DIC:

- Use of WinBUGS to compute $p_D$ and DIC (Example X.7)
- $\rho = p_D$ for linear regression analysis (Example X.7)
- Difference between conditional and marginal DIC (Example X.9)
- Impact of variability of random effects on DIC (Example X.10)
- Impact of scale of response on DIC (Example X.10)
- Negative $p_D$ (Example X.11)
Example X.8: Growth curve study – Model selection using $pD$ and DIC

- Well-known data set from Potthoff & Roy (1964)
- Dental growth measurements of a distance (mm)
- 11 girls and 16 boys at ages (years) 8, 10, 12, and 14
- Variables are gender (1=female, 0=male) and age
- Gaussian linear mixed models were fit to the longitudinal profiles
- WinBUGS: chapter 10 Potthoff-Roy growthcurves.odc
- Choice evaluated with $pD$ and DIC

Individual profiles:

Models:

Model $M_1$:

$$y_{ij} = \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{gender}_i + \beta_3 \text{gender}_i \times \text{age}_j + b_{0i} + b_{1i} \text{age}_j + \varepsilon_{ij}$$

- $y_{ij}$ = distance measurement
- $b_i = (b_{0i}, b_{1i})^T$ random intercept and slope with distribution $N((0, 0)^T, G)$
- $G = \begin{pmatrix} \sigma_0^2 & \rho \sigma_0 \sigma_1 \\ \rho \sigma_0 \sigma_1 & \sigma_1^2 \end{pmatrix}$
  - $\varepsilon_{ij} \sim N(0, \sigma_0^2)$ for boys
  - $\varepsilon_{ij} \sim N(0, \sigma_1^2)$ for girls

- The total number of parameters for model $M_1 = 63$
  - 4 fixed effects, 54 random effects, 3 + 2 variances (RE + ME)

Alternative models:

- Model $M_2$: model $M_1$, but assuming $\rho = 0$
- Model $M_3$: model $M_2$, but assuming $\sigma_0 = \sigma_1$
- Model $M_4$: model $M_1$, but assuming $\sigma_0 = \sigma_1$
- Model $M_5$: model $M_1$, but $b_{0i}, \varepsilon_{ij} \sim t_3$-(scaled) distributions
- Model $M_6$: model $M_1$, but $\varepsilon_{ij} \sim t_3$-(scaled) distribution
- Model $M_7$: model $M_1$, but $b_{0i} \sim t_3$-(scaled) distribution
- Nested model comparisons:
  - (1) $M_1$, $M_2$, $M_3$,
  - (2) $M_1$, $M_2$, $M_4$,
  - (3) $M_5$, $M_6$, $M_7$
**pD and DIC of the models:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dbar</th>
<th>Dhat</th>
<th>pD</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M₁</td>
<td>343.443</td>
<td>308.887</td>
<td>34.556</td>
<td>377.999</td>
</tr>
<tr>
<td>M₂</td>
<td>344.670</td>
<td>312.216</td>
<td>32.454</td>
<td>377.124</td>
</tr>
<tr>
<td>M₃</td>
<td>376.519</td>
<td>347.129</td>
<td>29.390</td>
<td>405.909</td>
</tr>
<tr>
<td>M₄</td>
<td>374.065</td>
<td>342.789</td>
<td>31.276</td>
<td>405.341</td>
</tr>
<tr>
<td>M₅</td>
<td>328.201</td>
<td>290.650</td>
<td>37.552</td>
<td>365.753</td>
</tr>
<tr>
<td>M₆</td>
<td>343.834</td>
<td>309.506</td>
<td>34.327</td>
<td>378.161</td>
</tr>
<tr>
<td>M₇</td>
<td>326.542</td>
<td>288.050</td>
<td>38.047</td>
<td>364.949</td>
</tr>
</tbody>
</table>

**Initial and final model:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Node Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>97.5%</th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>M₁</td>
<td>β₀</td>
<td>16.290</td>
<td>1.183</td>
<td>14.010</td>
<td>18.680</td>
<td>17.040</td>
<td>0.987</td>
<td>15.050</td>
</tr>
<tr>
<td></td>
<td>β₁</td>
<td>0.789</td>
<td>0.103</td>
<td>0.589</td>
<td>0.989</td>
<td>0.694</td>
<td>0.086</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>β₂</td>
<td>1.078</td>
<td>1.453</td>
<td>-1.844</td>
<td>3.883</td>
<td>-0.243</td>
<td>1.060</td>
<td>-0.454</td>
</tr>
<tr>
<td></td>
<td>β₃</td>
<td>-0.309</td>
<td>0.126</td>
<td>-0.549</td>
<td>-0.058</td>
<td>-0.243</td>
<td>0.106</td>
<td>-0.454</td>
</tr>
<tr>
<td></td>
<td>σ₀</td>
<td>1.786</td>
<td>0.734</td>
<td>0.385</td>
<td>3.381</td>
<td>1.312</td>
<td>0.541</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td>σ₁</td>
<td>0.139</td>
<td>0.065</td>
<td>0.021</td>
<td>0.280</td>
<td>0.095</td>
<td>0.052</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>ρ</td>
<td>-0.143</td>
<td>0.500</td>
<td>-0.829</td>
<td>0.931</td>
<td>-0.001</td>
<td>0.503</td>
<td>-0.792</td>
</tr>
<tr>
<td></td>
<td>σ₀</td>
<td>1.674</td>
<td>0.183</td>
<td>1.362</td>
<td>2.074</td>
<td>1.032</td>
<td>0.154</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td>σ₁</td>
<td>0.726</td>
<td>0.111</td>
<td>0.540</td>
<td>0.976</td>
<td>0.546</td>
<td>0.101</td>
<td>0.382</td>
</tr>
</tbody>
</table>

