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

March 9, 2022

Contents

1 Lecture 1: Classical vs Bayesian paradigms and conditional probability 2

1.1 Overview of the lecture . . . 2
1.2 Overview of the individual project . . . 3
1.3 History . . . 5
1.4 Probability calculus . . . 6
1.5 Example: Breast cancer and diagnostic tests . . . 8
1.6 Overview of the classical statistic . . . 8
1.6.1 Example: Primary outcome follows normal distribution . . . 9
1.6.2 Example: Primary outcome follows Bernoulli distribution . . . 10
1.7 Overview of the Bayesian methodology . . . 10
1.7.1 Bayes factors and p-values . . . 11
1.7.2 Priors . . . 13
1.8 Worksheet 1 . . . 13

2 Lecture 2: Conjugate Bayes, point estimates, and interval estimates 16

2.1 Binary data: Vision correction . . . 16
2.2 Bayes analysis of binary data . . . 17
2.2.1 Beta prior . . . 17
2.2.2 Likelihood . . . 18
2.2.3 Posterior distribution . . . 19
2.3 Bayes analysis of normal data . . . 21
2.4 Point estimates . . . 23
2.5 Credible intervals . . . 23
2.6 Worksheet 2 . . . 25

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 $A \to B$:

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

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 $A \to B$:
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: $P[T^- \mid D^-] = 0.968$

$$\text{Full bivariate distribution } P[T \cap D] = \begin{array}{c} \\ T^+_{\text{yes}} \\ T^-_{\text{no}} \end{array} \begin{array}{c} \begin{array}{cc} D^+_{\text{yes}} & D^-_{\text{no}} \end{array} \\ \begin{pmatrix} 0.03897 & 0.03056 \\ 0.00603 & 0.92444 \end{pmatrix} \end{array}$$

$$\text{marginal distribution for Test } P[T] = \begin{array}{c} T^+ \\ T^- \end{array} \begin{pmatrix} 0.06953 \\ 0.93047 \end{pmatrix}$$

$$\text{marginal distribution for Disease } P[D] = \begin{array}{c} D^+ \\ D^- \end{array} \begin{pmatrix} 0.045 \\ 0.955 \end{pmatrix}$$

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

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

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).

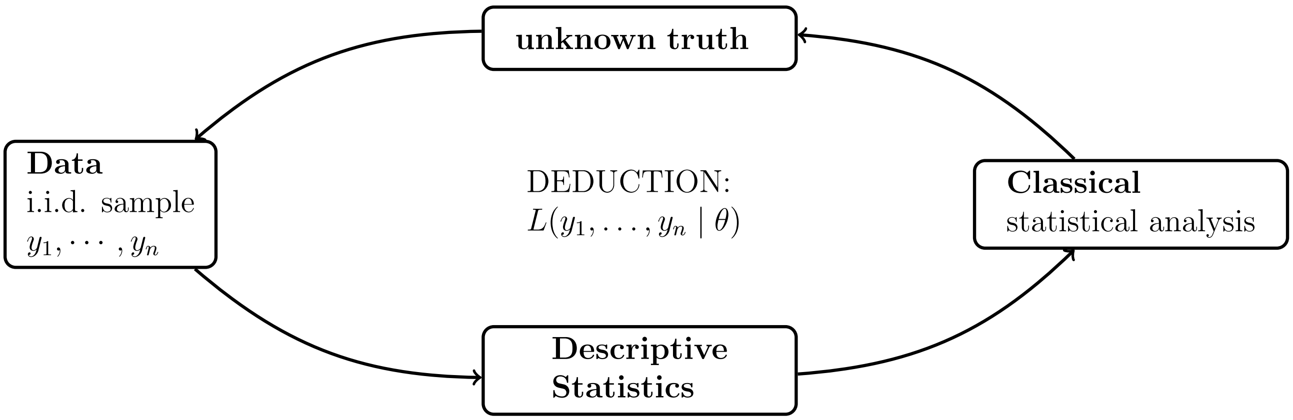

Figure 1.1: Overview of the classical statistic

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

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

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

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 $Y \overset{i.i.d.}{\sim} \mathrm{Be}(p)$ 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

