STA421
Foundations of Bayesian Methodology
FS22

Małgorzata Roos
Department of Biostatistics at Epidemiology, Biostatistics and Prevention Institute,
University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland
malgorzata.roos@uzh.ch

April 6, 2022

Contents

1 Lecture 1: Classical vs Bayesian paradigms and conditional probability 3
- 1.1 Overview of the lecture . . . 3
- 1.2 Overview of the individual project . . . 4
- 1.3 History . . . 6
- 1.4 Probability calculus . . . 7
- 1.5 Example: Breast cancer and diagnostic tests . . . 9
- 1.6 Overview of the classical statistic . . . 9
  - 1.6.1 Example: Primary outcome follows normal distribution . . . 10
  - 1.6.2 Example: Primary outcome follows Bernoulli distribution . . . 11
- 1.7 Overview of the Bayesian methodology . . . 11
  - 1.7.1 Bayes factors and p-values . . . 12
  - 1.7.2 Priors . . . 14
- 1.8 Worksheet 1 . . . 14

2 Lecture 2: Conjugate Bayes, point estimates, and interval estimates 17
- 2.1 Binary data: Vision correction . . . 18
- 2.2 Bayes analysis of binary data . . . 18
  - 2.2.1 Beta prior . . . 18
  - 2.2.2 Likelihood . . . 20
  - 2.2.3 Posterior distribution . . . 20
- 2.3 Bayes analysis of normal data . . . 23
- 2.4 Point estimates . . . 24
- 2.5 Credible intervals . . . 25
- 2.6 Worksheet 2 . . . 26

3 Lecture 3: Predictive distributions, asymptotics, and Monte Carlo simulations 29
- 3.1 Predictive distributions for binary data . . . 30
  - 3.1.1 Prior predictive distribution . . . 30
  - 3.1.2 Posterior predictive distribution . . . 31
- 3.2 Predictive distributions for normal data . . . 33
  - 3.2.1 Prior predictive distribution . . . 33
  - 3.2.2 Posterior predictive distribution . . . 33
- 3.3 Bayesian asymptotics . . . 36
- 3.4 Monte Carlo simulations . . . 38

3.5 Worksheet 3 40

4 Lecture 4: Markov chain Monte Carlo method 43
4.1 MCMC Sampling in R 43
4.2 Markov chain 44
4.3 The algorithm for MCMC sampling 44
4.4 General steps of MCMC algorithm 44
4.5 Gibbs sampler 45
4.6 Application of the Gibbs sampler 46
4.7 Metropolis-Hastings algorithm 48
4.8 Application of the Metropolis-Hastings sampler 49
4.9 Expert system 50
4.10 The meaning of priors for the slopes of centered and scaled (standardized) covariates 51
4.11 Worksheet 4 56

Chapter 1

Lecture 1: Classical vs Bayesian paradigms and conditional probability

1.1 Overview of the lecture

Bayesian methods combine prior knowledge with observed data and are powerful tools for data analysis in many domains of science. However, underlying concepts, derivations, and computations can be challenging. This lecture reviews fundamental concepts of Bayesian methodology and provides an accessible introduction to theoretical and practical tools with medical applications. A successful participant will be able to apply Bayesian methods in other areas of research.

Probability calculus	Distributions	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		Likelihood
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 1.1: Foundations of Bayesian Methodology: content of the lecture.

1.2 Overview of the individual project

Table 2 of Baeten et al. [2013] provides results of a Bayesian analysis of ASA20 responders at week 6 for Secukinumab and Placebo. This case-control study considers Ankylosing spondylitis in an experimental treatment with Secukinumab (monoclonal antibody) and uses historical controls. The primary binary endpoint ASAS20 indicates patients with a 20% response according to the Assessment of Spondylo Arthritis international Society criteria for improvement at week 6.

A classical clinical trial would for example use a 1:1 sampling with n=24 patients in the treatment group and n=24 patients in the placebo group. This Bayesian analysis uses a smaller number of patients. It applies a 4:1 study design with n=24 patients in the treatment group and only n=6 patients in the placebo group, but uses 8 similar historical placebo-controlled clinical trials to derive an informative prior for the placebo group instead.

Potential benefits of Bayesian analysis

- Reduces the number of placebo patients in the new trial
- Decreases costs
- Shortens trials duration ($\rightarrow$ faster decision)
- Facilitates recruitment ($\rightarrow$ faster decision)
- Can be more ethical in some situations

Secukinumab		**Placebo**
Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		Data
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 1.2: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

- An intermediate study can be conducted at any timepoint

Potential dangers of Bayesian analysis

- Posteriors hinge on the prior elicited for the placebo group
- The prior elicited for the placebo group depends on the prior for the between-study precision in a Bayesian meta-analysis

1.3 History

The history of both the Bayesian and the classical approaches to statistics is intertwined. This section reviews the most relevant historical facts.

Bayes

INDUCTIVE LOGIC
(θ) before $\longleftarrow$ after ($\mathbf{y}$)
before: possible, probable causes
after: effects, results

- James Bernoulli (1713)
- Reverend Thomas Bayes (1763)
- Laplace (1812)

Bayes Theorem

timeline $A \rightarrow B$:

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B]}$$

or

$$P[\theta \mid \mathbf{y}] = \frac{P[\mathbf{y} \mid \theta]P[\theta]}{P[\mathbf{y}]}$$

or

$$P[H \mid D] = \frac{P[D \mid H]P[H]}{P[D]}$$

- quantification of evidence
- Bayes factor

1940 Physics
1950 MCMC Metropolis Hastings
1980 Gibbs Sampling
1990 WinBUGS
... OpenBUGS, JAGS, Stan, INLA,
Variational Bayes, bayesmeta

Classical

DEDUCTIVE LOGIC
(θ) before $\longrightarrow$ after ($\mathbf{y}$)
before: causes
after: results
general rules, promises (θ) lead to
certain results and conclusions ($\mathbf{y}$)

- Pearson, Galton - 1890, 1900
- Gosset, Fisher - 1910, 1920
- Pearson, Neyman - 1930

Likelihood

timeline $A \rightarrow B$:
Only interested in $P[B \mid A]$ or $P[\mathbf{y} \mid \theta]$

- 95% confidence intervals
- tests
- p-values
- statistical programs

Nowadays, parallel usage of Bayesian and classical paradigms is quite common. See, for

example, your individual project motivated by Baeten et al. [2013].

Note that the Bayes approach is also based on the likelihood. Therefore, all problems for classical inference such as uncertainty about the sampling model, randomness of the data (outliers) and model complexity propagate.

However, Bayes needs more work. For example, priors $P[\theta]$ must be elicitated from contextual information. Contextual information is usually provided by mean, standard deviation, minimum, maximum (range). A good understanding of properties of different distributions is necessary in order to define a correct $P[\theta]$ prior. See, for example, distributions zoo: Leemis and McQueston [2008]. Moreover, good communication with experts is necessary to get the correct information. Bayesian computation comprises:

- Conjugate analyses
- MCMC sampling: R, JAGS, OpenBUGS, Stan
- Bayesian numerical approximations: INLA, bayesmeta

Recommended reading: Bayarri and Berger [2004], Martin et al. [2020], and Johnson et al. [2022]. You can also check interactive visualizations Seeing Theory `http://students.brown.edu/seeing-theory/index.html`.

1.4 Probability calculus

The probability calculus is based on three axioms:

$P[A] \geq 0$

$P[A] = 1$ if A is true

$P[A \text{ or } B] = P[A \cup B] = P[A] + P[B]$ if $A \cap B = \emptyset$ and $P[A \text{ and } B] = P[A \cap B] = 0$ (events A and B are mutually exclusive).

There are several important properties of probabilities.

Conditional probability

$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{P[A \cap B]}{P[B]}, \tag{1.1}$$

given that $P[B] > 0$.

Two events A and B are called independent if the occurrence of B does not change the probability of A

$$P[A \mid B] = P[A]$$

and vice versa

$$P[B \mid A] = P[B].$$

Thus,

$$P[A \text{ and } B] = P[A]P[B].$$

Note that from Equation (1.1)

$$P[A \text{ and } B] = P[A \mid B]P[B] = P[B \mid A]P[A].$$

This observation leads to the **Bayes theorem**

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B]}. \tag{1.2}$$

Assume that event A has a disjoint, complementary event A^c such that $P[A]+P[A^c] = 1$. Conditional probabilities behave like ordinary probabilities, so that we have

$$P[A \mid B] + P[A^c \mid B] = 1.$$

This leads to the simplest version of the law of total probability:

$$P[B] = P[B \mid A]P[A] + P[B \mid A^c]P[A^c].$$

Therefore, the **Bayes theorem** from Equation (1.2) can be rewritten as

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B \mid A]P[A] + P[B \mid A^c]P[A^c]}. \tag{1.3}$$

For formulas applying to more than two events see Held and Sabanés Bové [2020, Sections A.1.1–A.1.2].

Note that there is a link between probability P and odds O:

$$O = \frac{P}{1-P}$$

and

$$P = \frac{O}{1+O}.$$

We can obtain the odds form for the Bayes theorem

$$\frac{P[A \mid B]}{P[A^c \mid B]} = \frac{P[B \mid A]}{P[B \mid A^c]} \frac{P[A]}{P[A^c]}, \tag{1.4}$$

by dividing Equation (1.2) by the same equation applied to the disjoint, complementary event A^c instead of A.

The ratio

$$\frac{P[A \mid B]}{P[A^c \mid B]} = \frac{P[A \mid B]}{1 - P[A \mid B]}$$

is called posterior odds, and

$$\frac{P[A]}{P[A^c]} = \frac{P[A]}{1 - P[A]}$$

is called prior odds, and the ratio

$$\frac{P[B \mid A]}{P[B \mid A^c]}$$

is the Bayes factor (likelihood ratio).

Remark: This Bayes factor is a measure of evidence [Held and Ott, 2018] of the null hypothesis H_0 against an alternative hypothesis H_A, when we replace events A and A^c in Equation (1.4) by H_0 and H_A.

Remark: One can also derive a conditional version of the Bayes theorem

$$P[A \mid B, I] = \frac{P[A \mid I]P[B \mid A, I]}{P[B \mid I]},$$

where I is an additional piece of information.

Recommended reading: Held and Sabanés Bové [2020]: Sections 6.1 and 6.2, A1, A1.1, A1.2, A2.1, A2.2, A2.3. See also Rouder and Morey [2019] for a deeper insight into the meaning of the Bayes theorem.

1.5 Example: Breast cancer and diagnostic tests

This section demonstrates on one medical example that $P[D^+ \mid T^+] \neq P[T^+ \mid D^+]$.

Assumptions

Prevalence of breast cancer: $P[D^+] = 0.045$

Sensitivity: $P[T^+ \mid D^+] = 0.866$

Specificity: $P[T^- \mid D^-] = 0.968$

Full bivariate distribution $P[T \cap D] =$

$$\begin{array}{c} \\ T^+_{\text{yes}} \\ T^-_{\text{no}} \end{array} \begin{array}{c} \begin{array}{cc} D^+_{\text{yes}} & D^-_{\text{no}} \end{array} \\ \begin{pmatrix} 0.03897 & 0.03056 \\ 0.00603 & 0.92444 \end{pmatrix} \end{array}$$

marginal distribution for Test $P[T] =$

$$\begin{array}{c} \\ T^+ \\ T^- \end{array} \begin{array}{c} \\ \begin{pmatrix} 0.06953 \\ 0.93047 \end{pmatrix} \end{array}$$

marginal distribution for Disease $P[D] =$

$$\begin{array}{c} \\ D^+ \\ D^- \end{array} \begin{array}{c} \\ \begin{pmatrix} 0.045 \\ 0.955 \end{pmatrix} \end{array}$$

$$P[D^- \mid T^-] = \frac{P[D^- \cap T^-]}{P[T^-]} = \frac{0.92444}{0.93047} = 0.994$$

$$P[D^+ \mid T^+] = \frac{P[D^+ \cap T^+]}{P[T^+]} = \frac{0.03897}{0.06953} = 0.56 \neq P[T^+ \mid D^+]$$

Remark: Another way of computing

$$P[D^+ \mid T^+] = \frac{P[D^+]P[T^+ \mid D^+]}{P[T^+]} = \frac{P[D^+]P[T^+ \mid D^+]}{P[D^+]P[T^+ \mid D^+] + P[D^-]P[T^+ \mid D^-]}$$

1.6 Overview of the classical statistic

The classical statistic is based on the likelihood (Figure 1.1).

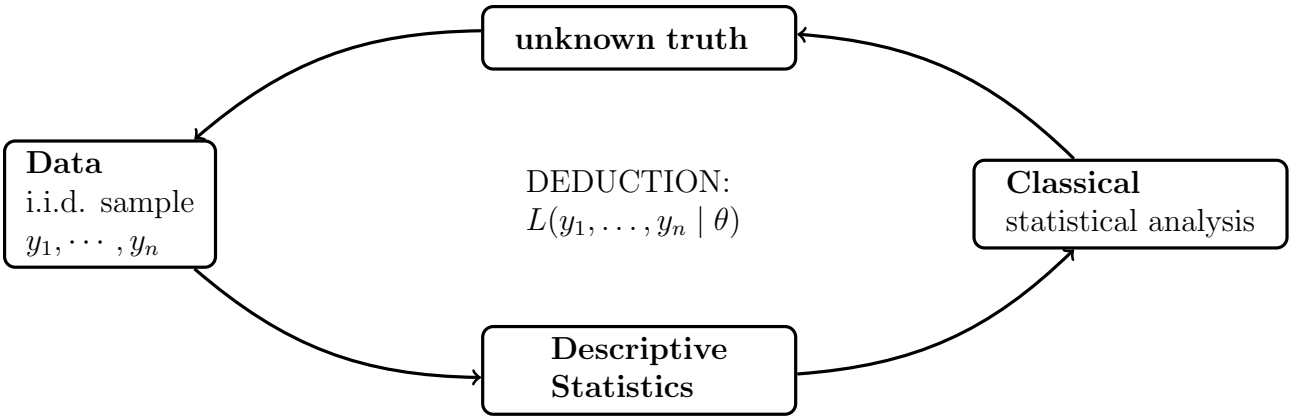

Figure 1.1: Overview of the classical statistic

1.6.1 Example: Primary outcome follows normal distribution

Data $Y \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$ and $\theta = \mu$.

Density $f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$.