**Example X.9: Growth curve study – Conditional and marginal DIC**

(Conditional) model $M_i$:  
$$y_{ij} \mid b_i \sim N(\mu_{ij}, \sigma^2) \quad (j = 1, \ldots, 4; i = 1, \ldots, 27)$$
with  
$$\mu_{ij} = \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{gender}_i + \beta_3 \text{gender}_i \times \text{age}_j + b_{0i} + b_{1i} \text{age}_j$$

▷ Deviance: $D_C(\mu, \sigma^2) \equiv D_C(\beta, \sigma^2) = -2 \sum_i \sum_j \log N(y_{ij} \mid \mu_{ij}, \sigma^2)$

▷ In addition: $b_i \sim N((0, 0)^T, G)$

▷ Interest (focus): current 27 $b_i$'s

▷ DIC = conditional DIC

▷ WinBUGS: $pD = 31.282$ and DIC = 405.444

Marginal(ized) $M_i$:  
$$y_i \sim N_4 \left( X_i \beta, ZGZ^T + R \right), (i = 1, \ldots, 27)$$
with $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^T$

$$X_i = \begin{pmatrix} 1 & 8 & \text{gender}_i & 8 \times \text{gender}_i \\ 1 & 10 & \text{gender}_i & 10 \times \text{gender}_i \\ 1 & 12 & \text{gender}_i & 12 \times \text{gender}_i \\ 1 & 14 & \text{gender}_i & 14 \times \text{gender}_i \end{pmatrix}, Z = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{pmatrix}, R = \sigma^2 I_4$$

▷ Deviance: $D_M(\beta, \sigma^2, G) = -2 \sum_i \log N_4 (y_i \mid X_i \beta, ZGZ^T + R)$

▷ Interest (focus): future $b_i$'s

▷ DIC = marginal DIC

▷ WinBUGS: $pD = 7.072$ and DIC = 442.572
**Frequentist analysis of $M_k$:** maximization of marginal likelihood

- SAS procedure MIXED: $p = 8$ and AIC = 443.8
- chapter 10 Potthoff-Roy growthcurves.sas

**DIC and $p_D$ – Additional remarks**

Some results on $p_D$ and DIC:

- $p_D$ and DIC depend on the parameterization of the model

- Several versions of DIC & $p_D$:
  - DIC in R2WinBUGS (classical definition) is overoptimistic (using data twice)
  - R2jags: $p_D = \text{var(deviance)} / 2$
  - rjags: classical DIC & DIC corrected for overoptimism (Plummer, 2008)
  - Except for (R2)WinBUGS, computation of DIC can be quite variable

**10.3 Model checking procedures**

**10.3.1 Introduction**

Selected model is not necessarily sensible nor does it guarantee a good fit to the data

Statistical model evaluation is needed:

1. Checking that inference from the chosen model is reasonable
2. Verifying that the model can reproduce the data
3. Sensitivity analyses by varying certain aspects of the model

Bayesian model evaluation:

- As frequentist approach
- Also prior needs attention
- Primarily based on sampling techniques
Practical procedure

Two procedures are suggested for Bayesian model checking

- **Sensitivity analysis:** varying statistical model & priors
- **Posterior predictive checks:** generate replicated data from assumed model, and compare discrepancy measure based on replicated and observed data

Classical approach:

- We measure difference of observed and predicted distribution via a statistic
- Then use the (large) sample distribution of the statistic under \( H_0 \) to see if the observed value of the statistic is extreme

But:

- In Bayesian approach, we condition on observed data \( \Rightarrow \) different route needs to be taken
- Now, take a summary measure from the observed distribution and see how extreme it is compared to the replicated summary measures sampled under assumed distribution

10.3.4 Posterior predictive check

Take Example III.8 (PPD for caries experience)

Observed dmft-distribution is quite different from PPD

(i) Idea behind posterior predictive check

- \( H_0 \): model \( M_0 \) holds for the data, i.e. \( y \sim p(y \mid \theta) \)
- Sample \( y = \{y_1, \ldots, y_n\} \) & \( \theta \) estimated from the data

Take, e.g.

- Goodness-of-fit test (GOF) statistic \( T(y) \) with large value = poor fit
- Discrepancy measure \( D(y, \theta) \) (GOF test depends on nuisance parameters)

Then

- Generate sample \( \tilde{y} = \{\tilde{y}_1, \ldots, \tilde{y}_n\} \) from \( p(y \mid \theta) \)
- Compare observed GOF statistic/discrepancy measure with that based on replicated data + take uncertainty about \( \theta \) into account
- This gives a proportion of times discrepancy measure of observed data is more extreme than that of replicated data = **Bayesian P-value**
(ii) Computation of the PPC

Computation of the PPP-value for $D(y, \theta)$ or $T(y)$:

1. Let $\theta^1, \ldots, \theta^K$ be a converged Markov chain from $p(\theta | y)$
2. Compute $D(y, \theta^k) (k = 1, \ldots, K)$ (for $T(y)$ only once)
3. Sample replicated data $\tilde{y}^k$ from $p(y|\theta^k)$ (each of size $n$)
4. Compute $D(\tilde{y}^k, \theta^k) (k = 1, \ldots, K)$
5. Estimate $p_D$ by $p_D = \frac{1}{K} \sum_{k=1}^{K} I[D(\tilde{y}^k, \theta^k) \geq D(y, \theta^k)]$