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

Log-likelihood

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

In order to derive an estimator of p, compute

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

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

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

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

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

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,

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

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

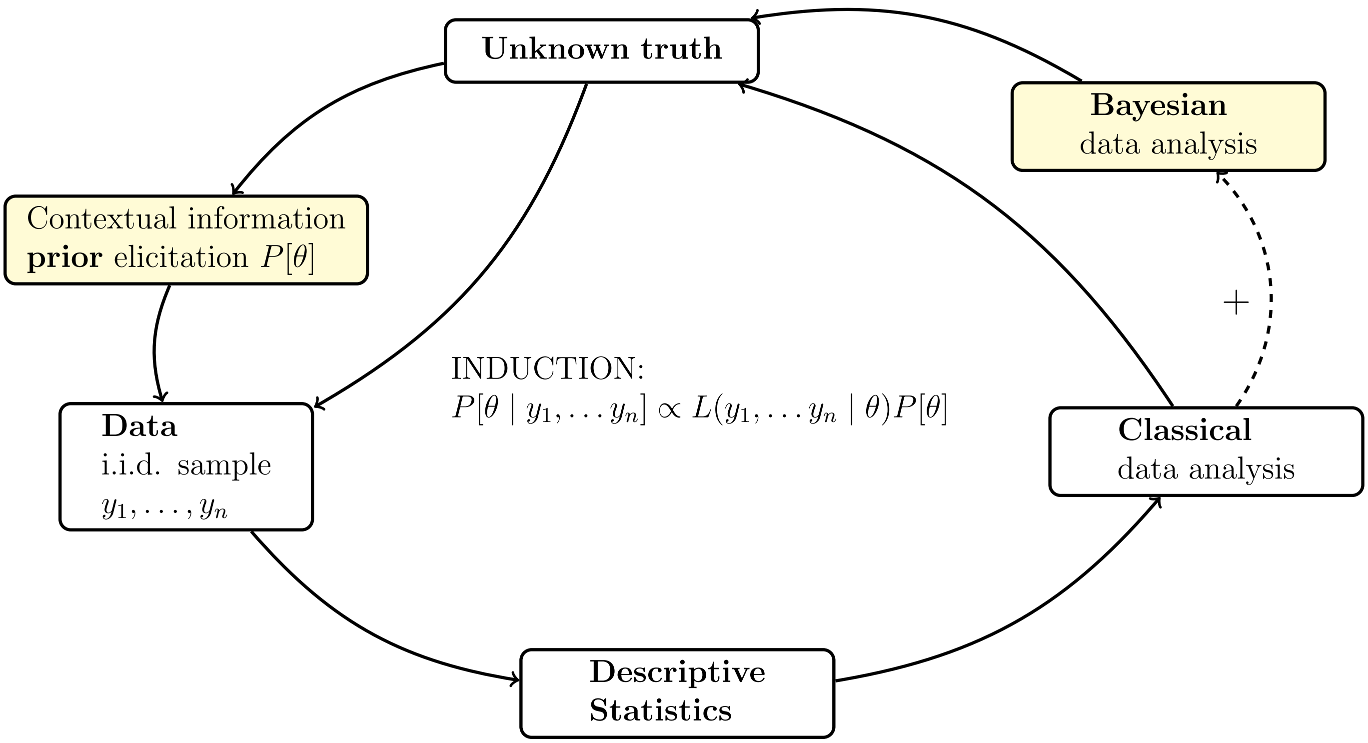

Figure 1.2: Overview of the Bayesian methodology. Fields with yellow background correspond to Bayes-specific steps.

