Bayesian Biostatistics

Emmanuel Lesaffre
Interuniversity Institute for Biostatistics and statistical Bioinformatics
KU Leuven & University Hasselt
emmanuel.lesaffre@med.kuleuven.be

Departamento de Ciências Exatas — University of Piracicaba
8 to 12 December 2014
## 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
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.1</td>
<td>The likelihood function</td>
<td>29</td>
</tr>
<tr>
<td>1.2.2</td>
<td>The likelihood principles</td>
<td>34</td>
</tr>
<tr>
<td>1.3</td>
<td>The Bayesian approach: some basic ideas</td>
<td>43</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Introduction</td>
<td>44</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Bayes theorem – Discrete version for simple events - 1</td>
<td>48</td>
</tr>
<tr>
<td>1.4</td>
<td>Outlook</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>Bayes theorem: computing the posterior distribution</td>
<td>59</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>60</td>
</tr>
<tr>
<td>2.2</td>
<td>Bayes theorem – The binary version</td>
<td>61</td>
</tr>
<tr>
<td>2.3</td>
<td>Probability in a Bayesian context</td>
<td>63</td>
</tr>
<tr>
<td>2.4</td>
<td>Bayes theorem – The categorical version</td>
<td>66</td>
</tr>
<tr>
<td>2.5</td>
<td>Bayes theorem – The continuous version</td>
<td>67</td>
</tr>
<tr>
<td>2.6</td>
<td>The binomial case</td>
<td>69</td>
</tr>
<tr>
<td>2.7</td>
<td>The Gaussian case</td>
<td>87</td>
</tr>
<tr>
<td>2.8</td>
<td>The Poisson case</td>
<td>102</td>
</tr>
</tbody>
</table>
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
3.6 A normal approximation to the posterior .......................................................... 158
3.7 Numerical techniques to determine the posterior ............................................. 159
  3.7.1 Numerical integration .................................................................................. 160
  3.7.2 Sampling from the posterior distribution ...................................................... 162
  3.7.3 Choice of posterior summary measures ...................................................... 175
3.8 Bayesian hypothesis testing .............................................................................. 176
  3.8.1 The Bayes factor ......................................................................................... 183
3.9 Medical literature on Bayesian methods in RCTs ............................................. 185

4 More than one parameter .................................................................................... 196
  4.1 Introduction .................................................................................................... 197
  4.2 Joint versus marginal posterior inference ...................................................... 199
  4.3 The normal distribution with $\mu$ and $\sigma^2$ unknown .................................. 201
    4.3.1 No prior knowledge on $\mu$ and $\sigma^2$ is available ................................. 202
    4.3.2 An historical study is available ................................................................. 214
    4.3.3 Expert knowledge is available .................................................................. 218
4.4 Multivariate distributions ................................................................. 220

4.5 Frequentist properties of Bayesian inference .................................... 227

4.6 The Method of Composition ............................................................ 228

4.7 Bayesian linear regression models ..................................................... 231
    4.7.1 The frequentist approach to linear regression ............................... 232
    4.7.2 A noninformative Bayesian linear regression model ....................... 235
    4.7.3 Posterior summary measures for the linear regression model .......... 236
    4.7.4 Sampling from the posterior distribution .................................... 239

4.8 Bayesian generalized linear models ................................................ 243
    4.8.1 More complex regression models .............................................. 244

5 Choosing the prior distribution ........................................................... 246

5.1 Introduction ...................................................................................... 247

5.2 The sequential use of Bayes theorem ............................................... 248

5.3 Conjugate prior distributions ........................................................... 249
    5.3.1 Conjugate priors for univariate data distributions ....................... 250
5.3.2 Conjugate prior for normal distribution – mean and variance unknown ........................................ 258
5.3.3 Multivariate data distributions .................................................................................................... 260
5.3.4 Conditional conjugate and semi-conjugate priors ................................................................. 264
5.3.5 Hyperpriors ............................................................................................................................... 265
5.4 Noninformative prior distributions ............................................................................................ 266
  5.4.1 Introduction ............................................................................................................................... 267
  5.4.2 Expressing ignorance ................................................................................................................. 268
  5.4.3 General principles to choose noninformative priors .............................................................. 271
  5.4.4 Improper prior distributions .................................................................................................... 273
  5.4.5 Weak/vague priors .................................................................................................................... 274
5.5 Informative prior distributions ..................................................................................................... 279
  5.5.1 Introduction ............................................................................................................................... 280
  5.5.2 Data-based prior distributions ................................................................................................. 285
  5.5.3 Elicitation of prior knowledge .................................................................................................. 286
  5.5.4 Archetypal prior distributions ................................................................................................. 294
5.6 Prior distributions for regression models
  5.6.1 Normal linear regression
  5.6.2 Generalized linear models
5.7 Modeling priors
5.8 Other regression models