$\triangleright$ When $p_D < 0.05/0.10$ or $p_D > 0.90/0.95$: “bad” fit of model to the data

$\triangleright$ Graphical checks:
- $T(y)$: Histogram of $T(\tilde{y}^k)$, $(k = 1, \ldots, K)$ with observed $T(y)$
- $D(y, \theta)$: X-Y plot of $D(\tilde{y}^k, \theta^k)$ versus $D(y, \theta^k) + 45^\circ$ line

PPP-values:

- $p_{T_{\text{skew}}} = 0.13$ & $p_{T_{\text{kurt}}} = 0.055$
- $p_{D_{\text{skew}}} = 0.27$ & $p_{D_{\text{kurt}}} = 0.26$

$\Rightarrow D_{\text{skew}}, D_{\text{kurt}}$ more conservative than $T_{\text{skew}}, T_{\text{kurt}}$

- Histogram + scatterplot (next page)

Example X.19: Osteoporosis study – Checking distributional assumptions

Checking the normality of $tbbmc$:

- (R2)WinBUGS: six PPCs in chapter 10 osteo study-PPC.R
- Skewness and kurtosis using $T(y)$ (fixing mean and variance):
  $$T_{\text{skew}}(y) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{s} \right)^3$$
  $$T_{\text{kurt}}(y) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{s} \right)^4 - 3$$
- Skewness and kurtosis using $D(y, \theta)$:
  $$D_{\text{skew}}(y, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \mu}{\sigma} \right)^3$$
  $$D_{\text{kurt}}(y, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \mu}{\sigma} \right)^4 - 3,$$
  with $\theta = (\mu, \sigma)^T$

PPP-plots:

- Histogram + scatterplot (next page)
Example X.19 – Checking distributional assumptions

Checking normality of \( r_i = (y_i - \bar{y}_i)/\sigma \) of regression \( \text{tb BMC} \) on \( \text{bmi} \):

- (R2)WinBUGS: six PPCs in chapter 10 osteo study2-PPC.R
- Kolmogorov-Smirnov discrepancy measure
  \[ D_{KS}(\mathbf{y}, \theta) = \max_{i \in \{1, \ldots, n\}} [\Phi(y_i \mid \mu, \sigma) - (i - 1)/n, i/n - \Phi(y_i \mid \mu, \sigma)] \]
- Gap discrepancy measure
  \[ D_{gap}(\mathbf{y}, \theta) = \max_{i \in \{1, \ldots, (n-1)\}} (y_{i+1} - y_i), \]
  with \( y(i) \) the \( i \)th ordered value of \( y \)
- Proportion of residuals greater than 2 in absolute value
  \[ D_{0.05}(\mathbf{y}, \theta) = \sum_{i=1}^{n} I(|y_i| > 2) \]

None showed deviation from normality

Potthoff & Roy growth curve study: Model M1 evaluated with PPCs

- R program using R2WinBUGS: LMM R2WB Potthoff Roy PPC model M1.R
- Four PPCs were considered:
  - PPC 1: sum of squared distances of data \((y_{ij} \text{ or } \bar{y}_{ij})\) to local mean value
    - Two versions \((m = 4)\):
      o Version 1: \( T_{SSS}(\mathbf{y}) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2 \)
      o Version 2: \( D_{SSS}(\mathbf{y}, \theta) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \mu_{ij})^2 \)
  - PPC 2 & PPC 3: skewness and kurtosis of residuals \( \theta = (\mu, \sigma)^T \):
    - Skewness: \( D_{skew}(\mathbf{y}, \theta) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{y_{ij} - \mu_{ij}}{\sigma} \right)^3 \)
    - Kurtosis: \( D_{kurt}(\mathbf{y}, \theta) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{y_{ij} - \mu_{ij}}{\sigma} \right)^4 - 3 \)

PPP 1 = 0.68

PPP 2 = 0.51, PPP 3 = 0.56, PPP 4 = 0.25
Take home messages

- The Bayesian approach offers the possibility to fit complex models
- But, model checking cannot and should not be avoided
- Certain aspects of model checking may take a (very long) time to complete
- The combination of WinBUGS/OpenBUGS with R is important
- R software should be looked for to do the model checking quicker

Part III

More advanced Bayesian modeling

Chapter 11

Advanced modeling with Bayesian methods

- MCMC techniques allow us to fit models to complex data
- This is for many statisticians the motivation to make use of the Bayesian approach
- In this final chapter we give 2 examples of more advanced Bayesian modeling:
  - Example 1: predicting a continuous response based on longitudinal profiles
  - Example 2: a complex dental model to relate the presence of caries on deciduous teeth to the presence of caries on adjacent permanent teeth
Predicting basal metabolic rate

- Study in suburb around Kuala Lumpur (Malaysia) on normal weight-for-age and height-for-age schoolchildren, performed between 1992 and 1995
- 70 boys (10.9 - 12.6 years) and 69 girls (10.0 - 11.6 years)
- Subjects were measured serially every 6 months for 3 years, dropouts: 23%
- Measurements: weight, height, (log) lean body mass (lnLBM) and basal metabolic rate (BMR)...
- Aim: Predict BMR from anthropometric measurements
- Program commands are in BMR prediction.R