Likelihood

$$\begin{aligned} L(y_1, \ldots, y_n \mid \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}, \end{aligned} \tag{1.5}$$

and the log-likelihood

$$\log L(y_1, \ldots, y_n \mid \mu) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

In order to derive an estimator of μ, compute

$$\frac{d \log L(y_1, \ldots, y_n \mid \mu)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \mu)(-1)\Big|_{\mu=\hat{\mu}} = 0,$$

$$\sum_{i=1}^{n} y_i - n\hat{\mu} = 0.$$

Thus,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

One can also derive that $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2}$.

1.6.2 Example: Primary outcome follows Bernoulli distribution

Let $Y \overset{i.i.d.}{\sim} \text{Be}(p)$ with $\theta = p$ and $P[Y = 0] = 1 - p$ and $P[Y = 1] = p$.

Density $f(y_i) = p^{y_i}(1-p)^{1-y_i}$.

Likelihood

$$\begin{aligned} L(y_1, \ldots, y_n \mid p) &= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}, \end{aligned} \tag{1.6}$$

Log-likelihood

$$\log L(y_1, \ldots, y_n \mid p) = \sum_{i=1}^{n} y_i \log p + \left(n - \sum_{i=1}^{n} y_i\right) \log(1-p).$$

In order to derive an estimator of p, compute

$$\frac{d \log L(y_1, \ldots, y_n \mid p)}{dp} = \sum_{i=1}^{n} y_i \frac{1}{p} + \left(n - \sum_{i=1}^{n} y_i\right) \frac{1}{1-p}(-1)\Bigg|_{p=\hat{p}} = 0,$$

$$\sum_{i=1}^{n} y_i \frac{1}{\hat{p}} - \left(n - \sum_{i=1}^{n} y_i\right) \frac{1}{1-\hat{p}} = 0,$$

$$\sum_{i=1}^{n} y_i (1-\hat{p}) = \left(n - \sum_{i=1}^{n} y_i\right) \hat{p},$$

$$\sum_{i=1}^{n} y_i = n\hat{p},$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

1.7 Overview of the Bayesian methodology

There is a consent that probability calculus leading to the Bayes formula in Equation (1.2) is objective. Bayesian methodology extends the classical approach based on the likelihood and considers

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

More specifically,

$$P[\theta \mid y_1, \ldots, y_n] \propto \text{L}(y_1, \ldots, y_n \mid \theta) \times P[\theta].$$

Figure 1.2 provides an overview of the Bayesian methodology and its relation to the classical statistics.

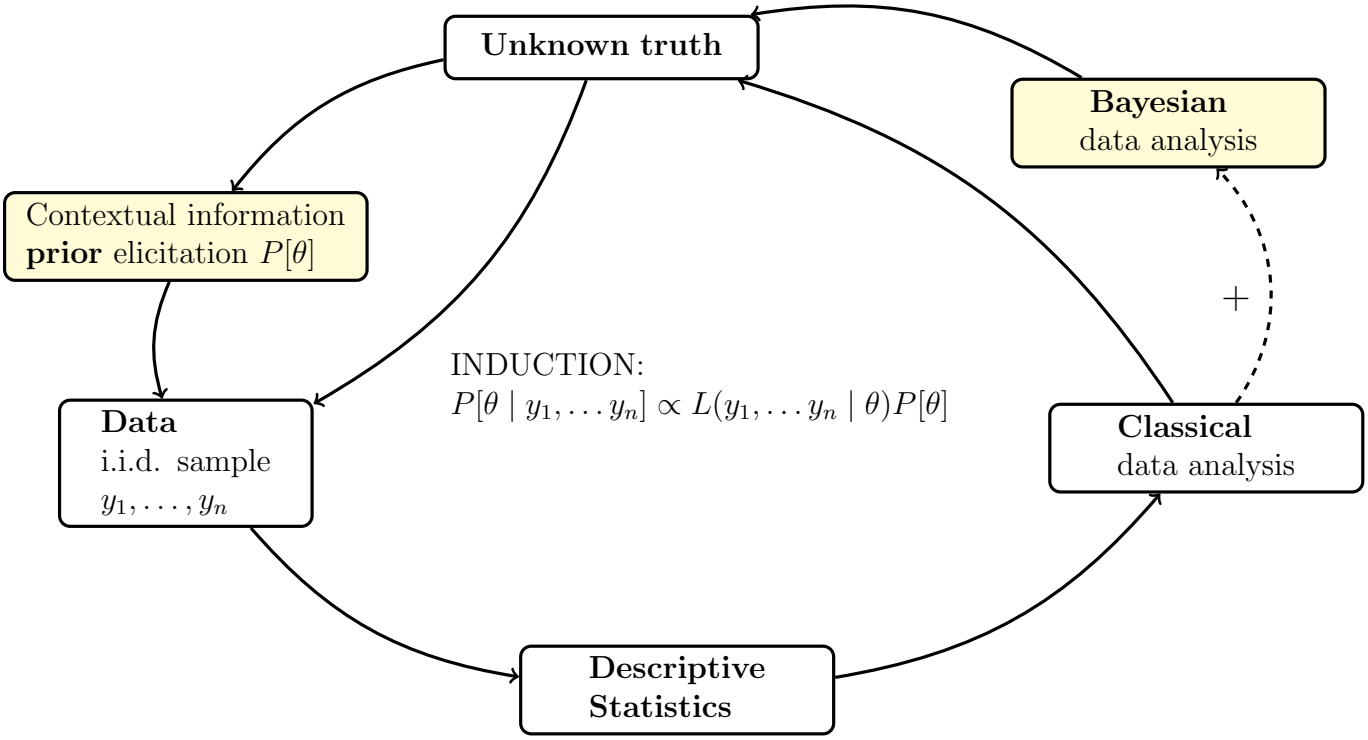

Figure 1.2: Overview of the Bayesian methodology. Fields with yellow background correspond to Bayes-specific steps.

1.7.1 Bayes factors and p-values

The Bayesian methodology enables an independent view of classical hypothesis testing. In applications, the estimation of the credibility of a conclusion expressed by the probability of H_0 given the data is usually of primary interest. The Bayes factor directly quantifies whether the data have increased or decreased the odds of H_0. Thus, Bayes factors facilitate direct conclusions about the probability of H_0 given the data, provided that both null H_0 and alternative H_1 hypotheses have been specified.

On page 70, Held and Sabanés Bové [2020] define the p-value: the probability, under the assumption of the null hypothesis H_0, of obtaining a result equal to or more extreme than what was actually observed. A p-value is computed under the assumption that the null hypothesis H_0 is true. It does not allow for conclusions about the probability of H_0 given the data. A particular p-value can be obtained either for a large study with a small effect or for a small study with a large effect. Thus, the p-value does not say anything about the actual effect or evidence that such an effect exists.

Consider a significance test with a point null hypothesis $H_0 : \theta = \theta_0$. The alternative hypothesis can be either simple $H_1 : \theta = \theta_1 \neq \theta_0$ or composite $H_1 : \theta \neq \theta_0$. For a composite H_1 a prior distribution $f(\theta \mid H_1)$ must be specified.

Note that $P[H_1] = 1 - P[H_0]$ and $P[y] = f(y \mid H_0)P[H_0] + f(y \mid H_1)P[H_1]$. The Bayes formula for H_0

$$P[H_0 \mid y] = \frac{f(y \mid H_0)P[H_0]}{P[y]} \tag{1.7}$$

divided by the Bayes formula for H_1

$$P[H_1 \mid y] = \frac{f(y \mid H_1)P[H_1]}{P[y]} \tag{1.8}$$

render

$$\frac{P[H_0 \mid y]}{P[H_1 \mid y]} = BF_{01}(y)\frac{P[H_0]}{P[H_1]}, \tag{1.9}$$

where

$$BF_{01}(y) = \frac{f(y \mid H_0)}{f(y \mid H_1)}. \tag{1.10}$$

Note that the Bayes factor $BF_{01}(y)$ transforms the prior odds $P[H_0]/P[H_1]$ into posterior odds $P[H_0 \mid y]/P[H_1 \mid y]$ in the light of the data y. $BF_{01}(y)$ is a direct quantitative measure of how data y have increased or decreased the odds of H_0 and is referred to as the strength of evidence for or against H_0. The evidence against the null hypothesis H_0 is provided by small Bayes factors $BF_{01}(y) < 1$. The evidence in favor of the null hypothesis H_0 is provided by large Bayes factors $BF_{01}(y) > 1$. Table 2 of Held and Ott [2018] provides a categorization of Bayes factors $BF_{01}(y) \leq 1$ into levels of evidence against H_0: weak (1 to 1/3), moderate (1/3 to 1/10), substantial (1/10 to 1/30), strong (1/30 to 1/100), very strong (1/100 to 1/300), and decisive ($< 1/300$).

$BF_{01}(y)$ is the ratio of the likelihood $f(y \mid H_0) = f(y \mid \theta = \theta_0)$ of the observed data y under the null hypothesis H_0 and the marginal likelihood

$$f(y \mid H_1) = \int f(y \mid \theta) f(\theta \mid H_1) d\theta \tag{1.11}$$

under the alternative hypothesis H_1. Equation (1.11) is useful for composite alternative hypotheses H_1. It is the average likelihood $f(y \mid \theta)$ with respect to the prior distribution $f(\theta \mid H_1)$ for θ under the alternative H_1, which is called marginal likelihood (prior predictive distribution at the observed data). For a simple alternative, Equation (1.11) reduces to the likelihood $f(y \mid H_1) = f(y \mid \theta = \theta_1)$ and the $BF_{01}(y)$ reduces to a likelihood ratio.

Once we know $BF_{01}(y)$, we can solve the formula in Equation (1.9) for the posterior probability of H_0. Note that

$$\frac{P[H_0 \mid y]}{1 - P[H_0 \mid y]} = BF_{01}(y)\frac{P[H_0]}{P[H_1]}.$$

Thus,

$$P[H_0 \mid y] = \frac{BF_{01}(y)\frac{P[H_0]}{P[H_1]}}{1 + BF_{01}(y)\frac{P[H_0]}{P[H_1]}}. \tag{1.12}$$

Note that Bayes factors facilitate multiple hypothesis comparisons because they can be updated sequentially:

$$BF_{01}(y)BF_{12}(y) = \frac{f(y \mid H_0)}{f(y \mid H_1)}\frac{f(y \mid H_1)}{f(y \mid H_2)} = \frac{f(y \mid H_0)}{f(y \mid H_2)} = BF_{02}(y).$$

The minimum Bayes factor is the smallest Bayes factor within a certain class of alternative hypotheses. Minimum Bayes factors are very interesting because they quantify the maximal evidence of a p-value against a point H_0 within a certain class of alternative hypotheses.

The Bayesian approach provides a way of transforming p-values to direct measures of evidence against the null hypothesis expressed by Bayes factors. This transformation is called calibration. Held and Ott [2018] consider different transformations of p-values to minimum Bayes factors and show that minimum Bayes factors provide less evidence against the null hypothesis than the corresponding p-value might suggest. They also demonstrate that many techniques have been proposed to calibrate p-values and there is no consensus which calibration is the optimal one.

Recommended reading: Held and Sabanés Bové [2020] Sections 3.3 and 7.2.1, Goodman [1999b], Goodman [1999a], Held and Ott [2018], and `pCalibrate` package.

Example: Discuss `pCalibrate` to show the calibration of p-values by Bayes factors on the border between the classical and the full Bayes analysis.

1.7.2 Priors

The use of prior distributions for Bayesian analysis can be controversial. Therefore, a good understanding of different distributions is very important.

- Discussion of different distributions Leemis and McQueston [2008].
- Monte Carlo (MC) simulations vs true parameters (expectation and variance).
- The Change-of-Variables Formula Held and Sabanés Bové [2020, Section A.2.3].

  Assume a one-to-one and differentiable transformation $g(\dot{})$. Assume that the random variable Y with probability density function $f_Y(y)$ is a transformation of a continuous random variable X with probability density function $f_X(x)$, where g is a one-to-one and differentiable transformation and $Y = g(X)$. Then

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{dg^{-1}(y)}{dy}\right|.$$

1.8 Worksheet 1

Probability calculus	**Distributions**	Change of variables formula
Priors	**MC sampling**	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		**Likelihood**
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 1.3: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 1.

Secukinumab		**Placebo**
Classical Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		**Data**
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 1.4: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 2

Lecture 2: Conjugate Bayes, point estimates, and interval estimates

This chapter deals with the conjugate Bayes argument and the resulting posterior Bayesian estimates. The Bayes theorem from Equation (1.2) can be rewritten in terms of densities

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)},$$

where $f(y) = \int f(y \mid \theta) f(\theta) d\theta$ is called the marginal likelihood or the prior predictive distribution at the observed data y. A conjugate Bayes emerges when both the likelihood and the prior are based on distributions that allow for a direct multiplication of distributional cores. This leads to a simple rule of computation

$$f(\theta \mid y) \propto f(y \mid \theta) f(\theta),$$

which will be demonstrated on two examples.

This simple rule can be perceived in two different ways. First, we start with the prior $f(\theta)$ which is then modified by the current data contained in the likelihood $f(y \mid \theta)$ to get the posterior $f(\theta \mid y)$. Second, we start with the current data summarized by the likelihood $f(y \mid \theta)$ and this likelihood is modified by the prior $f(\theta)$ to get the posterior $f(\theta \mid y)$. Depending on application, the first or the second way of reading this simple rule is more useful, which makes the Bayesian methodology very practicable.

Recommended reading: Held and Sabanés Bové [2020] Sections 6.3.1 and 6.4 and Hartnack and Roos [2021]. Note that Table 6.2 of Held and Sabanés Bové [2020] is very relevant. It shows which posterior distributions can be derived analytically given a likelihood and a conjugate prior.

Remark: Rouder and Morey [2019] discuss in detail the ratio form of the Bayes theorem

$$\frac{f(\theta \mid y)}{f(\theta)} = \frac{f(y \mid \theta)}{f(y)}$$

in terms of updating factors. Whereas the factor to the left denotes the strength of evidence form the data about θ, the factor to the right denotes the gain in predictive accuracy for θ, i.e. how well the data are predicted when conditioned on the value of θ relative to the marginal prediction. Thus, the strength of evidence is the relative predictive accuracy.

2.1 Binary data: Vision correction

We consider an example of vision correction. In a class with 20 – 30 years old students, 16 participants out of 22 required vision correction. What is the true probability π that young students need vision correction?

We know from the classical statistics that the best practice to analyze these data is to provide both a point estimate $\hat{\pi} = 16/22 = 0.727$ and an interval estimate in form of a 95% confidence interval for the true probability π (95%CI(π)) [Hartnack and Roos, 2021].

```
library(DescTools)

BinomCI(x = 16, n = 22, conf.level = 0.95, method = "wilson")

##            est    lwr.ci    upr.ci
## [1,] 0.7272727 0.5184827 0.8684924
```

The interpretation of a classical 95%CI is based on a repeated execution of the same experiment. A confidence interval with confidence $1 - \alpha$ provides limits T_l and T_u such that for the parameter of interest θ

$$P[T_l \leq \theta \leq T_u] = 1 - \alpha$$

holds. Classical confidence intervals aim to uniformly provide a prespecified coverage probability conditionally on any single point in parameter space.

Theoretical interpretation: For repeated random samples from a distribution with unknown parameter θ, a $(1 - \alpha)\dot{1}00\%$ confidence interval will cover θ in $(1 - \alpha)\dot{1}00\%$ of all cases (Held and Sabanés Bové [2020] on page 57 in Section 3.2.2).

Practical interpretation: In the vision correction example for the random sample at hand, we can say that the 95%CI for the true probability π of vision correction is 95%CI(π) (0.518, 0.868). Although we do not know whether this 95%CI(π) covers the true probability π of vision correction, we know that the procedure used for computation of the 95%CI(π) has a desirable characteristic. For repeated independent random samples drawn from the distribution of vision correction with the unkown true parameter π, 95%CI(π) intervals cover the true probability π in approximately 95% of cases.

2.2 Bayes analysis of binary data

2.2.1 Beta prior

A prior expresses contextual information or knowledge in form of a probability distribution. The Beta distribution is based on the observation that the integral $\int_0^1 u^{x-1}(1-u)^{y-1}du$ exists. This integral is called the Beta function $B(x, y)$.

- Beta function

$$\int_0^1 u^{x-1}(1-u)^{y-1}du = B(x, y),$$

where

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

and

$$\Gamma(x) = (x-1)!$$

- Beta distribution

  Let

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{2.1}$$

  be the density of the Beta(α, β) distribution with two shape parameters α and β. Then

$$\frac{1}{B(\alpha, \beta)} \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = 1.$$

Beta distribution is very flexible. It can attain different forms which can be symmetric and asymmetric. Elicitation of shape parameters α and β by moments matching is a convenient way to define a Beta prior (See your individual project Part 2B).

$$X \sim \mathrm{Beta}(\alpha, \beta)$$

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta}$$

$$\mathrm{Var}X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mathbb{E}(X)(1-\mathbb{E}X)}{(\alpha+\beta+1)}.$$

Prior effective sample size

$$\mathrm{PriESS} \approx \frac{1}{\mathrm{Var}(X)} \approx \frac{\mathbb{E}(X)(1-\mathbb{E}X) - \mathrm{Var}(X)}{\mathrm{Var}(X)} = \alpha + \beta \tag{2.2}$$

Therefore, it is convenient to think about $\alpha+\beta$ as a prior sample size. This number informs us about the weight of the prior.

Example Vision correction:
We consider three Beta priors, which are depicted in Figure 2.1.

- skeptical prior $\alpha + \beta = 0.5 + 0.5 = 1$ (equivalent to 1 observation)
- neutral prior $\alpha + \beta = 1 + 1 = 2$ (equivalent to 2 observations)
- enthusiastic prior $\alpha + \beta = 12 + 12 = 24$ (equivalent to more observations than in the final sample of 22)

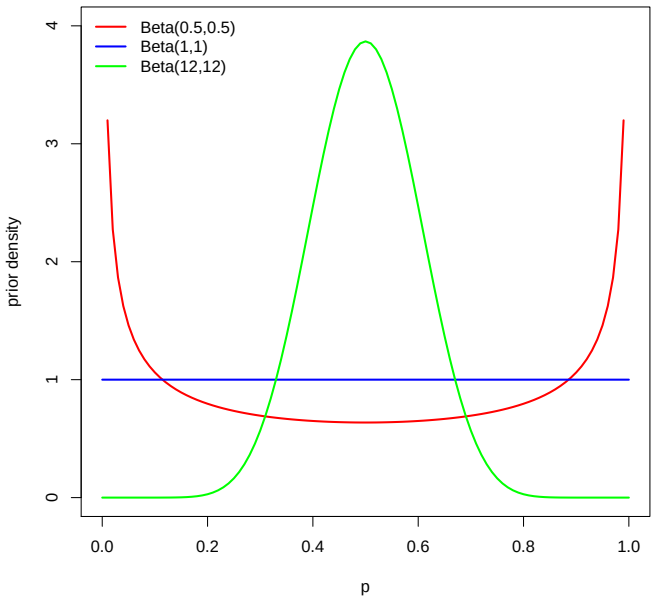

Figure 2.1: Beta priors in the vision correction example.

2.2.2 Likelihood

We assume that each binary observation y_i is a realization of independent identically distributed ($i.i.d.$) random variables, which follow the Bernoulli Be(p) distribution:

$$y_i \overset{i.i.d.}{\sim} \text{Be}(p) = \begin{cases} 1, & \text{with } p \\ 0, & \text{with } 1-p \end{cases} = p^{y_i}(1-p)^{1-y_i}, \quad i = 1, \cdots, n.$$

The likelihood is equal

$$L(y_1, \cdots, y_n \mid p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} = p^{n\bar{y}}(1-p)^{n-n\bar{y}}, \qquad (2.3)$$

where $n\bar{y} = n(\frac{1}{n}\sum_{i=1}^n y_i) = \sum_{i=1}^n y_i$ is the number of binary observations attaining value 1 in the sample of n observations. This likelihood is proportional to the binomial likelihood.

2.2.3 Posterior distribution

We begin with the computation of the posterior distribution with all constants. Note that due to conjugacy, the numerator of the Bayes formula, which is the multiplication of the likelihood $L(y_1, \cdots, y_n \mid p)$ from Equation (2.3) and the Beta(α, β) prior with density in Equation (2.1), is equal:

$$L(y_1, \cdots, y_n \mid p)f(p) = p^{n\bar{y}}(1-p)^{n-n\bar{y}}\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1} = \frac{1}{B(\alpha,\beta)}p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}.$$

Therefore,

$$
\begin{aligned}
f(p \mid y_1, \cdots, y_n) &= \frac{L(y_1, \cdots, y_n \mid p) f(p)}{\int_0^1 L(y_1, \cdots, y_n \mid p) f(p) dp} \\
&= \frac{\frac{1}{B(\alpha,\beta)} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}}{\frac{1}{B(\alpha,\beta)} B(\alpha+n\bar{y}, \beta+n-n\bar{y}) \underbrace{\int_0^1 \frac{1}{B(\alpha+n\bar{y}, \beta+n-n\bar{y})} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1} dp}_{=1}} \\
&= \frac{1}{B(\alpha+n\bar{y}, \beta+n-n\bar{y})} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}.
\end{aligned}
\tag{2.4}
$$

Thus, $p \mid y_1, \cdots, y_n \sim \text{Beta}(\alpha + n\bar{y}, \beta + n - n\bar{y})$.

Alternatively, we can identify the posterior distribution based on cores of distributions:

$$
\begin{aligned}
f(p \mid y_1, \cdots, y_n) &\propto L(y_1, \cdots, y_n \mid p) f(p) \\
&= \underbrace{p^{n\bar{y}}(1-p)^{n-n\bar{y}}}_{\text{likelihood kernel}} \underbrace{p^{\alpha-1}(1-p)^{\beta-1}}_{\text{prior kernel}} \\
&= \underbrace{p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}}_{\text{kernel of the posterior distribution Beta}(\alpha+n\bar{y}, \beta+n-n\bar{y})}.
\end{aligned}
\tag{2.5}
$$

Again, $p \mid y_1, \cdots, y_n \sim \text{Beta}(\alpha + n\bar{y}, \beta + n - n\bar{y})$.

The expectation of the posterior distribution is equal:

$$
\begin{aligned}
\mathbb{E}(p \mid y_1, \cdots, y_n) &= \frac{\alpha + n\bar{y}}{\alpha + n\bar{y} + \beta + n - n\bar{y}} = \frac{\alpha + n\bar{y}}{\alpha + \beta + n} \\
&= \frac{\alpha}{\alpha+\beta+n} + \frac{n\bar{y}}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n} \underbrace{\frac{\alpha}{\alpha+\beta}}_{\mathbb{E}(\text{prior})} + \big(1 - \frac{\alpha+\beta}{\alpha+\beta+n}\big) \underbrace{\bar{y}.}_{\text{MLE}}
\end{aligned}
\tag{2.6}
$$

It is a weighted average of the prior mean $\frac{\alpha}{\alpha+\beta}$ and the ML estimate $\bar{y}$. The relative prior sample size $\frac{\alpha+\beta}{\alpha+\beta+n}$ quantifies the weight of the prior mean in the posterior expectation. Note that this quantity decreases to 0 with data sample size increasing to ∞.

Posterior effective sample size (PostESS) of a $\text{Beta}(\alpha + n\bar{y}, \beta + n - n\bar{y})$ distribution is approximatively equal:

$$
PostESS \approx \overbrace{\frac{1}{\text{Var}(p \mid y_1, \cdots, y_n)}}^{\text{posterior precision}} \approx \alpha + n\bar{y} + \beta + n - n\bar{y} = \alpha + \beta + n. \tag{2.7}
$$

Note the analogy to the prior effective sample size.

Example: Vision correction

Data: $n\bar{y} = 16$ participants out of $n = 22$ required vision correction. Note that $n - n\bar{y} = 6$ participants did not require any vision correction. Figure 2.2 shows the standardized likelihood, the Beta(0.5, 0.5) prior and the resulting Beta(16.5, 6.5) posterior distributions.

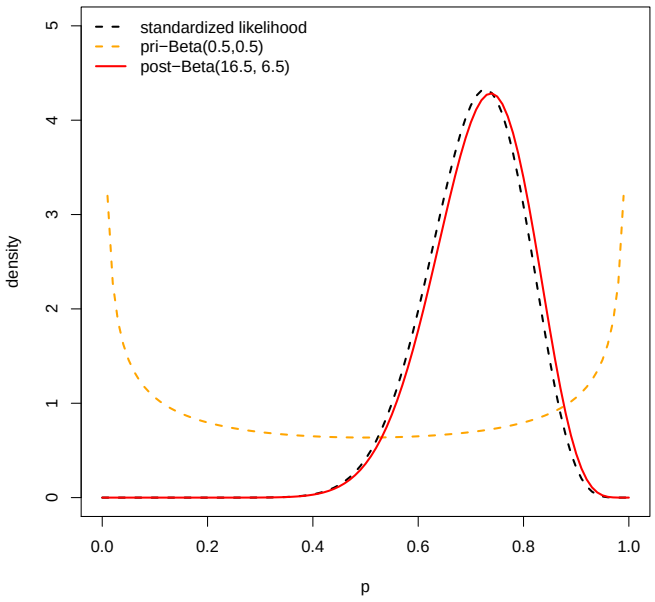

Figure 2.2: Likelihood, prior Beta(0.5, 0.5), posterior Beta(16.5, 6.5) in the vision correction example with $n\bar{y} = 16$ and $n = 22$.

Posterior distributions can be summarized by point and interval estimates. For example, for the Beta(16.5, 6.5) posterior we get the posterior mean $16.5/(16.5 + 6.5) = 0.717$. We can also compute an equi-tailed 95% credible interval (95%CrI):

```
qbeta(p = c(0.025, 0.975), shape1 = 16.5, shape2 = 6.5)

## [1] 0.5217688 0.8772947
```

The **interpretation of a 95%CrI** differs from that of confidence intervals. We can state that the posterior probability of vision correction p lies between 0.521 and 0.877 with probability 95%, when a Beta(0.5, 0.5) prior is assumed. This result addresses the actual question more directly and can be interpreted intuitively. Bayesian credible intervals account for the prior distribution and provide a coverage on average over the prior. They directly relate to the knowledge about the parameter after considering the data at hand.

For different priors, we get

- The skeptical prior Beta(0.5, 0.5) leads to a Beta(16.5, 6.5) posterior.
- The neutral prior Beta(1, 1) leads to a Beta(17, 7) posterior.
- The enthusiastic prior Beta(12, 12) leads to a Beta(28, 18) posterior.

These posterior distributions are shown in Figure 2.3.

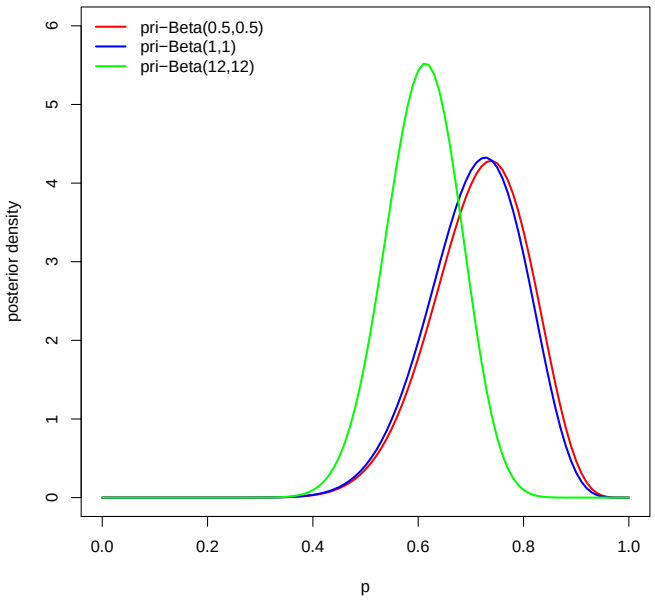

Figure 2.3: Posteriors in the vision correction example obtained for three different priors: Beta(0.5, 0.5), Beta(1, 1), and Beta(12, 12).

Note that posterior results depend on the prior. For the Beta(28, 18) posterior we get the posterior mean $28/(28+18) = 0.609$ and the equi-tailed 95%CrI:

```
qbeta(p = c(0.025, 0.975), shape1 = 28, shape2 = 18)

## [1] 0.4654101 0.7430241
```

2.3 Bayes analysis of normal data

Assume that $y_1, \ldots, y_n$ are realizations (observations) generated by *i.i.d.* random variables which follow a $\mathrm{N}(m, \kappa^{-1})$ distribution. Assume that the prior of m follows a $\mathrm{N}(\mu, \lambda^{-1})$ distribution, where κ, μ and λ are fixed (known) constants. We derive the posterior distribution of m given that $y_1, \ldots, y_n$ have been observed.

According to the Bayes formula in Equation (1.2), which has been rewritten in terms of densities, we get:

$$f(m \mid y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n \mid m) f(m)}{\int_{-\infty}^{\infty} f(y_1, \ldots, y_n \mid m) f(m) dm}. \tag{2.8}$$

The denominator is known as the marginal likelihood:

$$\int_{-\infty}^{\infty} f(y_1, \ldots, y_n \mid m) f(m) dm = \int_{-\infty}^{\infty} f(y_1, \ldots, y_n, m) dm = f(y_1, \ldots, y_n).$$

We derive the posterior based on kernels of distributions and use:

$$\underbrace{f(m \mid y_1, \ldots, y_n)}_{\text{posterior}} \propto \underbrace{f(y_1, \ldots, y_n \mid m)}_{\text{likelihood}} \underbrace{f(m)}_{\text{prior}}. \tag{2.9}$$

We combine the likelihood

$$f(y_1, \ldots, y_n \mid m) = \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i - m)^2\right\}$$

and the prior

$$f(m) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(m-\mu)^2\right)$$

and, following Equation (2.9), we get

$$f(m \mid y_1, \ldots, y_n) \propto \exp\left\{-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i - m)^2 - \frac{\lambda}{2}(m-\mu)^2\right\}.$$

At this stage, formulas for combining quadratic forms (Held and Sabanés Bové [2020] Section B.1.5) can be applied to show that

$$f(m \mid y_1, \ldots, y_n) \propto \exp\left\{-\frac{(n\kappa+\lambda)}{2}\Big(m - \frac{\kappa n\bar{y} + \lambda\mu}{n\kappa+\lambda}\Big)^2\right\}.$$

Thus, we get

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n\bar{y} + \lambda\mu}{n\kappa+\lambda}, (n\kappa+\lambda)^{-1}\right).$$

Note that the expectation of the posterior can be rewritten as

$$\frac{\lambda}{n\kappa+\lambda}\mu + \big(1 - \frac{\lambda}{n\kappa+\lambda}\big)\bar{y}.$$

The weight $\frac{\lambda}{n\kappa+\lambda}$ of the prior mean in the posterior expectation decreases to 0 with data sample size increasing to ∞. Moreover, the variance of the posterior distribution $(n\kappa+\lambda)^{-1}$ is smaller than the variance λ^{-1} of the prior distribution, because the precision is larger. The posterior variance decreases to 0 with data sample size increasing to ∞. In addition, the prior effective sample size is equal $PriESS \approx \lambda$ and the posterior effective sample size $PostESS \approx n\kappa + \lambda$ is larger than $PriESS$.

2.4 Point estimates

Bayesian point estimates such as mean, mode, and median have a deeper decision-theoretic meaning. They minimize an expected loss with respect to the posterior distribution. For example, the posterior mean minimizes the quadratic loss function $l(a, \theta) = (a-\theta)^2$, because the first derivative with respect to a of $\mathbb{E}(l(a,\theta) \mid y) = \int (a-\theta)^2 f(\theta \mid y) d\theta$ set to 0 results in $a = \int \theta f(\theta \mid y) d\theta = \mathbb{E}(\theta \mid y)$. For more details on point estimates and loss functions see Held and Sabanés Bové [2020, Section 6.4].

2.5 Credible intervals

There are at least two ways to compute Bayesian credible intervals $(1-\alpha)100\%\text{CrI}(\theta)$:

(a) equi-tailed credible intervals

(b) highest posterior density (HPD) intervals.

These Bayesian credible intervals allow for direct probability statements. However, they have different properties.

An equi-tailed $(1-\alpha)$ credible interval has $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$ quantiles of $\pi(\theta \mid y)$ at its endpoints. One discards equal amounts of posterior probability on either side of the interval. An equi-tailed credible interval is:

- Intuitively straightforward

- Easy to compute from MC and MCMC samples

- Has a nice invariance property

Let h be a monotone function (can be non-linear). A $(1-\alpha)$ equi-tailed credible interval for $h(\theta \mid y)$ can be obtained by applying $h()$ to the endpoints of the $(1-\alpha)$ equi-tailed credible interval for $\theta \mid y$. There is no concern about the chosen scale for inference. In fact, this property applies to all quantiles (also median) of the posterior.

Remark: The invariance property doesn't hold for expectations. In fact, there is functional non-invariance for a non-linear function $g()$:

$$\mathbb{E}(g(Y)) = \int_{\Omega} g(u) dP_Y \neq g(\mathbb{E}(Y)).$$

For example, if g is convex then $\mathbb{E}(g(Y)) \geq g(\mathbb{E}(Y))$ and if g is concave $\mathbb{E}(g(Y)) \leq g(\mathbb{E}(Y))$. See Jensen's inequality in [Held and Sabanés Bové, 2020] page 354, Section A.3.7.

The highest posterior probability HPD interval fulfills $\mathbb{P}[\theta : f(\theta \mid y) > c] = 1-\alpha$ and provides the shortest possible interval. For symmetric and unimodal posteriors, it coincides with $(1-\alpha)$ equi-tailed credible intervals. But for bimodal or multimodal posteriors this correspondence does not hold. Note that the invariance property for transformation $h()$ does not hold any more. For more details on Bayesian HPD credible intervals see Held and Sabanés Bové [2020, Section 6.4].

Remark: Sequential step by step vs pooled data in one step.

Assume a sequence of three measurements y_1, y_2, y_3. Then,

$$\begin{aligned}
\mathbb{P}[\theta \mid y_1, y_2, y_3, I] &\propto \mathbb{P}[y_3 \mid \theta, y_1, y_2, I]\mathbb{P}[\theta \mid y_1, y_2, I] \\
&\propto \mathbb{P}[y_3 \mid \theta, y_1, y_2, I]\mathbb{P}[y_2 \mid \theta, y_1, I] \underbrace{\mathbb{P}[y_1 \mid \theta, I] \overbrace{\mathbb{P}[\theta \mid I]}^{\text{prior}}}_{\propto\, \mathbb{P}[\theta \mid y_1, I]} \\
&\propto \underbrace{\prod_{i=1}^{3} \mathbb{P}[y_i \mid \theta, I]}_{\text{pooled likelihood}} \mathbb{P}[\theta \mid I].
\end{aligned}$$

Example:

Step by step

1.1 Prior Beta(α, β)

1.2 Data y_1

1.3 Posterior $(\alpha + y_1, \beta + 1 - y_1)$

2.1 Prior Beta$(\alpha + y_1, \beta + 1 - y_1)$

2.2 Data y_2

2.3 Posterior
Beta$(\alpha + y_1 + y_2, \beta + 2 - (y_1 + y_2))$

Pooled

a Prior Beta(α, β)

b Data y_1, y_2

c Posterior
Beta$(\alpha + y_1 + y_2, \beta + 2 - (y_1 + y_2))$

Note that $y_1 + y_2$ is a sufficient statistic with respect to the Binomial likelihood "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (p)".

2.6 Worksheet 2

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		Likelihood
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 2.1: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 2.

Secukinumab		**Placebo**
Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		**Beta(11, 32)**
Data (S)		**Data (P)**
	Classical analysis	
Posterior (S)		**Posterior (P)**
	Posterior probability of superiority	

Table 2.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 3

Lecture 3: Predictive distributions, asymptotics, and Monte Carlo simulations

This chapter further explores the strength and the beauty of the Bayes formula from Equation (1.2) rewritten in terms of densities:

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)}.$$

Bayesian analysis can be conducted at several stages.

Prior stage:

- prior distribution
- prior predictive distribution

Intermediate stage:

- intermediate posterior distribution
- intermediate posterior predictive distribution

Final stage:

- final posterior distribution
- final posterior predictive distribution

This chapter deals with prior predictive and posterior predictive distributions, independent Monte Carlo sampling, and Bayesian asymptotics.

Recommended reading: Held and Sabanés Bové [2020] Sections 9.3.1, 9.3.2, 8.3.1, and 6.6. See also Spiegelhalter et al. [1994] for an extensive and nuanced discussion of Bayesian approaches to randomized clinical trials.

3.1 Predictive distributions for binary data

We use notation from Section 2.2 and extend the argument to predictive distributions for binary data.

3.1.1 Prior predictive distribution

$$f(y_1, \ldots, y_k) = \int_0^1 f(y_1, \ldots, y_k, p)dp = \int_0^1 \underbrace{f(y_1, \ldots, y_k \mid p)}_{\text{Binomial likelihood}} \underbrace{f(p)}_{prior} dp$$

Let

$$\bar{y}^{(k)} = \frac{1}{k} \sum_{i=n+1}^{n+k} y_i$$

If

$$n = 0 \implies \bar{y}^{(k)} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

$$\begin{aligned}
&\int_0^1 \binom{k}{k\bar{y}^{(k)}} p^{k\bar{y}^{(k)}} (1-p)^{k-k\bar{y}^{(k)}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\
&= \binom{k}{k\bar{y}^{(k)}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k\bar{y}^{(k)})\Gamma(\beta+k-k\bar{y}^{(k)})}{\Gamma(\alpha+\beta+k)} \underbrace{\int_0^1 \frac{1}{\mathrm{B}(\alpha+k\bar{y}^{(k)})} p^{\alpha+k\bar{y}^{(k)}-1}(1-p)^{\beta+k-k\bar{y}^{(k)}-1} dp}_{=1} \\
&= \binom{k}{k\bar{y}^{(k)}} \frac{\mathrm{B}(\alpha+k\bar{y}^{(k)}, \beta+k-k\bar{y}^{(k)})}{\mathrm{B}(\alpha,\beta)}
\end{aligned} \tag{3.1}$$

Remark For $k = 1$, we get a Bernoulli distribution for one future observation y

$$\begin{aligned}
\binom{1}{y} \frac{\mathrm{B}(\alpha+y, \beta+1-y)}{\mathrm{B}(\alpha,\beta)} &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y)\Gamma(\beta+1-y)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{1}{\alpha+\beta} \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)} \frac{\Gamma(\beta+1-y)}{\Gamma(\beta)} \\
&= \begin{cases} \frac{\alpha}{\alpha+\beta}, & \text{if } y = 1 \\ \frac{\beta}{\alpha+\beta}, & \text{if } y = 0 \end{cases}.
\end{aligned} \tag{3.2}$$

Figure 3.1 demonstrates that the prior predictive distribution for binary data shows the probability of the number of events in a future sample based on k observations.

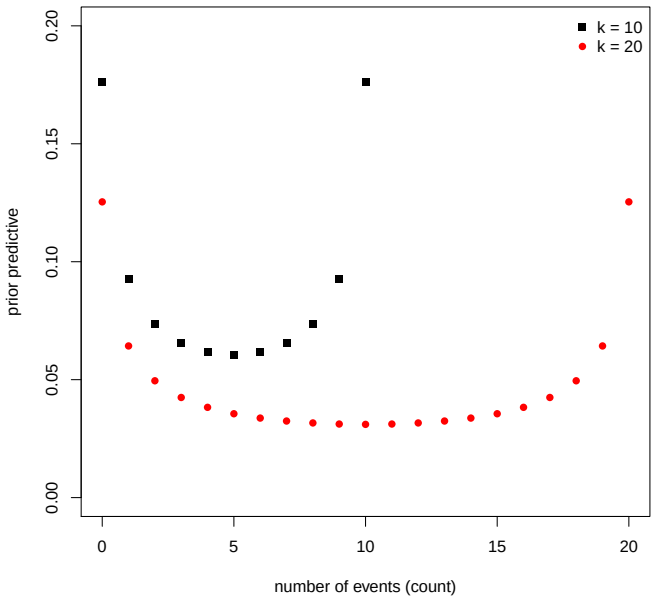

Figure 3.1: Prior predictive distribution for k future binary observations given a Beta(0.5, 0.5) prior.

3.1.2 Posterior predictive distribution

$$
\begin{aligned}
f(\underbrace{y_{n+1},\ldots,y_{n+k}}_{\text{future observations}} \mid \underbrace{y_1,\ldots,y_n}_{\text{known obserations}}) &= \int_0^1 f(y_{n+1},\ldots,y_{n+k},p \mid y_1,\ldots,y_n)dp \\
&\overset{\text{i.i.d. sample}}{=} \int_0^1 f(y_{n+1},\ldots,y_{n+k} \mid p)\,\underbrace{f(p \mid y_1,\ldots,y_n)}_{\text{posterior distribution}}\,dp \\
&= \text{compute again}\ldots \\
&= \text{or use the conjugacy and the formula derived for} \\
&\text{the prior predictive distribution}
\end{aligned}
\tag{3.3}
$$

Denote $\bar{y}^{(n)} = \frac{1}{n}\sum_{i=1}^n y_i$. Recall that

$$p \mid y_1,\ldots,y_n \sim \text{Beta}(\alpha + n\bar{y}^{(n)}, \beta + n + n\bar{y}^{(n)}).$$

Thus,

$$
f(\underbrace{y_{n+1},\ldots,y_{n+k}}_{\text{future observations}} \mid \underbrace{y_1,\ldots,y_n}_{\text{known obserations}}) = \binom{k}{k\bar{y}^{(k)}} \frac{\text{B}(\alpha + n\bar{y}^{(n)} + k\bar{y}^{(k)}, \beta + n - n\bar{y}^{(n)} + k - k\bar{y}^{(k)})}{\text{B}(\alpha + n\bar{y}^{(n)}, \beta + n - n\bar{y}^{(n)})}.
\tag{3.4}
$$

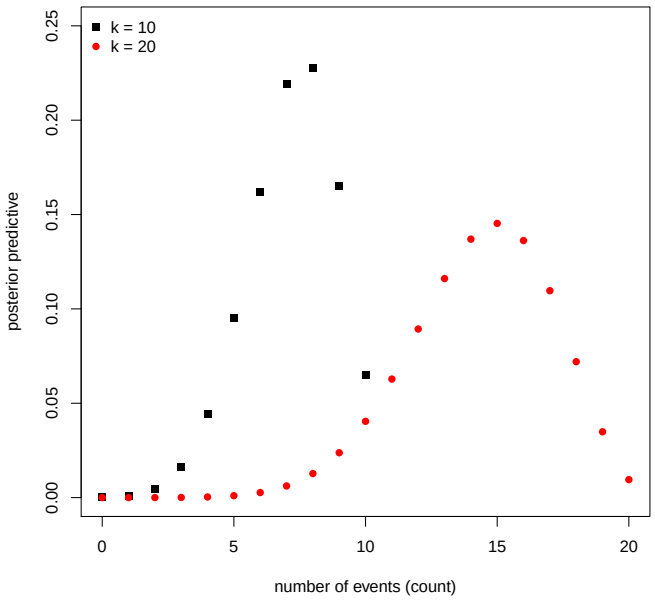

Figure 3.2: Posterior predictive distribution for k future observations given already observed data $n\bar{y}^{(n)} = 16$ in $n = 22$ observations and prior Beta(0.5, 0.5).