6 Markov chain Monte Carlo sampling
  6.1 Introduction
  6.2 The Gibbs sampler
    6.2.1 The bivariate Gibbs sampler
    6.2.2 The general Gibbs sampler
    6.2.3 Remarks
    6.2.4 Review of Gibbs sampling approaches
  6.3 The Metropolis(-Hastings) algorithm
    6.3.1 The Metropolis algorithm
    6.3.2 The Metropolis-Hastings algorithm
6.3.3 Remarks* .............................................................................................................................. 370

7 Assessing and improving convergence of the Markov chain .......................................................... 376
  7.1 Introduction .................................................................................................................................. 377
  7.2 Assessing convergence of a Markov chain ................................................................................... 378
    7.2.1 Definition of convergence for a Markov chain ...................................................................... 379
    7.2.2 Checking convergence of the Markov chain ......................................................................... 381
    7.2.3 Graphical approaches to assess convergence ...................................................................... 382
    7.2.4 Graphical approaches to assess convergence ...................................................................... 383
    7.2.5 Formal approaches to assess convergence .......................................................................... 394
    7.2.6 Computing the Monte Carlo standard error ...................................................................... 409
    7.2.7 Practical experience with the formal diagnostic procedures .................................................. 412
  7.3 Accelerating convergence ............................................................................................................. 414
  7.4 Practical guidelines for assessing and accelerating convergence ................................................. 418
  7.5 Data augmentation ...................................................................................................................... 419

8 Software ......................................................................................................................................... 436
8.1 WinBUGS and related software ................................................. 439
  8.1.1 A first analysis ............................................................. 440
  8.1.2 Information on samplers ................................................ 450
  8.1.3 Assessing and accelerating convergence ............................... 452
  8.1.4 Vector and matrix manipulations ..................................... 457
  8.1.5 Working in batch mode .................................................. 461
  8.1.6 Troubleshooting ......................................................... 463
  8.1.7 Directed acyclic graphs ................................................ 464
  8.1.8 Add-on modules: GeoBUGS and PKBUGS .............................. 468
  8.1.9 Related software ......................................................... 469

8.2 Bayesian analysis using SAS® .............................................. 472

8.3 Additional Bayesian software and comparisons .......................... 473
  8.3.1 Additional Bayesian software ......................................... 474
  8.3.2 Comparison of Bayesian software ..................................... 475
## Bayesian tools for statistical modeling

### 9 Hierarchical models

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>The Poisson-gamma hierarchical model</td>
<td>480</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Introduction</td>
<td>481</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Model specification</td>
<td>485</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Posterior distributions</td>
<td>491</td>
</tr>
<tr>
<td>9.1.4</td>
<td>Estimating the parameters</td>
<td>495</td>
</tr>
<tr>
<td>9.1.5</td>
<td>Posterior predictive distributions</td>
<td>502</td>
</tr>
<tr>
<td>9.2</td>
<td>Full versus Empirical Bayesian approach</td>
<td>505</td>
</tr>
<tr>
<td>9.3</td>
<td>Gaussian hierarchical models</td>
<td>509</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Introduction</td>
<td>510</td>
</tr>
<tr>
<td>9.3.2</td>
<td>The Gaussian hierarchical model</td>
<td>511</td>
</tr>
<tr>
<td>9.3.3</td>
<td>Estimating the parameters</td>
<td>512</td>
</tr>
<tr>
<td>9.4</td>
<td>Mixed models</td>
<td>521</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Introduction</td>
<td>522</td>
</tr>
</tbody>
</table>
9.4.2 The linear mixed model ................................................................. 523

9.4.3 The Bayesian generalized linear mixed model (BGLMM) ....................... 536

10 Model building and assessment ................................................................ 561

10.1 Introduction ...................................................................................... 562

10.2 Measures for model selection ............................................................ 564

10.2.1 The Bayes factor ......................................................................... 565

10.2.2 Information theoretic measures for model selection ......................... 566

10.3 Model checking procedures ............................................................... 584

10.3.1 Introduction .................................................................................... 585

III More advanced Bayesian modeling ......................................................... 599

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
IV Exercises

V Medical papers
Are you a Bayesian?
But ...
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

EMMANUEL LESAFFRE
and ANDREW LAWSON

WILEY

STATISTICS IN PRACTICE
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:
▷ Reflect on the ‘classical approach’ for statistical inference
▷ Look at a precursor of Bayesian inference: the likelihood approach
▷ 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 ($A: \text{Lamisil}$ and $B: \text{Itraconazol}$) 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 1.2
Example I.2: Graphical representation of $P$-value

$P$-value of RCT (Example I.1)
The $P$-value depends on the sample space

- $P$-value $=$ probability that observed or a more extreme result occurs if $H_0$ is true

$\Rightarrow$ Calculation of $P$-value depends on all possible samples (under $H_0$)

- The possible samples are similar in some characteristics to the observed sample (e.g. same sample size)

- Examples I.3 and I.4
Example 1.3: 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
Example 1.4: Kaldor et al’s case-control study

- Case-control study (Kaldor et al., 1990) to examine the impact of chemotherapy on leukaemia in Hodgkin’s survivors
- 149 cases (leukaemia) and 411 controls
- Question: Does chemotherapy induce excess risk of developing solid tumors, leukaemia and/or lymphomas?