Assessing relation between anthropometric measurements and BMR is useful:
- Determination of BMR is time consuming, prediction with anthropometric measurements is quicker
- Tool for epidemiological studies to verify if certain groups of schoolchildren has an unusual BMR

Prediction in longitudinal context:
- Based on measurements taken cross-sectionally
- Based on longitudinal measurements: joint modeling (here)

Covariate profiles and outcome histogram:

Joint modeling:

- Alternative approaches:
  - Classical regression using all anthropometric measurements at all ages: multicollinearity + problems with missing values
  - Ridge regression using all anthropometric measurements at all ages: problems with missing values
  - Two-stage approach: sampling error of intercepts and slopes is not taken into account
- Joint modeling: takes into account missing values and sampling error
- Bayesian approach takes sampling error automatically into account, but computer intensive
Analysis

The following longitudinal models were assumed:

- **Weight**:  
  
  \[ y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + b_{0i} + b_{1i} age_{ij} + \epsilon_{ij} \]

- **Height**:  
  
  \[ y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + b_{0i} + b_{1i} age_{ij} + \epsilon_{ij} \]

- **log(LBM)**:  
  
  \[ y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + b_{0i} + b_{1i} age_{ij} + \epsilon_{ij} \]

with RI + RS independent between models and of measurement error:

- \( b_{1i} \sim N(0, \Sigma_{b1}) \), \( b_{2i} \sim N(0, \Sigma_{b2}) \), \( b_{3i} \sim N(0, \Sigma_{b3}) \)

- \( \epsilon_{ij} \sim N(0, \sigma^2_{\epsilon}) \), \( \epsilon_{ij} \sim N(0, \sigma^2_{\epsilon}) \)

The following main model for BMR was assumed:

\[ y_i = \beta_0 + \beta_1 b_{0i} + \beta_2 b_{1i} + \beta_3 b_{02i} + \beta_4 b_{12i} + \beta_5 b_{03i} + \beta_6 b_{13i} + \beta_7 age_i + \beta_8 sex_i + \epsilon \]

- 3 chains, 1,500,000 iterations with 500,000 burn-in with thinning = 1,000

Results: Longitudinal models

<table>
<thead>
<tr>
<th>Param</th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>Rhat</th>
<th>n.eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-19.17</td>
<td>1.80</td>
<td>-22.65</td>
<td>-20.40</td>
<td>-19.19</td>
<td>-17.98</td>
<td>1.00</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>4.80</td>
<td>0.15</td>
<td>4.51</td>
<td>4.69</td>
<td>4.90</td>
<td>5.09</td>
<td>1.00</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-3.99</td>
<td>1.04</td>
<td>-6.06</td>
<td>-4.68</td>
<td>-3.28</td>
<td>-1.96</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Results: Main model for BMR

<table>
<thead>
<tr>
<th>Param</th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>Rhat</th>
<th>n.eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>59.80</td>
<td>97.21</td>
<td>-130.51</td>
<td>-6.10</td>
<td>60.28</td>
<td>124.82</td>
<td>1.00</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>17.26</td>
<td>8.55</td>
<td>0.10</td>
<td>11.45</td>
<td>17.43</td>
<td>22.85</td>
<td>34.06</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>375.31</td>
<td>82.15</td>
<td>215.49</td>
<td>319.77</td>
<td>373.65</td>
<td>431.70</td>
<td>534.91</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>10.15</td>
<td>7.41</td>
<td>-4.95</td>
<td>5.31</td>
<td>10.12</td>
<td>15.14</td>
<td>24.41</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>233.95</td>
<td>83.18</td>
<td>61.67</td>
<td>179.57</td>
<td>235.55</td>
<td>290.52</td>
<td>394.70</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>119.40</td>
<td>102.76</td>
<td>-30.94</td>
<td>52.80</td>
<td>123.35</td>
<td>186.43</td>
<td>316.20</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>1.49</td>
<td>99.83</td>
<td>-198.05</td>
<td>-67.33</td>
<td>69.44</td>
<td>188.71</td>
<td>479.20</td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>-0.23</td>
<td>0.25</td>
<td>-0.78</td>
<td>-0.40</td>
<td>-0.23</td>
<td>0.26</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Results: DIC and pD

<table>
<thead>
<tr>
<th>Model</th>
<th>Dbar</th>
<th>Dhat</th>
<th>pD</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmr6</td>
<td>1645.500</td>
<td>1637.260</td>
<td>8.233</td>
<td>1653.730</td>
</tr>
<tr>
<td>height</td>
<td>2298.300</td>
<td>2056.320</td>
<td>241.984</td>
<td>2540.290</td>
</tr>
<tr>
<td>weight</td>
<td>2592.730</td>
<td>2367.990</td>
<td>224.733</td>
<td>2817.460</td>
</tr>
<tr>
<td>lnLBM</td>
<td>-2885.930</td>
<td>-3133.980</td>
<td>248.054</td>
<td>-2637.870</td>
</tr>
<tr>
<td>total</td>
<td>3650.600</td>
<td>2927.590</td>
<td>723.004</td>
<td>4373.600</td>
</tr>
</tbody>
</table>

\[ Dbar = \text{post.mean of } -2\log L \]
\[ Dhat = \text{-2LogL at post.mean of stochastic nodes} \]

\[ DIC = Dbar - Dhat \]