Remark With $k = 1$ we get a Bernoulli distribution for the posterior predictive distribution of one future observation y_{n+1} given $y_1, \ldots, y_n$.

$$\text{Be}\left(\tfrac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}\right) = \begin{cases} \frac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1} = 1, \\ \frac{\beta+n-n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1} = 0. \end{cases}$$

Example: Vision correction

Data: $n\bar{y}^{(n)} = 16$ participants out of $n = 22$ required vision correction. Note that $n - n\bar{y}^{(n)} = 6$ participants did not require any vision correction. Data $n\bar{y}^{(n)} = 16$ and $n = 22$ combined with the prior Beta(0.5, 0.5) result in a posterior Beta(16.5, 6.5), which reflects our current knowledge. Figure 3.2 shows the posterior predictive distribution, which is based on this current posterior knowledge, for the number of events in future samples based on $k = 10$ and $k = 20$ observations.

This posterior predictive distribution can be used to provide prediction intervals [Hartnack and Roos, 2021].

```
PI(x = 16, n = 22, a = 0.5, b = 0.5, k = 10, conf.level = 0.95)

## lower upper
##     4    10

PI(x = 16, n = 22, a = 0.5, b = 0.5, k = 20, conf.level = 0.95)

## lower upper
##     9    19
```

3.2 Predictive distributions for normal data

This section is a continuation of the argument presented in Section 2.3. Assume that $y_1, \ldots, y_n$ are i.i.d. observations generated by a sampling distribution $\mathrm{N}(m, \kappa^{-1})$ and the prior for m follows a $\mathrm{N}(\mu, \lambda^{-1})$ distribution, where κ, μ and λ are fixed constants.

3.2.1 Prior predictive distribution

We derive analytically the prior predictive distribution for one future observation y assuming that no observations have been collected yet. The prior predictive distribution is defined by

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta) f(\theta) d\theta, \tag{3.5}$$

where $f(\theta)$ is the prior and $f(y \mid \theta)$ is the likelihood. Thus,

$$\begin{aligned}
f(y) &= \int_{-\infty}^{\infty} f(y \mid m) f(m) dm \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\kappa^{-1}}} \exp\left(-\frac{\kappa(y-m)^2}{2}\right) \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(-\frac{\lambda(m-\mu)^2}{2}\right) dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa(y^2 + m^2 - 2my) + \lambda(m^2 + \mu^2 - 2m\mu)\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa y^2 + \lambda\mu^2 - \frac{(\kappa y + \lambda\mu)^2}{\kappa + \lambda} + (\kappa + \lambda)\left(m - \frac{\kappa y + \lambda\mu}{\kappa + \lambda}\right)^2\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \exp\left\{-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right\} \sqrt{\frac{2\pi}{\kappa+\lambda}} \underbrace{\int_{-\infty}^{\infty} \sqrt{\frac{\kappa+\lambda}{2\pi}} \exp\left\{-\frac{\kappa+\lambda}{2}\left(m - \frac{\kappa y + \lambda\mu}{\kappa+\lambda}\right)^2\right\} dm}_{=1}
\end{aligned} \tag{3.6}$$

According to the property of probability density functions, the integral at the right hand side above is equal to 1. The remaining part of the equation reads:

$$f(y) = \sqrt{\frac{\kappa\lambda}{2\pi(\kappa+\lambda)}} \exp\left\{-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right\} = \sqrt{\frac{1}{2\pi\left(\frac{1}{\lambda} + \frac{1}{\kappa}\right)}} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{(\frac{1}{\lambda} + \frac{1}{\kappa})}\right\}$$

.

Hence, the prior predictive distribution of one future observation y is $\mathrm{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right)$.

3.2.2 Posterior predictive distribution

Derive analytically the posterior predictive distribution for one future observation y_{n+1} given $y_1, \ldots, y_n$ have been observed. According to the definition of the posterior predictive distribution in [Held and Sabanés Bové, 2020] Section 9.3.1, we have

$$
\begin{aligned}
f(y_{n+1} \mid y_1, \ldots, y_n) &= \int_{-\infty}^{\infty} f(y_{n+1}, m \mid y_1, \ldots, y_n) dm \\
&= \int_{-\infty}^{\infty} f(y_{n+1} \mid m, y_1, \ldots, y_n) f(m \mid y_1, \ldots, y_n) dm \\
&\overbrace{=}^{\text{cond. ind.}} \int_{-\infty}^{\infty} \underbrace{f(y_{n+1} \mid m)}_{\text{likelihood}} \underbrace{f(m \mid y_1, \ldots, y_n)}_{\text{posterior density}} dm
\end{aligned}
\tag{3.7}
$$

We have

$$
f(y_{n+1} \mid m) = \sqrt{\frac{\kappa}{2\pi}} \exp\left(\frac{\kappa(y_{n+1} - m)^2}{2}\right)
$$

We have already derived the posterior distribution in Section 2.3

$$
m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n \bar{y} + \lambda\mu}{n\kappa + \lambda}, \frac{1}{n\kappa + \lambda}\right).
$$

Denote

$$
\mu_{\text{post}} = \frac{\kappa n \bar{y} + \lambda\mu}{n\kappa + \lambda}
$$

and

$$
\lambda_{\text{post}} = n\kappa + \lambda.
$$

Hence,

$$
f(y_{n+1} \mid y_1, \ldots, y_n) = \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa}{2}(y_{n+1} - m)^2\right) \sqrt{\frac{\lambda_{\text{post}}}{2\pi}} \exp\left(-\frac{\lambda_{\text{post}}}{2}(m - \mu_{\text{post}})^2\right) dm.
$$

Following the argument for the prior predictive distribution in Section 3.2.1, we obtain the posterior predictive distribution:

$$
y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1}\right).
$$

Summary of results for the posterior distribution, the prior predictive distribution, and the posterior predictive distribution for normal observations. Frequently, these distributions are parametrized by variances or standard deviations instead of precisions.

Prerequisites:

Data: $y_1, y_2, \ldots, y_n \sim \mathrm{N}(m, \sigma^2) = \mathrm{N}(m, \kappa^{-1})$ with $\kappa = 1/\sigma^2$.

Prior: $m \sim \mathrm{N}(\mu, \tau^2) = \mathrm{N}(\mu, \lambda^{-1})$ with $\lambda = 1/\tau^2$.

Posterior distribution:

Posterior distribution parametrized by precisions:

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}, (n\kappa + \lambda)^{-1}\right). \tag{3.8}$$

Posterior distribution parametrized by variances:

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right) = \mathrm{N}\left(\frac{\tau^2 n\bar{y} + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right). \tag{3.9}$$

Note that posterior variance is smaller than the prior variance, because the precision gets larger. The impact of the prior decreases with increasing sample size.

Prior predictive distribution:

Prior predictive distribution parametrized by precisions:

$$y \sim \mathrm{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right) \tag{3.10}$$

Prior predictive distribution parametrized by variances:

$$y \sim \mathrm{N}\left(\mu, \sigma^2 + \tau^2\right) \tag{3.11}$$

The variance of the prior predictive distribution is larger than the variance of the prior and the data alone.

Posterior predictive distribution:

Posterior predictive distribution parametrized by precisions:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}, (n\kappa + \lambda)^{-1} + \kappa^{-1}\right) \tag{3.12}$$

Posterior predictive distribution parametrized by variances:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\tau^2 n\bar{y} + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2} + \sigma^2\right) \tag{3.13}$$

The variance of the posterior predictive distribution is larger than the variance of the data alone. The impact of the prior decreases with increasing sample size.

Remark: The Bayes theorem from Equation (1.2) can be rewritten in terms of densities

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)}, \tag{3.14}$$

where

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta) f(\theta) d\theta. \tag{3.15}$$

The prior predictive distribution $f(y)$ evaluated at observed data is called marginal likelihood. Note that Equation (3.14) can be rewritten to get

$$f(y) = \frac{f(y \mid \theta) f(\theta)}{f(\theta \mid y)}. \tag{3.16}$$

Thus, Equation (3.16) provides an alternative way to compute $f(y)$ defined in Equation (3.15) within the conjugate Bayes framework. Note that for this computation all constants play an important role.

Remark: An alternative proof of the prior predictive distribution for $Y \mid m \sim \mathrm{N}(m, \sigma^2)$ with $m \sim \mathrm{N}(\mu, \tau^2)$ is by rules of iterated expectation and total variance ([Spiegelhalter et al., 2002, p. 17, 84], [Held and Sabanés Bové, 2020, Section A.3.4]).

$$\mathbb{E}(Y) = \mathbb{E}_m[\mathbb{E}_Y(Y \mid m)] = \mathbb{E}_m(m) = \mu$$

$$\mathrm{Var}(Y) = \mathrm{Var}_m[\mathbb{E}_Y(Y \mid m)] + \mathbb{E}_m[\mathrm{Var}_Y(Y \mid m)] = \mathrm{Var}_m[m] + \mathbb{E}_m[\sigma^2] = \tau^2 + \sigma^2 = \lambda^{-1} + \kappa^{-1}.$$

Therefore,

$$Y \sim \mathrm{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right)$$

For predictions, we add variances and the uncertainty increases.

3.3 Bayesian asymptotics

Consider a model with n independent identically distributed (*i.i.d.*) random variables $Y_i, i = 1, \ldots, n$ and the resulting likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$, where $\theta \in \Theta$ and Θ is an open set in $\mathbb{R}^p$. Denote the maximum likelihood estimator (MLE) by $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$, the ordinary unit Fisher information of one observation Y_i by

$$I^*(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(Y_i \mid \theta)$$

and the expected unit Fisher information of one observation by

$$J^*(\theta) = -\mathbb{E}_\theta\Big(\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(Y_i \mid \theta)\Big).$$

We denote the counterparts of the full random sample by $J_{1:n}(\theta) = nJ^*(\theta)$ and $I_{1:n}(\theta) = nI^*(\theta)$ [Held and Sabanés Bové, 2020, Section 4.2.3, p.97] and use their informal notation.

Under regularity conditions (see Held and Sabanés Bové [2020] Definition 4.1 on p. 80), if θ_0 is a fixed true parameter, then

$$\hat{\theta}_n \overset{\text{approx}}{\approx} \text{N}(\theta_0, J_{1:n}(\theta_0)^{-1}) = \text{N}(\theta_0, \frac{1}{n} J^*(\theta_0)^{-1}).$$

The proof of Bayes asymptotic combines a Taylor expansion and the computation of the posterior in the normal model discussed in Section 2.3. This asymptotic says that, under regularity conditions, for any smooth prior which is strictly positive in a neighborhood of θ_0

1.
$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \text{N}(\hat{\theta}_n, I_{1:n}(\hat{\theta}_n)^{-1}) = \text{N}(\hat{\theta}_n, \frac{1}{n} I^*(\hat{\theta}_n)^{-1}), \tag{3.17}$$

where $I_{1:n}(\hat{\theta}_n)$ is the observed Fisher information of the whole sample and $I^*(\hat{\theta}_n)$ is the observed unit Fisher information of one observation. This result holds irrespective of the distributional form of the prior. Therefore, the influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{1/n}$ neighborhood of the MLE. See Held and Sabanés Bové [2020] Section 6.6.2 on page 206 for further details.

Note that there is a difference in what is considered fixed and what is random in classical and Bayesian asymptotics. For example, in Equation (3.17), $\theta \mid y_1, \ldots, y_n$ is a sequence of posterior distributions, which depend on n, and θ is a random variable. On the right hand side of Equation (3.17), the MLE $\hat{\theta}_n$ is a function of observations $y_i, i = 1, \ldots, n$ that are fixed.

In addition to Equation (3.17), there are three other approximations to this asymptotic result.

2. The observed Fisher information can be replaced by the expected Fisher information.

$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \text{N}(\hat{\theta}_n, J_{1:n}(\hat{\theta}_n)^{-1}), \tag{3.18}$$

3. With posterior mode $\text{Mod}(\theta \mid y_1, \ldots, y_n)$ and the negative curvature $C_{1:n}^{-1}$:

$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \text{N}(\text{Mod}(\theta \mid y_1, \ldots, y_n), C_{1:n}^{-1}), \tag{3.19}$$

4. With posterior expectation $\text{E}(\theta \mid y_1, \ldots, y_n)$ and posterior covariance $\text{Cov}(\theta \mid y_1, \ldots, y_n)$:

$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \text{N}(\text{E}(\theta \mid y_1, \ldots, y_n), \text{Cov}(\theta \mid y_1, \ldots, y_n)), \tag{3.20}$$

Formulas provided in Example 6.29 of Held and Sabanés Bové [2020] on page 208 can be applied to vision correction data. Figure 3.3 shows that asymptotic approximations 1–4 in Equations (3.17)–(3.20) work well for the Beta(16.5, 6.5) posterior obtained for 16 participants out of $n = 22$ who need vision correction combined with the Beta(0.5, 0.5) prior.

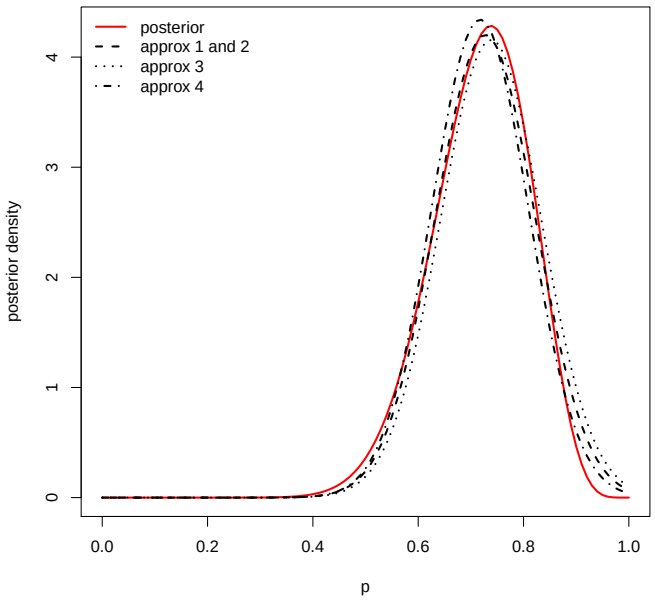

Figure 3.3: Asymptotic approximations of the posterior Beta(16.5, 6.5) obtained for vision correction data $n\bar{y} = 16$ and $n = 22$ combined with the Beta(0.5, 0.5) prior.

3.4 Monte Carlo simulations

The basis of the Monte Carlo (MC) method is the ability to generate independent identically distributed (*i.i.d.*) realizations from a uniform distribution in the unit interval $[0, 1]$. Inverse transform sampling takes this sample and uses it to generate *i.i.d.* realizations of another random variable, provided that we have an inverse of the cumulative distribution function of this random variable. This approach works, because $P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$.

Example Exponential distribution: $F(x) = 1 - \exp(-\lambda x)$ for $x \geq 0$. We can solve $F(x) = 1 - \exp(-\lambda x) = u$ for x and obtain $x = F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. Thus, if we have an *i.i.d.* sample from a uniform distribution (left panel of Figure 3.4), we can provide a sample from the exponential distribution (right panel of Figure 3.4).

The individual project in Worksheet 3 shows an interesting and in practice highly relevant application of the MC simulation. MC simulation is applied to posteriors in two different groups. Comparison of both samples directly demonstrates differences between both groups. The comparison can be expressed by differences, risk ratios, odds ratios or other measures. Based on these measures the posterior probability of superiority (PPS) can be computed. Remember to compute Monte Carlo standard errors when you use MC techniques (see Held and Sabanés Bové [2020] Section 8.3.1 and 8.3.2).

Monte Carlo sampling can also be used for computation of the area content of an object (for example of a unit circle or any other shape). This demonstrates that MC sampling can be used for integration. Methods such as importance sampling or rejection sampling can be used to explore a distribution. See Section 8.3 of Held and Sabanés Bové [2020] for more

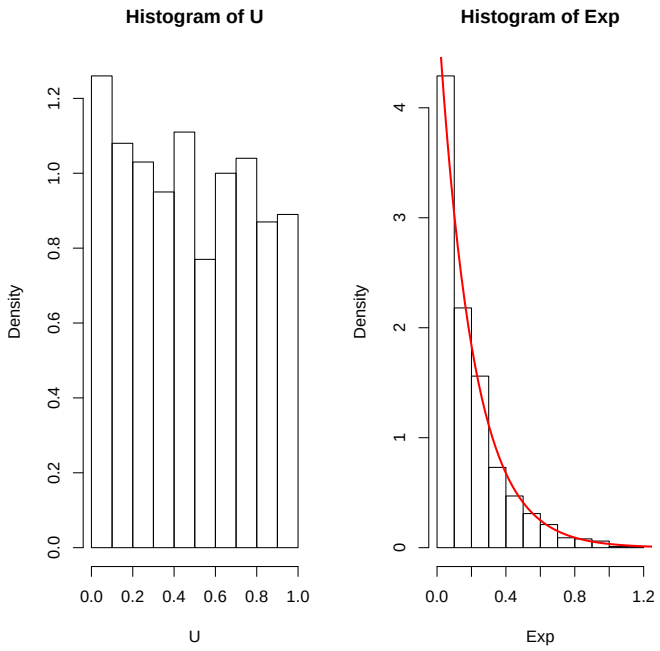

Figure 3.4: Monte Carlo simulation from a uniform distribution on the unit interval $[0, 1]$ (left panel) transformed to a random sample of the Exponential distribution with parameter $\lambda = 5$ by inverse transform sampling (right panel).

details on Monte Carlo methods.

3.5 Worksheet 3

Probability calculus	**Distributions**	**Change of variables formula**
Priors	**MC sampling**	**Asymptotics**

Bayes	**Classical**
Posterior $\propto$ Likelihood $\times$ Prior	Likelihood

Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 3.1: Foundations of Bayesian Methodology: content of the third lecture and material relevant for Worksheet 3.

Secukinumab		**Placebo**
Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		Data
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 3.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 4

Lecture 4: Markov chain Monte Carlo method

In Chapter 3, we showed that independent Monte Carlo simulations have manifold applications. In particular, they can approximate the area of a complicated object by dropping randomly points on it and computing

$$\widehat{Area} = \frac{\text{Number of points in the object}}{\text{Number of all points}}.$$

This lecture demonstrates that also dependent samples provided by Markov chain Monte Carlo (MCMC) can be useful in this respect. In particular, it overviews the theory behind the Gibbs sampler and the Metropolis-Hastings sampler, two famous techniques for Markov chain Monte Carlo (MCMC) sampling. The `R` code and applications are postponed to Worksheet 4.

Suggested reading: [Held and Sabanés Bové, 2014] Section 8.4, [Ntzoufras, 2009] Chapter 2, [Robert and Casella, 2010] Chapter 7.

4.1 MCMC Sampling in R

Stages of Bayesian analysis (an iterative process):

- Stage 1: model building ($\overbrace{\text{likelihood, parameters}}^{\text{classical}}, \overbrace{\text{priors}}^{\text{Bayesian}}$)
- Stage 2: calculation of the posterior distribution (or target distribution)
  - Analytical computation (Conjugate Bayes)
  - Bayesian numerical approximation (INLA, bayesmeta)
  - MCMC sampling: we get a sample either from own samplers or from JAGS, Stan, OpenBUGS, and WinBUGS
- Stage 3: **CODA convergence diagnostics are required for MCMC sampling.** If CODA indicates any problems with MCMC samples, go to Stage 1.

- Stage 4: Analysis of the posterior distribution. Compute descriptive statistics (mean, sd, quantiles, CrI, tail probabilities) of marginal posterior distributions.
- Stage 5: Description of the results for the client.

4.2 Markov chain

Construction of a Markov chain (an iterative procedure, where the value generated in one step depends on the value in the previous step) that eventually "converges" to the target distribution (stationary or equilibrium) $f(\theta \mid \mathbf{y})$. Although MCMC samples are no more independent, we can still approximate areas by MCMC samples.

4.3 The algorithm for MCMC sampling

Markov chain is a stochastic process $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$ such that $f(\theta^{(t+1)} \mid \theta^{(t)}, \ldots, \theta^{(1)}) = f(\theta^{(t+1)} \mid \theta^{(t)})$.

Markov chain must be

- irreducible
- aperiodic
- positive-recurrent

These properties will be discussed together with JAGS and convergence diagnostics (CODA) in Chapter **??** .

If $t \to \infty$ then $\theta^{(t)}$ converges to an equilibrium, which is also called the stationary distribution. This stationary distribution is independent of the initial value $\theta^{(0)}$. Stationarity is needed to guarantee that the sample is identically distributed.

The desirable properties of the algorithm for computation are:

- $f(\theta^{(t+1)} \mid \theta^{(t)})$ is easy to generate
- equilibrium of the Markov chain = $f(\theta \mid \mathbf{y})$ the target posterior distribution. Thus, $\theta^{(t)} \overset{\text{identically distributed}}{\sim} f(\theta \mid \mathbf{y})$, but $\theta^{(t)}$ are no more independent.

4.4 General steps of MCMC algorithm

1. Select an initial value $\theta^{(0)}$.
2. Generate T values until the equilibrium is reached.
3. Monitor convergence diagnostics (if CODA fails, generate more observations, that is, increase T).
4. Cutoff the first B observations (in BUGS it is called burn-in and in Stan warming up).

5. Consider $\theta^{(B+1)}, \theta^{(B+2)}, \ldots, \theta^{(T)}$ as the sample for the posterior analysis (possibly after some tuning).

6. Plot the posterior distribution (usually focus on univariate marginal distributions).

7. Obtain summaries of the posterior distribution (classical: sample mean, median, quantiles, MC-error), effective sample size (ESS) of a MCMC simulation.

Note that an iteration is a cycle of the algorithm that generates a full set of parameter values from the posterior distribution.

4.5 Gibbs sampler

The Gibbs sampler proposed by Geman and Geman [1984] is a special case of the single-component Metropolis-Hastings algorithm, which uses as proposal density $q(\theta' \mid \theta^{(t)})$ the full conditional posterior distribution $f(\theta_j \mid \boldsymbol{\theta}_{-j})$, where $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^T$. In each step of the Gibbs sampler, a candidate value θ_j' of the jth component of the vector of p parameters is proposed by $f(\theta_j' \mid \boldsymbol{\theta}_{-j}, \mathbf{y})$.

Algorithm

1. Set initial values $\theta^{(0)}$.

2. For $t = 1, \ldots, T$ repeat the following steps

   generate $\theta^{(t)}$ (new parameter values) given $\theta^{(t-1)}$ by

   $\theta_1^{(t)}$ from $f(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\theta_2^{(t)}$ from $f(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\vdots$

   $\theta_j^{(t)}$ from $f(\theta_j \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\vdots$

   $\theta_p^{(t)}$ from $f(\theta_p \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{p-1}^{(t)}, \mathbf{y})$

   which are univariate distributions and all other variables except θ_j are held constant at their given values.

3. Save the $\theta^{(t)}$ for the next iterations.

Advantages

- In each step, random values can be generated from univariate distributions for which a wide variety of computational tools exist.

- Frequently, the full conditional posterior distribution has a known form and random values can be easily simulated using standard functions in `R`.

- Gibbs sampler is always moving to a new value, because it accepts the new generated value with acceptance probability 1.
- Gibbs sampler does not require any tuning of proposal distribution.

Drawback Gibbs sampler can be ineffective when the parameter space is complicated or the parameters are highly correlated. It moves slowly.

4.6 Application of the Gibbs sampler

See Exercise 2 of Worksheet 4 (normal example) and the `R` code in the file `04GibbsSampler.R`.

Setting

1. Model: sampling distribution $y_1, \ldots, y_n \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$
2. Data: `set.seed(44566)`, $n = 30$ from $\mathrm{N}(\mu = 4, \sigma^2 = 16)$
3. Priors: independent mean and precision
   $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $\mu_0 = -3$, $\sigma_0^2 = 4$
   $\frac{1}{\sigma^2} \sim \mathrm{G}(a_0, b_0)$, $a_0 = 1.6$, $b_0 = 0.4$
   Mean $= \frac{a_0}{b_0} = \frac{1.6}{0.4} = 4$ and Var $= \frac{a_0}{b_0^2} = \frac{1.6}{0.4^2} = 10$.

Likelihood

$$\begin{aligned} f(y_1, \ldots, y_n \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned} \tag{4.1}$$

Posterior
$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n \mid \mu, \sigma^2) f(\mu, \sigma^2)$
Prior
$f(\mu, \sigma^2) = f(\mu) f(\sigma^2)$ informative and independent
$-\infty < \mu_0 < \infty$, $\sigma_0^2, a_0, b_0 > 0$
Prerequisite $X = \frac{1}{\sigma^2} \sim \mathrm{G}(a_0, b_0)$ [Held and Sabanés Bové, 2014, change of variables formula on page 336].
$f(x) = \frac{b_0^{a_0}}{\Gamma(a_0)} x^{a_0 - 1} \exp(-b_0 x)$
$y = \frac{1}{x} = g(x)$, $x = \frac{1}{y} = g^{-1}(y)$, $\frac{dg^{-1}(y)}{dy} = -\frac{1}{y^2}$

$$\begin{aligned} f(y) &= \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\frac{1}{y}\right)^{a_0 - 1} \exp\left\{-\frac{b_0}{y}\right\} \left|-\frac{1}{y^2}\right| \\ &= \frac{b_0^{a_0}}{\Gamma(a_0)} y^{-(a_0+1)} \exp\left\{-\frac{b_0}{y}\right\} \end{aligned} \tag{4.2}$$

Derivation of full conditional posterior distributions needed for Gibbs sampler.

Posterior

$$\begin{aligned} f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ \times (\sigma_0^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\ \times (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right) \end{aligned} \tag{4.3}$$

The full conditional posterior distribution for μ is obtained by treating the posterior in Equation (4.3) as a function of μ

$$\begin{aligned} f(\mu \mid \sigma^2, y_1, \ldots, y_n) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \times \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\ &= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n} y_i\right\} + \frac{1}{\sigma_0^2}\left\{\mu^2 + \mu_0^2 - 2\mu_0\mu\right\}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\mu^2\left\{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right\} - 2\mu\left\{\frac{\sum_{i=1}^{n} y_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right\}\right]\right) \end{aligned} \tag{4.4}$$

Exercise $p(\theta) \propto \exp\left(-\frac{1}{2}\left(a\theta^2 - 2b\theta\right)\right)$, then $\theta \sim \mathrm{N}\left(\frac{b}{a}, \frac{1}{a}\right)$.

Therefore,

$$\mu^{(t)} \mid (\sigma^2)^{(t-1)}, y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{n} y_i}{(\sigma^2)^{(t-1)}} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}\right)$$

The full conditional distribution for σ^2 is obtained by treating the posterior in Equation (4.3) as a function of σ^2.

$$\begin{aligned} f(\sigma^2 \mid \mu, y_1, \ldots, y_n) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right) \\ &= (\sigma^2)^{-\frac{n}{2}-(a_0+1)} \exp\left(-\frac{1}{\sigma^2}\left[b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]\right), \end{aligned} \tag{4.5}$$

which is the kernel of an inverse Gamma distribution.

$$(\sigma^2)^{(t)} \mid \mu^{(t)}, y_1, \ldots, y_n \sim \mathrm{InvG}\left(\frac{n}{2} + a_0, b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu^{(t)})^2\right).$$

Gibbs algorithm $\boldsymbol{\theta} = (\mu, \sigma^2)$, $p = 2$

- Set initial values
- For t=1, ..., T

  μ - step: calculate mean $= \frac{\frac{\sum_{i=1}^{n} y_i}{(\sigma^2)^{(t-1)}} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}$, var $= \frac{1}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}$

  generate one random value μ from N(mean, var), set $\mu^{(t)} = \mu$

  σ^2 - step: calculate shape $= \frac{n}{2} + a_0$, scale $= b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu^{(t)})^2$,

  generate σ^2 from InvG(shape, scale), set $(\sigma^2)^{(t)} = \sigma^2$.

4.7 Metropolis-Hastings algorithm

The goal of the Metropolis-Hastings algorithm [Metropolis (1953), Hastings (1970)] is to explore target function $f(\theta \mid \mathbf{y})$. See Metropolis [1987] for a historical overview and Hill and Spall [2019] for an easily accessible introduction.

Algorithm

- Set initial values $\theta^{(0)}$.
- For $t = 1, \ldots, T$ repeat the following steps.
  1. Set $\theta = \theta^{(t-1)}$
  2. Generate new candidate parameter values θ' from a proposal distribution $q(\theta^{(t-1)} \mid \theta)$
  3. Calculate the acceptance probability

     $$A = \min\left(1, \frac{f(\theta' \mid \mathbf{y})q(\theta \mid \theta')}{f(\theta \mid \mathbf{y})q(\theta' \mid \theta)}\right) = \frac{f(\mathbf{y} \mid \theta')f(\theta')q(\theta \mid \theta')}{f(\mathbf{y} \mid \theta)f(\theta)q(\theta' \mid \theta)}$$

     The chain moves towards direction where $\frac{f(\theta'|\mathbf{y})}{q(\theta'|\theta)}$ is higher than at the present position (has a higher probability).
  4. Decide randomly whether to accept the proposed value and update $\theta^{(t)} = \theta'$ with probability A otherwise set $\theta^{(t)} = \theta$

     In other words throw a coin, $U \sim \text{Unif}[0, 1]$

     accept if $U < A$

     reject if $U \geq A$

The MH algorithm converges to its equilibrium distribution regardless of whatever proposal distribution q is selected. In practice, the choice of proposal is important. MH can be inefficient if independent proposals are used. To increase the efficiency, use bivariate (multivariate) proposals. Tune also the spread of the proposal to get an acceptance rate of ca. 25%.

Special cases of MH

- random-walk Metropolis
- single component MH
- Gibbs sampler

4.8 Application of the Metropolis-Hastings sampler

See Exercise 3 of Worksheet 4 (mice example with logistic regression) and `R` code in the file `04MHSampler.R`.

1. Data
   - Suppose we have N binomial observations from $\frac{y_i}{n_i}$, $i = 1, \ldots, N$
   - Expectation $\mathbb{E}(y_i) = n_i p_i$, where p_i is the corresponding response probability ($\hat{p}_i$ estimated relative frequency of deaths)
2. Logistic model

   Transformation to make linear: $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$

   $p_i = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$
3. Likelihood

$$\begin{aligned} f((\mathbf{y}_i, \mathbf{n}_i, \mathbf{x}_i) \mid \alpha, \beta) &= \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \\ &= \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{n_i - y_i} \end{aligned} \tag{4.6}$$

4. Priors independent

   $f(\alpha) = \text{N}(0, \sigma^2)$, $f(\beta) = \text{N}(0, \sigma^2)$, $\sigma^2 = 10^4$.
5. Posterior distribution

$$f(\alpha, \beta \mid (\mathbf{y}_i, \mathbf{n}_i, \mathbf{x}_i)) \propto \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{n_i - y_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\beta^2}{2\sigma^2}}$$

6. Random walk univariate proposal $\alpha' \sim \text{N}(\alpha, \sigma_\alpha^2)$, $q(\alpha \mid \alpha') = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha - \alpha')^2\right)$,
   $q(\alpha' \mid \alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha - \alpha')^2\right)$.

Log-acceptance

$$\begin{aligned}\ln(A^{\alpha}) &= \ln\left(\frac{f(\alpha',\beta \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\alpha \mid \alpha')}{f(\alpha,\beta \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\alpha' \mid \alpha)}\right) \\ &= \ln(f(\alpha',\beta \mid \mathbf{y},\mathbf{n},\mathbf{x})) - \ln(f(\alpha,\beta \mid \mathbf{y},\mathbf{n},\mathbf{x}))\end{aligned} \tag{4.7}$$

If $\ln(\text{runif}(1)) \leq \ln A^{\alpha}$ then $\alpha \leftarrow \alpha'$, accept α' with probability $\ln(A^{\alpha})$.

7. Random walk univariate proposal. $\beta' \sim \text{N}(\beta, \sigma^2_{\beta})$

$$\ln(A^{\beta}) = \ln\left(\frac{f(\alpha,\beta' \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\beta \mid \beta')}{f(\alpha,\beta \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\beta' \mid \beta)}\right) \tag{4.8}$$

If $\log(\text{runif}(1)) \leq \log A^{\beta}$ then $\beta \leftarrow \beta'$.

Remarks:

- The user is responsible for tuning of σ^2_{α} and σ^2_{β}.
- Random walk bivariate proposal. $\binom{\alpha'}{\beta'} \leftarrow \texttt{rmvnorm}\left(\binom{\alpha}{\beta}, \sigma\right)$

$$\ln(A)^{\alpha,\beta} = \ln\left(\frac{f(\alpha',\beta' \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\alpha,\beta \mid \alpha',\beta')}{f(\alpha,\beta \mid \mathbf{y},\mathbf{n},\mathbf{x})q(\alpha',\beta' \mid \alpha,\beta)}\right) \tag{4.9}$$

If $\log(\text{runif}(A)) \leq \log A^{\alpha,\beta}$ then $\binom{\alpha}{\beta} \leftarrow \binom{\alpha'}{\beta'}$. Motivation adjust the shape of the probing proposal density to the shape of the posterior.

In case of a Gibbs sampler, we get $A = 1$. To see this, consider a single component Metropolis algorithm with

$$A = \min\left(1, \frac{f(\theta'_j \mid \boldsymbol{\theta}_{-j},\mathbf{y})q(\theta_j \mid \theta'_j,\boldsymbol{\theta}_j)}{f(\theta_j \mid \boldsymbol{\theta}_{-j},\mathbf{y})q(\theta'_j \mid \theta_j,\boldsymbol{\theta}_{-j})}\right). \tag{4.10}$$

Take $q(\theta_j \mid \theta'_j,\boldsymbol{\theta}_{-j}) = f(\theta_j \mid \boldsymbol{\theta}_{-j},\mathbf{y})$ the full conditional posterior distribution, and $q(\theta'_j \mid \theta_j,\boldsymbol{\theta}_{-j}) = f(\theta'_j \mid \boldsymbol{\theta}_{-j},\mathbf{y})$. Then,

$$\frac{f(\theta'_j \mid \theta_{-\mathbf{j}},\mathbf{y})f(\theta_j \mid \theta_{-\mathbf{j}},\mathbf{y})}{f(\theta_j \mid \theta_{-\mathbf{j}},\mathbf{y})f(\theta'_j \mid \theta_{-\mathbf{j}},\mathbf{y})} = 1, \tag{4.11}$$

so that the Gibbs proposal will be always accepted.

4.9 Expert system

Figure 4.1 demonstrates an expert system that is used by OpenBUGS for an automatic choice of MCMC samplers.

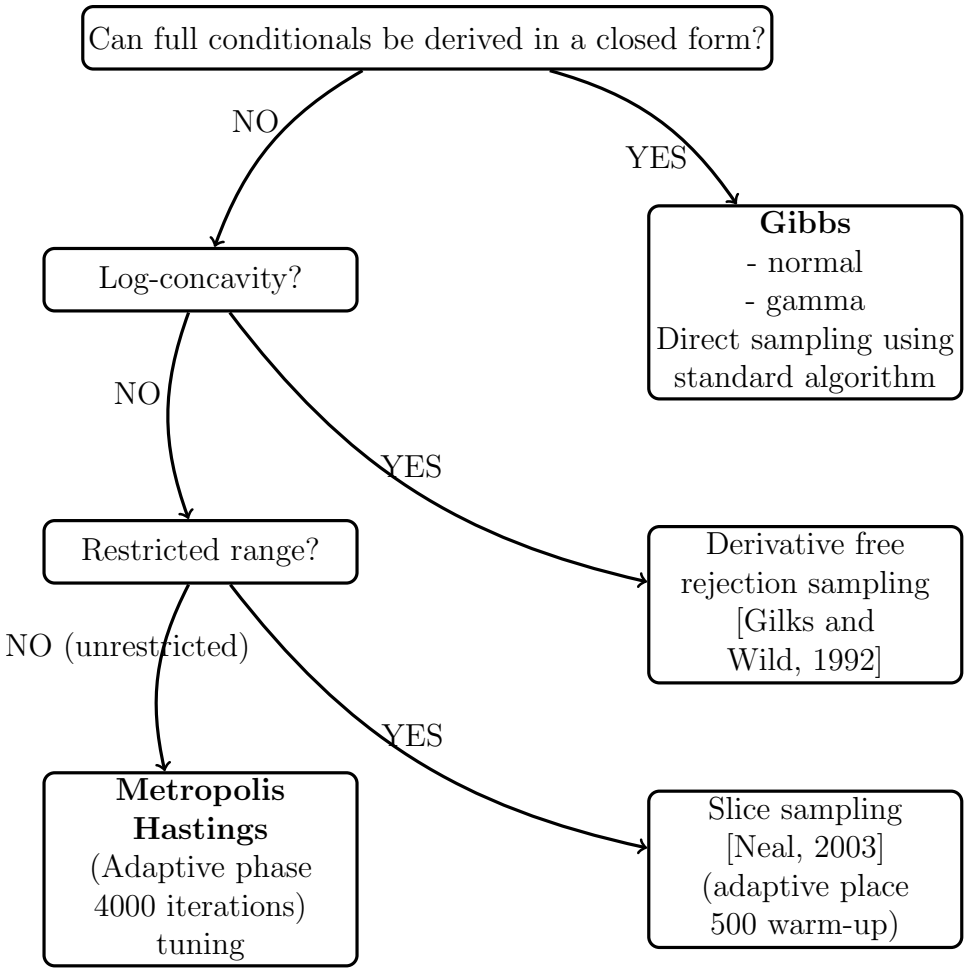

Figure 4.1: Expert system for the choice of MCMC samplers in OpenBUGS and WinBUGS

4.10 The meaning of priors for the slopes of centered and scaled (standardized) covariates

To stabilize numerical estimation and to express the estimated regression coefficients in common units, covariates in regression models are frequently centered and scaled, i.e. standardized. Centering of a covariate x means that we subtract its mean $\bar{x}$ and use $x - \bar{x}$ for analysis. Scaling of a covariate x means that we divide it by its standard deviation $\hat{\sigma}$ and use $x/\hat{\sigma}$ for analysis. Standardization means that we both center and scale the covariate to get $(x - \bar{x})/\hat{\sigma}$. The mean of both a centered and a standardized covariate is equal to 0. Moreover, the standard deviation of both a scaled and a standardized covariate is equal to 1. Note that if one prefers to scale by $2\hat{\sigma}$, then the standard deviation of the standardized covariate is equal to 0.5 [Gelman, 2008]. Below, we discuss the impact of standardization through centering and scaling on the values of parameter estimates. Moreover, we investigate the meaning of priors in these settings.

Classical paremeter estimates of the linear regression equation expressed in terms of

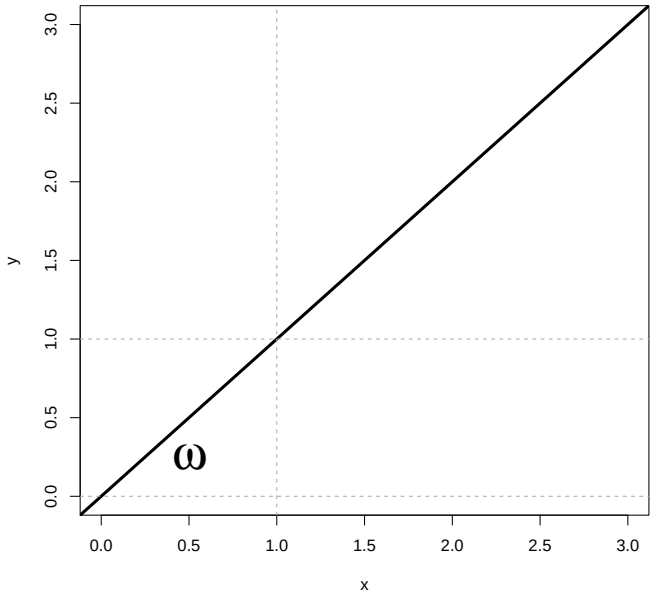

Figure 4.2: Regression line for a non-scaled covariate with the angle $\omega = \pi/4$ (45 degrees) and the slope $\beta = \tan(\pi/4) = 1/1 = 1$.

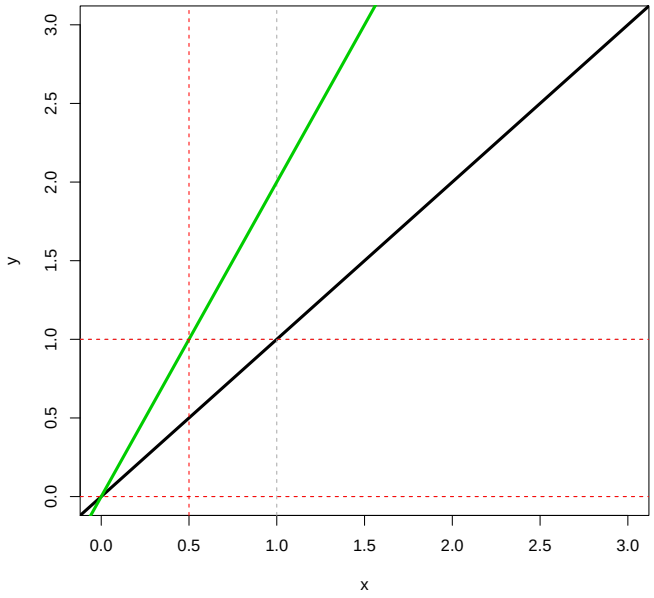

Figure 4.3: Regression lines for non-scaled (black) and scaled with $\hat{\sigma} = 2 > 1$ (green) covariates.

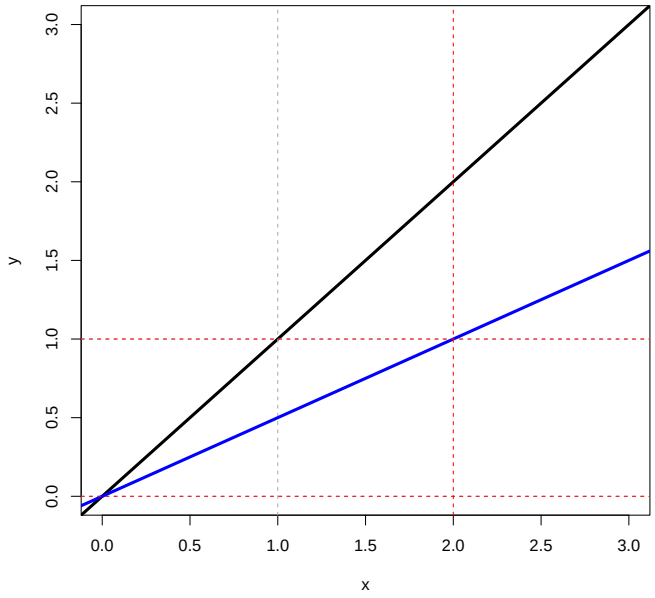

Figure 4.4: Regression lines for non-scaled (black) and scaled with $\hat{\sigma} = 0.5 < 1$ (blue) covariates.

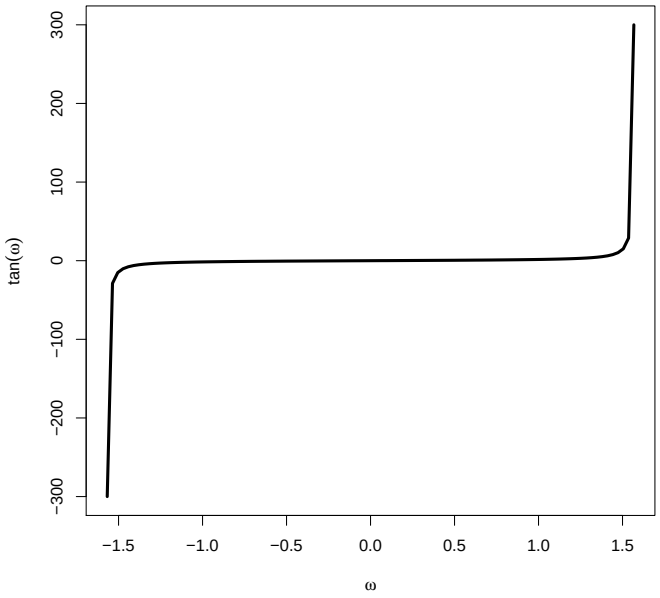

Figure 4.5: Tangens of angles ω ranging between -1.57 and 1.57 radians (89.8 and -89.8 degrees).

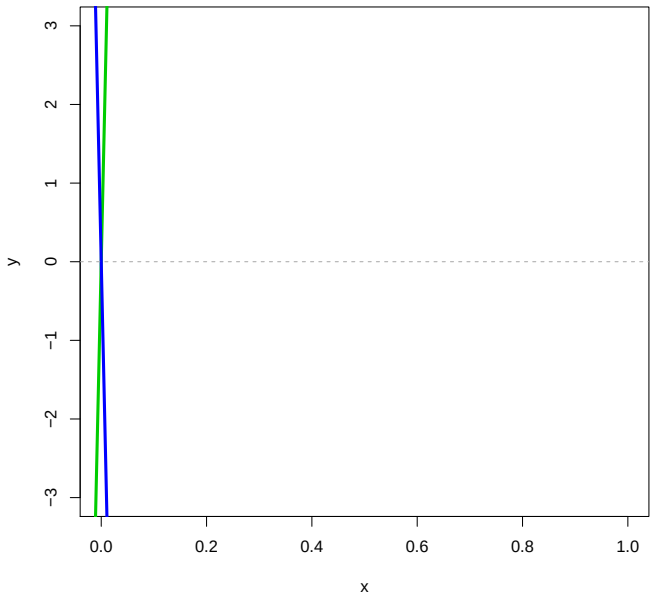

Figure 4.6: Slopes 300 and -300 for angles $\omega = 89.8$ and $\omega = -89.8$ degrees.

original covariates

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$$

are equal to

$$\hat{\alpha} = \bar{y}_\cdot - \hat{\beta}_1 \bar{x}_{1\cdot} - \hat{\beta}_2 \bar{x}_{2\cdot}, \quad \hat{\beta}_1, \quad \text{and} \quad \hat{\beta}_2.$$

In contrast, classical parameters estimates of the linear regression equation expressed in terms of centered covariates

$$y_i = \gamma + \beta_1 (x_{1i} - \bar{x}_{1\cdot}) + \beta_2 (x_{2i} - \bar{x}_{2\cdot})$$

are equal to

$$\hat{\gamma} = \bar{y}_\cdot, \quad \hat{\beta}_1, \quad \text{and} \quad \hat{\beta}_2.$$

This demonstrates that centering of covariates does not affect the values of slope estimates.

As we show below, the scaling of covariates x_1 and x_2 by their standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$ impacts the values of slope estimates. Indeed, classical paremeter estimates of the linear regression equation expressed in terms of non-centered but scaled covariates

$$y_i = \alpha + \delta_1 x_{1i}/\hat{\sigma}_1 + \delta_2 x_{2i}/\hat{\sigma}_2$$

are equal to

$$\hat{\alpha} = \bar{y}_\cdot - \hat{\delta}_1 \bar{x}_{1\cdot}/\hat{\sigma}_1 - \hat{\delta}_2 \bar{x}_{2\cdot}/\hat{\sigma}_2, \quad \hat{\delta}_1 = \hat{\beta}_1 \hat{\sigma}_1, \quad \text{and} \quad \hat{\delta}_2 = \hat{\beta}_2 \hat{\sigma}_2.$$

Moreover, classical parameters estimates of the linear regression equation expressed in terms of centered and scaled (standardized) covariates

$$y_i = \gamma + \delta_1(x_{1i}/\hat{\sigma}_1 - \bar{x}_{1\cdot}/\hat{\sigma}_1) + \delta_2(x_{2i}/\hat{\sigma}_2 - \bar{x}_{2\cdot}/\hat{\sigma}_2)$$

are equal to

$$\hat{\gamma} = \bar{y}_\cdot, \quad \hat{\delta}_1 = \hat{\beta}_1\hat{\sigma}_1, \quad \text{and} \quad \hat{\delta}_2 = \hat{\beta}_2\hat{\sigma}_2.$$

Whereas centering of covariates does not impact the values of slope estimates, scaling greatly affects their values. The estimates of slopes $\hat{\delta}_1$ and $\hat{\delta}_2$ for scaled covariates are either smaller ($\hat{\sigma} < 1$) or larger ($\hat{\sigma} > 1$) than the slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained for non-scaled covariates (Figures 4.2–4.4).

To see the impact of scaling on the slope values, we consider the same problem from a different point of view. The slope β in a simple linear regression measures the tangens of the angle ω the line makes with the positive x-axis

$$\beta = \tan(\omega) = \frac{\Delta y}{\Delta x}.$$

If the scale of the covariate x changes, then the slope will be affected by this scaling

$$\delta = \frac{\Delta y}{\Delta x/\hat{\sigma}} = \beta\hat{\sigma}.$$

Again, the estimate of the slope δ for the scaled covariateis either smaller ($\hat{\sigma} < 1$) or larger ($\hat{\sigma} > 1$) than the original coefficient β based on the non-scaled covariate (Figures 4.2–4.4).

Let us focus on the meaning of the prior for the slope in a Bayesian regression model. A prior $\mathrm{N}(0, 100^2)$ assumes that with probability 0.997 the slope can range between $-3 \times 100 = -300$ and $3 \times 100 = 300$. The prior for the slope is actually a prior that we put on the tangens of the angle ω. Slopes 300 and -300 correspond to angles with $\omega = 89.8$ and $\omega = -89.8$ degrees (1.57 and -1.57 radians) (Figures 4.5 and 4.6).

Usually, similar priors are assumed for slopes of Bayesian regression models. Now, we show that the meaning of these priors depends on the meaning of covariates. For example,

$$\beta_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \beta_2 \sim \mathrm{N}(0, 100^2)$$

priors assumed for slopes based on non-scaled covariates demonstrates that we assume that both slopes have equal properties disregarding the original units of covariates x_1 and x_2. In contrast, priors

$$\delta_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \delta_2 \sim \mathrm{N}(0, 100^2)$$

assumed for slopes based on scaled covariates correspond to

$$\delta_1 = \beta_1\hat{\sigma}_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \delta_2 = \beta_2\hat{\sigma}_2 \sim \mathrm{N}(0, 100^2)$$

and to two different priors

$$\beta_1 \sim \mathrm{N}(0, (100/\hat{\sigma}_1)^2) \quad \text{and} \quad \beta_2 \sim \mathrm{N}(0, (100/\hat{\sigma}_2)^2)$$

for non-scaled covariates x_1 and x_2. This demonstrates that the effective meaning of priors always depends on the effective scale of covariates included in the equation of the multiple linear regression.

To standardize or not to standardize, that is the question. There are two different approaches to standardize covariates. First, the function `scale` in `R` can center, scale, and standardize (center and scale) covariates. Second, functions `rescale` and `standardize` in the `arm` package in `R` implement an approach proposed by [Gelman, 2008]. This approach scales continuous covariats by twice standard deviation, so that the standard deviation of the scaled covariate is equal to 0.5. This is an attempt to adjust the meaning of binary and continuous covariates, so that the slope of scaled continuous and binary covariates is defined on an equal footing in terms of a unit change, which is equal to twice the standard deviation of scaled covariates ($2 \times 0.5 = 1$). Moreover, in this setting the same generic default prior can be assumed for all slopes of standardized covariates in a Bayesian regression model for an automatic use [Gelman et al., 2008]. For original covariates, these generic priors for standardized covariates induce covariate-specific priors that highly depend on standard deviations that were used for scaling.

By default, the `scale` function in `R` scales covariates x_1 and x_2 by their corresponding standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$. Note that we can use this function to scale the covariates by a different value. For example, we can modify the call of the `scale` function slightly

```
scale(X, center = FALSE, scale = 2*apply(X, 2, sd, na.rm = TRUE)), 2, summary))
```

to scale covariates x_1 and x_2 stored in a matrix `X` by $2\hat{\sigma}_1$ and $2\hat{\sigma}_2$. Note that the slope γ of a covariate scaled by $2\hat{\sigma}$ is twice as large as the slope δ of a covariate scaled by $\hat{\sigma}$:

$$\gamma = \frac{\Delta y}{\Delta x/(2\hat{\sigma})} = \beta 2\hat{\sigma} = 2\delta, \quad \text{where} \quad \delta = \frac{\Delta y}{\Delta x/\hat{\sigma}} = \beta\hat{\sigma}.$$

For original non-transformed covariates, there are sevaral approaches that use information specific to a particular analysis to elicit covariate-specific priors of slopes [Gelman et al., 2008, Section 1.2]. For example, one can set a prior distribution by eliciting the possible values of outcomes given different combinations of regression imputs. Alternatively, one can elicit a prior distribution by characterizing expecet effects in informativeness ranges.

Although centering of covariates is unnecessary if we are interested in the slope [Gelman, 2008], centering can facilitate interpretation of interactions. Belsley [1991] warns that centering of covariates does not solve the problem of collinearity but rather hides it, so that a potentially harmful impact of collinearity cannot any longer be clearly detected and diagnosed. In any case, we should be aware of the impact of centering and scaling on the values of parameter estimates in a multiple linear regession in both the classical and the Bayesian setting.

4.11 Worksheet 4

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		Likelihood
Conjugate Bayes	**MCMC sampling**	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 4.1: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 4.

Acknowledgement:

We thank Sona Hunanyan for typing an earlier version of this script for the Bayesian Data Analysis lecture in the spring term 2018.

Secukinumab		**Placebo**
Bayesian sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		Data
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 4.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Bibliography

D. Baeten, X. Baraliakos, J. Braun, J. Sieper, P. Emery, D. van der Heijde, I. McInnes, J. van Laar, R. Landewé, P. Wordsworth, J. Wollenhaupt, H. Kellner, J. Paramarta, J. Wei, A. Brachat, S. Bek, D. Laurent, Y. Li, Y.A. Wang, A.P. Bertolino, S. Gsteiger, A.M. Wright, and W. Hueber. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *The Lancet*, 382:1705–1713, 2013.

M.J. Bayarri and J.O. Berger. An interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

D.A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley & Sons Inc, 1991.

A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, 2008. URL `https://doi.org/10.1002/sim.3107`.

A. Gelman, A. Jakulin, M.G. Pittau, and Y-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2 (4):1360–1383, 2008.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. doi: 10.1109/TPAMI.1984.4767596.

W.R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

S.N. Goodman. Toward evidence-based medical statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999a. URL `https://doi.org/10.7326/0003-4819-130-12-199906150-00019`.

S.N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999b. URL `https://doi.org/10.7326/0003-4819-130-12-199906150-00008`.

S. Hartnack and M. Roos. Teaching: confidence, prediction and tolerance intervals in scientific practice: a tutorial on binary variables. *Emerging Themes in Epidemiology*, 18(17), 2021. URL `https://doi.org/10.1186/s12982-021-00108-1`.

L Held and M. Ott. On p-values and Bayes Factors. *Annual Review of Statistics and Its Application*, 5:393–419, 2018. URL https://doi.org/10.1146/annurev-statistics-031017-100307.

L. Held and D. Sabanés Bové. *Applied Statistical Inference. Likelihood and Bayes.* Springer, 2014.

L. Held and D. Sabanés Bové. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine.* Springer, 2020. URL https://www.springer.com/gp/book/9783662607916.

S.D. Hill and J.C. Spall. Stationarity and convergence of the Metropolis-Hastings algorithm: Insights into theoretical aspects. *IEEE Control Systems Magazine*, 39(1):56–67, 2019. doi: 10.1109/MCS.2018.2876959. URL https://ieeexplore.ieee.org/document/8616920.

A.A. Johnson, M.Q. Ott, and M. Dogucu. *Bayes Rules! An Introduction to Applied Bayesian Modeling.* Chapman and Hall/CRC, 2022. URL https://www.bayesrulesbook.com.

L.M. Leemis and J.T. McQueston. Univariate Distribution Relationships. *The American Statistician*, 62(1):45–53, 2008. doi: 10.1198/000313008X270448. URL https://doi.org/10.1198/000313008X270448.

G.M. Martin, D.T. Frazier, and C.P. Robert. Computing Bayes: Bayesian computation from 1763 to the 21st century. *arXiv:2004.06425*, 2020. URL https://arxiv.org/abs/2004.06425.

N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science*, 15:125–130, 1987.

R.M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

I. Ntzoufras. *Bayesian Modeling Using WinBUGS.* John Wiley & Sons, Inc., 2009.

C.P. Robert and G. Casella. *Introducing Monte Carlo Methods with R.* Springer, 2010.

J.N. Rouder and R.D. Morey. Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73(2):186–190, 2019. URL https://doi.org/10.1080/00031305.2017.1341334.

D.J. Spiegelhalter, L.S. Freedman, and K.B. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):357–416, 1994. URL http://www.jstor.com/stable/2983527.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B.*, 64(4):583–616, 2002. ISSN 1369-7412. URL https://doi.org/10.1111/1467-9868.00353.