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Controls</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Chemo</td>
<td>160</td>
<td>11</td>
</tr>
<tr>
<td>Chemo</td>
<td>251</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>411</td>
<td>149</td>
</tr>
</tbody>
</table>
• 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>Treatment</th>
<th>Controls</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Chemo</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Chemo</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>2</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 \( = p(y_i) = \theta \)

\[ \Rightarrow \text{Binomial distribution:} \]
Expresses probability of \( s \) successes out of \( n \) experiments.
Binomial distribution

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

- \( \theta \) fixed & function of \( s \):

  \( f_\theta(s) \) is discrete distribution with \( \sum_{s=0}^{n} f_\theta(s) = 1 \)

- \( s \) fixed & function of \( \theta \):

  \( \Rightarrow \) binomial likelihood function \( L(\theta|s) \)
Binomial distribution

- Maximum likelihood estimate (MLE) $\hat{\theta}$ maximizes $L(\theta|s)$
- Maximizing $L(\theta|s)$ equivalent to maximizing $\log[L(\theta|s)] \equiv \ell(\theta|s)$
Example 1.7 – 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 $n = 12 \Rightarrow \hat{\theta} = 0.75$
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.
Likelihood principle 1

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.
Example 1.7 (continued)

- Maximal evidence for $\theta = 0.75$

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

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

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

- Interval of ($\geq 1/2$ maximal) evidence
Inference on the likelihood function:

Bayesian Biostatistics - Piracicaba 2014

37
**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

\[ L_1(\theta|s) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} \]

▷ **Surgeon 2**: \( s = \sum_{i=1}^{n} y_i \) has a negative binomial (Pascal) distribution

\[ L_2(\theta|s) = \binom{s+k-1}{s} \theta^s (1 - \theta)^k \]
Inference on the likelihood function:

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

  $$
  p [s \geq 9 | \theta = 0.5] = \sum_{s=9}^{\infty} \binom{2 + s}{s} 0.5^s (1 - 0.5)^3
  $$

Frequentist conclusion $\neq$ Likelihood conclusion
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 types of “glasses”
1.3.1 Introduction

- **Examples 1.7 and 1.8**: combination of information from a similar historical surgical technique could be used in the evaluation of current technique = **Bayesian exercise**

- **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??**
• **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??**
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 1.9: 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)

• $p(A, B) = p(A) \cdot p(B \mid A) = p(B) \cdot p(A \mid B)$

• Bayes theorem = Theorem on Inverse Probability

\[
p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}
\]
• Bayes theorem - version II:

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

- $B =$ “diseased”, $A =$ “positive diagnostic test”

- Characteristics of diagnostic test:
  - Sensitivity ($S_e$) = $p(A \mid B)$
  - Specificity ($S_p$) = $p(A^C \mid B^C)$
  - 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 $S_e$, $S_p$ and prev using Bayes theorem
- 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>

- $S_e = \frac{56}{70} = 0.80$
- $S_p = \frac{461}{510} = 0.90$
- $\text{prev} = \frac{70}{580} = 0.12$
• 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 \( S_e, S_p \) and \( prev \):

\[
pred^+ = \frac{S_e \cdot prev}{S_e \cdot prev + (1 - S_p) \cdot (1 - prev)}
\]

• 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.
For $p(B) = 0.03 \Rightarrow \text{pred}^+ = 0.20 \& \text{pred}^- = 0.99$

For $p(B) = 0.30 \Rightarrow \text{pred}^+ = ?? \& \text{pred}^- = ??$

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 $\Rightarrow$ posterior = 0.47
1.4 Outlook

- Bayes theorem will be further developed in the next chapter \( \Rightarrow \) 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 I.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)
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
**Bayes 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, ..., 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

$\Rightarrow$ Bayes theorem for categorical parameter:

$$p(\theta_k \mid y) = \frac{p(y \mid \theta_k) p(\theta_k)}{\sum_{k=1}^{K} p(y \mid \theta_k) p(\theta_k)}$$
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: $p(y, \theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y)$

$\Rightarrow$ Bayes theorem for continuous parameters:

$$p(\theta|y) = \frac{L(\theta|y)p(\theta)}{p(y)} = \frac{L(\theta|y)p(\theta)}{\int L(\theta|y)p(\theta)d\theta}$$
• Shorter: $p(\theta | y) \propto L(\theta | y)p(\theta)$