\( \Delta DIC \) and \( pD \) are given for each model: 3 longitudinal models and main model

Averaged relative error: 11%
Results: Random effects structure: presence of correlation

Further modeling:
- Looking for correct scale in longitudinal models and main model
- Possible alternative: model 4-dimensional response jointly over time
- Involve 23 other anthropometric measurements in prediction

11.2 Example 2: relating caries on deciduous teeth with caries on permanent teeth

See 2012 presentation JASA paper
For the exercises you need the following software:

- Rstudio + R
- WinBUGS
- OpenBUGS

### Pre-course exercises

Software: **Rstudio + R**

Run the following R programs making use of R-studio to get used to R and to check your software

- dmft-score and poisson distribution.r
- gamma distributions.r

### Exercises Chapter 1

Software: **Rstudio + R**

**Exercise 1.1** Plot the binomial likelihood and the log-likelihood of the *surgery example*. Use chapter 1 lik and llik binomial example.R for this. What happens when $x=90$ and $n=120$, or when $x=0$ and $n=12$?

**Exercise 1.2** Compute the positive predictive value (pred+) of the Folin-Wu blood test for prevalences that range from 0.05 to 0.50. Plot pred+ as a function of the prevalence.
Exercises Chapter 2

Software: Rstudio + R

Exercise 2.1 In the (fictive) ECASS 2 study on 8 out of the 100 rt-PA treated stroke patients suffered from SICH. Take a vague prior for \( \theta \) = probability of suffering from SICH with rt-PA. Derive the posterior distribution with R. Plot the posterior distribution, together with the prior and the likelihood. What is the posterior probability that the incidence of SICH is lower than 0.10? Change program chapter 2 binomial lik + prior + post.R that is used to take the ECASS 2 study data as prior. So now the ECASS 2 data make up the likelihood.

Exercise 2.2 Plot the Poisson likelihood and log-likelihood based on the first 100 subjects from the dmft data set (dmft.txt). Use program chapter 2 Poisson likelihood dmft.R

Exercise 2.3 Suppose that in the surgical example of Chapter 1, the 9 successes out of 12 surgeries were obtained as follows: SSFSSFSSFSSSSS. Start with a uniform prior on the probability of success, \( \theta \), and show that how the posterior changes when the results of the surgeries become available. Then use a Beta(0.5,0.5) as prior. How do the results change? Finally, take a skeptical prior for the probability of success. Use program chapter 2 binomial lik + prior + post2.R and chapter 2 binomial lik + prior + post3.R.

Exercise 2.4 Suppose that the prior distribution for the proportion of subjects with an elevated serum cholesterol (>200 mg/dl), say \( \theta \), in a population is centered around 0.10 with 95% prior uncertainty (roughly) given by the interval [0.035,0.17]. In a study based on 200 randomly selected subjects in that population, 30% suffered from hypercholesteremia. Establish the best fitting beta prior to the specified prior information and combine this with the data to arrive at the posterior distribution of \( \theta \). Represent the solutions graphically. What do you conclude? Use program chapter 2 binomial lik + prior + post cholesterol.R.

Exercise 2.5 Take the first ten subjects in the dmft.txt file. Assume a Poisson distribution for the dmft-index, combine first with a Gamma(3,1) prior and then with a Gamma(30,10) prior for \( \theta \) (mean of the Poisson distribution). Establish the posterior(s) with R. Then take all dmft-indexes and combine them with the two gamma priors. Repeat the computations. Represent your solutions graphically. What do you conclude? Use program chapter 2 poisson + gamma dmft.R.

Exercises Chapter 3

Software: Rstudio + R

Exercise 3.1 Verify with R the posterior summary measures reported in Examples III.2 and III.5. Try first yourself to write the program. But, if you prefer, you can always look at the website of the book.

Exercise 3.2 Take the first ten subjects in the dmft.txt file, take a Gamma(3,1) prior and compute the PPD of a future count in a sample of size 100 with R. What is the probability that a child will have caries? Use program chapter 3 negative binomial distribution.R.

Exercise 3.3 Take the sequence of successes and failures reported in Exercise 2.3. Compute the PPD for the next 30 results. See program chapter 3 exercise 3-3.R.
Exercise 3.4 In Exercise 2.1, determine the probability that in the first interim analysis there will be more than 5 patients with SICH.

Exercise 3.5 The GUSTO-1 study is a mega-sized randomized controlled clinical trial comparing two thrombolytics: streptokinase (SK) and recombinant plasminogen activator (rt-PA) for the treatment of patients with an acute myocardial infarction. The outcome of the study is 30-day mortality, which is a binary indicator whether the treated patient died after 30 days or not. The study recruited 41021 acute infarct patients from 15 countries and 1081 hospitals in the period December 1990 - February 1993. The basic analysis was reported in Gusto (1993) and found a statistically significant lower 30-day mortality rate for rt-PA compared with SK. Brophy and Joseph (1995) reanalyzed the GUSTO-1 trial results from a Bayesian point of view, leaving out the subset of patients who were treated with both rt-PA and SK. As prior information the data from two previous studies have been used: GISSI-2 and ISIS-3. In the table on next page data from the GUSTO-1 study and the two historical studies are given. The authors compared the results of streptokinase and rt-PA using the difference of the two observed proportions, equal to absolute risk reduction (ar), of death, nonfatal stroke and the combined endpoint of death or nonfatal stroke. Use the asymptotic normality of ar in the calculations below.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Drug</th>
<th>N of Patients</th>
<th>N of Deaths</th>
<th>N of Nonfatal Strokes</th>
<th>N of Deaths or Strokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUSTO</td>
<td>SK</td>
<td>20,173</td>
<td>1,473</td>
<td>101</td>
<td>1,574</td>
</tr>
<tr>
<td></td>
<td>rt-PA</td>
<td>10,343</td>
<td>652</td>
<td>62</td>
<td>714</td>
</tr>
<tr>
<td>GISSI-2</td>
<td>SK</td>
<td>10,396</td>
<td>929</td>
<td>56</td>
<td>985</td>
</tr>
<tr>
<td></td>
<td>rt-PA</td>
<td>10,372</td>
<td>993</td>
<td>74</td>
<td>1,067</td>
</tr>
<tr>
<td>ISIS-3</td>
<td>SK</td>
<td>13,780</td>
<td>1,455</td>
<td>75</td>
<td>1,530</td>
</tr>
<tr>
<td></td>
<td>rt-PA</td>
<td>13,746</td>
<td>1,418</td>
<td>95</td>
<td>1,513</td>
</tr>
</tbody>
</table>