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

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

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.
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 $\hat{p} = 16/22 = 0.727$ and an interval estimate in form of a 95% confidence interval (95%CI).

```
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.

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).

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).

$$\begin{aligned} X &\sim \text{Beta}(\alpha,\beta) \\ \mathbb{E}X &= \frac{\alpha}{\alpha+\beta} \\ \text{Var}X &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mathbb{E}(X)(1-\mathbb{E}X)}{(\alpha+\beta+1)}. \end{aligned}$$

Prior effective sample size

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

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)

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

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

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.

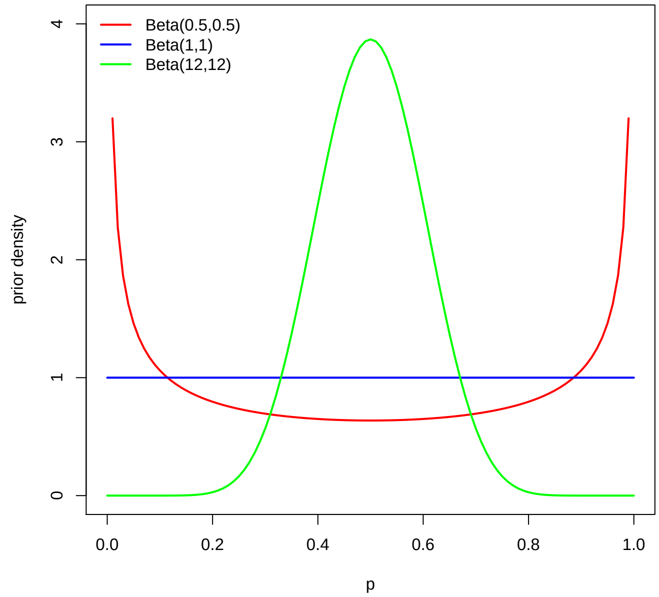

Figure 2.1: Beta priors in the vision correction example.

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:

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

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.

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

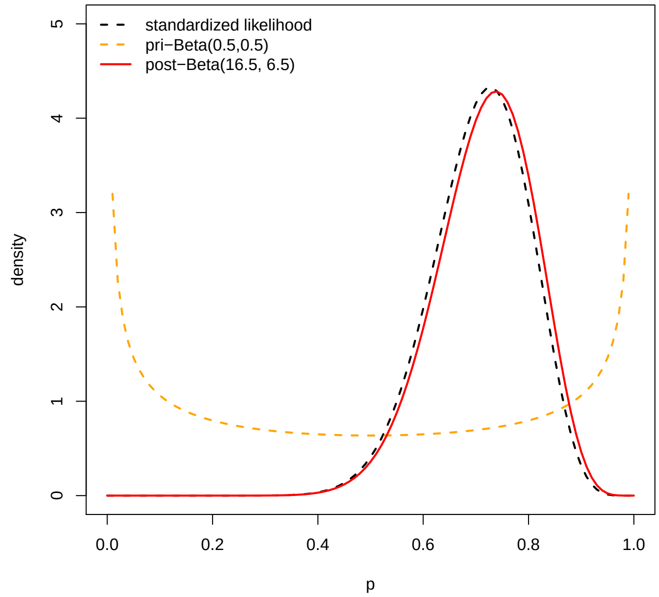

Figure 2.2: Likelihood, prior Beta(0.5, 0.5), posterior Beta(16.5, 6.5) in the vision correction example with $n\bar{y} = 16$ and $n = 22$.

- 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.

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 $y_1, \dots, y_n$ 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 $y_1, \dots, y_n$ have been observed.

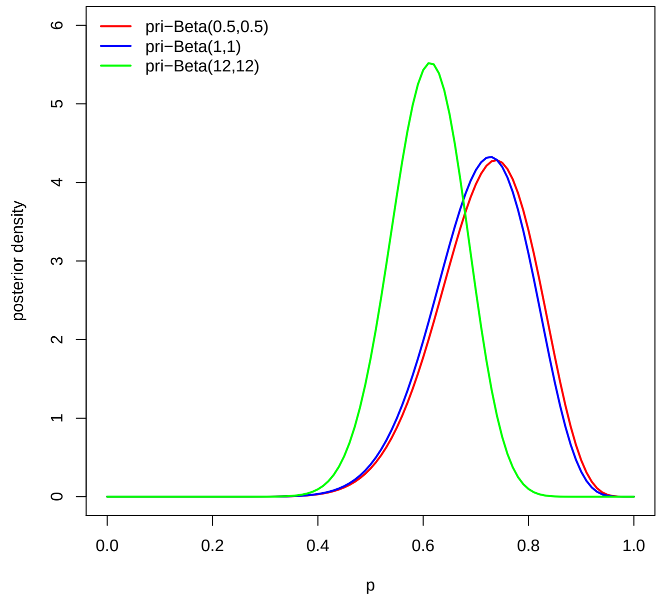

Figure 2.3: Posteriors in the vision correction example obtained for three different priors: Beta(0.5, 0.5), Beta(1, 1), and Beta(12, 12).

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

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

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

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

Thus, we get

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

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 $(1-\alpha)100\%\mathrm{CrI}(\theta)$:

(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 $\pi(\theta \mid \mathbf{y})$ 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 $\theta \mid \mathbf{y}$. 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.

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**
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.

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.

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

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`.

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`.