• $\int L(\theta | y)p(\theta)d\theta = \text{averaged likelihood}$

• 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!
2.6 The binomial case

Example II.1: Stroke study – Monitoring safety

▷ Rt-PA: thrombolytic for ischemic stroke
▷ Historical studies ECASS 1 and ECASS 2: complication SICH
▷ ECASS 3 study: patients with ischemic stroke (Tx between 3 & 4.5 hours)
▷ 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 ⇒ 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 $y_i = 0$

• $y = \sum_{1}^{n} y_i$ 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?
• **Kernel** of binomial likelihood \( \theta^{y_0}(1 - \theta)^{(n_0 - y_0)} \propto \text{beta density } \text{Beta}(\alpha, \beta): \)

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

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

\( \Gamma(\cdot) \) gamma function

• \( \alpha_0(= 9) \equiv y_0 + 1 \)

\( \beta_0(= 100 - 8 + 1) \equiv n_0 - y_0 + 1 \)
Some beta densities

\[ \alpha, \beta = (3, 2) \]

\[ \alpha, \beta = (3, 3) \]

\[ \alpha, \beta = (10, 10) \]

\[ \alpha, \beta = (1, 1) \]

\[ \alpha, \beta = (0.5, 0.5) \]

\[ \alpha, \beta = (1, 0.5) \]
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
  
  \[
  L(\theta|y)p(\theta) = \binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0+y-1}(1 - \theta)^{\beta_0+n-y-1}
  \]

- Denominator of Bayes theorem = averaged likelihood
  
  \[
  p(y) = \int L(\theta|y)p(\theta)d\theta = \binom{n}{y} \frac{B(\alpha_0 + y, \beta_0 + n - y)}{B(\alpha_0, \beta_0)}
  \]
Posterior distribution = \[
\frac{\binom{n}{y} \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0 + y - 1} (1 - \theta)^{\beta_0 + n - y - 1}}{\left(\frac{n}{y}\right) B(\alpha_0 + y, \beta_0 + n - y) / B(\alpha_0, \beta_0)}
\]

⇒ Posterior distribution = 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

\[
\overline{\alpha} = \alpha_0 + y \\
\overline{\beta} = \beta_0 + n - y
\]
2. Prior, likelihood & posterior
3. Characteristics of the posterior distribution

- Posterior = compromise between prior & likelihood
- Posterior mode: \( \hat{\theta}_M = \frac{n_0}{n_0+n} \theta_0 + \frac{n}{n_0+n} \hat{\theta} \)
- Shrinkage: \( \theta_0 \leq \hat{\theta}_M \leq \hat{\theta} \) when \( \frac{y_0}{n_0} \leq \frac{y}{n} \)
- 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 \( \theta = \) MLE of combined ECASS 2 data & interim data ECASS 3 (!!!), i.e.

\[
\hat{\theta}_M = \frac{y_0 + y}{n_0 + n}
\]
Example dominance of likelihood

Likelihood based on 5,000 subjects with 800 “successes”
4. Equivalence of prior information and extra data

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

\[ \Rightarrow \text{Prior} \]

\[ \approx \text{extra data to observed data set: } (\alpha_0 - 1) \text{ successes and } (\beta_0 - 1) \text{ failures} \]
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 ⇒ posterior inference is the same

• The neurologists could also combine their qualitative prior belief with ECASS 2 data to construct a prior distribution ⇒ 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 = $p(\theta) = I_{[0,1]} = \text{flat prior on } [0,1]$

- Uniform prior on [0,1] = Beta(1,1)
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
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
1. Specifying the (IBBENS) prior distribution

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

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

\[
f(y) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} \right]
\]

- Sample \( y_1, \ldots, y_n \Rightarrow \text{likelihood} \)

\[
L(\mu|y) \propto \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \propto \exp \left[ -\frac{1}{2} \left( \frac{\mu - \bar{y}}{\sigma/\sqrt{n}} \right)^2 \right] \equiv L(\mu|\bar{y})
\]
Histogram and likelihood IBBENS study:

(a) Histogram of cholesterol (mg/day) with bars representing frequency distribution.