Questions:
1. Determine the normal prior for ar based on the data from (a) the GISSI-2 study, (b) the ISIS-3 study and (c) the combined data from the GISSI-2 and ISIS-3 studies.
2. Determine the posterior for ar for the GUSTO-1 study based on the above priors and a noninformative normal prior.
3. Determine the posterior belief in the better performance of streptokinase or rt-PA based on the above derived posterior distributions.
4. Compare the Bayesian analyses to the classical frequentist analyses of comparing the proportions in the two treatment arms. What do you conclude?
5. Illustrate your results graphically.

Hint: Adapt the program chapter 3 Gusto.R.

Exercise 3.6 Sample from the PPD of Example III.6 making use of the analytical results and the obtained posterior summary measures for a non-informative normal prior for μ.

Exercise 3.7 In Exercise 2.4, derive analytically the posterior distribution of logit(θ), if you dare. Or . . . , compute the posterior distribution of logit(θ) via sampling and compute the descriptive statistics of this posterior distribution.

Exercise 3.8 Apply the R program chapter 3 Bayesian P-value cross-over.R to compute the contour probability of Example III.14, but with x=17 and n=30.

Exercise 3.9 Berry (Nature Reviews/Drug Discovery, 2006, 5, 27-36) describes an application of Bayesian predictive probabilities in monitoring RCTs. The problem is as follows: After 16 control patients have been treated for breast cancer, 4 showed a complete response at a first DSMB interim analysis, but with an experimental treatment 12 out of 18 patients showed a complete response. It was planned to treat in total 164 patients. Using a Bayesian predictive calculation, it was estimated that the probability to get a statistical significant result at the end was close to 95%. This was the reason for the DSMB to advice to stop the trial. Apply the R program chapter 3 predictive significance.R to compute this probability. In a second step apply the program with different numbers.
Exercises Chapter 4

Software: Rstudio + R

Exercise 4.1 Run the R program chapter 4 regression analysis osteoporosis.R, which is based on the Method of Composition.

Exercise 4.2 Sample from a $t_3(\mu, \sigma^2)$-distribution with $\mu = 5$ and $\sigma = 2$ using the Method of Composition (extract relevant part from previous program).

Exercise 4.3 Sample from a mixture of three Gamma distributions with:

- Mixture weights: $\pi_1 = 0.15$, $\pi_2 = 0.35$ and $\pi_3 = 0.55$
- Parameters gamma distributions: $\alpha_1 = 2.5, \beta_1 = 1.75$, $\alpha_2 = 3.8, \beta_2 = 0.67$ and $\alpha_3 = 6, \beta_3 = 0.3$.

Exercises Chapter 5

Software: Rstudio + R, WinBUGS

Exercise 5.1 Establish a prior for the prevalence of HIV patients in your country. Make use of the internet (or any source that you would like to use) to establish your prior. Combine this prior information with the results of a(n hypothetical) survey based on 5,000 randomly chosen individuals from your population with 10% HIV-positives.

Exercise 5.2 Theoretical question: show that the Jeffreys prior for the mean parameter $\lambda$ of a Poisson likelihood is given by $\sqrt{\lambda}$.

Exercise 5.3 In chapter 5 sceptical and enthusiastic prior.R we apply a skeptical and an enthusiastic prior to interim results comparing to responder rates of two cancer treatments. Run the program, interpret the results and apply the program to the final results.

Exercise 5.4 Example VI.4: What types of priors have been used?

Exercises Chapter 6

Software: Rstudio + R

Exercise 6.1 Example VI.1: With chapter 6 gibbs sampling normal distribution.R, determine posterior summary measures from the sampled histograms.

Exercise 6.2 Example VI.2: Use R program chapter 6 beta-binomial Gibbs.R to sample from the distribution $f(x)$. 
Exercises Chapter 7

Software: Rstudio + R, WinBUGS/OpenBUGS

Exercise 7.1 The file chapter 7 osteoporosis convergence.R contains the commands to perform the diagnostic plots and tests based on three CODA output files with names osteop1.txt, osteop2.txt, osteop3.txt and one administrative file with name osteoind.txt. These files were obtained from a WinBUGS run. Run this R program and evaluate the output.

Exercise 7.2 In chapter 7 caries.odc a Bayesian logistic regression is performed. Use and adapt the R program in the previous exercise to evaluate the convergence of the model parameters (hint: produce three chains as for osteoporosis study). Compare the performance of the convergence of the model parameters when the regressors are centered and standardized. Perform a Bayesian probit regression model.

Exercise 7.3 Apply some the acceleration techniques to the program of Exercise 7.2.

Exercise 7.4 Fit a Bayesian Poisson model on the dmft-counts of the first 20 children in ‘dmft.txt’. Then fit a negative binomial model to the counts. Check the convergence of your sampler, and compute posterior summary statistics. This exercise requires that you build up a WinBUGS program from scratch (hint: use an existing program as start).

Exercise 7.5 Run the WinBUGS Mice example - Volume I. This is an example of a Bayesian Weibull regression analysis.

Exercise 7.6 Use the Mice example, as a guiding example to solve the following problem. The data is taken from the book of Duchateau and Janssen (The Frailty Model, Springer). It consists in the time from last time of giving birth to the time of insemination of a cow. It is of interest to find predictors of the insemination time. The data can be found in WinBUGS format in insemination.odc. The following variables are studied for their predictor effect:

- PARITY: the number of times that the cow gave birth
- Dichotomized version of PARITY: FIRST = 1 if the cow gave birth only once, FIRST = 0 when the cow gave birth multiple times.
- UREUM: Ureum percentage in the milk
- PROTEIN: Protein percentage in the milk

The cows belong to herds (cluster). Some cows are killed before the next insemination and are then censored at the time of death. If a cow is not inseminated after 300 days, it is censored at that time. The data is restricted to the first 15 clusters. The time and cluster information are captured in the following variables:

- TIME: time from birth to insemination (or censoring)
- STATUS: = 1 for events, = 0 for censored
- CLUSTERID: cluster identifier

The questions/tasks are:

- Model the time to insemination using a Weibull distribution, taking censoring into account but ignoring the clustering in the data. Use a vague prior for the parameters. Then:
  - Check the convergence of sampler and the precision of your estimates.
  - Compute posterior summary statistics for each parameter.
- What is the increase of risk of insemination in cows with multiple births (hazard ratio)?
- Evaluate the sensitivity of the results on the choice of priors.
Exercises Chapter 8

Software: WinBUGS/OpenBUGS

Exercise 8.1 In the analysis of the osteoporosis data (file chapter8osteoporosis.odc), switch off the blocking mode and compare the solution to that when the blocking mode option is turned on.

Exercise 8.2 Adapt the above WinBUGS program (blocking option switched off) to analyze the osteoporosis data and:
- perform a simple linear regression analysis with bmi centered
- add the quadratic term \( bmi^2 \)
- explore how convergence can be improved
- give 4 starting values, apply BGR diagnostics, export the chains and explore the posterior distribution in R.

Exercise 8.3 Run WinBUGS with chapter8osteomultipleregression.odc. Suppose that we wish to predict for a new individual of 65 years old, with length = 165 cm, weight = 70 kg, (fill in BMI yourself) and strength variable equal to 96.25 her tbbmc. Perform the prediction and give the 95% CI.

Exercise 8.4 Vary in Exercise 8.3 the distribution of the error term. Check WinBUGS manual which distributions are available.

Exercise 8.5 Program Example II.1 in WinBUGS.

Exercise 8.6 Simulate with WinBUGS prior distributions. Take informative and non-informative priors.

Exercise 8.7 In WinBUGS the command \( y \sim \text{dnorm}(\mu, \sigma)(a,b) \) means that the random variable \( y \) is interval censored in the interval \([a,b]\). WinBUGS does not have a command to specify that the random variable is truncated in that interval. In OpenBUGS the censoring command is \( C(a,b) \), while the truncation command is \( T(a,b) \). Look into chapter8openbugscenstrunc.odc and try out the various models to see the difference between censoring and truncation.

Exercise 8.8 Example VI.4: In chapter6coalmine.odc a WinBUGS program is given to analyze the coalmine disaster data. Particular priors have been chosen to simplify sampling. Derive the posterior of \( \theta - \lambda \) and \( \theta/\lambda \). Take other priors to perform what is called a sensitivity analysis.

Exercise 8.9 In chapter5prevalenceAgElisatestCysticercosis.odc you find the program used in Example V.6. Run the program and remove the constraint on the prevalence. What happens?

Exercise 8.10 Check the samplers in the program chapter6coalmine.odc. Note the change in samplers in the 3 programs in chapter6coalminesamplers.odc.
Exercises Chapter 9

Software: Rstudio + R, WinBUGS/OpenBUGS, R2WinBUGS

Exercise 9.1 A meta-analysis on 22 clinical trials on beta-blockers for reducing mortality after acute myocardial infarction was performed by Yusuf et al. ‘Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985;27:335-71’. The authors performed a classical analysis. Here the purpose is to perform a Bayesian meta-analysis. In the WinBUGS file chapter 9 beta-blocker.odc two Bayesian meta-analyses are performed. Tasks:

- Use the program in Model 1 (and Data 1, Inits 1/1) to evaluate the overall mean effect of the beta-blockers and the individual effects of each of the beta-blockers. Evaluate also the heterogeneity in the effect of the beta-blockers. Rank the studies according to their baseline risk. Suppose a new study with twice 100 patients are planned to set up. What is the predictive distribution of the number of deaths in each arm? For this use Data 2 and Inits 1/2.