(b) Likelihood function with a peak at MLE = 328.
- Denote sample $n_0$ IBBENS data: $\mathbf{y}_0 \equiv \{y_{0,1}, \ldots, y_{0,n_0}\}$ with mean $\bar{y}_0$

- Likelihood $\propto N(\mu_0, \sigma_0^2)$

\begin{align*}
\mu_0 &\equiv \bar{y}_0 = 328 \\
\sigma_0 &= \sigma / \sqrt{n_0} = 120.3 / \sqrt{563} = 5.072
\end{align*}

- 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 $\mu_0 \equiv \bar{y}_0$
2. Constructing the posterior distribution

- **IBBENS-2 study:**
  
  sample $y$ with $n=50$
  
  $\bar{y} = 318$ mg/day & $s = 119.5$ mg/day
  
  95% confidence interval = $[284.3, 351.9]$ mg/day $\Rightarrow$ wide

- Combine IBBENS prior distribution IBBENS-2 normal likelihood:
  
  - IBBENS-2 likelihood: $L(\mu \mid \bar{y})$
  
  - IBBENS prior density: $p(\mu) = N(\mu \mid \mu_0, \sigma_0^2)$

- Posterior distribution $\propto p(\mu)L(\mu \mid \bar{y})$:

  $$p(\mu \mid y) \equiv p(\mu \mid \bar{y}) \propto \exp \left\{ -\frac{1}{2} \left[ \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 + \left( \frac{\mu - \bar{y}}{\sigma / \sqrt{n}} \right)^2 \right] \right\}$$
Integration constant to obtain density?

Recognize standard distribution: exponent (quadratic function of $\mu$)

Posterior distribution:

$$p(\mu|y) = N(\mu, \sigma^2),$$

with

$$\mu = \frac{1/\sigma_0 \cdot \mu_0 + n/\sigma^2 \cdot \bar{y}}{1/\sigma_0^2 + n/\sigma^2} \quad \text{and} \quad \sigma^2 = \frac{1}{1/\sigma_0^2 + n/\sigma^2}$$

Here: $\bar{\mu} = 327.2$ and $\bar{\sigma} = 4.79$. 
IBBENS-2 posterior distribution:
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

\[ w_0 = \frac{1}{\sigma_0^2} \quad \text{and} \quad w_1 = \frac{1}{\sigma^2/n} \]

▷ The posterior precision = 1/posterior variance:

\[ \frac{1}{\sigma^2} = w_0 + w_1 \]

with \( w_0 = 1/\sigma_0^2 = \text{prior precision} \) and \( w_1 = 1/(\sigma^2/n) = \text{sample precision} \)
• Posterior is always more peaked than prior and likelihood

• When $n \to \infty$ or $\sigma_0 \to \infty$: $p(\mu|y) = \mathcal{N}(\bar{y}, \sigma^2/n)$

$\Rightarrow$ When sample size increases the likelihood dominates the prior

• Posterior = normal = prior $\Rightarrow$ conjugacy
4. Equivalence of prior information and extra data

- Prior variance $\sigma_0^2 = \sigma^2$ (unit information prior) $\Rightarrow \bar{\sigma}^2 = \sigma^2/(n + 1)$

  $\Rightarrow$ Prior information = adding one extra observation to the sample

- General: $\sigma_0^2 = \sigma^2/n_0$, with $n_0$ general

  $\bar{\mu} = \frac{n_0}{n_0 + n} \mu_0 + \frac{n}{n_0 + n} \bar{y}$

  and

  $\bar{\sigma}^2 = \frac{\sigma^2}{n_0 + n}$
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:

(a) Discounted prior

(b) Discounted prior + shift
Bayesian approach: no prior information is available

- Non-informative prior: $\sigma_0^2 \rightarrow \infty$

$\Rightarrow$ Posterior: $N(\bar{y}, \sigma^2/n)$
Non-informative prior:
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:

$$L(\theta|y) \equiv \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \left(\frac{\theta^{y_i}}{y_i!}\right) e^{-n\theta}$$
Example II.6: Describing caries experience in Flanders

The Signal-Tandmobiel® (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
- $E(\theta) = \alpha_0 / \beta_0$ & $var(\theta) = \alpha_0 / \beta_0^2$

- **STM study:** $\alpha_0 = 3$ & $\beta_0 = 1$
Gamma prior for STM study:

\[ \text{Gamma}(3,1) \]
Some gamma densities

\[ \alpha, \beta = (3, 1) \]  
\[ \alpha, \beta = (3, 3) \]  
\[ \alpha, \beta = (5, 5) \]  
\[ \alpha, \beta = (1, 1) \]  
\[ \alpha, \beta = (0.1, 0.1) \]  
\[ \alpha, \beta = (1, 0.5) \]
2. Constructing the posterior distribution

• Posterior

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

\[
\propto \theta^{(\sum y_i + \alpha_0) - 1} e^{-(n+\beta_0)\theta}
\]

• Recognize kernel of a Gamma(\(\sum y_i + \alpha_0, n + \beta_0\)) distribution

\[
\Rightarrow p(\theta | y) \equiv p(\theta | y) = \frac{\bar{\beta}^{\bar{\alpha}}}{\Gamma(\bar{\alpha})} \theta^{\bar{\alpha} - 1} e^{-\bar{\beta} \theta}
\]

with \(\bar{\alpha} = \sum y_i + \alpha_0 = 9758 + 3 = 9761\) and \(\bar{\beta} = n + \beta_0 = 4351 + 1 = 4352\)

\[
\Rightarrow \text{STM study: effect of prior is minimal}
\]
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 \( p(h^{-1}(\psi)) \left( \left| \frac{dh^{-1}(\psi)}{d\psi} \right| \right) \)

- 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 $\psi = \log(\theta)$

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

with $\bar{\alpha} = 19$ and $\bar{\beta} = 133$. 
Posterior and transformed posterior:
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:

\[
p_2(\theta \mid y) = \frac{L_2(\theta \mid y)p(\theta)}{\int L_2(\theta \mid y)p(\theta)d\theta}
= c \frac{L_1(\theta \mid y)p(\theta)}{\int c L_1(\theta \mid y)p(\theta)d\theta}
= p_1(\theta \mid y)
\]

- 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 (**data fixed**), uncertainty about \( \theta \) (**\( \theta \) stochastic**)
  - No asymptotic arguments are needed, all inference depends on posterior
  - Integration is key tool
  - Does not depend on stopping rules
• Frequentist and Bayesian approach can give the same numerical output (with different interpretation), but may give quite a different inference

• Frequentist ideas in Bayesian approaches (MCMC)
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.

- The 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 (1749-1827).

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: \( P(a < \theta < b \mid y) \) 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
  \[ p(\theta \mid y) = \text{Beta}(19, 133)-distribution \]

- \( P(a < \theta < b \mid y) \):
  - \( a = 0.2, b = 1.0 \): \( P(0.2 < \theta < 1 \mid y) = 0.0062 \)
  - \( a = 0.0, b = 0.08 \): \( P(0 < \theta < 0.08 \mid y) = 0.033 \)
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. \( 0.5 = \int_{\theta_M} p(\theta|y) d\theta \)

- 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: \( \text{Beta}(\bar{\alpha}, \bar{\beta}) \) (\( \bar{\alpha} = 19 \) & \( \bar{\beta} = 133 \))

- Posterior mode: \textbf{maximize} \( (\bar{\alpha} - 1) \ln(\theta) + (\bar{\beta} - 1) \ln(1 - \theta) \) wrt \( \theta \)
  \[ \Rightarrow \hat{\theta}_M = (\bar{\alpha} - 1)/((\bar{\alpha} + \bar{\beta} - 2)) = 18/150 = 0.12 \]

- Posterior mean: \textbf{integrate} \( \frac{1}{B(\bar{\alpha}, \bar{\beta})} \int_0^1 \theta^{\bar{\alpha}-1}(1 - \theta)^{\bar{\beta}-1} \, d\theta \)
  \[ \Rightarrow \bar{\theta} = B(\bar{\alpha} + 1, \bar{\beta})/B(\bar{\alpha}, \bar{\beta}) = \bar{\alpha}/(\bar{\alpha} + \bar{\beta}) = 19/152 = 0.125 \]

- Posterior median: \textbf{solve} \( 0.5 = \frac{1}{B(\bar{\alpha}, \bar{\beta})} \int_{\theta_M}^1 \theta^{\bar{\alpha}-1}(1 - \theta)^{\bar{\beta}-1} \, d\theta \) for \( \theta \)
  \[ \Rightarrow \bar{\theta}_M = = 0.122 \ (R\text{-function } qbeta) \]

- Posterior variance: \textbf{calculate} also \( \frac{1}{B(\bar{\alpha}, \bar{\beta})} \int_0^1 \theta^2 \theta^{\bar{\alpha}-1}(1 - \theta)^{\bar{\beta}-1} \, d\theta \)
  \[ \Rightarrow \sigma^2 = \frac{\bar{\alpha} \bar{\beta}}{[(\bar{\alpha} + \bar{\beta})^2(\bar{\alpha} + \bar{\beta} + 1)]} = 0.0267^2 \]
Graphical representation measures:
Posterior summary measures:

- Posterior for $\mu$ (based on IBBENS prior):
  
  Gaussian with $\hat{\mu}_M \equiv \mu \equiv \bar{\mu}_M$

- 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\% \text{ credible interval (CI) for } \theta \text{ if } P(a \leq \theta \leq b \mid y) = 0.95\]

- Two types of 95\% credible interval:
  - 95\% equal tail CI \([a, b]\):
    \(\text{AUC} = 0.025 \text{ is left to } a \& \text{AUC} = 0.025 \text{ is right to } b\)
  - 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)
  - \(h(\text{HPD CI}) \neq \text{HPD CI}\), but \(h(\text{equal-tail CI}) = \text{equal-tail CI}\)
  - 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 = $\text{Beta}(19, 133)$-distribution

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

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

- Computations HPD interval: use $R$-function $\text{optimize}$
Equal tail credible interval:

\[ \theta \sim \text{BETA}(19,133) \]

95% EQUAL TAIL CI

0.025

0.025
HPD interval:

HPD interval is not-invariant to (monotone) transformations
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) = $8/100 = 0.08$ 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
    $\Rightarrow$ observed result of 10 SICH patients out of 50 is extreme

- But, 8% is likely not the true value
Binomial predictive distribution

Ignores uncertainty in $\theta$
Example III.7: Unknown incidence rate

- Uncertainty of $\theta$ expressed by posterior (ECASS 2 prior) = $\text{Beta}(\bar{\alpha}, \bar{\beta})$-distribution
  with $\bar{\alpha} = 9$ and $\bar{\beta} = 93$

- 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 = $\text{Beta-binomial distribution } \text{BB}(m, \bar{\alpha}, \bar{\beta})$

$$p(\tilde{y}|y) = \int_0^1 \binom{m}{\tilde{y}} \theta^{\tilde{y}} (1 - \theta)^{m-\tilde{y}} \frac{\theta^{\bar{\alpha}-1}(1 - \theta)^{\bar{\beta}-1}}{B(\bar{\alpha}, \bar{\beta})} \, d\theta$$

$$= \binom{m}{\tilde{y}} \frac{B(\tilde{y} + \bar{\alpha}, m - \tilde{y} + \bar{\beta})}{B(\bar{\alpha}, \bar{\beta})}$$
Result:

- $BB(m, \bar{\alpha}, \bar{\beta})$ shows more variability than $\text{Bin}(m, \hat{\theta})$

- 94.4% PPS is $\{0, 1, \ldots, 9\} \Rightarrow 10$ SICH patients out of 50 less extreme
Binomial and beta-binomial predictive distribution

\[ \text{BB}(50, 9, 93) \quad \text{Bin}(50, 0.08) \]

# future RT-PA patients with SICH

Bayesian Biostatistics - Piracicaba 2014
3.4.2 Posterior predictive distribution: General case

- Central idea: Take the posterior uncertainty into account

- Three cases:
  ▶ All mass (AUC \( \approx 1 \)) of \( p(\theta \mid y) \) at \( \hat{\theta}_M \) ⇒ distribution of \( \tilde{y} \): \( p(\tilde{y} \mid \hat{\theta}_M) \)
  ▶ All mass at \( \theta^1, \ldots, \theta^K \) ⇒ distribution of \( \tilde{y} \): \( \sum_{k=1}^{K} p(\tilde{y} \mid \theta^k)p(\theta^k \mid y) \)
  ▶ General case: posterior predictive distribution (PPD)

  ⇒ distribution of \( \tilde{y} \)

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

- 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

• $y_i = 100 / \sqrt{alp_i}$ ($i = 1, \ldots, 250$) $\approx$ 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 $alp = [104.33, 510.18]$
Histogram \( \text{alp} + 95\% \) normal range
Example III.6: Bayesian approach

- PPD: \( \tilde{y} | y \sim N(\mu, \sigma^2 + \sigma_0^2) \)

- 95% normal range for \( y \): \([\mu - 1.96 \sqrt{\sigma^2 + \sigma_0^2}, \mu + 1.96 \sqrt{\sigma^2 + \sigma_0^2}]\)

- Prior variance \( \sigma_0^2 \) large \( \Rightarrow \bar{\sigma}^2 = \sigma^2 / n \)

\( \Rightarrow \) Bayesian 95% normal range = frequentist 95% normal range

\( \Rightarrow \) 95% normal range for \( \alpha/p \) = \([104.33, 510.18]\)

- 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(α, β)

\[
p(\tilde{y}|y) = \frac{\Gamma(\alpha + \tilde{y})}{\Gamma(\alpha) \tilde{y}!} \left( \frac{\beta}{\beta + 1} \right)^\alpha \left( \frac{1}{\beta + 1} \right)^{\tilde{y}}
\]
Example III.8: Caries study – PPD for caries experience

OBSERVED DISTRIBUTION

<table>
<thead>
<tr>
<th>dmft-index</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 5</td>
<td>0.4</td>
</tr>
<tr>
<td>5 - 10</td>
<td>0.3</td>
</tr>
<tr>
<td>10 - 15</td>
<td>0.2</td>
</tr>
<tr>
<td>15 - 20</td>
<td>0.1</td>
</tr>
</tbody>
</table>

PPD
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, . . .
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(\(\theta\)) with \(\theta = \text{mean dmft-index}\)
- With prior: Gamma(3, 1) \(\Rightarrow\) Posterior: Gamma(29, 11)

▷ Replace gamma prior by lognormal prior, then posterior

\[
\propto \theta^{\sum_{i=1}^{n} y_i - 1} e^{-n\theta - \left(\frac{\log(\theta) - \mu_0}{2\sigma_0}\right)^2}, (\theta > 0)
\]

▷ 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:

$$\int \theta p(\theta|y) \, d\theta \approx \hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\theta}_k,$$

for $K$ 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

▷ 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)

black box = true posterior mean & red box = sampled posterior mean
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

AR algorithm:
only product of likelihood and prior is needed
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 $p(\theta \mid y)$ (target)

- Assumption: $p(\theta \mid y) < A q(\theta)$ 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 $u \leq p(\tilde{\theta} \mid y)/Aq(\tilde{\theta})$
  - **Reject**: when $u > p(\tilde{\theta} \mid y)/Aq(\tilde{\theta})$

- **Properties AR algorithm**:
  - Produces a sample from the posterior
  - **Only needs** $p(y \mid \theta)p(\theta)$
  - Probability of acceptance $= 1/A$
Example III.13: Caries study – Sampling from posterior with lognormal prior

Accept-reject algorithm

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

\[ Aq(\theta) = \theta \sum_{i=1}^{n} y_i^{-1} e^{-n\theta} \propto \text{Gamma}(\sum_i y_i, n)\text{-distribution} \]

• 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]
Sampling approximation
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

Tangent ARS

\[ \log \text{posterior} \]

\( \theta_1 \quad \theta_2 \quad \theta \quad \theta_3 \)

SQUEEZING

TANGENT

Derivative-free ARS

\[ \log \text{posterior} \]

\( \theta_1 \quad \theta_2 \quad \theta \quad \theta_3 \)

SQUEEZING

DERIVATIVE-FREE
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:
  - Direct use of 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 \( p_B \):
    \[
P[p(\theta | y) > p(\theta_0 | y)] \equiv (1 - p_B)
\]

- 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 $A$)

• $\theta = P(A$ 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 ($p_B = 0.023$)
Graphical representation of $p_B$: 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{graph.png}
\end{figure}

$p_B$ against $\theta = 0.5$
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:** \( \text{power} = \Phi \left( \sqrt{\frac{n\Delta^2}{2\sigma^2}} + z_{\alpha/2} \right) \)
  - **Sample size given power:** \( n = \frac{2\sigma^2}{\Delta^2} \left( z_\beta - z_{\alpha/2} \right)^2 \)

- 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 − \( \beta \)
Comparison of means – continued

▷ 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: $\Delta \sim N(0.5, 0.1)$ & $\sigma \sim N(1, 0.3) | [0, \infty)$

▷ Predictive distributions of power and $n$ with WinBUGS
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 $H_0$

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

- Bayes factor: $BF(y)$

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

- 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
Bayes factor - Jeffreys classification

- **Jeffreys classification** for favoring $H_0$ and against $H_a$
  - Decisive ($BF(y) > 100$)
  - Very strong ($32 < BF(y) \leq 100$)
  - Strong ($10 < BF(y) \leq 32$)
  - Substantial ($3.2 < BF(y) \leq 10$)
  - ‘Not worth more than a bare mention’ ($1 < BF(y) \leq 3.2$)

- **Jeffreys classification** for favoring $H_a$ and against $H_0$: replace criteria by $1/criteria$
  - Decisive ($BF(y) > 1/100$)
  - . . .
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 Exhale 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

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 (Pt) was superior to the posterior probability of responding to treatment for the sham control arm (Pc), with probability (Pr) of 96.5%. Trial success requires that \( \pi > 0.965 \), where \( \pi = \text{Pr}[\text{Pt}>\text{Pc}] \).

- Phase II neoadjuvant HER2/neu-positive breast cancer trial: planned accrual was 164 patients randomized to chemotherapy with and without trastuzumab
- 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

\[ P(\text{success}) = 0.94 \]
Example 3: Cannon et al. (Am Heart J, 2009, 158: 513-9.e3)

**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-based 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 prestated level of confidence is required to define how much the true incidence rate in CV events may vary versus...
Example 3:

Design of the DEFINE trial: Determining the EFFicacy and Tolerability of CETP INhibition with AnaceTrapib

The estimated event rate, based upon the observed number of CV events, this Bayesian-based approach provides the mechanism to compute confidence levels to dismiss a deleterious safety signal of the magnitude of that observed for torcetrapib (ie, a 25% increase in CV events). Of course, a point-estimate of fewer comparative events in the anacetrapib group would help provide greater levels of confidence in dismissing a torcetrapib-like deleterious safety signal.
### Example 3: Decision table

#### Table II. Confidence in dismissing a taraetrapib-type adverse effect on CV events* (assuming 16 events; pT < 1.25* pC)

<table>
<thead>
<tr>
<th>16 total events (based on 1.1% 24 wk-event rate)</th>
<th>Confidence (pT &lt; 1.25* pC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anacetrapib (pT)</td>
<td>Placebo (pC)</td>
</tr>
<tr>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>

The green confidence values of >30% suggest study continuation. The red confidence values of <30% suggest study termination.

*Prospective CV events include CV death, nonfatal MI, stroke, and hospitalization for unstable angina.
**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