- Use the program in Model 2 (and Data 2, Inits 2/2) to determine the predictive distribution of the number of deaths in each arm of the new study (with 2x100 patients recruited). What is the difference with the previous analysis? Why is there a difference? Rank the studies according to their baseline risk. Illustrate the shrinkage effect of the estimated treatment effects compared to the observed treatment effects.

Exercise 9.2 Perform a Bayesian meta-analysis using the WinBUGS program in chapter 9 sillero meta-analysis.odc. Perform a sensitivity analysis by varying the prior distributions and the normal distribution of the log odds ratio. Run the program also in OpenBUGS.

Exercise 9.3 In chapter 9 toenail RI BGLMM.odc a logistic random intercept model is given to relate the presence of moderate to severe onycholysis to time and treatment. Run the program and evaluate the convergence of the chain. Vary also the assumed distribution of the random intercept.

Exercise 9.4 In chapter 9 dietary study-R2WB.R an R2WinBUGS analysis is performed on the IBBENS data. Run this program and perform some diagnostic tests in R to check the convergence of the chain.

Exercise 9.5 Use chapter 9 dietary study chol2.odc to examine the dependence of the posterior of $\sigma_\theta$ on the choice of $\varepsilon$ in the inverse gamma ‘noninformative’ prior IG($\varepsilon, \varepsilon$).

Exercise 9.6 Example IX.12: Check the impact of choosing an Inverse Wishart prior for the covariance matrix of the random intercept and slope on the estimation of the model parameters, by varying the parameters of the prior. Compare this solution with the solution obtained from taking a product of three uniform priors on the standard deviation of the random intercept and slope, and the correlation, respectively. Use chapter 9 toenail LMM.odc.

Exercise 9.7 Medical device trials are nowadays often designed and analyzed in a Bayesian way. The Bayesian approach allows to include prior information from previous studies. It is important to reflect on the way subjective information is included into a new trial. A too strong prior may influence the current results too heavily and sometimes render the planning of a new study obsolete. A suggested approach is to incorporate the past trials with the current trial into an hierarchical model. In chapter 9 prior MACE.odc the fictive data are given on 6 historical studies on the effect of a cardiovascular device in terms of a 30-day MACE rate (composite score of major cardiovascular events), see Pennello and Thompson (J Biopharm Stat, 2008, 18:1, 81-115). For the new device the claim is that the probability of 30-day MACE, $p_1$ is less than 0.249. The program evaluates what the prior probability of this claim is based on the historical studies. Run this program and check also, together with the paper (pages 84-87), how this prior claim can be adjusted.
Exercises Chapter 10

Software: Rstudio + R, WinBUGS/OpenBUGS

Exercise 10.1 Replay the growth curve analyses in the WinBUGS program chapter 10 Pothoff-Roy growthcurves.odc and evaluate with DIC and $p_D$ the well performing models.

Exercise 10.2 In chapter 10 Poisson models on dmft scores.odc several models are fit to the dmft-index of 276 children of the Signal-Tandmobiel® study. Evaluate the best performing model. Confirm that the Poisson-gamma model has a different DIC and $p_D$ from the negative binomial model. Perform a chi-squared type of posterior-predictive check.

Exercise 10.3 Example X.19: In chapter 10 osteo study-PPC.R the response $tbbmc$ is checked for normality with 6 PPCs. The program uses R2WinBUGS. Run the program and check the results. Perform also the gap test. Apply this approach to test the normality of the residuals when regressing $tbbmc$ on age and bmi.

Exercise 10.4 In chapter 10 caries with residuals.odc a logistic regression model is fitted to the CE data of the Signal-Tandmobiel® study of Example VI.9 with some additional covariates, i.e. gender of the child (1=girl), age of the child, whether or not the child takes sweets to the school and whether or not the children frequently brushed their teeth. Apply the program and use the option Compare to produce various plots based on $\varepsilon_i$. Are there outliers? Do the plots indicate a special pattern of the residuals? In the same program also a Poisson model is fitted to the dmft-indices. Apply a PPC with the $\chi^2$-discrepancy measure to check the Poisson assumption.

Medical papers

Bayesian methodology is slowly introduced into clinical trial and epidemiological research. Below we provide some references to medical/epidemiological papers where the Bayesian methodology has been used. Methodological papers to highlight the usefulness of methodological papers in medical research are:

Bayesian methods in clinical trials

Evaluation of safety during the study conduct in a Bayesian manner.

Evaluation of safety during the study conduct in a Bayesian manner.

Bayesian methods were used to vary parameters when computing the cost-effectiveness.

Bayesian methods were used for trial design and formal primary analysis using a Bayesian hierarchical model approach to decrease the sample size.

Bayesian methods were used for trial design and formal primary analysis using a Bayesian hierarchical model approach to decrease the sample size.

A Bayesian adaptive design was used to set up the study, the primary efficacy and safety endpoints were analysed using a Bayesian approach.

A Bayesian design was used to set up the study (allowing for stopping for futility), the primary efficacy endpoint was analysed using a Bayesian approach. In addition frequentist methods were applied for confirmation.

A preplanned Bayesian analysis was specified to evaluate the sparse CV event in the light of previous studies. Other analyses have a frequentist nature.

item[Guidance document:]* The FDA guidance document for the planning, conduct and analysis of medical device trials is also a good introductory text for the use of Bayesian analysis in clinical trials in general (google ‘FDA medical devices guidance’)

Bayesian methods in epidemiology

Two relatively simple papers on the application of hierarchical models in epidemiology:
