Solutions for Exercises in:

Applied Statistical Inference: Likelihood and Bayes

Leonhard Held and Daniel Sabanés Bové

Solutions provided by:
Leonhard Held, Daniel Sabanés Bové, Andrea Kraus and Manuela Ott

March 31, 2017

Corrections and other feedback are appreciated. Please send them to
manuela.ott@uzh.ch

Contents

2 Likelihood 1

3 Elements of frequentist inference 13

4 Frequentist properties of the likelihood 29

5 Likelihood inference in multiparameter models 51

6 Bayesian inference 115

7 Model selection 143

8 Numerical methods for Bayesian inference 157

9 Prediction 195

Bibliography 205

Index 207

2 Likelihood

1. Examine the likelihood function in the following examples.

a) In a study of a fungus which infects wheat, 250 wheat seeds are disseminated after contaminating them with the fungus. The research question is how large the probability θ is that an infected seed can germinate. Due to technical problems, the exact number of germinated seeds cannot be evaluated, but we know only that less than 25 seeds have germinated. Write down the likelihood function for θ based on the information available from the experiment.

▶ *Since the seeds germinate independently of each other with fixed probability θ, the total number X of germinated seeds has a binomial distribution, $X \sim \mathrm{Bin}(250, \theta)$. The event we know that happened is $X \leq 24$. Thus, the likelihood function for θ is here the cdf of the binomial distribution with parameter θ evaluated at 24:*

$$\begin{aligned} L(\theta) &= \mathsf{Pr}(X \leq 24; \theta) \\ &= \sum_{x=0}^{24} f(x; n = 250, \theta) \\ &= \sum_{x=0}^{24} \binom{250}{x} \theta^x (1-\theta)^{250-x}. \end{aligned}$$

b) Let $X_{1:n}$ be a random sample from a $\mathrm{N}(\theta, 1)$ distribution. However, only the largest value of the sample, $Y = \max(X_1, \dots, X_n)$, is known. Show that the density of Y is

$$f(y) = n \left\{\Phi(y-\theta)\right\}^{n-1} \varphi(y-\theta), \quad y \in \mathbb{R},$$

where $\Phi(\cdot)$ is the distribution function and $\varphi(\cdot)$ is the density function of the standard normal distribution $\mathrm{N}(0, 1)$. Derive the distribution function of Y and the likelihood function $L(\theta)$.

▶ *We have $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{N}(\theta, 1)$ and their maximum is Y. Hence, the cdf of Y is*

$$\begin{aligned} F(y) = \Pr(Y \leq y) &= \\ &= \Pr\{\max(X_1, \ldots, X_n) \leq y\} \\ &= \prod_{i=1}^{n} \Pr(X_i \leq y) \\ &= \{\Phi(y-\theta)\}^n \end{aligned}$$

because $X_i \leq y$ is equivalent to $X_i - \theta \leq y - \theta$, and $X_i - \theta$ follows a standard normal distribution for all $i = 1, \ldots, n$. The probability density function of Y follows:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= n\{\Phi(y-\theta)\}^{n-1} \phi(y-\theta) \end{aligned}$$

and the likelihood function $L(\theta)$ is exactly this density function, but seen as a function of θ for fixed y.

c) Let $X_{1:3}$ denote a random sample of size $n = 3$ from a Cauchy $\mathrm{C}(\theta, 1)$ distribution, *cf*. Appendix A.5.2. Here $\theta \in \mathbb{R}$ denotes the location parameter of the Cauchy distribution with density

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2}, \quad x \in \mathbb{R}.$$

Derive the likelihood function for θ.

▶ *Since we have an iid sample from the $\mathrm{C}(\theta, 1)$ distribution, the likelihood function is given by the product of the densities $f(x_i; \theta)$:*

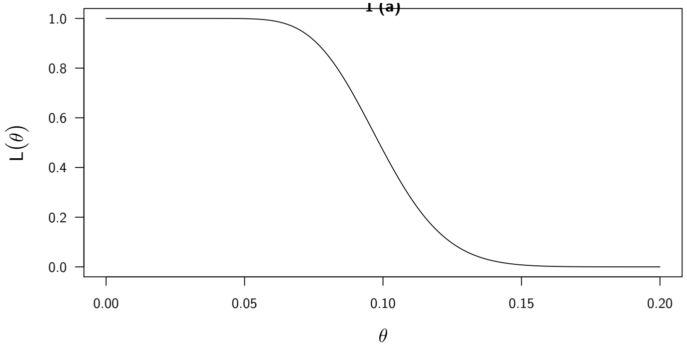
$$\begin{aligned} L(\theta) &= \prod_{i=1}^{3} f(x_i; \theta) \\ &= \prod_{i=1}^{3} \frac{1}{\pi} \cdot \frac{1}{1 + (x_i - \theta)^2} \\ &= \frac{1}{\pi^3} \frac{1}{\{1 + (x_1-\theta)^2\}\{1 + (x_2-\theta)^2\}\{1 + (x_3-\theta)^2\}}. \end{aligned}$$

d) Using `R`, produce a plot of the likelihood functions:

i. $L(\theta)$ in 1a).

▶ *We can use the implemented cdf (`pbinom`) of the binomial distribution:*

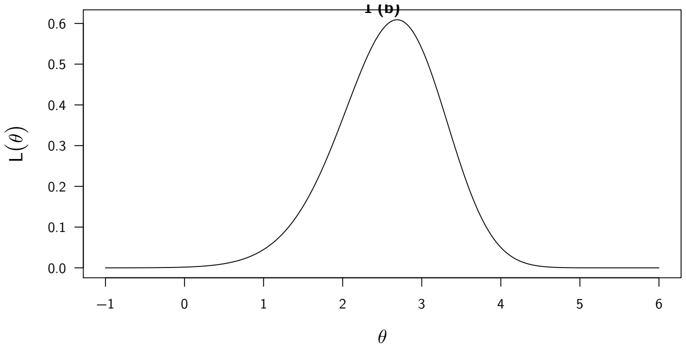
```
> ## likelihood function in partially observed
> ## binomial experiment:
> likelihood1 <- function(theta,     # probability
                          n,         # sample size
                          x)         # number in interval [0, x] was observed
  {
      pbinom(x, size = n, p = theta)      # use existing function
  }
> theta <- seq(from=0, to=0.2,            # range for theta
               length=1000)               # determines resolution
> plot(theta, likelihood1(theta, n=250, x=24),
       type="l",
       xlab=expression(theta), ylab=expression(L(theta)),
       main="1 (a)")
```



ii. $L(\theta)$ in 1b) if the observed sample is $x = (1.5, 0.25, 3.75, 3.0, 2.5)$.

▶ *Here too we can use the implemented functions from* `R`*:*

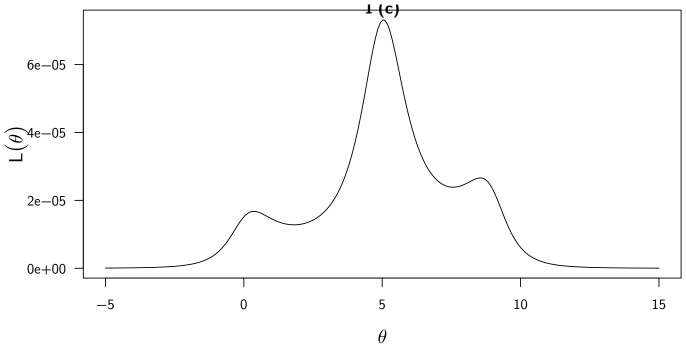
```
> ## likelihood for mean of normal distribution if only the
> ## maximum is known from the sample
> likelihood2 <- function(theta,         # mean of iid normal distributions
                          n,             # sample size
                          y)             # observed maximum
  {
      n * pnorm(y - theta)^(n-1) * dnorm(y - theta)
  }
> x <- c(1.5,0.25, 3.75, 3.0, 2.5)
> theta <- seq(-1, 6, length = 1000)
> plot(theta, likelihood2(theta, y=max(x), n=length(x)),
       type="l",
       xlab=expression(theta), ylab=expression(L(theta)),
       main="1 (b)")
```



iii. $L(\theta)$ in 1c) if the observed sample is $x = (0, 5, 9)$.

▶ *Here the vector computations are very useful:*

```
> ## likelihood for location parameter of Cauchy in
> ## random sample
> likelihood3 <- function(theta,          # location parameter
                          x)              # observed data vector
  {
      1/pi^3 / prod((1+(x-theta)^2))
  }
> ## In order to plot the likelihood, the function must be able
> ## to take not only one theta value but a theta vector. We can use the
> ## following trick to get a vectorised likelihood function:
> likelihood3vec <- Vectorize(likelihood3,
                              vectorize.args="theta")
> x <- c(0, 5, 9)
> theta <- seq(from=-5, to=15,
               length = 1000)
> plot(theta, likelihood3vec(theta, x=x),
       type="l",
       xlab=expression(theta), ylab=expression(L(theta)),
       main="1 (c)")
```



2. A first-order autoregressive process $X_0, X_1, \dots, X_n$ is specified by the conditional distribution

$$X_i \,|\, X_{i-1} = x_{i-1}, \dots, X_0 = x_0 \sim \mathrm{N}(\alpha \cdot x_{i-1}, 1), \quad i = 1, 2, \dots, n$$

and some initial distribution for X_0. This is a popular model for time series data.

a) Consider the observation $X_0 = x_0$ as fixed. Show that the log-likelihood kernel for a realization $x_1, \dots, x_n$ can be written as

$$l(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2.$$

▶ *The likelihood is given by*

$$\begin{aligned} L(\alpha) &= f(x_1, \dots, x_n \,|\, x_0; \alpha) \\ &= f(x_n \,|\, x_{n-1}, \dots, x_1, x_0; \alpha) f(x_{n-1} \,|\, x_{n-2}, \dots, x_0; \alpha) \cdots f(x_1 \,|\, x_0; \alpha) \\ &= \prod_{i=1}^{n} f(x_i \,|\, x_{i-1}, \dots, x_0; \alpha) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \alpha x_{i-1})^2\right\} \\ &= \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(x_i - \alpha x_{i-1})^2\right\} \\ &= \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}(x_i - \alpha x_{i-1})^2\right\}. \end{aligned}$$

The log-likelihood kernel is thus

$$l(\alpha) = \log L(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2.$$

b) Derive the score equation for α, compute $\hat{\alpha}_{\mathrm{ML}}$ and verify that it is really the maximum of $l(\alpha)$.

▶ *The score function for α is*

$$\begin{aligned} S(\alpha) &= \frac{dl(\alpha)}{d\alpha} \\ &= -\frac{1}{2}\sum_{i=1}^{n} 2(x_i - \alpha x_{i-1}) \cdot (-x_{i-1}) \\ &= \sum_{i=1}^{n} x_i x_{i-1} - \alpha x_{i-1}^2 \\ &= \sum_{i=1}^{n} x_i x_{i-1} - \alpha \sum_{i=1}^{n} x_{i-1}^2, \end{aligned}$$

so the score equation $S(\alpha) = 0$ is solved by

$$\hat{\alpha}_{\text{ML}} = \frac{\sum_{i=1}^{n} x_i x_{i-1}}{\sum_{i=1}^{n} x_{i-1}^2}.$$

This is really a local maximum of the log-likelihood function, because the latter is (strictly) concave, which can easily be verified from

$$\frac{dS(\alpha)}{d\alpha} = -\sum_{i=1}^{n} x_{i-1}^2 < 0.$$

The Fisher information

$$I(\alpha) = -\frac{dS(\alpha)}{d\alpha} = \sum_{i=1}^{n} x_{i-1}^2 > 0$$

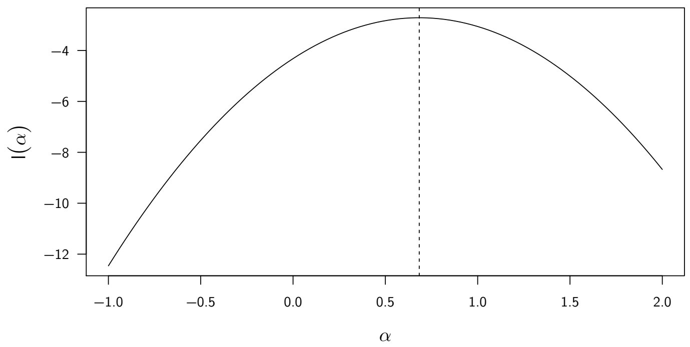
is here independent of the parameter and thus equals the observed Fisher information $I(\hat{\alpha}_{\text{ML}})$. Since there are no other local maxima or restrictions of the parameter space ($\alpha \in \mathbb{R}$), $\hat{\alpha}_{\text{ML}}$ is really the global maximum of $l(\alpha)$.

c) Create a plot of $l(\alpha)$ and compute $\hat{\alpha}_{\text{ML}}$ for the following sample:

$$(x_0, \ldots, x_6) = (-0.560, -0.510, 1.304, 0.722, 0.490, 1.960, 1.441).$$

▶ *Note that in the R-code `x[1]` corresponds to x_0 in the text because indexing starts at 1 in R.*

```
> ## implement log-likelihood kernel
> loglik <- function(alpha,                # parameter for first-order term
                     x)                    # observed data vector
  {
      i <- 2:length(x)
      - 1/2 * sum((x[i] - alpha * x[i-1])^2)
  }
> ## and plot for the given data
> x <- c(-0.560, -0.510, 1.304, 0.722, 0.490, 1.960, 1.441)
> alpha <- seq(-1, 2, length = 100)
> plot(x=alpha,
       y=sapply(alpha, function(alpha) loglik(alpha, x)),
       type = "l",
       xlab = expression(alpha),
       ylab = expression(l(alpha)))
> ## then calculate the MLE and plot it
> i <- seq(along = x)[-1]              # again an indexing vector is necessary here
> alphaMl <- sum(x[i] * x[i-1]) / sum(x[i-1]^2)
> alphaMl
[1] 0.6835131
> abline(v = alphaMl, lty = 2)
```



3. Show that in Example 2.2 the likelihood function $L(N)$ is maximised at $\hat{N} = \lfloor \frac{M \cdot n}{x} \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is smaller than x. To this end, analyse the monotonic behaviour of the ratio $L(N)/L(N-1)$. In which cases is the MLE not unique? Give a numeric example.

▶ *For $N \in \Theta = \{\max(n, M+n-x), \max(n, M+n-x)+1, \ldots\}$, the likelihood function is*

$$L(N) \propto \frac{\binom{N-M}{n-x}}{\binom{N}{n}}.$$

The ratio $R(N) = L(N)/L(N-1)$ is thus

$$\begin{aligned} R(N) &= \frac{\binom{N-M}{n-x}}{\binom{N}{n}} \cdot \frac{\binom{N-1}{n}}{\binom{N-1-M}{n-x}} \\ &= \frac{(N-M)!n!(N-n)!}{(n-x)!(N-M-n+x)!N!} \cdot \frac{(N-1)!(n-x)!(N-1-M-n+x)!}{n!(N-1-n)!(N-1-M)!} \\ &= \frac{(N-M)(N-n)}{(N-M-n+x)N}. \end{aligned}$$

As the local maximum of $L(N)$, the MLE $\hat{N}_{\text{ML}}$ has to satisfy both that $L(\hat{N}_{\text{ML}}) \geq L(\hat{N}_{\text{ML}}-1)$ (i. e. $R(\hat{N}_{\text{ML}}) \geq 1$) and that $L(\hat{N}_{\text{ML}}) \geq L(\hat{N}_{\text{ML}}+1)$ (i. e. $R(\hat{N}_{\text{ML}}+1) \leq 1$). From the equation above we can see that $R(N) \geq 1$ if and only if $N \leq Mn/x$. Hence, $R(N+1) \leq 1$ if and only if $N+1 \geq Mn/x$, and, equivalently, if $N \geq Mn/x - 1$. It follows that each integer in the interval $[Mn/x - 1, Mn/x]$ is an MLE. If the right endpoint Mn/x is not integer, the MLE $\hat{N}_{\text{ML}} = \lfloor Mn/x \rfloor$ is unique. However, if Mn/x is integer, we have two solutions and MLE is not unique.
For example, if we change the sample size in the numerical example in Figure 2.2 from $n = 63$ to $n = 65$, we obtain the highest likelihood for both $26 \cdot 65/5 = 1690$ and 1689.

4. Derive the MLE of π for an observation x from a geometric Geom(π) distribution. What is the MLE of π based on a realization $x_{1:n}$ of a random sample from this distribution?

▶ *The log-likelihood function is*

$$l(\pi) = \log f(x;\pi) = \log(\pi) + (x-1)\log(1-\pi),$$

so the score function is

$$S(\pi) = \frac{d}{d\pi} l(\pi) = \frac{1}{\pi} - \frac{x-1}{1-\pi}.$$

Solving the score equation $S(\pi) = 0$ *yields the MLE* $\hat{\pi}_{\text{ML}} = 1/x$. *The Fisher information is*

$$I(\pi) = -\frac{d}{d\pi} S(\pi) = \frac{1}{\pi^2} + \frac{x-1}{(1-\pi)^2},$$

which is positive for every $0 < \pi < 1$, *since* $x \geq 1$ *by definition. Thus,* $1/x$ *indeed maximises the likelihood.*

For a realisation $x_{1:n}$ *of a random sample from this distribution, the quantities calculated above become*

$$\begin{aligned} l(\pi) &= \sum_{i=1}^{n} \log f(x_i;\pi) \\ &= \sum_{i=1}^{n} \log(\pi) + (x_i - 1)\log(1-\pi) \\ &= n\log(\pi) + n(\bar{x}-1)\log(1-\pi), \\ S(\pi) &= \frac{d}{d\pi} l(\pi) \\ &= \frac{n}{\pi} - \frac{n(\bar{x}-1)}{1-\pi}, \\ \text{and} \quad I(\pi) &= -\frac{d}{d\pi} S(\pi) \\ &= \frac{n}{\pi^2} + \frac{n(\bar{x}-1)}{(1-\pi)^2}. \end{aligned}$$

The Fisher information is again positive, thus the solution $1/\bar{x}$ *of the score equation is the MLE.*

5. A sample of 197 animals has been analysed regarding a specific phenotype. The number of animals with phenotypes AB, Ab, aB and ab, respectively, turned out to be

$$\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\top = (125, 18, 20, 34)^\top.$$

A genetic model now assumes that the counts are realizations of a multinomially distributed multivariate random variable $\boldsymbol{X} \sim \mathrm{M}_4(n, \boldsymbol{\pi})$ with $n = 197$ and probabilities $\pi_1 = (2+\phi)/4$, $\pi_2 = \pi_3 = (1-\phi)/4$ and $\pi_4 = \phi/4$ (Rao, 1973, p. 368).

a) What is the parameter space of ϕ? See Table A.3 in the Appendix for details on the multinomial distribution and the parameter space of $\boldsymbol{\pi}$.

▶ *The first requirement for the probabilities is satisfied for all* $\phi \in \mathbb{R}$:

$$\sum_{j=1}^{4} \pi_j = \frac{1}{4}\big(2 + \phi + 2(1-\phi) + \phi\big) = \frac{1}{4}(4 + 2\phi - 2\phi) = 1.$$

Moreover, each probability π_j $(j = 1, \ldots, 4)$ *must lie in the interval* $(0,1)$. *We thus have*

$$\begin{aligned} 0 < \frac{2+\phi}{4} < 1 &\iff 0 < 2 + \phi < 4 \iff -2 < \phi < 2, \\ 0 < \frac{1-\phi}{4} < 1 &\iff 0 < 1 - \phi < 4 \iff -3 < \phi < 1, \qquad (2.1) \\ 0 < \frac{\phi}{4} < 1 &\iff 0 < \phi < 4. \qquad (2.2) \end{aligned}$$

Hence, (2.2) *and* (2.1) *imply the lower and upper bounds, respectively, for the range* $0 < \phi < 1$. *This is the intersection of the sets suitable for the probabilities.*

b) Show that the likelihood kernel function for ϕ, based on the observation $\boldsymbol{x}$, has the form

$$L(\phi) = (2+\phi)^{m_1}(1-\phi)^{m_2}\phi^{m_3}$$

and derive expressions for m_1, m_2 and m_3 depending on $\boldsymbol{x}$.

▶ *We derive the likelihood kernel function based on the probability mass function:*

$$\begin{aligned} L(\phi) &= \frac{n!}{\prod_{j=1}^{4} x_j!} \prod_{j=1}^{4} \pi_j^{x_j} \\ &\propto \left(\frac{2+\phi}{4}\right)^{x_1} \left(\frac{1-\phi}{4}\right)^{x_2} \left(\frac{1-\phi}{4}\right)^{x_3} \left(\frac{\phi}{4}\right)^{x_4} \\ &= \left(\frac{1}{4}\right)^{x_1+x_2+x_3+x_4} (2+\phi)^{x_1}(1-\phi)^{x_2+x_3}\phi^{x_4} \\ &\propto (2+\phi)^{m_1}(1-\phi)^{m_2}\phi^{m_3} \end{aligned}$$

with $m_1 = x_1$, $m_2 = x_2 + x_3$ *and* $m_3 = x_4$.

c) Derive an explicit formula for the MLE $\hat{\phi}_{\text{ML}}$, depending on m_1, m_2 and m_3. Compute the MLE given the data given above.

▶ *The log-likelihood kernel is*

$$l(\phi) = m_1 \log(2+\phi) + m_2 \log(1-\phi) + m_3 \log(\phi),$$

so the score function is

$$\begin{aligned} S(\phi) &= \frac{dl(\phi)}{d\phi} \\ &= \frac{m_1}{2+\phi} + \frac{m_2}{1-\phi}(-1) + \frac{m_3}{\phi} \\ &= \frac{m_1(1-\phi)\phi - m_2(2+\phi)\phi + m_3(2+\phi)(1-\phi)}{(2+\phi)(1-\phi)\phi}. \end{aligned}$$

The score equation $S(\phi) = 0$ is satisfied if and only if the numerator in the expression above equals zero, i. e. if

$$\begin{aligned} 0 &= m_1(1-\phi)\phi - m_2(2+\phi)\phi + m_3(2+\phi)(1-\phi) \\ &= m_1\phi - m_1\phi^2 - 2m_2\phi - m_2\phi^2 + m_3(2 - 2\phi + \phi - \phi^2) \\ &= \phi^2(-m_1 - m_2 - m_3) + \phi(m_1 - 2m_2 - m_3) + 2m_3. \end{aligned}$$

This is a quadratic equation of the form $a\phi^2 + b\phi + c = 0$, with $a = -(m_1 + m_2 + m_3)$, $b = (m_1 - 2m_2 - m_3)$ and $c = 2m_3$, which has two solutions $\phi_{0/1} \in \mathbb{R}$ given by

$$\phi_{0/1} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There is no hope for simplifying this expression much further, so we just implement it in R, and check which of $\phi_{0/1}$ is in the parameter range $(0, 1)$:

```
> mle.phi <- function(x)
  {
      m <- c(x[1], x[2] + x[3], x[4])
      a <- - sum(m)
      b <- m[1] - 2 * m[2] - m[3]
      c <- 2 * m[3]
      phis <- (- b + c(-1, +1) * sqrt(b^2 - 4 * a * c)) / (2 * a)
      correct.range <- (phis > 0) & (phis < 1)
      return(phis[correct.range])
  }
> x <- c(125, 18, 20, 34)
> (phiHat <- mle.phi(x))
[1] 0.6268215
```

Note that this example is also used in the famous EM algorithm paper (Dempster et al., 1977, p. 2), producing the same result as we obtained by using the EM algorithm (cf. Table 2.1 in Subsection 2.3.2).

d) What is the MLE of $\theta = \sqrt{\phi}$?

▶ *From the invariance property of the MLE we have*

$$\hat{\theta}_{\text{ML}} = \sqrt{\hat{\phi}_{\text{ML}}},$$

which in the example above gives $\hat{\theta}_{\text{ML}} \approx 0.792$:

```
> (thetaHat <- sqrt(phiHat))
[1] 0.7917206
```

6. Show that $h(X) = \max_i(X_i)$ is sufficient for θ in Example 2.18.

▶ *From Example 2.18, we know that the likelihood function of θ is*

$$L(\theta) = \begin{cases} \prod_{i=1}^n f(x_i;\theta) = \frac{1}{\theta^n} & \text{for } \theta \geq \max_i(x_i), \\ 0 & \text{otherwise} \end{cases}.$$

We also know that $L(\theta) = f(x_{1:n};\theta)$, the density of the random sample. The density can thus be rewritten as

$$f(x_{1:n};\theta) = \frac{1}{\theta^n} \mathsf{I}_{[0,\theta]}(\max_i(x_i)).$$

Hence, we can apply the Factorisation theorem (Result 2.2) with $g_1(t;\theta) = \frac{1}{\theta^n}\mathsf{I}_{[0,\theta]}(t)$ and $g_2(x_{1:n}) = 1$ to conclude that $T = \max_i(X_i)$ is sufficient for θ.

7. a) Let $X_{1:n}$ be a random sample from a distribution with density

$$f(x_i;\theta) = \begin{cases} \exp(i\theta - x_i) & x_i \geq i\theta \\ 0 & x_i < i\theta \end{cases}$$

for X_i, $i = 1, \dots, n$. Show that $T = \min_i(X_i/i)$ is a sufficient statistic for θ.

▶ *Since $x_i \geq i\theta$ is equivalent to $x_i/i \geq \theta$, we can rewrite the density of the i-th observation as*

$$f(x_i;\theta) = \exp(i\theta - x_i)\mathsf{I}_{[\theta,\infty)}(x_i/i).$$

The joint density of the random sample then is

$$\begin{aligned} f(x_{1:n};\theta) &= \prod_{i=1}^n f(x_i) \\ &= \exp\left\{\theta\left(\sum_{i=1}^n i\right) - n\bar{x}\right\} \prod_{i=1}^n \mathsf{I}_{[\theta,\infty)}(x_i/i) \\ &= \underbrace{\exp\left\{\theta\frac{n(n+1)}{2}\right\}\mathsf{I}_{[\theta,\infty)}(\min_i(x_i/i))}_{=g_2(h(x_{1:n})=\min_i(x_i/i);\theta)} \cdot \underbrace{\exp\{-n\bar{x}\}}_{=g_2(x_{1:n})}. \end{aligned}$$

The result now follows from the Factorisation Theorem (Result 2.2). The crucial step is that $\prod_{i=1}^n \mathsf{I}_{[\theta,\infty)}(x_i/i) = \mathsf{I}_{[\theta,\infty)}(\min_i(x_i/i)) = \mathsf{I}_{[\theta,\infty)}(h(x_{1:n}))$.

Now we will show minimal sufficiency (required for next item). Consider the likelihood ratio

$$\Lambda_{x_{1:n}}(\theta_1,\theta_2) = \exp\bigl\{(\theta_1 - \theta_2)n(n+1)/2\bigr\}\mathsf{I}_{[\theta_1,\infty)}(h(x_{1:n}))/\mathsf{I}_{[\theta_2,\infty)}(h(x_{1:n})).$$

If $\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \Lambda_{\tilde{x}_{1:n}}(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in \mathbb{R}$ for two realisations $x_{1:n}$ and $\tilde{x}_{1:n}$, then necessarily

$$\frac{\mathsf{I}_{[\theta_1,\infty)}(h(x_{1:n}))}{\mathsf{I}_{[\theta_2,\infty)}(h(x_{1:n}))} = \frac{\mathsf{I}_{[\theta_1,\infty)}(h(\tilde{x}_{1:n}))}{\mathsf{I}_{[\theta_2,\infty)}(h(\tilde{x}_{1:n}))}. \tag{2.3}$$

Now assume that $h(x_{1:n}) \neq h(\tilde{x}_{1:n})$, and, without loss of generality, that $h(x_{1:n}) > h(\tilde{x}_{1:n})$. Then for $\theta_1 = \{h(x_{1:n}) + h(\tilde{x}_{1:n})\}/2$ and $\theta_2 = h(x_{1:n})$, we obtain 1 on the left-hand side of (2.3) an 0 on the right-hand side. Hence, $h(x_{1:n}) = h(\tilde{x}_{1:n})$ must be satisfied for the equality to hold for all $\theta_1, \theta_2 \in \mathbb{R}$, and so the statistic $T = h(X_{1:n})$ is minimal sufficient.

b) Let $X_{1:n}$ denote a random sample from a distribution with density

$$f(x;\theta) = \exp\{-(x-\theta)\}, \quad \theta < x < \infty, \quad -\infty < \theta < \infty.$$

Derive a minimal sufficient statistic for θ.

▶ *We have a random sample from the distribution of X_1 in (7a), hence we proceed in a similar way. First we rewrite the above density as*

$$f(x;\theta) = \exp(\theta - x)\mathsf{I}_{[\theta,\infty)}(x)$$

and second we write the joint density as

$$f(x_{1:n};\theta) = \exp(n\theta - n\bar{x})\mathsf{I}_{[\theta,\infty)}(\min_i(x_i)).$$

By the Factorisation Theorem (Result 2.2), the statistic $T = \min_i(X_i)$ is sufficient for θ. Its minimal sufficiency can be proved in the same way as in (7a).

8. Let $T = h(X_{1:n})$ be a sufficient statistic for θ, $g(\cdot)$ a one-to-one function and $\tilde{T} = \tilde{h}(X_{1:n}) = g\{h(X_{1:n})\}$. Show that $\tilde{T}$ is sufficient for θ.

▶ *By the Factorisation Theorem (Result 2.2), the sufficiency of $T = h(X_{1:n})$ for θ implies the existence of functions g_1 and g_2 such that*

$$f(x_{1:n};\theta) = g_1\{h(x_{1:n});\theta\} \cdot g_2(x_{1:n}).$$

If we set $\tilde{g}_1 := g_1 \circ g^{-1}$, we can write

$$f(x_{1:n};\theta) = g_1(g^{-1}[g\{h(x_{1:n})\}];\theta) \cdot g_2(x_{1:n}) = \tilde{g}_1\{\tilde{h}(x_{1:n});\theta\} \cdot g_2(x_{1:n}),$$

which shows the sufficiency of $\tilde{T} = \tilde{h}(X_{1:n})$ for θ.

9. Let X_1 and X_2 denote two independent exponentially $\operatorname{Exp}(\lambda)$ distributed random variables with parameter $\lambda > 0$. Show that $h(X_1, X_2) = X_1 + X_2$ is sufficient for λ.

▶ *The likelihood $L(\lambda) = f(x_{1:2}, \lambda)$ is*

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{2} \lambda \exp(-\lambda x_i) \\ &= \underbrace{\lambda^2 \exp\{-\lambda(x_1 + x_2)\}}_{g_1\{h(x_{1:2}) = x_1 + x_2\,;\lambda\}} \cdot \underbrace{1}_{g_2(x_{1:n})}, \end{aligned}$$

and the result follows from the Factorisation Theorem (Result 2.2).

3 Elements of frequentist inference

1. Sketch why the MLE

$$\hat{N}_{\text{ML}} = \left\lfloor \frac{M \cdot n}{x} \right\rfloor$$

in the capture-recapture experiment (*cf*. Example 2.2) cannot be unbiased. Show that the alternative estimator

$$\hat{N} = \frac{(M+1)\cdot(n+1)}{(x+1)} - 1$$

is unbiased if $N \leq M + n$.

▶ *If $N \geq n + M$, then X can equal zero with positive probability. Hence, the MLE*

$$\hat{N}_{\text{ML}} = \left\lfloor \frac{M \cdot n}{X} \right\rfloor$$

can be infinite with positive probability. It follows that the expectation of the MLE is infinite if $N \geq M + n$ and so cannot be equal to the true parameter value N. We have thus shown that for some parameter values, the expectation of the estimator is not equal to the true parameter value. Hence, the MLE is not unbiased.

To show that the alternative estimator is unbiased if $N \leq M + n$, we need to compute its expectation. If $N \leq M + n$, the smallest value in the range $\mathcal{T}$ of the possible values for X is $\max\{0, n - (N - M)\} = n - (N - M)$. The expectation of the statistic $g(X) = (M+1)(n+1)/(X+1)$ can thus be computed as

$$\begin{aligned} \mathsf{E}\{g(X)\} &= \sum_{x \in \mathcal{T}} g(x) \Pr(X = x) \\ &= \sum_{x=n-(N-M)}^{\min\{n,M\}} \frac{(M+1)(n+1)}{x+1} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \\ &= (N+1) \sum_{x=n-(N-M)}^{\min\{n,M\}} \frac{\binom{M+1}{x+1}\binom{(N+1)-(M+1)}{n-x}}{\binom{N+1}{n+1}}. \end{aligned}$$

We may now shift the index in the sum, so that the summands containing x in the expression above contain $x-1$. Of course, we need to change the range of summation accordingly. By doing so, we obtain that

$$
\begin{aligned}
&\sum_{x=n-(N-M)}^{\min\{n,M\}} \frac{\binom{M+1}{x+1}\binom{(N+1)-(M+1)}{n-x}}{\binom{N+1}{n+1}} \\
&= \sum_{x=(n+1)-((N+1)-(M+1))}^{\min\{n+1,M+1\}} .
\end{aligned}
$$

Note that the sum above is a sum of probabilities corresponding to a hypergeometric distribution with different parameters, namely $\mathrm{HypGeom}(n+1, N+1, M+1)$*, i. e.*

$$
\begin{aligned}
&= \sum_{x=(n+1)-((N+1)-(M+1))}^{\min\{n+1,M+1\}} \frac{\binom{M+1}{x}\binom{(N+1)-(M+1)}{(n+1)-x}}{\binom{N+1}{n+1}} \\
&= \sum_{x\in\mathcal{T}^*} \mathsf{Pr}(X^* = x) \\
&= 1,
\end{aligned}
$$

where X^* is a random variable, $X^* \sim \mathrm{HypGeom}(n+1, N+1, M+1)$. It follows that

$$\mathsf{E}(\hat{N}) = \mathsf{E}\{g(X)\} - 1 = N + 1 - 1 = N.$$

Note however that the alternative estimator is also not unbiased for $N > M + n$. Moreover, its values are not necessarily integer and thus not necessarily in the parameter space. The latter property can be remedied by rounding. This, however, would lead to the loss of unbiasedness even for $N \leq M + n$.

2. Let $X_{1:n}$ be a random sample from a distribution with mean μ and variance $\sigma^2 > 0$. Show that

$$\mathsf{E}(\bar{X}) = \mu \quad \text{and} \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

▶ *By linearity of the expectation, we have that*

$$\mathsf{E}(\bar{X}) = n^{-1}\sum_{i=1}^{n} \mathsf{E}(X_i) = n^{-1} n \cdot \mu = \mu.$$

Sample mean $\bar{X}$ is thus unbiased for expectation μ.

The variance of a sum of uncorrelated random variables is the sum of the respective variances; hence,

$$\mathrm{Var}(\bar{X}) = n^{-2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = n^{-2} n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

3. Let $X_{1:n}$ be a random sample from a normal distribution with mean μ and variance $\sigma^2 > 0$. Show that the estimator

$$\hat{\sigma} = \sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} S$$

is unbiased for σ, where S is the square root of the sample variance S^2 in (3.1).

▶ *It is well known that for $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$,*

$$Y := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

see e. g. Davison (2003, page 75). For the expectation of the statistic $g(Y) = \sqrt{Y}$ we thus obtain that

$$
\begin{aligned}
\mathsf{E}\{g(Y)\} &= \int_0^\infty g(y) f_Y(y)\, dy \\
&= \int_0^\infty \frac{(\frac{1}{2})^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})} y^{\frac{n-1}{2}-1} \exp(-y/2) y^{\frac{1}{2}}\, dy \\
&= \left(\frac{1}{2}\right)^{-\frac{1}{2}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \int_0^\infty \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-y/2)\, dy.
\end{aligned}
$$

The integral on the most-right-hand side is the integral of the density of the $\chi^2(n)$ distribution over its support and therefore equals one. It follows that

$$\mathsf{E}(\sqrt{Y}) = \sqrt{2}\Gamma\left(\frac{n}{2}\right) / \Gamma\left(\frac{n-1}{2}\right),$$

and

$$\mathsf{E}(\hat{\sigma}) = \mathsf{E}\left\{\sigma \frac{\sqrt{Y}}{\sqrt{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\right\} = \sigma.$$

4. Show that the sample variance S^2 can be written as

$$S^2 = \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} (X_i - X_j)^2.$$

Use this representation to show that

$$\mathrm{Var}(S^2) = \frac{1}{n}\left\{c_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right\},$$

where $c_4 = \mathsf{E}\left[\{X - \mathsf{E}(X)\}^4\right]$ is the fourth central moment of X.

▶ *We start with showing that the estimator S^2 can be rewritten as $T := \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} (X_i - X_j)^2$:*

$$
\begin{aligned}
(n-1)T &= \frac{1}{2n} \cdot \sum_{i,j=1}^{n} (X_i^2 - 2X_iX_j + X_j^2) = \\
&= \frac{1}{2n} \cdot \left(n\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j + n\sum_{j=1}^{n} X_j^2 \right) = \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = (n-1)S^2.
\end{aligned}
$$

It follows that we can compute the variance of S^2 from the pairwise correlations between the terms $(X_i - X_j)^2$, $i, j = 1, \ldots, n$ as

$$
\begin{aligned}
\operatorname{Var}(S^{2Auf:arithmetischesMittel}) &= \big\{2n(n-1)\big\}^{-2} \operatorname{Var} \sum_{i,j} (X_i - X_j)^2 = \\
&= \big\{2n(n-1)\big\}^{-2} \sum_{i,j,k,l} \operatorname{Cov}\big\{(X_i - X_j)^2, (X_k - X_l)^2\big\}.
\end{aligned}
\tag{3.1}
$$

Depending on the combination of indices, the covariances in the sum above take one of the three following values:

- $\operatorname{Cov}\big\{(X_i - X_j)^2, (X_k - X_l)^2\big\} = 0$ *if $i = j$ and/or $k = l$ (in this case either the first or the second term is identically zero) or if i, j, k, l are all different (in this case the result follows from the independence between the different X_i).*
- *For $i \neq j$, $\operatorname{Cov}\big\{(X_i - X_j)^2, (X_i - X_j)^2\big\} = 2\mu_4 + 2\sigma^4$. To show this, we proceed in two steps. We denote $\mu := \mathsf{E}(X_1)$, and, using the independence of X_i and X_j, we obtain that*

  $$
  \begin{aligned}
  \mathsf{E}\big\{(X_i - X_j)^2\big\} &= \mathsf{E}\big\{(X_i - \mu)^2\big\} + \mathsf{E}\big\{(X_j - \mu)^2\big\} - 2\,\mathsf{E}\big\{(X_i - \mu)(X_j - \mu)\big\} = \\
  &= 2\sigma^2 - 2\big\{\mathsf{E}(X_i) - \mu\big\}\big\{E(X_j) - \mu\big\} = 2\sigma^2.
  \end{aligned}
  $$

  In an analogous way, we can show that

  $$
  \begin{aligned}
  \mathsf{E}\big((X_i - X_j)^4\big) &= \mathsf{E}\big((X_i - \mu + \mu - X_j)^4\big) = \\
  &= \mathsf{E}\big((X_i - \mu)^4\big) - 4\,\mathsf{E}\big((X_i - \mu)^3(X_j - \mu)\big) + 6\,\mathsf{E}\big((X_i - \mu)^2(X_j - \mu)^2\big) \\
  &\quad - 4\,\mathsf{E}\big((X_i - \mu)(X_j - \mu)^3\big) + \mathsf{E}\big((X_j - \mu)^4\big) = \\
  &= \mu_4 - 4 \cdot 0 + 6 \cdot (\sigma^2)^2 - 4 \cdot 0 + \mu_4 = 2\mu_4 + 6\sigma^4.
  \end{aligned}
  $$

  It follows that

  $$
  \begin{aligned}
  \operatorname{Cov}\big((X_i - X_j)^2, (X_i - X_j)^2\big) &= \operatorname{Var}\big((X_i - X_j)^2\big) = \\
  &= \mathsf{E}\big((X_i - X_j)^4\big) - \big\{\mathsf{E}\big((X_i - X_j)^2\big)\big\}^2 = \\
  &= 2\mu_4 + 6\sigma^4 - (2\sigma^2)^2 = 2\mu_4 + 2\sigma^4.
  \end{aligned}
  $$

  Note that since $(X_i - X_j)^2 = (X_j - X_i)^2$, there are $2 \cdot n(n-1)$ such terms in the sum (3.1).
- *In an analogous way, we may show that if i, j, k are all different, $\operatorname{Cov}\big\{(X_i - X_j)^2, (X_k - X_j)^2\big\} = \mu_4 - \sigma^4$. We can form $n(n-1)(n-2)$ different triplets (i, j, k) of i, j, k that are all different elements of $\{1, \ldots, n\}$. For each of these triplets, there are four different terms in the sum (3.1): $\operatorname{Cov}\big\{(X_i - X_j)^2, (X_k - X_j)^2\big\}$ $\operatorname{Cov}\big\{(X_i - X_j)^2, (X_j - X_k)^2\big\}$, $\operatorname{Cov}\big\{(X_j - X_i)^2, (X_j - X_k)^2\big\}$, and $\operatorname{Cov}\big\{(X_j - X_i)^2, (X_k - X_j)^2\big\}$, each with the same value. In total, we thus have $4 \cdot n(n-1)(n-2)$ terms in (3.1) with the value of $\mu_4 - \sigma^4$.*

By combining these intermediate computations, we finally obtain that

$$
\begin{aligned}
\operatorname{Var}(S^2) &= \frac{1}{\big\{2n(n-1)\big\}^2} \big\{2n(n-1)(2\mu_4 + 2\sigma^4) + 4n(n-1)(n-2)(\mu_4 - \sigma^4)\big\} = \\
&= \frac{1}{n(n-1)} \big\{\mu_4 + \sigma^4 + (n-2)(\mu_4 - \sigma^4)\big\} = \\
&= \frac{1}{n} \left\{ \mu_4 - \Big(\frac{n-3}{n-1}\Big)\sigma^4 \right\}.
\end{aligned}
$$

5. Show that the confidence interval defined in Example 3.6 indeed has coverage probability 50% for all values $\theta \in \Theta$.

▶ *To prove the statement, we need to show that $\mathsf{Pr}\{\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)\} = 0.5$ for all $\theta \in \Theta$. This follows by the simple calculation:*

$$
\begin{aligned}
\mathsf{Pr}\{\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)\} &= \mathsf{Pr}(X_1 \leq \theta \leq X_2) + \mathsf{Pr}(X_2 \leq \theta \leq X_1) \\
&= \mathsf{Pr}(X_1 \leq \theta)\,\mathsf{Pr}(X_2 \geq \theta) + \mathsf{Pr}(X_2 \leq \theta)\,\mathsf{Pr}(X_1 \geq \theta) \\
&= 0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5.
\end{aligned}
$$

6. Consider a random sample $X_{1:n}$ from the uniform model $\mathrm{U}(0, \theta)$, *cf*. Example 2.18. Let $Y = \max(X_1, \ldots, X_n)$ denote the maximum of the random sample $X_{1:n}$. Show that the confidence interval for θ with limits

$$
Y \quad \text{and} \quad (1 - \gamma)^{-1/n} Y
$$

has coverage γ.

▶ *Recall that the density function of the uniform distribution* $\mathrm{U}(0,\theta)$ *is* $f(x) = \frac{1}{\theta}\mathsf{I}_{[0,\theta)}(x)$*. The corresponding distribution function is*

$$\begin{aligned} F_x(x) &= \int_{-\infty}^{x} f(u)\,du \\ &= \begin{cases} 0 & \textit{for } x \le 0,\\ \int_0^x \frac{1}{\theta}\,du = \frac{x}{\theta} & \textit{for } 0 \le x \le \theta,\\ 1 & \textit{for } x \ge \theta. \end{cases} \end{aligned}$$

To prove the coverage of the confidence interval with limits Y *and* $(1-\gamma)^{-1/n}Y$*, we need to show that* $\mathsf{Pr}\{Y \le \theta \le (1-\gamma)^{-1/n}Y\} = \gamma$ *for all* $\theta \in \Theta$*. We first derive the distribution of the random variable* Y*. For its distribution function* F_Y*, we obtain that*

$$\begin{aligned} F_Y(y) &= \mathsf{Pr}(Y \le y) \\ &= \mathsf{Pr}\{\max(X_1,\dots,X_n) \le y\} \\ &= \mathsf{Pr}(X_1 \le y,\dots,X_n \le y) \\ &= \{F(y)\}^n. \end{aligned}$$

For its density $f_Y(y)$*, it follows that*

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= n\,F(y)^{n-1}f(y) \\ &= \begin{cases} \frac{n}{\theta^n}\,y^{n-1} & \textit{for } 0 \le y \le \theta,\\ 0 & \textit{otherwise.} \end{cases} \end{aligned}$$

The coverage of the confidence interval can now be calculated as

$$\begin{aligned} \mathsf{Pr}\{Y \le \theta \le (1-\gamma)^{-1/n}Y\} &= \mathsf{Pr}\{\theta\,(1-\gamma)^{1/n} \le Y \le \theta\} \\ &= \int_{\theta(1-\gamma)^{1/n}}^{\theta} f_Y(y)\,dy \\ &= \frac{n}{\theta^n}\int_{\theta(1-\gamma)^{1/n}}^{\theta} y^{n-1}\,dy \\ &= \frac{n}{\theta^n}\left[\frac{\theta^n}{n} - \frac{\{\theta(1-\gamma)^{1/n}\}^n}{n}\right] \\ &= \{1-(1-\gamma)\} = \gamma. \end{aligned}$$

7. Consider a population with mean μ and variance σ^2. Let $X_1,\dots,X_5$ be independent draws from this population. Consider the following estimators for μ:

$$\begin{aligned} T_1 &= \frac{1}{5}(X_1+X_2+X_3+X_4+X_5),\\ T_2 &= \frac{1}{3}(X_1+X_2+X_3),\\ T_3 &= \frac{1}{8}(X_1+X_2+X_3+X_4)+\frac{1}{2}X_5,\\ T_4 &= X_1+X_2\\ \text{and}\quad T_5 &= X_1. \end{aligned}$$

a) Which estimators are unbiased for μ?

▶ *The estimators* T_1*,* T_2*, and* T_5 *are sample means of sizes 5, 3, and 1, respectively, and as such are unbiased for* μ*, cf. Exercise 2. Further,* T_3 *is also unbiased, as*

$$\mathsf{E}(T_3) = \frac{1}{8}\cdot 4\mu + \frac{1}{2}\mu = \mu.$$

On the contrary, T_4 *is not unbiased, as*

$$\mathsf{E}(T_4) = \mathsf{E}(X_1) + \mathsf{E}(X_2) = 2\mu.$$

Note, however, that T_4 *is unbiased for* 2μ*.*

b) Compute the MSE of each estimator.

▶ *Since the biased of an unbiased estimator is zero, the* MSE *of an unbiased estimator is equal to its variance, cf. 3.5. For the sample means* T_1, T_2*, and* T_5*, we therefore directly have that*

$$MSE(T_1) = \frac{\sigma^2}{5},\quad MSE(T_2) = \frac{\sigma^2}{3}\quad \textit{und}\quad MSE(T_5) = \sigma^2,$$

cf. Exercise 2. For the MSE *of* T_3*, we have that*

$$MSE(T_3) = \mathrm{Var}(T_3) = \frac{1}{8^2}\cdot 4\sigma^2 + \frac{1}{2^2}\sigma^2 = \frac{5}{16}\sigma^2.$$

Finally, the MSE *of* T_4*, we have that*

$$MSE(T_4) = \{\mathsf{E}(T_4)\}^2 + \mathrm{Var}(T_4) = \mu^2 + 2\sigma^2.$$

8. The distribution of a multivariate random variable $\boldsymbol{X}$ belongs to an exponential family of order p, if the logarithm of its probability mass or density function can be written as

$$\log\{f(\boldsymbol{x};\boldsymbol{\tau})\} = \sum_{i=1}^{p}\eta_i(\boldsymbol{\tau})T_i(\boldsymbol{x}) - B(\boldsymbol{\tau}) + c(\boldsymbol{x}). \tag{3.2}$$

Here $\boldsymbol{\tau}$ is the p-dimensional parameter vector and T_i, η_i, B and c are real-valued functions. It is assumed that the set $\{1, \eta_1(\boldsymbol{\tau}), \dots, \eta_p(\boldsymbol{\tau})\}$ is linearly independent. Then we define the canonical parameters $\theta_1 = \eta_1(\tau_1), \dots, \theta_p = \eta_p(\tau_p)$. With $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ and $T(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \dots, T_p(\boldsymbol{x}))^\top$ we can write the log density in canonical form:

$$\log\{f(\boldsymbol{x};\boldsymbol{\theta})\} = \boldsymbol{\theta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\theta}) + c(\boldsymbol{x}). \tag{3.3}$$

Exponential families are interesting because most of the commonly used distributions, such as the Poisson, geometric, binomial, normal and gamma distribution, are exponential families. Therefore it is worthwhile to derive general results for exponential families, which can then be applied to many distributions at once. For example, two very useful results for the exponential family of order one in canonical form are $\mathsf{E}\{T(X)\} = dA/d\theta(\theta)$ and $\operatorname{Var}\{T(X)\} = d^2A/d\theta^2(\theta)$.

a) Show that $T(\boldsymbol{X})$ is minimal sufficient for $\boldsymbol{\theta}$.

▶ *Consider two realisations $\boldsymbol{x}$ and $\boldsymbol{y}$ with corresponding likelihood ratios $\Lambda_{\boldsymbol{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\Lambda_{\boldsymbol{y}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ being equal, which on the log scale gives the equation*

$$\log f(\boldsymbol{x};\boldsymbol{\theta}_1) - \log f(\boldsymbol{x};\boldsymbol{\theta}_2) = \log f(\boldsymbol{y};\boldsymbol{\theta}_1) - \log f(\boldsymbol{y};\boldsymbol{\theta}_2).$$

Plugging in (3.3) *we can simplify it to*

$$\begin{aligned}\boldsymbol{\theta}_1^\top T(\boldsymbol{x}) - A(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_2^\top T(\boldsymbol{x}) + A(\boldsymbol{\theta}_2) &= \boldsymbol{\theta}_1^\top T(\boldsymbol{y}) - A(\boldsymbol{\theta}_1) - \boldsymbol{\theta}_2^\top T(\boldsymbol{y}) + A(\boldsymbol{\theta}_2)\\ (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top T(\boldsymbol{x}) &= (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top T(\boldsymbol{y}).\end{aligned}$$

If $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then this equation holds for all possible values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Therefore $T(\boldsymbol{x})$ is sufficient for $\boldsymbol{\theta}$. On the other hand, if this equation holds for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, then $T(\boldsymbol{x})$ must equal $T(\boldsymbol{y})$. Therefore $T(\boldsymbol{x})$ is also minimal sufficient for $\boldsymbol{\theta}$.

b) Show that the density of the Poisson distribution $\operatorname{Po}(\lambda)$ can be written in the forms (3.2) and (3.3), respectively. Thus derive the expectation and variance of $X \sim \operatorname{Po}(\lambda)$.

▶ *For the density of a random variable $X \sim \operatorname{Po}(\lambda)$, we have that*

$$\log f(x;\lambda) = \log\left\{\frac{\lambda^x}{x!}\exp(-\lambda)\right\} = \log(\lambda)x - \lambda - \log(x!),$$

so $p = 1$, $\theta = \eta(\lambda) = \log(\lambda)$, $T(x) = x$, $B(\lambda) = \lambda$ and $c(x) = -\log(x!)$. For the canonical representation, we have $A(\theta) = B\{\eta^{-1}(\theta)\} = B\{\exp(\theta)\} = \exp(\theta)$. Hence, both the expectation $\mathsf{E}\{T(X)\} = dA/d\theta(\theta)$ and the variance $\operatorname{Var}\{T(X)\} = d^2A/d\theta^2(\theta)$ of X are $\exp(\theta) = \lambda$.

c) Show that the density of the normal distribution $\mathrm{N}(\mu, \sigma^2)$ can be written in the forms (3.2) and (3.3), respectively, where $\boldsymbol{\tau} = (\mu, \sigma^2)^\top$. Hence derive a minimal sufficient statistic for $\boldsymbol{\tau}$.

▶ *For $X \sim \mathrm{N}(\mu, \sigma^2)$, we have $\boldsymbol{\tau} = (\mu, \sigma^2)^\top$. We can rewrite the log density as*

$$\begin{aligned}\log f(x;\mu,\sigma^2) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{x^2 - 2x\mu + \mu^2}{\sigma^2}\\ &= -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log(2\pi)\\ &= \eta_1(\boldsymbol{\tau})T_1(x) + \eta_2(\boldsymbol{\tau})T_2(x) - B(\boldsymbol{\tau}) + c(x),\end{aligned}$$

where

$$\begin{aligned}\theta_1 = \eta_1(\boldsymbol{\tau}) &= -\frac{1}{2\sigma^2} & T_1(x) &= x^2\\ \theta_2 = \eta_2(\boldsymbol{\tau}) &= \frac{\mu}{\sigma^2} & T_2(x) &= x\\ B(\boldsymbol{\tau}) &= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2) & &\\ \textit{and}\quad c(x) &= -\frac{1}{2}\log(2\pi). & &\end{aligned}$$

We can invert the canonical parametrisation $\boldsymbol{\theta} = \eta(\boldsymbol{\tau}) = (\eta_1(\boldsymbol{\tau}), \eta_2(\boldsymbol{\tau}))^\top$ by

$$\begin{aligned}\sigma^2 &= -2\theta_1,\\ \mu &= \theta_2\sigma^2 = -2\theta_1\theta_2,\end{aligned}$$

so for the canonical form we have the function

$$\begin{aligned}A(\boldsymbol{\theta}) &= B\{\eta^{-1}(\boldsymbol{\theta})\}\\ &= \frac{(-2\theta_1\theta_2)^2}{2(-2\theta_1)} + \frac{1}{2}\log(-2\theta_1)\\ &= -\frac{\theta_1^2\theta_2^2}{\theta_1} + \frac{1}{2}\log(-2\theta_1)\\ &= -\theta_1\theta_2^2 + \frac{1}{2}\log(-2\theta_1).\end{aligned}$$

Finally, from above, we know that $T(x) = (x^2, x)^\top$ is minimal sufficient for $\boldsymbol{\tau}$.

d) Show that for an exponential family of order one, $I(\hat{\tau}_{\mathrm{ML}}) = J(\hat{\tau}_{\mathrm{ML}})$. Verify this result for the Poisson distribution.

▶ *Let X be a random variable with density from the exponential family of order one. By taking the derivative of the log likelihood, we obtain the score function*

$$S(\tau) = \frac{d\eta(\tau)}{d\tau}T(x) - \frac{dB(\tau)}{d\tau},$$

so that the MLE $\hat{\tau}_{\text{ML}}$ satisfies the equation

$$T(x) = \frac{\frac{dB(\hat{\tau}_{\text{ML}})}{d\tau}}{\frac{d\eta(\hat{\tau}_{\text{ML}})}{d\tau}}.$$

We obtain the observed Fisher information from the Fisher information

$$I(\tau) = \frac{d^2B(\tau)}{d\tau^2} - \frac{d^2\eta(\tau)}{d\tau^2}T(x)$$

by plugging in the MLE:

$$I(\hat{\tau}_{\text{ML}}) = \frac{d^2A(\hat{\tau}_{\text{ML}})}{d\tau^2} - \frac{d^2\eta(\hat{\tau}_{\text{ML}})}{d\tau^2}\frac{\frac{dB(\hat{\tau}_{\text{ML}})}{d\tau}}{\frac{d\eta(\hat{\tau}_{\text{ML}})}{d\tau}}.$$

Further, we have that

$$\mathsf{E}\{T(X)\} = \frac{d}{d\theta}(B \circ \eta^{-1})(\theta) = \frac{dB(\eta^{-1}(\theta))}{d\tau} \cdot \frac{d\eta^{-1}(\theta)}{d\theta} = \frac{\frac{dB(\tau)}{d\tau}}{\frac{d\eta(\tau)}{d\tau}},$$

where $\theta = \eta(\tau)$ is the canonical parameter. Hence

$$J(\tau) = \frac{d^2B(\tau)}{d\tau^2} - \frac{d^2\eta(\tau)}{d\tau^2}\frac{\frac{dB(\tau)}{d\tau}}{\frac{d\eta(\tau)}{d\tau}}$$

follows. If we now plug in $\hat{\tau}_{\text{ML}}$, we obtain the same formula as for $I(\hat{\tau}_{\text{ML}})$.
For the Poisson example, we have $I(\lambda) = x/\lambda^2$ and $J(\lambda) = 1/\lambda$. Plugging in the MLE $\hat{\lambda}_{\text{ML}} = x$ leads to $I(\hat{\lambda}_{\text{ML}}) = J(\hat{\lambda}_{\text{ML}}) = 1/x$.

e) Show that for an exponential family of order one in canonical form, $I(\theta) = J(\theta)$. Verify this result for the Poisson distribution.

▶ *In the canonical parametrisation* (3.3),

$$S(\theta) = T(x) - \frac{dA(\theta)}{d\theta}$$
$$\text{and} \quad I(\theta) = \frac{d^2A(\theta)}{d\theta^2},$$

where the latter is independent of the observation x, and therefore obviously $I(\theta) = J(\theta)$.
For the Poisson example, the canonical parameter is $\theta = \log(\lambda)$. Since $A(\theta) = \exp(\theta)$, also the second derivative equals $\exp(\theta) = I(\theta) = J(\theta)$.

f) Suppose $X_{1:n}$ is a random sample from a one-parameter exponential family with canonical parameter θ. Derive an expression for the log-likelihood $l(\theta)$.

▶ *Using the canonical parametrisation of the density, we can write the log-likelihood of a single observation as*

$$\log\{f(x;\theta)\} = \theta\, T(x) - A(\theta) + c(x).$$

The log-likelihood $l(\theta)$ of the random sample $X_{1:n}$ is thus

$$l(\theta) = \sum_{i=1}^{n} \log\{f(x_i;\theta)\} = \sum_{i=1}^{n}\{\theta\, T(x_i) - A(\theta) + c(x_i)\} \propto \theta \sum_{i=1}^{n} T(x_i) - n\, A(\theta).$$

9. Assume that survival times $X_{1:n}$ form a random sample from a gamma distribution $\mathrm{G}(\alpha, \alpha/\mu)$ with mean $\mathsf{E}(X_i) = \mu$ and shape parameter α.

a) Show that $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is a consistent estimator of the mean survival time μ.

▶ *The sample mean $\bar{X}$ is unbiased for μ and has variance $\operatorname{Var}(\bar{X}) = \operatorname{Var}(X_i)/n = \mu^2/(n\alpha)$, cf. Exercise 2 and Appendix A.5.2. It follows that its mean squared error MSE $= \mu^2/(n\alpha)$ goes to zero as $n \to \infty$. Thus, the estimator is consistent in mean square and hence also consistent.*
Note that this holds for all random samples where the individual random variables have finite expectation and variance.

b) Show that $X_i/\mu \sim \mathrm{G}(\alpha, \alpha)$.

▶ *From Appendix A.5.2, we know that by multiplying a random variable with $\mathrm{G}(\alpha, \alpha/\mu)$ distribution by μ^{-1}, we obtain a random variable with $\mathrm{G}(\alpha, \alpha)$ distribution.*

c) Define the approximate pivot from Result 3.1,

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Using the result from above, show that the distribution of Z does not depend on μ.

▶ *We can rewrite Z as follows:*

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \\ &= \frac{\bar{X}/\mu - 1}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i/\mu - \bar{X}/\mu)^2}} \\ &= \frac{\bar{Y} - 1}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \end{aligned}$$

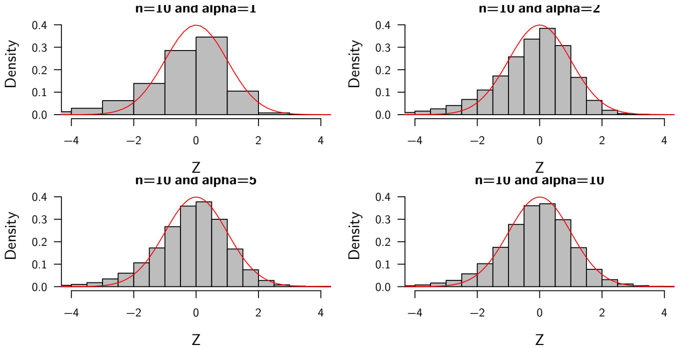
where $Y_i = X_i/\mu$ and $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i = \bar{X}/\mu$. From above, we know that $Y_i \sim \mathrm{G}(\alpha, \alpha)$, so its distribution depends only on α and not on μ. Therefore, Z is a function of random variables whose distributions do not depend on μ. It follows that the distribution of Z does not depend on μ either.

d) For $n = 10$ and $\alpha \in \{1, 2, 5, 10\}$, simulate 100 000 samples from Z, and compare the resulting 2.5% and 97.5% quantiles with those from the asymptotic standard normal distribution. Is Z a good approximate pivot?

▶

```
> ## simulate one realisation
> z.sim <- function(n, alpha)
  {
      y <- rgamma(n=n, alpha, alpha)

      yq <- mean(y)
      sy <- sd(y)

      z <- (yq - 1) / (sy / sqrt(n))
      return(z)
  }
> ## fix cases:
> n <- 10
> alphas <- c(1, 2, 5, 10)
> ## space for quantile results
> quants <- matrix(nrow=length(alphas),
                   ncol=2)
> ## set up graphics space
> par(mfrow=c(2, 2))
> ## treat every case
> for(i in seq_along(alphas))
  {
      ## draw 100000 samples
      Z <- replicate(n=100000, expr=z.sim(n=n, alpha=alphas[i]))

      ## plot histogram
      hist(Z,
           prob=TRUE,
           col="gray",
           main=paste("n=", n, " and alpha=", alphas[i], sep=""),
           nclass=50,
           xlim=c(-4, 4),
           ylim=c(0, 0.45))

      ## compare with N(0, 1) density
      curve(dnorm(x),
            from=min(Z),
            to=max(Z),
            n=201,
            add=TRUE,
            col="red")

      ## save empirical quantiles
      quants[i, ] <- quantile(Z, prob=c(0.025, 0.975))
  }
> ## so the quantiles were:
> quants
          [,1]     [,2]
[1,] -4.095855 1.623285
[2,] -3.326579 1.741014
```

```
[3,] -2.841559 1.896299
[4,] -2.657800 2.000258
> ## compare with standard normal ones:
> qnorm(p=c(0.025, 0.975))
[1] -1.959964  1.959964
```



We see that the distribution of Z is skewed to the left compared to the standard normal distribution: the 2.5% quantiles are clearly lower than -1.96, and also the 97.5% quantiles are slightly lower than 1.96. For increasing α (and also for increasing n of course), the normal approximation becomes better. Altogether, the normal approximation does not appear too bad, given the fact that $n = 10$ is a rather small sample size.

e) Show that $\bar{X}/\mu \sim \mathrm{G}(n\alpha, n\alpha)$. If α was known, how could you use this quantity to derive a confidence interval for μ?

▶ *We know from above that the summands X_i/μ in $\bar{X}/\mu$ are independent and have $\mathrm{G}(\alpha, \alpha)$ distribution. From Appendix A.5.2, we obtain that $\sum_{i=1}^{n} X_i/\mu \sim \mathrm{G}(n\alpha, \alpha)$. From the same appendix, we also have that by multiplying the sum by n^{-1} we obtain $\mathrm{G}(n\alpha, n\alpha)$ distribution.*

If α was known, then $\bar{X}/\mu$ would be a pivot for μ and we could derive a 95% confidence interval as follows:

$$\begin{aligned}0.95 &= \Pr\{q_{0.025}(n\alpha) \leq \bar{X}/\mu \leq q_{0.975}(n\alpha)\}\\ &= \Pr\{1/q_{0.975}(n\alpha) \leq \mu/\bar{X} \leq 1/q_{0.025}(n\alpha)\}\\ &= \Pr\{\bar{X}/q_{0.975}(n\alpha) \leq \mu \leq \bar{X}/q_{0.025}(n\alpha)\}\end{aligned}$$

where $q_\gamma(\beta)$ denotes the γ quantile of $\mathrm{G}(\beta, \beta)$. So the confidence interval would be

$$\left[\bar{X}/q_{0.975}(n\alpha), \bar{X}/q_{0.025}(n\alpha)\right]. \tag{3.4}$$

f) Suppose α is unknown, how could you derive a confidence interval for μ?

▶ *If α is unknown, we could estimate it and then use the confidence interval* *from (3.4). Of course we could also use $Z \overset{a}{\sim} \mathrm{N}(0,1)$ and derive the standard Wald interval*

$$[\bar{X} \pm 1.96 \cdot S/\sqrt{n}] \tag{3.5}$$

from that. A third possibility would be to simulate from the exact distribution of Z as we have done above, using the estimated α value, and use the empirical quantiles from the simulation instead of the ± 1.96 values from the standard normal distribution to construct a confidence interval analogous to (3.5).

10. All beds in a hospital are numbered consecutively from 1 to $N > 1$. In one room a doctor sees $n \leq N$ beds, which are a random subset of all beds, with (ordered) numbers $X_1 < \cdots < X_n$. The doctor now wants to estimate the total number of beds N in the hospital.

a) Show that the joint probability mass function of $\mathbf{X} = (X_1, \dots, X_n)$ is

$$f(\mathbf{x}; N) = \binom{N}{n}^{-1} \mathsf{I}_{\{n,\dots,N\}}(x_n).$$

▶ *There are $\binom{N}{n}$ possibilities to draw n values without replacement out of N values. Hence, the probability of one outcome $\mathbf{x} = (x_1, \dots, x_n)$ is the inverse, $\binom{N}{n}^{-1}$. Due to the nature of the problem, the highest number x_n cannot be larger than N, nor can it be smaller than n. Altogether, we thus have*

$$\begin{aligned} f(\mathbf{x}; N) &= \mathsf{Pr}(X_1 = x_1, \dots, X_n = x_n; N) \\ &= \binom{N}{n}^{-1} \mathsf{I}_{\{n,\dots,N\}}(x_n). \end{aligned}$$

b) Show that X_n is minimal sufficient for N.

▶ *We can factorise the probability mass function as follows:*

$$\begin{aligned} f(\mathbf{x}; N) &= \left\{ \frac{N!}{(N-n)!n!} \right\}^{-1} \mathsf{I}_{\{n,\dots,N\}}(x_n) \\ &= \underbrace{n!}_{=g_2(\mathbf{x})} \underbrace{\frac{(N-n)!}{N!} \mathsf{I}_{\{n,\dots,N\}}(x_n)}_{=g_1\{h(\mathbf{x})=x_n; N\}}, \end{aligned}$$

so from the Factorization Theorem (Result 2.2), we have that X_n is sufficient for N. In order to show the minimal sufficiency, consider two data sets $\mathbf{x}$ and $\mathbf{y}$ such that for every two parameter values N_1 and N_2, the likelihood ratios are identical, i. e.

$$\Lambda_{\mathbf{x}}(N_1, N_2) = \Lambda_{\mathbf{y}}(N_1, N_2).$$

This can be rewritten as

$$\frac{\mathsf{I}_{\{n,\dots,N_1\}}(x_n)}{\mathsf{I}_{\{n,\dots,N_2\}}(x_n)} = \frac{\mathsf{I}_{\{n,\dots,N_1\}}(y_n)}{\mathsf{I}_{\{n,\dots,N_2\}}(y_n)}. \tag{3.6}$$

Now assume that $x_n \neq y_n$. Without loss of generality, let $x_n < y_n$. Then we can choose $N_1 = x_n$ and $N_2 = y_n$, and equation (3.6) gives us

$$\frac{1}{1} = \frac{0}{1},$$

which is not true. Hence, $x_n = y_n$ must be fulfilled. It follows that X_n is minimal sufficient for N.

c) Confirm that the probability mass function of X_n is

$$f_{X_n}(x_n; N) = \frac{\binom{x_n - 1}{n-1}}{\binom{N}{n}} \mathsf{I}_{\{n,\dots,N\}}(x_n).$$

▶ *For a fixed value $X_n = x_n$ of the maximum, there are $\binom{x_n-1}{n-1}$ possibilities how to choose the first $n-1$ values. Hence, the total number of possible draws giving a maximum of x is $\binom{N}{n}/\binom{x_n-1}{n-1}$. Considering also the possible range for x_n, this leads to the probability mass function*

$$f_{X_n}(x_n; N) = \frac{\binom{x_n - 1}{n-1}}{\binom{N}{n}} \mathsf{I}_{\{n,\dots,N\}}(x_n).$$

d) Show that

$$\hat{N} = \frac{n+1}{n} X_n - 1$$

is an unbiased estimator of N.

▶ *For the expectation of X_n, we have*

$$\begin{aligned} \mathsf{E}(X_n) &= \binom{N}{n}^{-1} \sum_{x=n}^{N} x \cdot \binom{x-1}{n-1} \\ &= \binom{N}{n}^{-1} \sum_{x=n}^{N} n \cdot \binom{x}{n} \\ &= \binom{N}{n}^{-1} n \cdot \sum_{x=n}^{N} \binom{x+1-1}{n+1-1} \\ &= \binom{N}{n}^{-1} n \cdot \sum_{x=n+1}^{N+1} \binom{x-1}{(n+1)-1}. \end{aligned}$$

Since $\sum_{x=n}^{N} \binom{x-1}{n-1} = \binom{N}{n}$, we have

$$\begin{aligned} \mathsf{E}(X_n) &= \binom{N}{n}^{-1} n \cdot \binom{N+1}{n+1} \\ &= \frac{n!(N-n)!}{N!} n \cdot \frac{(N+1)!}{(n+1)!(N-n)!} \\ &= \frac{n}{n+1}(N+1). \end{aligned}$$

Altogether thus

$$\begin{aligned} E(\hat{N}) &= \frac{n+1}{n} E(X_n) - 1 \\ &= \frac{n+1}{n} \frac{n}{n+1}(N+1) - 1 \\ &= N. \end{aligned}$$

So $\hat{N}$ is unbiased for N.

e) Study the ratio $L(N+1)/L(N)$ and derive the ML estimator of N. Compare it with $\hat{N}$.

▶ *The likelihood ratio of $N \geq x_n$ relative to $N+1$ with respect to* $\mathbf{x}$ *is*

$$\frac{f(\mathbf{x}; N+1)}{f(\mathbf{x}; N)} = \frac{\binom{N}{n}}{\binom{N+1}{n}} = \frac{N+1-n}{N+1} < 1,$$

so N must be as small as possible to maximise the likelihood, i. e. $\hat{N}_{\text{ML}} = X_n$. From above, we have $\mathsf{E}(X_n) = \frac{n}{n+1}(N+1)$, so the bias of the MLE is

$$\begin{aligned} \mathsf{E}(X_n) - N &= \frac{n}{n+1}(N+1) - N \\ &= \frac{n(N+1) - (n+1)N}{n+1} \\ &= \frac{nN + n - nN - N}{n+1} \\ &= \frac{n-N}{n+1} < 0. \end{aligned}$$

This means that $\hat{N}_{\text{ML}}$ systematically underestimates N, in contrast to $\hat{N}$.

4 Frequentist properties of the likelihood

1. Compute an approximate 95% confidence interval for the true correlation ρ based on the MLE $r = 0.7$, a sample of size of $n = 20$ and Fisher's z-transformation.

▶ *Using Example 4.16, we obtain the transformed correlation as*

$$z = \tanh^{-1}(0.7) = 0.5 \log\left(\frac{1+0.7}{1-0.7}\right) = 0.867.$$

Using the more accurate approximation $1/(n-3)$ for the variance of $\zeta = \tanh^{-1}(\rho)$, we obtain the standard error $1/\sqrt{n-3} = 0.243$. The 95%-Wald confidence interval for ζ is thus

$$[z \pm 1.96 \cdot \operatorname{se}(z)] = [0.392, 1.343].$$

By back-transforming using the inverse Fisher's z-transformation, we obtain the following confidence interval for ρ:

$$[\tanh(0.392), \tanh(1.343)] = [0.373, 0.872].$$

2. Derive a general formula for the score confidence interval in the Poisson model based on the Fisher information, *cf*. Example 4.9.

▶ *We consider a random sample $X_{1:n}$ from Poisson distribution* $\operatorname{Po}(e_i\lambda)$ *with known offsets $e_i > 0$ and unknown rate parameter λ. As in Example 4.8, we can see that if we base the score statistic for testing the null hypothesis that the true rate parameter equals λ on the Fisher information $I(\lambda; x_{1:n})$, we obtain*

$$T_2(\lambda; x_{1:n}) = \frac{S(\lambda; x_{1:n})}{\sqrt{I(\lambda; x_{1:n})}} = \sqrt{n} \cdot \frac{\bar{x} - \bar{e}\lambda}{\sqrt{\bar{x}}}.$$

We now determine the values of λ for which the score test based on the asymptotic distribution of T_2 would not reject the null hypothesis at level α. These are the values for which we have $|T_2(\lambda; x_{1:n})| \leq q := z_{1-\alpha/2}$:

$$\left|\sqrt{n} \cdot \frac{\bar{x} - \bar{e}\lambda}{\sqrt{\bar{x}}}\right| \leq q$$

$$\left|\frac{\bar{x}}{\bar{e}} - \lambda\right| \leq \frac{q}{\bar{e}}\sqrt{\frac{\bar{x}}{n}}$$

$$\lambda \in \left[\frac{\bar{x}}{\bar{e}} \pm \frac{q}{\bar{e}}\sqrt{\frac{\bar{x}}{n}}\right]$$

Note that this score confidence interval is symmetric around the MLE $\bar{x}/\bar{e}$, unlike the one based on the expected Fisher information derived in Example 4.9.

3. A study is conducted to quantify the evidence against the null hypothesis that less than 80 percent of the Swiss population have antibodies against the human herpesvirus. Among a total of 117 persons investigated, 105 had antibodies.

a) Formulate an appropriate statistical model and the null and alternative hypotheses. Which sort of P-value should be used to quantify the evidence against the null hypothesis?

▶ *The researchers are interested in the frequency of herpesvirus antibodies occurrence in the Swiss population, which is very large compared to the $n = 117$ probands. Therefore, the binomial model, actually assuming infinite population, is appropriate. Among the total of $n = 117$ draws, $x = 105$ "successes" were obtained, and the proportion π of these successes in the theoretically infinite population is of interest. We can therefore suppose that the observed value $x = 105$ is a realisation of a random variable $X \sim \text{Bin}(n, \pi)$.*

The null hypothesis is $H_0 : \pi < 0.8$, while the alternative hypothesis is $H_1 : \pi \geq 0.8$. Since this is a one-sided testing situation, we will need the corresponding one-sided P-value to quantify the evidence against the null hypothesis.

b) Use the Wald statistic (4.12) and its approximate normal distribution to obtain a P-value.

▶ *The Wald statistic is $z(\pi) = \sqrt{I(\hat{\pi}_{\text{ML}})}(\hat{\pi}_{\text{ML}} - \pi)$. As in Example 4.10, we have $\hat{\pi}_{\text{ML}} = x/n$. Further, $I(\pi) = x/\pi^2 + (n-x)/(1-\pi)^2$, and so*

$$I(\hat{\pi}_{\text{ML}}) = \frac{x}{(x/n)^2} + \frac{n-x}{(1-x/n)^2} = \frac{n^2}{x} + \frac{n^2(n-x)}{(n-x)^2} = \frac{n^2}{x} + \frac{n^2}{n-x} = \frac{n^3}{x(n-x)}.$$

We thus have

$$z(\pi) = \sqrt{\frac{n^3}{x(n-x)}}\left(\frac{x}{n} - \pi\right) = \sqrt{n}\,\frac{x - n\pi}{\sqrt{x(n-x)}}. \qquad (4.1)$$

To obtain an approximate one-sided P-value, we calculate its realisation $z(0.8)$ and compare it to the approximate normal distribution of the score statistic under the null hypothesis that the true proportion is 0.8. Since a more extreme result in the direction of the alternative H_1 corresponds to a larger realisation x and hence a larger observed value of the score statistic, the approximate one-sided P-value is the probability that a standard normal random variable is greater than $z(0.8)$:

```
> ## general settings
> x <- 105
> n <- 117
> pi0 <- 0.8
> ## the first approximate pivot
> z.pi <- function(x, n, pi)
  {
      sqrt(n) * (x - n * pi) / sqrt( (x * (n - x)) )
  }
> z1 <- z.pi(x, n, pi0)
> (p1 <- pnorm(z1, lower.tail=FALSE))
[1] 0.0002565128
```

$$\Pr\{Z(0.8) > z(0.8)\} \approx 1 - \Phi\{z(0.8)\} = 1 - \Phi(3.47) \approx 0.00026.$$

c) Use the logit-transformation (compare Example 4.22) and the corresponding Wald statistic to obtain a P-value.

▶ *We can equivalently formulate the testing problem as $H_0 : \phi < \phi_0 = \text{logit}(0.8)$ versus $H_1 : \phi > \phi_0$ after parametrising the binomial model with $\phi = \text{logit}(\pi)$ instead of π. Like in Example 4.22, we obtain the test statistic*

$$Z_\phi(\phi) = \frac{\log\{X/(n-X)\} - \phi}{\sqrt{1/X + 1/(n-X)}}, \qquad (4.2)$$

which, by the delta method, is asymptotically normally distributed. To compute the corresponding P-value, we may proceed as follows:

```
> ## the second approximate pivot
> z.phi <- function(x, n, phi)
  {
      (log(x / (n - x)) - phi) / sqrt(1/x + 1/(n-x))
  }
> (phi0 <- qlogis(pi0))
[1] 1.386294
> z2 <- z.phi(x, n, phi0)
> (p2 <- pnorm(z2, lower.tail=FALSE))
[1] 0.005103411
```

$$\Pr[Z_\phi\{\text{logit}(0.8)\} > z_\phi\{\text{logit}(0.8)\}] \approx 1 - \Phi\{z_\phi(1.3863)\} = 1 - \Phi(2.57) \approx 0.0051.$$

d) Use the score statistic (4.2) to obtain a P-value. Why do we not need to consider parameter transformations when using this statistic?

▶ *By Result 4.5, the score statistic* $V(\pi) = S(\pi; X_{1:n})/\sqrt{J_{1:n}(\pi)}$ *asymptotically follows the standard normal distribution under the Fisher regularity assumptions. In our case, we may use that a binomial random variable* $X \sim \text{Bin}(n, \pi)$ *can be viewed as the sum of* n *independent random variables with Bernoulli distribution* $\text{B}(\pi)$, *so the asymptotic results apply to the score statistic corresponding to* X *as* $n \to \infty$. *The score function corresponding to the binomial variable is* $S(\pi; X) = X/\pi - (n - X)/(1 - \pi)$ *and the expected Fisher information is* $J(\pi) = n/\{\pi(1 - \pi)\}$, *cf. Example 4.10. To calculate a third approximate* P*-value, we may therefore proceed as follows:*

```
> ## and the third
> v.pi <- function(x, n, pi)
  {
      (x/pi - (n - x)/(1 - pi)) / sqrt(n/pi/(1-pi))
  }
> v <- v.pi(x, n, pi0)
> (p3 <- pnorm(v, lower.tail=FALSE))
[1] 0.004209022
```

$$\Pr\{V(0.8) > v(0.8)\} \approx 1 - \Phi(2.63) \approx 0.00421.$$

We do not need to consider parameter transformations when using the score statistic, because it is invariant to one-to-one transformations. That is, the test statistic does not change if we choose a different parametrisation. This is easily seen from Result 4.3 and is written out in Section 4.1 in the context of the corresponding confidence intervals.

e) Use the exact null distribution from your model to obtain a P-value. What are the advantages and disadvantages of this procedure in general?

▶ *Of course we can also look at the binomial random variable* X *itself and consider it as a test statistic. Then the one-sided* P*-value is*

```
> ## now the "exact" p-value
> (p4 <- pbinom(x-1, size=n, prob=pi0, lower.tail=FALSE))
[1] 0.003645007
> ## note the x-1 to get
> ## P(X >= x) = P(X > x-1)
```

$$\Pr(X \geq x; \pi_0) = \sum_{w=x}^{n} \binom{n}{w} \pi_0^w (1 - \pi_0)^{n-w} \approx 0.00365.$$

The advantage of this procedure is that it does not require a large sample size n *for a good fit of the approximate distribution (normal distribution in the above cases) and a correspondingly good* P*-value. However, the computation of the* P*-value is difficult without a computer (as opposed to the easy use of standard normal tables for the other statistics). Also, only a finite number of* P*-values can be obtained, which corresponds to the discreteness of* X.

Note that the z*-statistic on the* ϕ*-scale and the score statistic produce* P*-values which are closer to the exact* P*-value than that from the* z*-statistic on the* π*-scale. This is due to the bad quadratic approximation of the likelihood on the* π*-scale.*

4. Suppose $X_{1:n}$ is a random sample from an $\text{Exp}(\lambda)$ distribution.

a) Derive the score function of λ and solve the score equation to get $\hat{\lambda}_{\text{ML}}$.

▶ *From the log-likelihood*

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} \log(\lambda) - \lambda x_i \\ &= n \log(\lambda) - n\lambda\bar{x} \end{aligned}$$

we get the score function

$$S(\lambda; x) = \frac{n}{\lambda} - n\bar{x},$$

which has the root

$$\hat{\lambda}_{\text{ML}} = 1/\bar{x}.$$

Since the Fisher information

$$\begin{aligned} I(\lambda) &= -\frac{d}{d\lambda} S(\lambda; x) \\ &= -\{(-1)n\lambda^{-2}\} \\ &= n/\lambda^2 \end{aligned}$$

is positive, we indeed have the MLE.

b) Calculate the observed Fisher information, the standard error of $\hat{\lambda}_{\text{ML}}$ and a 95% Wald confidence interval for λ.

▶ *By plugging the MLE* $\hat{\lambda}_{\text{ML}}$ *into the Fisher information, we get the observed Fisher information*

$$I(\hat{\lambda}_{\text{ML}}) = n\bar{x}^2$$

and hence the standard error of the MLE,

$$\text{se}(\hat{\lambda}_{\text{ML}}) = I(\hat{\lambda}_{\text{ML}})^{-1/2} = \frac{1}{\bar{x}\sqrt{n}}.$$

The 95% Wald confidence interval for λ *is thus given by*

$$\left[\hat{\lambda}_{\text{ML}} \pm z_{0.975}\, \text{se}(\hat{\lambda}_{\text{ML}})\right] = \left[\frac{1}{\bar{x}} \pm z_{0.975}/(\bar{x}\sqrt{n})\right].$$

c) Derive the expected Fisher information $J(\lambda)$ and the variance stabilizing transformation $\phi = h(\lambda)$ of λ.

▶ *Because the Fisher information does not depend on x in this case, we have simply*

$$J(\lambda) = \mathsf{E}\{I(\lambda; X)\} = n/\lambda^2.$$

Now we can derive the variance stabilising transformation:

$$\begin{aligned}\phi = h(\lambda) &\propto \int^{\lambda} J_\lambda(u)^{1/2}\, du \\ &\propto \int^{\lambda} u^{-1}\, du \\ &= \log(u)\big|_{u=\lambda} \\ &= \log(\lambda).\end{aligned}$$

d) Compute the MLE of ϕ and derive a 95% confidence interval for λ by back-transforming the limits of the 95% Wald confidence interval for ϕ. Compare with the result from 4b).

▶ *Due to the invariance of ML estimation with respect to one-to-one transformations we have*

$$\hat{\phi}_{\text{ML}} = \log \hat{\lambda}_{\text{ML}} = -\log \bar{x}$$

as the MLE of $\phi = \log(\lambda)$. Using the delta method we can get the corresponding standard error as

$$\begin{aligned}\operatorname{se}(\hat{\phi}_{\text{ML}}) &= \operatorname{se}(\hat{\lambda}_{\text{ML}}) \left| \frac{d}{d\lambda} h(\hat{\lambda}_{\text{ML}}) \right| \\ &= \frac{1}{\bar{x}\sqrt{n}} \left| 1/\hat{\lambda}_{\text{ML}} \right| \\ &= \frac{1}{\bar{x}\sqrt{n}} \bar{x} \\ &= n^{-1/2}.\end{aligned}$$

So the 95% Wald confidence interval for ϕ is

$$\left[-\log \bar{x} \pm z_{0.975} \cdot n^{-1/2}\right],$$

and transformed back to the λ-space we have the 95% confidence interval

$$\begin{aligned}&\left[\exp(-\log \bar{x} - z_{0.975} n^{-1/2}), \exp(-\log \bar{x} + z_{0.975} n^{-1/2})\right] = \\ &\qquad = \left[\bar{x}^{-1}/\exp(z_{0.975}/\sqrt{n}), \bar{x}^{-1} \cdot \exp(z_{0.975}/\sqrt{n})\right],\end{aligned}$$

which is not centred around the MLE $\hat{\lambda}_{\text{ML}} = \bar{x}^{-1}$, unlike the original Wald confidence interval for λ.

e) Derive the Cramér-Rao lower bound for the variance of unbiased estimators of λ.

▶ *If $T = h(X)$ is an unbiased estimator for λ, then Result 4.8 states that*

$$\operatorname{Var}(T) \geq J(\lambda)^{-1} = \frac{\lambda^2}{n},$$

which is the Cramér-Rao lower bound.

f) Compute the expectation of $\hat{\lambda}_{\text{ML}}$ and use this result to construct an unbiased estimator of λ. Compute its variance and compare it to the Cramér-Rao lower bound.

▶ *By the properties of exponential distribution we know that $\sum_{i=1}^n X_i \sim \operatorname{G}(n, \lambda)$, cf. Appendix A.5.2. Next, by the properties of Gamma distribution we that get that $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \operatorname{G}(n, n\lambda)$, and $\hat{\lambda}_{\text{ML}} = 1/\bar{X} \sim \operatorname{IG}(n, n\lambda)$, cf. Appendix A.5.2. It follows that*

$$\mathsf{E}(\hat{\lambda}_{\text{ML}}) = \frac{n\lambda}{n-1} > \lambda,$$

cf. again Appendix A.5.2. Thus, $\hat{\lambda}_{\text{ML}}$ is a biased estimator of λ. However, we can easily correct it by multiplying with the constant $(n-1)/n$. This new estimator $\hat{\lambda} = (n-1)/(n\bar{X})$ is obviously unbiased, and has variance

$$\begin{aligned}\operatorname{Var}(\hat{\lambda}) &= \frac{(n-1)^2}{n^2} \operatorname{Var}(1/\bar{X}) \\ &= \frac{(n-1)^2}{n^2} \frac{n^2\lambda^2}{(n-1)^2(n-2)} \\ &= \frac{\lambda^2}{n-2},\end{aligned}$$

cf. again Appendix A.5.2. This variance only asymptotically reaches the Cramér-Rao lower bound λ^2/n. Theoretically there might be other unbiased estimators which have a smaller variance than $\hat{\lambda}$.

5. An alternative parametrization of the exponential distribution is

$$f_X(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) \mathsf{I}_{\mathbb{R}^+}(x), \quad \theta > 0.$$

Let $X_{1:n}$ denote a random sample from this density. We want to test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$.

a) Calculate both variants T_1 and T_2 of the score test statistic.

▶ *Recall from Section 4.1 that*

$$T_1(x_{1:n}) = \frac{S(\theta_0; x_{1:n})}{\sqrt{J_{1:n}(\theta_0)}} \quad \textit{and} \quad T_2(x_{1:n}) = \frac{S(\theta_0; x_{1:n})}{\sqrt{I(\theta_0; x_{1:n})}}.$$

Like in the previous exercise, we can compute the log-likelihood

$$\begin{aligned} l(\theta) &= -\sum_{i=1}^{n} \log(\theta) - \frac{x_i}{\theta} \\ &= -n\log(\theta) - \frac{n\bar{x}}{\theta} \end{aligned}$$

and derive the score function

$$S(\theta; x_{1:n}) = (n\theta - n\bar{x}) \cdot \left(-\frac{1}{\theta^2}\right) = \frac{n(\bar{x}-\theta)}{\theta^2},$$

the Fisher information

$$I(\theta; x_{1:n}) = -\frac{d}{d\theta} S(\theta; x_{1:n}) = n\frac{2\bar{x}-\theta}{\theta^3},$$

and the expected Fisher information

$$J_{1:n}(\theta) = n\frac{2\,\mathsf{E}(\bar{X})-\theta}{\theta^3} = \frac{n}{\theta^2}.$$

The test statistics can now be written as

$$\begin{aligned} T_1(x_{1:n}) &= \frac{n(\bar{x}-\theta_0)}{\theta_0^2}\cdot\frac{\theta_0}{\sqrt{n}} = \sqrt{n}\frac{\bar{x}-\theta_0}{\theta_0} \\ \textit{and}\quad T_2(x_{1:n}) &= \frac{n(\bar{x}-\theta_0)}{\theta_0^2}\cdot\frac{\theta_0^{3/2}}{\sqrt{n(2\bar{x}-\theta_0)}} = T_1(\theta_0)\sqrt{\frac{\theta_0}{2\bar{x}-\theta_0}}. \end{aligned}$$

b) A sample of size $n = 100$ gave $\bar{x} = 0.26142$. Quantify the evidence against $H_0 : \theta_0 = 0.25$ using a suitable significance test.

▶ *By plugging these numbers into the formulas for $T_1(x_{1:n})$ and $T_2(x_{1:n})$, we obtain*

$$T_1(x_{1:n}) = 0.457 \quad \textit{and} \quad T_2(x_{1:n}) = 0.437.$$

Under the null hypothesis, both statistics follow asymptotically the standard normal distribution. Hence, to test at level α, we need to compare the observed values with the $(1-\alpha/2)\cdot 100\%$ quantile of the standard normal distribution. For $\alpha = 0.05$, we compare with $z_{0.975} \approx 1.96$. As neither of the observed values is larger than the critical value, the null hypothesis cannot be rejected.

6. In a study assessing the sensitivity π of a low-budget diagnostic test for asthma, each of n asthma patients is tested repeatedly until the first positive test result is obtained. Let X_i be the number of the first positive test for patient i. All patients and individual tests are independent, and the sensitivity π is equal for all patients and tests.

a) Derive the probability mass function $f(x;\pi)$ of X_i.

▶ *X_i can only take one of the values $1, 2, \dots,$ so it is a discrete random variable supported on natural numbers $\mathbb{N}$. For a given $x \in \mathbb{N}$, the probability that X_i equals x is*

$$\begin{aligned} f(x;\pi) &= \mathsf{Pr}(\textit{First test negative}, \dots, (x-1)\textit{-st test negative}, x\textit{-th test positive}) \\ &= \underbrace{(1-\pi)\cdots(1-\pi)}_{x-1\ \textit{times}}\cdot\pi \\ &= (1-\pi)^{x-1}\pi, \end{aligned}$$

since the results of the different tests are independent. This is the probability mass function of the geometric distribution $\mathrm{Geom}(\pi)$ *(cf. Appendix A.5.1), i. e. we have $X_i \overset{iid}{\sim} \mathrm{Geom}(\pi)$ for $i = 1, \dots, n$.*

b) Write down the log-likelihood function for the random sample $X_{1:n}$ and compute the MLE $\hat{\pi}_{\mathrm{ML}}$.

▶ *For a realisation $x_{1:n} = (x_1, \dots, x_n)$, the likelihood is*

$$\begin{aligned} L(\pi) &= \prod_{i=1}^{n} f(x_i;\pi) \\ &= \prod_{i=1}^{n} \pi(1-\pi)^{x_i-1} \\ &= \pi^n(1-\pi)^{\sum_{i=1}^{n} x_i - n} \\ &= \pi^n(1-\pi)^{n(\bar{x}-1)}, \end{aligned}$$

yielding the log-likelihood

$$l(\pi) = n\log(\pi) + n(\bar{x}-1)\log(1-\pi).$$

The score function is thus

$$S(\pi; x_{1:n}) = \frac{d}{d\pi} l(\pi) = \frac{n}{\pi} - \frac{n(\bar{x}-1)}{1-\pi}$$

and the solution of the score equation $S(\pi; x_{1:n}) = 0$ is

$$\hat{\pi}_{\mathrm{ML}} = 1/\bar{x}.$$

The Fisher information is

$$\begin{aligned} I(\pi) &= -\frac{d}{d\pi} S(\pi) \\ &= \frac{n}{\pi^2} + \frac{n(\bar{x}-1)}{(1-\pi)^2}, \end{aligned}$$

yielding the observed Fisher information

$$
\begin{aligned}
I(\hat{\pi}_{\mathrm{ML}}) &= n\left\{\bar{x}^2 + \frac{\bar{x}-1}{(\frac{\bar{x}-1}{\bar{x}})^2}\right\} \\
&= \frac{n\bar{x}^3}{\bar{x}-1},
\end{aligned}
$$

which is positive, as $x_i \geq 1$ *for every* i *by definition. It follows that* $\hat{\pi}_{\mathrm{ML}} = 1/\bar{x}$ *indeed is the MLE.*

c) Derive the standard error $\mathrm{se}(\hat{\pi}_{\mathrm{ML}})$ of the MLE.

▶ *The standard error is*

$$
\mathrm{se}(\hat{\pi}_{\mathrm{ML}}) = I(\hat{\pi}_{\mathrm{ML}})^{-1/2} = \sqrt{\frac{\bar{x}-1}{n\bar{x}^3}}.
$$

d) Give a general formula for an approximate 95% confidence interval for π. What could be the problem of this interval?

▶ *A general formula for an approximate 95% confidence interval for* π *is*

$$
[\hat{\pi}_{\mathrm{ML}} \pm z_{0.975} \cdot \mathrm{se}(\hat{\pi}_{\mathrm{ML}})] = \left[1/\bar{x} \pm z_{0.975}\sqrt{\frac{\bar{x}-1}{n\bar{x}^3}}\right]
$$

where $z_{0.975} \approx 1.96$ *is the 97.5% quantile of the standard normal distribution. The problem of this interval could be that it might contain values outside the range* $(0, 1)$ *of the parameter* π*. That is,* $1/\bar{x} - z_{0.975}\sqrt{\frac{\bar{x}-1}{n\bar{x}^3}}$ *could be smaller than 0 or* $1/\bar{x} + z_{0.975}\sqrt{\frac{\bar{x}-1}{n\bar{x}^3}}$ *could be larger than 1.*

e) Now we consider the parametrization with $\phi = \mathrm{logit}(\pi) = \log\{\pi/(1-\pi)\}$. Derive the corresponding MLE $\hat{\phi}_{\mathrm{ML}}$, its standard error and associated approximate 95% confidence interval. What is the advantage of this interval?

▶ *By the invariance of the MLE with respect to one-to-one transformations, we have*

$$
\begin{aligned}
\hat{\phi}_{\mathrm{ML}} &= \mathrm{logit}(\hat{\pi}_{\mathrm{ML}}) \\
&= \log\left(\frac{1/\bar{x}}{1-1/\bar{x}}\right) \\
&= \log\left(\frac{1}{\bar{x}-1}\right) \\
&= -\log(\bar{x}-1).
\end{aligned}
$$

By the delta method, we further have

$$
\mathrm{se}(\hat{\phi}_{\mathrm{ML}}) = \mathrm{se}(\hat{\pi}_{\mathrm{ML}})\left|\frac{d}{d\pi}\,\mathrm{logit}(\hat{\pi}_{\mathrm{ML}})\right|.
$$

We therefore compute

$$
\begin{aligned}
\frac{d}{d\pi}\,\mathrm{logit}(\pi) &= \frac{d}{d\pi}\log\left(\frac{\pi}{1-\pi}\right) \\
&= \left(\frac{\pi}{1-\pi}\right)^{-1}\frac{1(1-\pi)-(-1)\pi}{(1-\pi)^2} \\
&= \frac{1-\pi}{\pi}\cdot\frac{1-\pi+\pi}{(1-\pi)^2} \\
&= \frac{1}{\pi(1-\pi)},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{d}{d\pi}\,\mathrm{logit}(\hat{\pi}_{\mathrm{ML}}) &= \frac{1}{1/\bar{x}(1-1/\bar{x})} \\
&= \frac{\bar{x}}{\frac{\bar{x}-1}{\bar{x}}} \\
&= \frac{\bar{x}^2}{\bar{x}-1}.
\end{aligned}
$$

The standard error of $\hat{\phi}_{\mathrm{ML}}$ *is hence*

$$
\begin{aligned}
\mathrm{se}(\hat{\phi}_{\mathrm{ML}}) &= \mathrm{se}(\hat{\pi}_{\mathrm{ML}})\left|\frac{d}{d\pi}\,\mathrm{logit}(\hat{\pi}_{\mathrm{ML}})\right| \\
&= \sqrt{\frac{\bar{x}-1}{n\bar{x}^3}}\,\frac{\bar{x}^2}{\bar{x}-1} \\
&= n^{-1/2}(\bar{x}-1)^{1/2-1}(\bar{x})^{-3/2+2} \\
&= n^{-1/2}(\bar{x}-1)^{-1/2}(\bar{x})^{1/2} \\
&= \sqrt{\frac{\bar{x}}{n(\bar{x}-1)}}.
\end{aligned}
$$

The associated approximate 95% confidence interval for ϕ *is given by*

$$
\left[-\log(\bar{x}-1) \pm z_{0.975}\cdot\sqrt{\frac{\bar{x}}{n(\bar{x}-1)}}\right].
$$

The advantage of this interval is that it is for a real-valued parameter $\phi \in \mathbb{R}$*, so its bounds are always contained in the parameter range.*

f) $n = 9$ patients did undergo the trial and the observed numbers were $x = (3, 5, 2, 6, 9, 1, 2, 2, 3)$. Calculate the MLEs $\hat{\pi}_{\mathrm{ML}}$ and $\hat{\phi}_{\mathrm{ML}}$, the confidence intervals from 6d) and 6e) and compare them by transforming the latter back to the π-scale.

▶ *To compute the MLEs* $\hat{\pi}_{\mathrm{ML}}$ *and* $\hat{\phi}_{\mathrm{ML}}$*, we may proceed as follows.*

```
> ## the data:
> x <- c(3, 5, 2, 6, 9, 1, 2, 2, 3)
> xq <- mean(x)
> ## the MLE for pi:
> (mle.pi <- 1/xq)
[1] 0.2727273
> ## The logit function is the quantile function of the
> ## standard logistic distribution, hence:
> (mle.phi <- qlogis(mle.pi))
[1] -0.9808293
> ## and this is really the same as
> - log(xq - 1)
[1] -0.9808293
```

We obtain that $\hat{\pi}_{\text{ML}} = 0.273$ *and that* $\hat{\phi}_{\text{ML}} = -0.981$. *To compute the confidence intervals, we proceed as follows.*

```
> n <- length(x)
> ## the standard error for pi:
> (se.pi <- sqrt((xq-1) / (n * xq^3)))
[1] 0.07752753
> ## the standard error for phi:
> (se.phi <- sqrt(xq / (n * (xq - 1))))
[1] 0.390868
> ## the CIs:
> (ci.pi <- mle.pi + c(-1, +1) * qnorm(0.975) * se.pi)
[1] 0.1207761 0.4246784
> (ci.phi <- mle.phi + c(-1, +1) * qnorm(0.975) * se.phi)
[1] -1.7469164 -0.2147421
```

Now we can transform the bounds of the confidence interval for ϕ *back to the* π*-scale. Note that the logit transformation, and hence also its inverse, are strictly monotonically increasing, which is easily seen from*

$$\frac{d}{d\pi}\operatorname{logit}(\pi) = \frac{1}{\pi(1-\pi)} > 0.$$

We can work out the inverse transformation by solving $\phi = \operatorname{logit}(\pi)$ *for* π*:*

$$\begin{aligned}
\phi &= \log\left(\frac{\pi}{1-\pi}\right)\\
\exp(\phi) &= \frac{\pi}{1-\pi}\\
\exp(\phi) - \pi\exp(\phi) &= \pi\\
\exp(\phi) &= \pi\{1+\exp(\phi)\}\\
\pi &= \frac{\exp(\phi)}{1+\exp(\phi)}.
\end{aligned}$$

So we have $\operatorname{logit}^{-1}(\phi) = \frac{\exp(\phi)}{1+\exp(\phi)}$. *This is also the cdf of the standard logistic distribution. To transform the confidence interval in* R*, we may therefore proceed as follows.*

```
> (ci.pi.2 <- plogis(ci.phi))
[1] 0.1484366 0.4465198
> ## This is identical to:
> invLogit <- function(phi)
  {
      exp(phi) / (1 + exp(phi))
  }
> invLogit(ci.phi)
[1] 0.1484366 0.4465198
```

We thus obtain two confidence intervals for π*:* $(0.121, 0.425)$ *and* $(0.148, 0.447)$*. We now compare their lengths.*

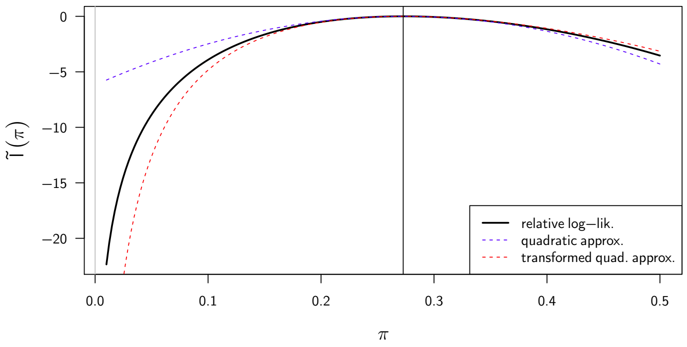
```
> ## compare the lengths of the two pi confidence intervals:
> ci.pi.2[2] - ci.pi.2[1]
[1] 0.2980833
> ci.pi[2] - ci.pi[1]
[1] 0.3039023
```

Compared to the first confidence interval for π*, the new interval is slightly shifted to the right, and smaller. An advantage is that it can never lie outside the* $(0, 1)$ *range.*

g) Produce a plot of the relative log-likelihood function $\tilde{l}(\pi)$ and two approximations in the range $\pi \in (0.01, 0.5)$: The first approximation is based on the direct quadratic approximation of $\tilde{l}_\pi(\pi) \approx q_\pi(\pi)$, the second approximation is based on the quadratic approximation of $\tilde{l}_\phi(\phi) \approx q_\phi(\phi)$, *i. e.* $q_\phi\{\operatorname{logit}(\pi)\}$ values are plotted. Comment the result.

▶ *We produce a plot of the relative log-likelihood function* $\tilde{l}(\pi)$ *and two (quadratic) approximations:*

```
> ## functions for pi:
> loglik.pi <- function(pi)
  {
      n * log(pi) + n * (xq - 1) * log(1 - pi)
  }
> rel.loglik.pi <- function(pi)
  {
      loglik.pi(pi) - loglik.pi(mle.pi)
  }
> approx.rel.loglik.pi <- function(pi)
  {
      - 0.5 * se.pi^(-2) * (pi - mle.pi)^2
  }
> ## then for phi:
> loglik.phi <- function(phi)
  {
      loglik.pi(plogis(phi))
  }
> rel.loglik.phi <- function(phi)
  {
      loglik.phi(phi) - loglik.phi(mle.phi)
  }
> approx.rel.loglik.phi <- function(phi)
  {
      - 0.5 * se.phi^(-2) * (phi - mle.phi)^2
  }
> ## and the plot
> piGrid <- seq(0.01, 0.5, length=201)
> plot(piGrid, rel.loglik.pi(piGrid),
       type="l",
       xlab=expression(pi),
       ylab = expression(tilde(l)(pi)),
       lwd=2)
> abline(v=0, col="gray")
> lines(piGrid, approx.rel.loglik.pi(piGrid),
        lty=2,
        col="blue")
> lines(piGrid, approx.rel.loglik.phi(qlogis(piGrid)),
        lty=2,
        col="red")
> abline(v=mle.pi)
> legend("bottomright",
         legend=
         c("relative log-lik.",
           "quadratic approx.",
           "transformed quad. approx."),
         col=
         c("black",
           "blue",
           "red"),
         lty=
         c(1,
           2,
           2),
         lwd=
         c(2,
           1,
           1))
```



The transformed quadratic approximation $q_\phi(\text{logit}(\pi))$ is closer to the true relative log-likelihood $\tilde{l}(\pi)$ than the direct quadratic approximation $q_\pi(\pi)$. This corresponds to a better performance of the second approximate confidence interval.

7. A simple model for the drug concentration in plasma over time after a single intravenous injection is $c(t) = \theta_2 \exp(-\theta_1 t)$, with $\theta_1, \theta_2 > 0$. For simplicity we assume here that $\theta_2 = 1$.

a) Assume that n probands had their concentrations c_i, $i = 1, \dots, n$, measured at the same single time-point t and assume that the model $c_i \overset{\text{iid}}{\sim} \mathrm{N}(c(t), \sigma^2)$ is appropriate for the data. Calculate the MLE of θ_1.

▶ *The likelihood is*

$$L(\theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left\{c_i - \exp(-\theta_1 t)\right\}^2\right],$$

yielding the log-likelihood

$$l(\theta_1) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\{c_i - \exp(-\theta_1 t)\right\}^2\right].$$

For the score function we thus have

$$\begin{aligned} S(\theta_1; c_{1:n}) &= -\frac{\exp(-\theta_1 t)\, t}{\sigma^2} \sum_{i=1}^{n} \left\{c_i - \exp(-\theta_1 t)\right\} \\ &= \frac{\exp(-\theta_1 t)\, nt}{\sigma^2} \left\{\exp(-\theta_1 t) - \bar{c}\right\}, \end{aligned}$$

and for the Fisher information

$$I(\theta_1) = \frac{\exp(-\theta_1 t)\, nt^2}{\sigma^2} \left\{2\exp(-\theta_1 t) - \bar{c}\right\}.$$

The score equation is solved as

$$\begin{aligned} 0 &= S(\theta_1; c_{1:n}) \\ \exp(-\theta_1 t) &= \bar{c} \\ \hat{\theta}_1 &= -\frac{1}{t}\log(\bar{c}). \end{aligned}$$

The observed Fisher information

$$I(\hat{\theta}_1) = \frac{\bar{c}^2 nt^2}{\sigma^2}$$

is positive; thus, $\hat{\theta}_1$ is indeed the MLE.

b) Calculate the asymptotic variance of the MLE.

▶ *By Result 4.10, the asymptotic variance of $\hat{\theta}_1$ is the inverse of the expected Fisher information*

$$\begin{aligned} J_{1:n}(\theta_1) &= \mathsf{E}\{I(\theta_1; C_{1:n})\} \\ &= \frac{\exp(-2\theta_1 t)\, nt^2}{\sigma^2}. \end{aligned}$$

c) In pharmacokinetic studies one is often interested in the area under the concentration curve, $\alpha = \int_0^\infty \exp(-\theta_1 t)\,dt$. Calculate the MLE for α and its variance estimate using the delta theorem.

▶ *By the invariance of the MLE with respect to one-to-one transformations, we obtain that*

$$\begin{aligned}\hat{\alpha}_{\text{ML}} &= \int\limits_0^\infty \exp(-\hat{\theta}_1 t)\,dt \\ &= \frac{1}{\hat{\theta}_1} \\ &= -\frac{t}{\log(\bar{c})}.\end{aligned}$$

Further, by the delta method, we obtain that

$$\begin{aligned}\text{se}(\hat{\alpha}_{\text{ML}}) = \text{se}(\hat{\theta}_1)\left|\frac{d}{d\theta_1}\frac{1}{\hat{\theta}_1}\right| \\ = \frac{\sigma}{\exp(-\hat{\theta}_1 t)\sqrt{n}t\theta_1^2}.\end{aligned}$$

Thus, the asymptotic variance of $\hat{\alpha}_{\text{ML}}$ is

$$\frac{\sigma^2}{\exp(-2\hat{\theta}_1 t)nt^2\theta_1^4}.$$

d) We now would like to determine the optimal time point for measuring the concentrations c_i. Minimise the asymptotic variance of the MLE with respect to t, when θ_1 is assumed to be known, to obtain an optimal time point t_{opt}.

▶ *We take the derivative of the asymptotic variance of $\hat{\theta}_1$ with respect to t:*

$$\frac{d}{dt}\left\{\frac{\sigma^2}{n}\cdot\frac{\exp(2\theta_1 t)}{t^2}\right\} = \frac{\sigma^2}{n}\left\{\frac{2\theta_1\exp(2\theta_1 t)}{t^2} - \frac{2\exp(2\theta_1 t)}{t^3}\right\},$$

and find that it is equal zero for t_{opt} satisfying that

$$\frac{2\theta_1\exp(2\theta_1 t)}{t_{\text{opt}}^2} = \frac{2\exp(2\theta_1 t)}{t_{\text{opt}}^3},$$

i. e. for

$$t_{\text{opt}} = \frac{1}{\theta_1}.$$

In order to verify that t_{opt} minimises the asymptotic variance, we compute the second derivative of the asymptotic variance with respect to t:

$$\frac{2\sigma^2}{n}\left(\frac{2\theta_1^2\exp(2\theta_1 t)}{t^2} - \frac{4\theta_1\exp(2\theta_1 t)}{t^3} + \frac{3\exp(2\theta_1 t)}{t^4}\right)$$

If we plug in $t_{\text{opt}} = 1/\theta_1$ for t, we obtain

$$\frac{2\sigma^2}{n}\exp(2)(2\theta_1^4 - 4\theta_1^4 + 3\theta_1^4) = \frac{2\theta_1^4\sigma^2\exp(2)}{n},$$

which is positive. Thus, t_{opt} indeed minimises the variance.

8. Assume the gamma model $\text{G}(\alpha, \alpha/\mu)$ for the random sample $X_{1:n}$ with mean $\mathsf{E}(X_i) = \mu > 0$ and shape parameter $\alpha > 0$.

a) First assume that α is known. Derive the MLE $\hat{\mu}_{\text{ML}}$ and the observed Fisher information $I(\hat{\mu}_{\text{ML}})$.

▶ *The log-likelihood kernel for μ is*

$$l(\mu) = -\alpha n\log(\mu) - \frac{\alpha}{\mu}\sum_{i=1}^n x_i,$$

and the score function is

$$S(\mu; x) = \frac{d}{d\mu}l(\mu) = -\frac{\alpha n}{\mu} + \alpha\mu^{-2}\sum_{i=1}^n x_i.$$

The score equation $S(\mu; x) = 0$ can be written as

$$n = \frac{1}{\mu}\sum_{i=1}^n x_i$$

and is hence solved by $\hat{\mu}_{\text{ML}} = \bar{x}$. The ordinary Fisher information is

$$\begin{aligned}I(\mu) &= -\frac{d}{d\mu}S(\mu; x) \\ &= -\left(\alpha n\mu^{-2} - 2\alpha\mu^{-3}\sum_{i=1}^n x_i\right) \\ &= \frac{2\alpha}{\mu^3}\sum_{i=1}^n x_i - \frac{\alpha n}{\mu^2},\end{aligned}$$

so the observed Fisher information equals

$$\begin{aligned}I(\hat{\mu}_{\text{ML}}) &= I(\bar{x}) \\ &= \frac{2\alpha n\bar{x}}{(\bar{x})^3} - \frac{\alpha n}{(\bar{x})^2} \\ &= \frac{\alpha n}{(\bar{x})^2} \\ &= \frac{\alpha n}{\hat{\mu}_{\text{ML}}^2}.\end{aligned}$$

As it is positive, we have indeed found the MLE.

b) Use the p^* formula to derive an asymptotic density of $\hat{\mu}_{\text{ML}}$ depending on the true parameter μ. Show that the kernel of this approximate density is exact in this case, *i. e.* it equals the kernel of the exact density known from Exercise 9 from Chapter 3.

▶ *The p^* formula gives us the following approximate density of the MLE:*

$$\begin{aligned}
f^*(\hat{\mu}_{\text{ML}}) &= \sqrt{\frac{I(\hat{\mu}_{\text{ML}})}{2\pi}}\frac{L(\mu)}{L(\hat{\mu}_{\text{ML}})} \\
&= \sqrt{\frac{I(\hat{\mu}_{\text{ML}})}{2\pi}}\exp\{l(\mu) - l(\hat{\mu}_{\text{ML}})\} \\
&= \sqrt{\frac{\alpha n}{\hat{\mu}_{\text{ML}}^2 2\pi}}\exp\left\{-\alpha n\log(\mu) - \alpha/\mu\cdot n\hat{\mu}_{\text{ML}} + \alpha n\log(\hat{\mu}_{\text{ML}}) + \alpha/\hat{\mu}_{\text{ML}}\cdot n\hat{\mu}_{\text{ML}}\right\} \\
&= \sqrt{\frac{\alpha n}{2\pi}}\mu^{-\alpha n}\exp(\alpha n)\cdot\hat{\mu}_{\text{ML}}^{\alpha n-1}\exp\left(-\frac{\alpha n}{\mu}\hat{\mu}_{\text{ML}}\right). \qquad (4.3)
\end{aligned}$$

From Appendix A.5.2 we know that $\sum_{i=1}^n X_i \sim \mathrm{G}(n\alpha, \alpha/\mu)$, and $\bar{X} = \hat{\mu}_{\text{ML}} \sim \mathrm{G}(n\alpha, n\alpha/\mu)$; cf. Exercise 9 in Chapter 3. The corresponding density function has the kernel

$$f(\hat{\mu}_{\text{ML}}) \propto \hat{\mu}_{\text{ML}}^{n\alpha-1}\exp\left(-\frac{\alpha n}{\mu}\hat{\mu}_{\text{ML}}\right),$$

which is the same as the kernel in (4.3).

c) Stirling's approximation of the gamma function is

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}\frac{x^x}{\exp(x)}. \qquad (4.4)$$

Show that approximating the normalising constant of the exact density with (4.4) gives the normalising constant of the approximate p^* formula density.

▶ *The normalising constant of the exact distribution* $\mathrm{G}(n\alpha, n\alpha/\mu)$ *is:*

$$\begin{aligned}
\frac{\left(\frac{\alpha n}{\mu}\right)^{\alpha n}}{\Gamma(\alpha n)} &\approx \left(\frac{\alpha n}{\mu}\right)^{\alpha n}\sqrt{\frac{\alpha n}{2\pi}\frac{\exp(\alpha n)}{(\alpha n)^{\alpha n}}} \\
&= \mu^{-\alpha n}\exp(\alpha n)\sqrt{\frac{\alpha n}{2\pi}},
\end{aligned}$$

which equals the normalising constant of the approximate density in (4.3).

d) Now assume that μ is known. Derive the log-likelihood, score function and Fisher information of α. Use the digamma function $\psi(x) = \frac{d}{dx}\log\{\Gamma(x)\}$ and the trigamma function $\psi'(x) = \frac{d}{dx}\psi(x)$.

▶ *The log-likelihood kernel of α is*

$$l(\alpha) = \alpha n\log\left(\frac{\alpha}{\mu}\right) - n\log\{\Gamma(\alpha)\} + \alpha\sum_{i=1}^n\log(x_i) - \frac{\alpha}{\mu}\sum_{i=1}^n x_i,$$

and the score function reads

$$\begin{aligned}
S(\alpha; x) &= \frac{d}{d\alpha}l(\alpha) \\
&= n\log\left(\frac{\alpha}{\mu}\right) + \alpha n\frac{\mu}{\alpha}\frac{1}{\mu} - n\psi(\alpha) + \sum_{i=1}^n\log(x_i) - \frac{1}{\mu}\sum_{i=1}^n x_i \\
&= n\log\left(\frac{\alpha}{\mu}\right) + n - n\psi(\alpha) + \sum_{i=1}^n\log(x_i) - \frac{1}{\mu}\sum_{i=1}^n x_i.
\end{aligned}$$

Hence, the Fisher information is

$$\begin{aligned}
I(\alpha) &= -\frac{d}{d\alpha}S(\alpha; x) \\
&= -\left\{\frac{n}{\alpha} - n\psi'(\alpha)\right\} \\
&= n\left\{\psi'(\alpha) - \frac{1}{\alpha}\right\}.
\end{aligned}$$

e) Show, by rewriting the score equation, that the MLE $\hat{\alpha}_{\text{ML}}$ fulfils

$$-n\psi(\hat{\alpha}_{\text{ML}}) + n\log(\hat{\alpha}_{\text{ML}}) + n = -\sum_{i=1}^n\log(x_i) + \frac{1}{\mu}\sum_{i=1}^n x_i + n\log(\mu). \qquad (4.5)$$

Hence show that the log-likelihood kernel can be written as

$$l(\alpha) = n\big[\alpha\log(\alpha) - \alpha - \log\{\Gamma(\alpha)\} + \alpha\psi(\hat{\alpha}_{\text{ML}}) - \alpha\log(\hat{\alpha}_{\text{ML}})\big].$$

▶ *The score equation $S(\hat{\alpha}_{\text{ML}}; x) = 0$ can be written as*

$$\begin{aligned}
n\log(\hat{\alpha}_{\text{ML}}) - n\log(\mu) + n - n\psi(\hat{\alpha}_{\text{ML}}) + \sum_{i=1}^n\log(x_i) - \frac{1}{\mu}\sum_{i=1}^n x_i &= 0 \\
-n\psi(\hat{\alpha}_{\text{ML}}) + n\log(\hat{\alpha}_{\text{ML}}) + n &= -\sum_{i=1}^n\log(x_i) + \frac{1}{\mu}\sum_{i=1}^n x_i + n\log(\mu).
\end{aligned}$$

Hence, we can rewrite the log-likelihood kernel as follows:

$$\begin{aligned}
l(\alpha) &= \alpha n\log\left(\frac{\alpha}{\mu}\right) - n\log\{\Gamma(\alpha)\} + \alpha\sum_{i=1}^n\log(x_i) - \frac{\alpha}{\mu}\sum_{i=1}^n x_i \\
&= \alpha n\log(\alpha) - n\log\{\Gamma(\alpha)\} - \alpha\left\{n\log(\mu) - \sum_{i=1}^n\log(x_i) + \frac{1}{\mu}\sum_{i=1}^n x_i\right\} \\
&= \alpha n\log(\alpha) - n\log\{\Gamma(\alpha)\} - \alpha\left\{-n\psi(\hat{\alpha}_{\text{ML}}) + n\log(\hat{\alpha}_{\text{ML}}) + n\right\} \\
&= n[\alpha\log(\alpha) - \alpha - \log\{\Gamma(\alpha)\} + \alpha\psi(\hat{\alpha}_{\text{ML}}) - \alpha\log(\hat{\alpha}_{\text{ML}})].
\end{aligned}$$

f) Implement an R-function of the p^* formula, taking as arguments the MLE value(s) $\hat{\alpha}_{\text{ML}}$ at which to evaluate the density, and the true parameter α. For numerical reasons, first compute the approximate log-density

$$\log f^*(\hat{\alpha}_{\text{ML}}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\{I(\hat{\alpha}_{\text{ML}})\} + l(\alpha) - l(\hat{\alpha}_{\text{ML}}),$$

and then exponentiate it. The R-functions `digamma`, `trigamma` and `lgamma` can be used to calculate $\psi(x)$, $\psi'(x)$ and $\log\{\Gamma(x)\}$, respectively.

▶ *We first rewrite the relative log-likelihood $l(\alpha) - l(\hat{\alpha}_{\text{ML}})$ as*

$$\begin{aligned}
&n[\alpha\log(\alpha) - \hat{\alpha}_{\text{ML}}\log(\hat{\alpha}_{\text{ML}}) - (\alpha - \hat{\alpha}_{\text{ML}}) - \log\{\Gamma(\alpha)\} + \log\{\Gamma(\hat{\alpha}_{\text{ML}})\} \\
&\qquad + (\alpha - \hat{\alpha}_{\text{ML}})\psi(\hat{\alpha}_{\text{ML}}) - (\alpha - \hat{\alpha}_{\text{ML}})\log(\hat{\alpha}_{\text{ML}})] \\
&= n[\alpha\{\log(\alpha) - \log(\hat{\alpha}_{\text{ML}})\} - (\alpha - \hat{\alpha}_{\text{ML}}) - \log\{\Gamma(\alpha)\} + \log\{\Gamma(\hat{\alpha}_{\text{ML}})\} + (\alpha - \hat{\alpha}_{\text{ML}})\psi(\hat{\alpha}_{\text{ML}})].
\end{aligned}$$

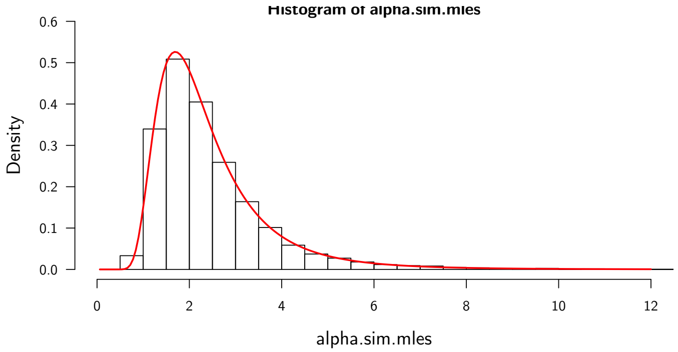
Now we are ready to implement the approximate density of the MLE, as described by the p^* formula:

```
> approx.mldens <- function(alpha.mle, alpha.true)
  {
      relLogLik <- n * (alpha.true * (log(alpha.true) - log(alpha.mle)) -
                        (alpha.true - alpha.mle) - lgamma(alpha.true) +
                        lgamma(alpha.mle) + (alpha.true - alpha.mle) *
                        digamma(alpha.mle))
      logObsFisher <- log(n) + log(trigamma(alpha.mle) - 1 / alpha.mle)
      logret <- - 0.5 * log(2 * pi) + 0.5 * logObsFisher + relLogLik
      return(exp(logret))
  }
```

g) In order to illustrate the quality of this approximation, we consider the case with $\alpha = 2$ and $\mu = 3$. Simulate 10 000 data sets of size $n = 10$, and compute the MLE $\hat{\alpha}_{\text{ML}}$ for each of them by numerically solving (4.5) using the R-function `uniroot` (*cf*. Appendix C.1.1). Plot a histogram of the resulting 10 000 MLE samples (using `hist` with option `prob=TRUE`). Add the approximate density derived above to compare.

▶ *To illustrate the quality of this approximation by simulation, we may run the following code.*

```
> ## the score function
> scoreFun.alpha <- function(alpha,
                             x,            # the data
                             mu)           # the mean parameter
  {
      n <- length(x)
      ret <- n * log(alpha / mu) + n - n * digamma(alpha) +
          sum(log(x)) - sum(x) / mu
      ## be careful that this is vectorised in alpha!
      return(ret)
  }
> ## this function computes the MLE for alpha
> getMle.alpha <- function(x,                    # the data
                           mu)                   # the mean parameter
  {
      ## solve the score equation
      uniroot(f=scoreFun.alpha,
              interval=c(1e-10, 1e+10),
              x=x,                            # pass additional parameters
              mu=mu)$root                     # to target function
  }
> ## now simulate the datasets and compute the MLE for each
> nSim <- 10000
> alpha <- 2
> mu <- 3
> n <- 10
> alpha.sim.mles <- numeric(nSim)
> set.seed(93)
> for(i in seq_len(nSim))
  {
      alpha.sim.mles[i] <- getMle.alpha(x=
                                        rgamma(n=n,
                                               alpha,
                                               alpha / mu),
                                        mu=mu)
  }
> ## compare the histogram with the p* density
> hist(alpha.sim.mles,
       prob=TRUE,
       nclass=50,
       ylim=c(0, 0.6),
       xlim=c(0, 12))
> curve(approx.mldens(x, alpha.true=alpha),
        add=TRUE,
        lwd=2,
        n=201,
        col="red")
```



We may see that we have a nice agreement between the sampling distribution and the approximate density.

5 Likelihood inference in multiparameter models

1. In a cohort study on the incidence of ischaemic heart disease (IHD) 337 male probands were enrolled. Each man was categorised as non-exposed (group 1, daily energy consumption $\geq 2750\,\text{kcal}$) or exposed (group 2, daily energy consumption $< 2750\,\text{kcal}$) to summarise his average level of physical activity. For each group, the number of person years ($Y_1 = 2768.9$ and $Y_2 = 1857.5$) and the number of IHD cases ($D_1 = 17$ and $D_2 = 28$) was registered thereafter.

We assume that $D_i \,|\, Y_i \overset{\text{ind}}{\sim} \text{Po}(\lambda_i Y_i), i = 1, 2$, where $\lambda_i > 0$ is the group-specific incidence rate.

a) For each group, derive the MLE $\hat{\lambda}_i$ and a corresponding 95% Wald confidence interval for $\log(\lambda_i)$ with subsequent back-transformation to the λ_i-scale.

▶ *The log-likelihood kernel corresponding to a random variable X with Poisson distribution* $\text{Po}(\theta)$ *is*

$$l(\theta) = -\theta + x \log(\theta),$$

implying the score function

$$S(\theta; x) = \frac{d}{d\theta} l(\theta) = -1 + \frac{x}{\theta}$$

and the Fisher information

$$I(\theta) = -\frac{d}{d\theta} S(\theta; x) = \frac{x}{\theta^2}.$$

The score equation is solved by $\hat{\theta}_{\text{ML}} = x$ and since the observed Fisher information $I(\hat{\theta}_{\text{ML}}) = 1/x$ is positive, $\hat{\theta}_{\text{ML}}$ indeed is the MLE of θ.

The rates of the Poisson distributions for D_i can therefore be estimated by the maximum likelihood estimators $\hat{\theta}_i = D_i$. Now,

$$\lambda_i = \frac{\theta_i}{Y_i},$$

so, by the invariance of the MLE, we obtain that

$$\hat{\lambda}_i = \frac{\theta_i}{Y_i} = \frac{D_i}{Y_i}.$$

With the given data we have the results $\hat{\lambda}_1 = D_1/Y_1 = 6.14 \cdot 10^{-3}$ *and* $\hat{\lambda}_2 = D_2/Y_2 = 1.51 \cdot 10^{-2}$.

We can now use the fact that Poisson distribution $\mathrm{Po}(\theta)$ *with* θ *a natural number may be seen as the distribution of a sum of* θ *independent random variables, each with distribution* $\mathrm{Po}(1)$*. As such,* $\mathrm{Po}(\theta)$ *for reasonably large* θ *is approximately* $\mathrm{N}(\theta, \theta)$*. In our case,* $D_1 = \hat{\theta}_1 = 17$ *and* $D_2 = \hat{\theta}_2 = 28$ *might be considered approximately normally distributed,* $\hat{\theta}_i \sim \mathrm{N}(\theta_i, \theta_i)$*. It follows that* $\hat{\lambda}_i$ *are, too, approximately normal,* $\hat{\lambda}_i \sim \mathrm{N}(\lambda_i, \lambda_i/Y_i)$*. The standard errors of* $\hat{\lambda}_i$ *can be estimated as* $\mathrm{se}(\hat{\lambda}_i) = \sqrt{D_i}/Y_i$.

Again by the invariance of the MLE, the MLEs of $\psi_i = \log(\lambda_i) = f(\lambda_i)$ *are*

$$\hat{\psi}_i = \log\left(\frac{D_i}{Y_i}\right).$$

By the delta method, the standard errors of $\hat{\psi}_i$ *are*

$$\begin{aligned}\mathrm{se}(\hat{\psi}_i) &= \mathrm{se}(\hat{\lambda}_i) \cdot \left|f'(\hat{\lambda}_i)\right| \\ &= \frac{\sqrt{D_i}}{Y_i} \cdot \left|\frac{1}{\hat{\lambda}_i}\right| \\ &= \frac{1}{\sqrt{D_i}}.\end{aligned}$$

Therefore, the back-transformed limits of the 95% Wald confidence intervals with log-transformation for λ_i *equal*

$$\left[\exp\left(\hat{\psi}_i - z_{0.975}/\sqrt{D_i}\right), \exp\left(\hat{\psi}_i + z_{0.975}/\sqrt{D_i}\right)\right],$$

and with the data we get:

```
> (ci1 <- exp(log(d[1]/y[1]) + c(-1, 1) * qnorm(0.975) / sqrt(d[1])))
[1] 0.003816761 0.009876165
> (ci2 <- exp(log(d[2]/y[2]) + c(-1, 1) * qnorm(0.975) / sqrt(d[2])))
[1] 0.01040800 0.02183188
```

i. e. the confidence interval for λ_1 is $(0.00382, 0.00988)$ and for λ_2 it is $(0.01041, 0.02183)$.

b) In order to analyse whether $\lambda_1 = \lambda_2$, we reparametrise the model with $\lambda = \lambda_1$ and $\theta = \lambda_2/\lambda_1$. Show that the joint log-likelihood kernel of λ and θ has the following form:

$$l(\lambda, \theta) = D\log(\lambda) + D_2\log(\theta) - \lambda Y_1 - \theta\lambda Y_2,$$

where $D = D_1 + D_2$.

▶ *First, note that* θ *now has a different meaning than in the solution of 1a). By the independence of* D_1 *and* D_2*, the joint log-likelihood kernel in the original parametrisation is*

$$l(\lambda_1, \lambda_2) = D_1\log(\lambda_1) - \lambda_1 Y_1 + D_2\log(\lambda_2) - \lambda_2 Y_2.$$

In terms of the new parametrisation,

$$\lambda_1 = \lambda \quad \text{and} \quad \lambda_2 = \lambda\theta,$$

so if, moreover, we denote $D = D_1 + D_2$*, we have*

$$\begin{aligned} l(\lambda, \theta) &= D_1\log(\lambda) - \lambda Y_1 + D_2\log(\lambda) + D_2\log(\theta) - \lambda\theta Y_2 \\ &= D\log(\lambda) + D_2\log(\theta) - \lambda Y_1 - \lambda\theta Y_2. \end{aligned} \tag{5.1}$$

c) Compute the MLE $(\hat{\lambda}, \hat{\theta})$, the observed Fisher information matrix $\boldsymbol{I}(\hat{\lambda}, \hat{\theta})$ and derive expressions for both profile log-likelihood functions $l_p(\lambda) = l\{\lambda, \hat{\theta}(\lambda)\}$ and $l_p(\theta) = l\{\hat{\lambda}(\theta), \theta\}$.

▶ *The score function is*

$$\boldsymbol{S}(\lambda, \theta) = \begin{pmatrix} \frac{d}{d\lambda} l(\lambda, \theta) \\ \frac{d}{d\theta} l(\lambda, \theta) \end{pmatrix} = \begin{pmatrix} \frac{D}{\lambda} - Y_1 - \theta Y_2 \\ \frac{D_2}{\theta} - \lambda Y_2 \end{pmatrix},$$

and the Fisher information is

$$\boldsymbol{I}(\lambda, \theta) = -\begin{pmatrix} \frac{d^2}{d\lambda^2} l(\lambda, \theta) & \frac{d^2\, l(\lambda,\theta)}{d\lambda\, d\theta} \\ \frac{d^2\, l(\lambda,\theta)}{d\lambda\, d\theta} & \frac{d^2}{d\theta^2} l(\lambda, \theta) \end{pmatrix} = \begin{pmatrix} \frac{D}{\lambda^2} & Y_2 \\ Y_2 & \frac{D_2}{\theta^2} \end{pmatrix}.$$

The score equation $\boldsymbol{S}(\lambda, \theta) = \boldsymbol{0}$ *is solved by*

$$(\hat{\lambda}, \hat{\theta}) = \left(\frac{D_1}{Y_1}, \frac{D_2 Y_1}{D_1 Y_2}\right),$$

and, as the observed Fisher information

$$\boldsymbol{I}(\hat{\lambda}, \hat{\theta}) = \begin{pmatrix} \frac{D Y_1^2}{D_1^2} & Y_2 \\ Y_2 & \frac{D_1^2 Y_2^2}{D_2 Y_1^2} \end{pmatrix}$$

is positive definite, $(\hat{\lambda}, \hat{\theta})$ *indeed is the MLE.*

In order to derive the profile log-likelihood functions, one first has to compute the maxima of the log-likelihood with fixed λ *or* θ*, which we call here* $\hat{\theta}(\lambda)$ *and* $\hat{\lambda}(\theta)$*. This amounts to solving the score equations* $\frac{d}{d\theta} l(\lambda, \theta) = 0$ *and* $\frac{d}{d\lambda} l(\lambda, \theta) = 0$ *separately for* θ *and* λ*, respectively. The solutions are*

$$\hat{\theta}(\lambda) = \frac{D_2}{\lambda Y_2} \quad \text{and} \quad \hat{\lambda}(\theta) = \frac{D}{Y_1 + \theta Y_2}.$$

The strictly positive diagonal entries of the Fisher information show that the log-likelihoods are strictly concave, so $\hat{\theta}(\lambda)$ *and* $\hat{\lambda}(\theta)$ *indeed are the maxima. Now we can obtain the profile log-likelihood functions by plugging in* $\hat{\theta}(\lambda)$ *and* $\hat{\lambda}(\theta)$ *into the log-likelihood (5.1). The results are (after omitting additive constants not depending on the arguments* λ *and* θ*, respectively)*

$$\begin{aligned} &l_p(\lambda) = D_1\log(\lambda) - \lambda Y_1 \\ \text{and} \quad &l_p(\theta) = -D\log(Y_1 + \theta Y_2) + D_2\log(\theta). \end{aligned}$$

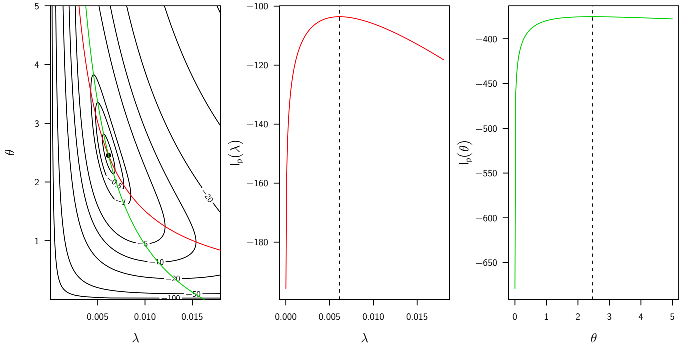
d) Plot both functions $l_p(\lambda)$ and $l_p(\theta)$, and also create a contour plot of the relative log-likelihood $\tilde{l}(\lambda, \theta)$ using the R-function `contour`. Add the points $\{\lambda, \hat{\theta}(\lambda)\}$ and $\{\hat{\lambda}(\theta), \theta\}$ to the contour plot, analogously to Figure 5.3a).

▶ *The following* R *code produces the desired plots.*

```
> ## log-likelihood of lambda, theta:
> loglik <- function(param)
  {
      lambda <- param[1]
      theta <- param[2]
      return(dtot * log(lambda) + d[2] * log(theta) -
             lambda * y[1] - theta * lambda * y[2])
  }
> ## the MLE
> (mle <- c(d[1]/y[1], (y[1] * d[2]) / (y[2] * d[1])))
```

```
[1] 0.006139622 2.455203864
```

```
> ## relative log-likelihood
> rel.loglik <- function(param)
  {
      return(loglik(param) - loglik(mle))
  }
> ## set up parameter grids
> lambda.grid <- seq(1e-5, 0.018, length = 200)
> theta.grid <- seq(1e-5, 5, length = 200)
> grid <- expand.grid(lambda = lambda.grid, theta = theta.grid)
> values <- matrix(data = apply(grid, 1, rel.loglik),
                   nrow = length(lambda.grid))
> ## set up plot frame
> par(mfrow = c(1,3))
> ## contour plot of the relative loglikelihood:
> contour(lambda.grid, theta.grid, values,
          xlab = expression(lambda), ylab = expression(theta),
          levels = -c(0.1, 0.5, 1, 5, 10, 20, 50, 100, 500, 1000, 1500, 2000),
          xaxs = "i", yaxs = "i")
> points(mle[1], mle[2], pch = 19)
> ## add the profile log-likelihood points:
> lines(lambda.grid, d[2] / (lambda.grid * y[2]), col = 2)
> lines(dtot/(y[1] + theta.grid * y[2]), theta.grid, col = 3)
> ## the profile log-likelihood functions:
> prof.lambda <- function(lambda){
      return(d[1] * log(lambda) - lambda * y[1])
  }
> prof.theta <- function(theta){
      return(-dtot * log(y[1] + theta*y[2]) + d[2] * log(theta))
  }
> ## plot them separately:
> plot(lambda.grid, prof.lambda(lambda.grid), xlab = expression(lambda),
       ylab = expression(l[p](lambda)), col = 2, type = "l")
> abline(v=mle[1], lty=2)
> plot(theta.grid, prof.theta(theta.grid), xlab = expression(theta),
       ylab = expression(l[p](theta)), col = 3, type = "l")
> abline(v=mle[2], lty=2)
```



e) Compute a 95% Wald confidence interval for $\log(\theta)$ based on the profile log-likelihood. What can you say about the P-value for the null hypothesis $\lambda_1 = \lambda_2$?

▶ *First we derive the standard error of $\hat{\theta}$, by computing the negative curvature of the profile log-likelihood $l_p(\theta)$:*

$$I_p(\theta) = -\frac{d}{d\theta}\left\{\frac{d}{d\theta} l_p(\theta)\right\} = -\frac{d}{d\theta}\left(-\frac{DY_2}{Y_1 + \theta Y_2} + \frac{D_2}{\theta}\right) = -\frac{DY_2^2}{(Y_1 + \theta Y_2)^2} + \frac{D_2}{\theta^2}.$$

The profile likelihood is maximised at the MLE $\hat{\theta} = D_2 Y_1/(D_1 Y_2)$, and the negative curvature there is

$$I_p(\hat{\theta}) = \frac{D_1^3 Y_2^2}{D Y_1^2 D_2}.$$

Note that, by Result 5.1, we could have obtained the same expression by inverting the observed Fisher information matrix $\boldsymbol{I}(\hat{\lambda}, \hat{\theta})$ and taking the reciprocal of the second diagonal value. The standard error of $\hat{\theta}$ is thus

$$\operatorname{se}(\hat{\theta}) = \{I_p(\hat{\theta})\}^{-\frac{1}{2}} = \frac{Y_1 \sqrt{D_2 D}}{Y_2 D_1 \sqrt{D_1}}$$

and by the delta method we get the standard error for $\hat{\phi} = \log(\hat{\theta})$:

$$\operatorname{se}(\hat{\phi}) = \operatorname{se}(\hat{\theta}) \cdot (\hat{\theta})^{-1} = \frac{Y_1 \sqrt{D_2 D}}{Y_2 D_1 \sqrt{D_1}} \cdot \frac{D_1 Y_2}{D_2 Y_1} = \sqrt{\frac{D}{D_1 D_2}}.$$

The 95% Wald confidence interval for ϕ is hence given by

$$\left[\log\left(\frac{D_2 Y_1}{D_1 Y_2}\right) - z_{0.975} \cdot \sqrt{\frac{D}{D_1 D_2}}, \log\left(\frac{D_2 Y_1}{D_1 Y_2}\right) + z_{0.975} \cdot \sqrt{\frac{D}{D_1 D_2}}\right]$$

and for our data equals:

```
> (phiCi <- log(d[2] * y[1] / d[1] / y[2]) +
             c(-1, +1) * qnorm(0.975) * sqrt(dtot / d[1] / d[2]))
```

```
[1] 0.2955796 1.5008400
```

Since ϕ is the log relative incidence rate, and zero is not contained in the 95% confidence interval $(0.296, 1.501)$, the corresponding P-value for testing the null hypothesis $\phi = 0$, which is equivalent to $\theta = 1$ and $\lambda_1 = \lambda_2$, must be smaller than $\alpha = 5\%$.

Note that the use of (asymptotic) results from the likelihood theory can be justified here by considering the Poisson distribution $\mathrm{Po}(n)$ *as the distribution of a sum of n independent Poisson random variables with unit rates, as in the solution to 1a).*

2. Let $\boldsymbol{Z}_{1:n}$ be a random sample from a bivariate normal distribution $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = \boldsymbol{0}$ and covariance matrix

$$\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

a) Interpret σ^2 and ρ. Derive the MLE $(\hat{\sigma}^2_{\mathrm{ML}}, \hat{\rho}_{\mathrm{ML}})$.

▶ *σ^2 is the variance of each of the components X_i and Y_i of the bivariate vector $\boldsymbol{Z}_i$. The components have correlation ρ.*

To derive the MLE, we first compute the log-likelihood kernel

$$l(\boldsymbol{\Sigma}) = \sum_{i=1}^{n} -\frac{1}{2}\left\{\log|\boldsymbol{\Sigma}| + (x_i, y_i)\boldsymbol{\Sigma}^{-1}\begin{pmatrix} x_i \\ y_i \end{pmatrix}\right\}.$$

In our case, since $|\boldsymbol{\Sigma}| = \sigma^4(1-\rho^2)$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2(1-\rho^2)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix},$$

we obtain

$$l(\sigma^2, \rho) = -\frac{n}{2}\log\{\sigma^4(1-\rho^2)\} - \frac{1}{2\sigma^2(1-\rho^2)}Q(\rho),$$

where $Q(\rho) = \sum_{i=1}^n (x_i^2 - 2\rho x_i y_i + y_i^2)$. The score function thus has the components

$$\frac{d}{d\sigma^2}l(\sigma^2, \rho) = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4(1-\rho^2)}Q(\rho)$$

$$\text{and} \quad \frac{d}{d\rho}l(\sigma^2, \rho) = \frac{n\rho}{1-\rho^2} + \frac{1}{\sigma^2(1-\rho^2)}\left\{\sum_{i=1}^{n} x_i y_i - \frac{\rho}{1-\rho^2}Q(\rho)\right\}.$$

The first component of the score equation can be rewritten as

$$\sigma^2 = \frac{1}{2n(1-\rho^2)}Q(\rho),$$

which, plugged into the second component of the score equation, yields

$$\frac{2\sum_{i=1}^n x_i y_i}{Q(\rho)} = \frac{\rho}{1-\rho^2}.$$

The equations are solved by

$$\hat{\rho}_{\mathrm{ML}} = \frac{\sum_{i=1}^n x_i y_i}{\frac{1}{2}\sum_{i=1}^n (x_i^2 + y_i^2)}$$

$$\text{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{2n}\sum_{i=1}^{n}(x_i^2 + y_i^2).$$

As the observed Fisher information matrix shown below is positive definite, the above estimators are indeed the MLEs.

b) Show that the Fisher information matrix is

$$\boldsymbol{I}(\hat{\sigma}^2_{\mathrm{ML}}, \hat{\rho}_{\mathrm{ML}}) = \begin{pmatrix} \frac{n}{\hat{\sigma}^4_{\mathrm{ML}}} & -\frac{n\hat{\rho}_{\mathrm{ML}}}{\hat{\sigma}^2_{\mathrm{ML}}(1-\hat{\rho}^2_{\mathrm{ML}})} \\ -\frac{n\hat{\rho}_{\mathrm{ML}}}{\hat{\sigma}^2_{\mathrm{ML}}(1-\hat{\rho}^2_{\mathrm{ML}})} & \frac{n(1+\hat{\rho}^2_{\mathrm{ML}})}{(1-\hat{\rho}^2_{\mathrm{ML}})^2} \end{pmatrix}.$$

▶ *The components of the Fisher information matrix $\boldsymbol{I}(\sigma^2, \rho)$ are computed as*

$$-\frac{d^2}{d(\sigma^2)^2}l(\sigma^2, \rho) = \frac{Q(\rho) - n\sigma^2(1-\rho^2)}{\sigma^6(1-\rho^2)},$$

$$-\frac{d^2}{d\sigma^2\, d\rho}l(\sigma^2, \rho) = \frac{\sum_{i=1}^n x_i y_i(1-\rho^2) - \rho Q(\rho)}{\sigma^4(1-\rho^2)^2},$$

$$\text{and} \quad -\frac{d^2}{d\rho^2}l(\sigma^2, \rho) = \frac{(1-\rho^2)Q(\rho) - n\sigma^2(1-\rho^4) - 4\rho\sum_{i=1}^n(\rho y_i - x_i)(\rho x_i - y_i)}{\sigma^2(1-\rho^2)^3}.$$

and those of the observed Fisher information matrix $\boldsymbol{I}(\hat{\sigma}^2_{\mathrm{ML}}, \hat{\rho}_{\mathrm{ML}})$ are obtained by plugging in the MLEs. The wished-for expressions can be obtained by simple algebra, using that

$$\sum_{i=1}^{n}\{(\hat{\rho}_{\mathrm{ML}}y_i - x_i)(\hat{\rho}_{\mathrm{ML}}x_i - y_i)\} = n\hat{\rho}_{\mathrm{ML}}\hat{\sigma}^2_{\mathrm{ML}}(\hat{\rho}^2_{\mathrm{ML}} - 1).$$

The computations can also be performed in a suitable software.

c) Show that

$$\mathrm{se}(\hat{\rho}_{\mathrm{ML}}) = \frac{1-\hat{\rho}^2_{\mathrm{ML}}}{\sqrt{n}}.$$

▶ *Using the expression for the inversion of a 2×2 matrix, cf. Appendix B.1.1, we obtain that the element I^{22} of the inversed observed Fisher information matrix $\boldsymbol{I}(\hat{\sigma}^2_{\mathrm{ML}}, \hat{\rho}_{\mathrm{ML}})^{-1}$ is*

$$\left\{\frac{n^2(1+\hat{\rho}^2_{\mathrm{ML}})}{\hat{\sigma}^4_{\mathrm{ML}}(1-\hat{\rho}^2_{\mathrm{ML}})^2} - \frac{n^2\hat{\rho}^2_{\mathrm{ML}}}{\hat{\sigma}^4_{\mathrm{ML}}(1-\hat{\rho}^2_{\mathrm{ML}})^2}\right\}^{-1} \cdot \frac{n}{\hat{\sigma}^4_{\mathrm{ML}}} = \frac{(1-\hat{\rho}^2_{\mathrm{ML}})^2}{n}.$$

The standard error of $\hat{\rho}_{\mathrm{ML}}$ is the square root of this expression.

3. Calculate again the Fisher information of the profile log-likelihood in Result 5.2, but this time without using Result 5.1. Use instead the fact that $\hat{\alpha}_{\text{ML}}(\delta)$ is a point where the partial derivative of $l(\alpha, \delta)$ with respect to α equals zero.

▶ *Suppose again that the data are split in two independent parts (denoted by 0 and 1), and the corresponding likelihoods are parametrised by α and β, respectively. Then the log-likelihood decomposes as*

$$l(\alpha, \beta) = l_0(\alpha) + l_1(\beta).$$

We are interested in the difference $\delta = \beta - \alpha$. Obviously $\beta = \alpha + \delta$, so the joint log-likelihood of α and δ is

$$l(\alpha, \delta) = l_0(\alpha) + l_1(\alpha + \delta).$$

Furthermore,

$$\begin{aligned}\frac{d}{d\alpha} l(\alpha, \delta) &= \frac{d}{d\alpha} l_0(\alpha) + \frac{d}{d\alpha} l_1(\alpha + \delta)\\ &= S_0(\alpha) + S_1(\alpha + \delta),\end{aligned}$$

where S_0 and S_1 are the score functions corresponding to l_0 and l_1, respectively. For the profile log-likelihood $l_p(\delta) = l\{\hat{\alpha}_{\text{ML}}(\delta), \delta\}$, we need the value $\hat{\alpha}_{\text{ML}}(\delta)$ for which $\frac{d}{d\alpha} l\{\hat{\alpha}_{\text{ML}}(\delta), \delta\} = 0$, hence it follows that $S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\} = -S_0\{\hat{\alpha}_{\text{ML}}(\delta)\}$. This is also the derivative of the profile log-likelihood, because

$$\begin{aligned}\frac{d}{d\delta} l_p(\delta) &= \frac{d}{d\delta} l\{\hat{\alpha}_{\text{ML}}(\delta), \delta\}\\ &= \frac{d}{d\delta} l_0\{\hat{\alpha}_{\text{ML}}(\delta)\} + \frac{d}{d\delta} l_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\\ &= S_0\{\hat{\alpha}_{\text{ML}}(\delta)\} \frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta) + S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\left\{\frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta) + 1\right\}\\ &= [S_0\{\hat{\alpha}_{\text{ML}}(\delta)\} + S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}] \frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta) + S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\\ &= S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}.\end{aligned}$$

The Fisher information (negative curvature) of the profile log-likelihood is given by

$$\begin{aligned}I_p(\delta) &= -\frac{d^2}{d\delta^2} l_p(\delta)\\ &= -\frac{d}{d\delta} S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\\ &= I_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\left\{\frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta) + 1\right\},\end{aligned} \tag{5.2}$$

which is equal to

$$\begin{aligned}I_p(\delta) &= \frac{d}{d\delta} S_0\{\hat{\alpha}_{\text{ML}}(\delta)\}\\ &= -I_0\{\hat{\alpha}_{\text{ML}}(\delta)\} \frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta)\end{aligned}$$

as $S_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\} = -S_0\{\hat{\alpha}_{\text{ML}}(\delta)\}$. Here I_0 and I_1 denote the Fisher information corresponding to l_0 and l_1, respectively. Hence, we can solve

$$I_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}\left\{\frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta) + 1\right\} = -I_0\{\hat{\alpha}_{\text{ML}}(\delta)\} \frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta)$$

for $\frac{d}{d\delta}\hat{\alpha}_{\text{ML}}(\delta)$, and plug the result into (5.2) to finally obtain

$$\frac{1}{I_p(\delta)} = \frac{1}{I_0\{\hat{\alpha}_{\text{ML}}(\delta)\}} + \frac{1}{I_1\{\hat{\alpha}_{\text{ML}}(\delta) + \delta\}}. \tag{5.3}$$

4. Let $X \sim \text{Bin}(m, \pi_x)$ and $Y \sim \text{Bin}(n, \pi_y)$ be independent binomial random variables. In order to analyse the null hypothesis $H_0 : \pi_x = \pi_y$ one often considers the *relative risk* $\theta = \pi_x / \pi_y$ or the *log relative risk* $\psi = \log(\theta)$.

a) Compute the MLE $\hat{\psi}_{\text{ML}}$ and its standard error for the log relative risk estimation. Proceed as in Example 5.8.

▶ *As in Example 5.8, we may use the invariance of the MLEs to conclude that*

$$\hat{\theta}_{\text{ML}} = \frac{\hat{\pi}_x}{\hat{\pi}_y} = \frac{x_1/m}{x_2/n} = \frac{n x_1}{m x_2} \quad \textit{and} \quad \hat{\psi}_{\text{ML}} = \log\left(\frac{n x_1}{m x_2}\right).$$

Further, $\psi = \log(\theta) = \log(\pi_x) - \log(\pi_y)$, so we can use Result 5.2 to derive the standard error of $\hat{\psi}_{\text{ML}}$. In Example 2.10, we derived the observed Fisher information corresponding to the MLE $\hat{\pi}_{\text{ML}} = x/n$ as

$$I(\hat{\pi}_{\text{ML}}) = \frac{n}{\hat{\pi}_{\text{ML}}(1 - \hat{\pi}_{\text{ML}})}.$$

Using Result 2.1, we obtain that

$$I\{\log(\hat{\pi}_{\text{ML}})\} = I(\hat{\pi}_{\text{ML}}) \cdot \left\{\frac{1}{\hat{\pi}_{\text{ML}}}\right\}^{-2} = \frac{n}{\hat{\pi}_{\text{ML}}(1 - \hat{\pi}_{\text{ML}})} \cdot \hat{\pi}_{\text{ML}}^2.$$

By Result 5.2, we thus have that

$$\text{se}(\hat{\phi}_{\text{ML}}) = \sqrt{I\{\log(\hat{\pi}_x)\}^{-1} + I\{\log(\hat{\pi}_y)\}^{-1}} = \sqrt{\frac{1 - \hat{\pi}_x}{m\hat{\pi}_x} + \frac{1 - \hat{\pi}_y}{n\hat{\pi}_y}}.$$

b) Compute a 95% confidence interval for the relative risk θ given the data in Table 3.1.

▶ *The estimated risk for preeclampsia in the Diuretics group is $\hat{\pi}_x = 6/108 = 0.056$, and in the Placebo group it is $\hat{\pi}_y = 2/103 = 0.019$. The log-relative risk is thus estimated by*

$$\hat{\psi}_{\mathrm{ML}} = \log(6/108) - \log(2/103) = 1.051,$$

with the 95% confidence interval

$$[1.051 - 1.96 \cdot 0.648, 1.051 + 1.96 \cdot 0.648] = [-0.218, 2.321].$$

Back-transformation to the relative risk scale by exponentiating gives the following 95% confidence interval for the relative risk $\theta = \pi_1/\pi_2$: $[0.804, 10.183]$.

c) Also compute the profile likelihood and the corresponding 95% profile likelihood confidence interval for θ.

▶ *The joint log-likelihood for $\theta = \pi_x/\pi_y$ and π_y is*

$$l(\theta, \pi_y) = x\log(\theta) + (m - x)\log(1 - \theta\pi_y) + (x + y)\log(\pi_y) + (n - y)\log(1 - \pi_y).$$

In order to compute the profile likelihood, we need to maximise $l(\theta, \pi_y)$ with respect to π_y for fixed θ. To do this, we look for the points where

$$\frac{d}{d\pi_y} l(\theta, \pi_y) = \frac{\theta(m - x)}{\theta\pi_y - 1} + \frac{x + y}{\pi_y} + \frac{n - y}{\pi_y - 1}$$

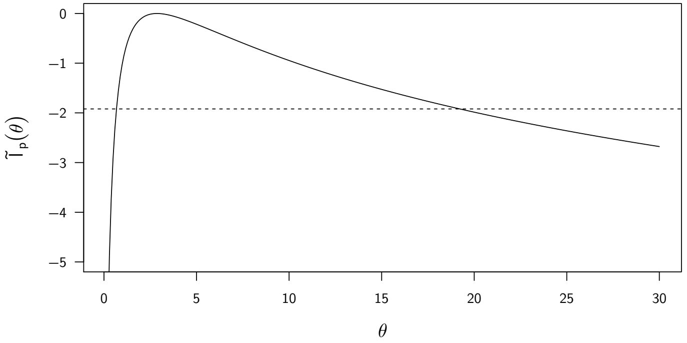
equals zero. To this aim, we need to solve the quadratic equation

$$\pi_y^2\theta(m + n) - \pi_y\big\{\theta(m + y) + n + x\big\} + x + y = 0$$

for π_y. In this case, it is easier to perform the numerical maximisation.

```
> ## data
> x <- c(6, 2)
> n <- c(108, 103)
> ## MLE
> piMl <- x/n
> thetaMl <- piMl[1] / piMl[2]
> ## log-likelihood of theta and pi2
> loglik <- function(theta, pi2)
  {
      pi <- c(theta * pi2, pi2)
      sum(dbinom(x, n, pi, log = TRUE))
  }
> ## implementing the gradient in pi2
> grad <- function(theta, pi2)
  {
      (theta * (n[1] - x[1])) / (theta * pi2 - 1) +
          sum(x) / pi2 +
              (n[2] - x[2]) / (pi2 - 1)
  }
```

```
> ## profile log-likelihood of theta:
> profilLoglik <- function(theta)
  {
      res <- theta
      eps <- sqrt(.Machine$double.eps)

      for(i in seq_along(theta)){      # the function can handle vectors
          optimResult <-
              optim(par = 0.5,
                    fn = function(pi2) loglik(theta[i], pi2),
                    gr = function(pi2) grad(theta[i], pi2),
                    method = "L-BFGS-B", lower = eps, upper = 1/theta[i] - eps,
                    control = list(fnscale = -1))
          if(optimResult$convergence == 0){ # has the algorithm converged?
              res[i] <- optimResult$value   # only then save the value (not the parameter!)
          } else {
              res[i] <- NA                  # otherwise return NA
          }
      }

      return(res)
  }
> ## plot the normed profile log-likelihood:
> thetaGrid <- seq(from = 0.1, to = 30, length = 309)
> normProfVals <- profilLoglik(thetaGrid) - profilLoglik(thetaMl)
> plot(thetaGrid, normProfVals,
       type = "l", ylim = c(-5, 0),
       xlab = expression(theta), ylab = expression(tilde(l)[p](theta))
       )
> ## show the cutpoint:
> abline(h = - 1/2 * qchisq(0.95, 1), lty = 2)
```



The downward outliers are artefacts of numerical optimisation. We therefore compute the profile likelihood confidence intervals using the function programmed in Example 4.19. Note that they can be approximated from the positions of the cut-off values in the graph above.

```
> ## general function that takes a given likelihood
> likelihoodCi <- function(
                           alpha = 0.05,  # 1-alpha ist the level of the interval
                           loglik,        # log-likelihood (not normed)
                           thetaMl,       # MLE
                           lower,         # lower boundary of the parameter space
                           upper,         # upper boundary of the parameter space
                           ...            # additional arguments for the loglik (e.g.
                                          # data)
                           )
  {
      ## target function
      f <- function(theta, ...)
          loglik(theta, ...) - loglik(thetaMl, ...) + 1/2*qchisq(1-alpha, df=1)

      ## determine the borders of the likelihood interval
      eps <- sqrt(.Machine$double.eps)      # stay a little from the boundaries
      lowerBound <- uniroot(f, interval = c(lower + eps, thetaMl), ...)$root
      upperBound <- uniroot(f, interval = c(thetaMl, upper - eps), ...)$root

      return(c(lower = lowerBound, upper = upperBound))
  }
> thetaProfilCi <-
      likelihoodCi(alpha = 0.05, loglik = profilLoglik, thetaMl = thetaMl,
                   lower = 0.01, upper = 20)
> thetaProfilCi
     lower      upper
 0.6766635 19.2217991
```

In comparison to the Wald confidence intervals with logarithmic transformation, the interval is almost twice as wide and implies therefore bigger uncertainty in the estimation of θ. Also this time it does not appear (at the 5% level) that the use of diuretics is associated with a higher risk for preeclampsia.

5. Suppose that ML estimates of *sensitivity* π_x and *specificity* π_y of a diagnostic test for a specific disease are obtained from independent binomial samples $X \sim \text{Bin}(m, \pi_x)$ and $Y \sim \text{Bin}(n, \pi_y)$, respectively.

a) Use Result 5.2 to compute the standard error of the logarithm of the *positive* and *negative likelihood ratio*, defined as $\text{LR}^+ = \pi_x/(1-\pi_y)$ and $\text{LR}^- = (1-\pi_x)/\pi_y$. Suppose $m = n = 100$, $x = 95$ and $y = 90$. Compute a point estimate and the limits of a 95% confidence interval for both LR^+ and LR^-, using the standard error from above.

▶ *As in Example 5.8, we may use the invariance of the MLEs to conclude that*

$$\widehat{LR^+} = \frac{\hat{\pi}_x}{1-\hat{\pi}_y} = \frac{x_1/m}{1-x_2/n} = \frac{nx_1}{nm - mx_2}$$

$$\textit{and} \quad \widehat{LR^-} = \frac{1-\hat{\pi}_x}{\hat{\pi}_y} = \frac{1-x_1/m}{x_2/n} = \frac{nm - nx_1}{mx_2}.$$

Now,

$$\log(LR^+) = \log(\pi_x) - \log(1-\pi_y) \quad \textit{and} \quad \log(LR^-) = \log(1-\pi_x) - \log(\pi_y),$$

so we can use Result 5.2 to derive the standard errors of $\widehat{LR^+}$ and $\widehat{LR^-}$. By Example 2.10, the observed Fisher information corresponding to the MLE $\hat{\pi}_{\text{ML}} = x/n$ is

$$I(\hat{\pi}_{\text{ML}}) = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}.$$

Using Result 2.1, we obtain that

$$I\{\log(\hat{\pi}_{\text{ML}})\} = I(\hat{\pi}_{\text{ML}}) \cdot \left\{\frac{1}{\hat{\pi}_{\text{ML}}}\right\}^{-2} = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})} \cdot \hat{\pi}_{\text{ML}}^2,$$

and

$$I\{\log(1-\hat{\pi}_{\text{ML}})\} = I(\hat{\pi}_{\text{ML}}) \cdot \left\{-\frac{1}{1-\hat{\pi}_{\text{ML}}}\right\}^{-2} = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})} \cdot (1-\hat{\pi}_{\text{ML}})^2.$$

By Result 5.2, we thus have that

$$\text{se}\{\log(\widehat{LR^+})\} = \sqrt{I\{\log(\hat{\pi}_x)\}^{-1} + I\{\log(1-\hat{\pi}_y)\}^{-1}} = \sqrt{\frac{1-\hat{\pi}_x}{m\hat{\pi}_x} + \frac{\hat{\pi}_y}{n(1-\hat{\pi}_y)}}$$

and

$$\text{se}\{\log(\widehat{LR^-})\} = \sqrt{I\{\log(1-\hat{\pi}_x)\}^{-1} + I\{\log(\hat{\pi}_y)\}^{-1}} = \sqrt{\frac{\hat{\pi}_x}{m(1-\hat{\pi}_x)} + \frac{1-\hat{\pi}_y}{n\hat{\pi}_y}}.$$

The estimated positive likelihood ratio is $\widehat{LR^+} = (95/100)/(1-90/100) = 9.5$, and the estimated negative likelihood ratio is $\widehat{LR^-} = (1-95/100)/(90/100) = 0.056$. Their logarithms are thus estimated by $\log(\widehat{LR^+}) = \log(9.5) = 2.251$, $\log(\widehat{LR^-}) = \log(0.056) = -2.89$, with the 95% confidence intervals

$$[2.251 - 1.96 \cdot 0.301, 2.251 + 1.96 \cdot 0.301] = [1.662, 2.841].$$

and

$$[-2.89 - 1.96 \cdot 0.437, -2.89 + 1.96 \cdot 0.437] = [-3.747, -2.034].$$

Back-transformation to the positive and negative likelihood ratio scale by exponentiating gives the following 95% confidence intervals $[5.268, 17.133]$ and $[0.024, 0.131]$.

b) The *positive predictive value* PPV is the probability of disease, given a positive test result. The equation

$$\frac{\text{PPV}}{1-\text{PPV}} = \text{LR}^+ \cdot \omega$$

relates PPV to LR^+ and to the pre-test odds of disease ω. Likewise, the following equation holds for the *negative predictive value* NPV, the probability to be disease-free, given a negative test result:

$$\frac{1-\text{NPV}}{\text{NPV}} = \text{LR}^- \cdot \omega.$$

Suppose $\omega = 1/1000$. Use the 95% confidence interval for LR$^+$ and LR$^-$, obtained in 5a), to compute the limits of a 95% confidence interval for both PPV and NPV.

▶ *By multiplying the endpoints of the confidence intervals for* LR^+ *and* LR^- *obtained in 5a) by* ω*, we obtain confidence intervals for* $PPV/(1-PPV)$ *and* $(1-NPV)/NPV$*, respectively. It remains to transform these intervals to the scales of PPV and NPV. Now, if the confidence interval for* $PPV/(1-PPV)$ *is* $[l,u]$*, then the confidence interval for PPV is* $[l/(1+l), u/(1+u)]$*; and if the confidence interval for* $(1-NPV)/NPV$ *is* $[l,u]$*, then the confidence interval for NPV is* $[1/(1+u), 1/(1+l)]$*. With our data, we obtain the following confidence intervals for* $PPV/(1-PPV)$ *and* $(1-NPV)/NPV$*, respectively:* $[0.00527, 0.01713]$ *and* $[2.4e-05, 0.000131]$*; and the following confidence intervals for PPV and NPV, respectively:* $[0.00524, 0.01684]$ *and* $[0.999869, 0.999976]$.

6. In the placebo-controlled clinical trial of diuretics during pregnancy to prevent preeclampsia by Fallis *et al.* (*cf*. Table 1.1), 6 out of 38 treated women and 18 out of 40 untreated women got preeclampsia.

a) Formulate a statistical model assuming independent binomial distributions in the two groups. Translate the null hypothesis “there is no difference in preeclampsia risk between the two groups” into a statement on the model parameters.

▶ *We consider two independent random variables* $X_1 \sim \mathrm{Bin}(n_1, \pi_1)$ *and* $X_2 \sim \mathrm{Bin}(n_2, \pi_2)$ *modelling the number of preeclampsia cases in the treatment and control group, respectively. Here the sample sizes are* $n_1 = 38$ *and* $n_2 = 40$*; and we have observed the realisations* $x_1 = 6$ *and* $x_2 = 18$*. No difference in preeclampsia risk between the two groups is expressed in the hypothesis* $H_0: \pi_1 = \pi_2$*.*

b) Let θ denote the risk difference between treated and untreated women. Derive the MLE $\hat{\theta}_{\mathrm{ML}}$ and a 95% Wald confidence interval for θ. Also give the MLE for the number needed to treat (NNT) which is defined as $1/\theta$.

▶ *In terms of the model parameters, we have* $\theta = \pi_1 - \pi_2$*. By the invariance of the MLEs, we obtain the MLE of* θ *as*

$$\hat{\theta}_{\mathrm{ML}} = \hat{\pi}_1 - \hat{\pi}_2 = x_1/n_1 - x_2/n_2.$$

To derive the standard error, we may use Result 5.2 and Example 2.10 to conclude that

$$\widehat{\mathrm{se}}(\hat{\theta}_{\mathrm{ML}}) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

It follows that a 95% Wald confidence interval for θ *is given by*

$$\left[\hat{\pi}_1 - \hat{\pi}_2 - z_{0.975}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}, \hat{\pi}_1 - \hat{\pi}_2 + z_{0.975}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}\right].$$

We can also get the MLE for the number needed to treat as

$$\widehat{NNT} = \frac{1}{\left|\hat{\theta}_{\mathrm{ML}}\right|} = \frac{1}{|x_1/n_1 - x_2/n_2|}.$$

For the concrete data example:

```
> ## the data
> x <- c(6, 18)
> n <- c(38, 40)
> ## MLEs
> pi1Hat <- x[1] / n[1]
> pi2Hat <- x[2] / n[2]
> (thetaHat <- pi1Hat - pi2Hat)
[1] -0.2921053
> (nntHat <- 1 / abs(thetaHat))
[1] 3.423423
> ## Wald CI
> seTheta <- sqrt(pi1Hat * (1 - pi1Hat) / n[1] +
                  pi2Hat * (1 - pi2Hat) / n[2])
> (thetaWald <- thetaHat + c(-1, 1) * qnorm(0.975) * seTheta)
[1] -0.48500548 -0.09920505
```

c) Write down the joint log-likelihood kernel $l(\pi_1, \theta)$ of the risk parameter π_1 in the treatment group and θ. In order to derive the profile log-likelihood of θ, consider θ as fixed and write down the score function for π_1,

$$S_{\pi_1}(\pi_1, \theta) = \frac{\partial}{\partial \pi_1} l(\pi_1, \theta).$$

Which values are allowed for π_1 when θ is fixed?

▶ *The joint likelihood kernel of* π_1 *and* π_2 *is*

$$L(\pi_1, \pi_2) = \pi_1^{x_1}(1-\pi_1)^{n_1-x_1} \cdot \pi_2^{x_2}(1-\pi_2)^{n_2-x_2}.$$

We can rewrite the risk in the control group as $\pi_2 = \pi_1 - \theta$*. Plugging this into the log-likelihood kernel of* π_1 *and* π_2 *gives the joint log-likelihood kernel of risk in the treatment group* π_1 *and the risk difference* θ *as*

$$l(\pi_1, \theta) = x_1 \log(\pi_1) + (n_1 - x_1)\log(1-\pi_1) + x_2 \log(\pi_1 - \theta) + (n_2 - x_2)\log(1 - \pi_1 + \theta).$$

The score function for π_1 *is thus given by*

$$\begin{aligned} S_{\pi_1}(\pi_1, \theta) &= \frac{d}{d\pi_1} l(\pi_1, \theta) \\ &= \frac{x_1}{\pi_1} + \frac{x_1 - n_1}{1 - \pi_1} + \frac{x_2}{\pi_1 - \theta} + \frac{x_2 - n_2}{1 - \pi_1 + \theta}. \end{aligned}$$

If we want to solve the score equation $S_{\pi_1}(\pi_1, \theta) = 0$*, we must think about the allowed range for* π_1*: Of course, we have* $0 < \pi_1 < 1$*. And we also have this for the second proportion* π_2*, giving* $0 < \pi_1 - \theta < 1$ *or* $\theta < \pi_1 < 1 + \theta$*. Altogether we have the range* $\max\{0, \theta\} < \pi_1 < \min\{1, 1 + \theta\}$*.*

d) Write an R-function which solves $S_{\pi_1}(\pi_1, \theta) = 0$ (use `uniroot`) and thus gives an estimate $\hat{\pi}_1(\theta)$. Hence write an R-function for the profile log-likelihood $l_p(\theta) = l\{\hat{\pi}_1(\theta), \theta\}$.

▶

```
> ## the score function for pi1:
> pi1score <- function(pi1, theta)
  {
      x[1] / pi1 + (x[1] - n[1]) / (1 - pi1) +
          x[2] / (pi1 - theta) + (x[2] - n[2]) / (1 - pi1 + theta)
  }
> ## get the MLE for pi1 given fixed theta:
> getPi1 <- function(theta)
  {
      eps <- 1e-9
      uniroot(pi1score,
              interval=
              c(max(0, theta) + eps,
                min(1, 1 + theta) - eps),
              theta=theta)$root
  }
> ## the joint log-likelihood kernel:
> loglik <- function(pi1, theta)
  {
      x[1] * log(pi1) + (n[1] - x[1]) * log(1 - pi1) +
          x[2] * log(pi1 - theta) + (n[2] - x[2]) * log(1 - pi1 + theta)
  }
> ## so we have the profile log-likelihood for theta:
> profLoglik <- function(theta)
  {
      pi1Hat <- getPi1(theta)
      loglik(pi1Hat, theta)
  }
```

e) Compute a 95% profile likelihood confidence interval for θ using numerical tools. Compare it with the Wald interval. What can you say about the P-value for the null hypothesis from 6a)?

▶

```
> ## now we need the relative profile log-likelihood,
> ## and for that the value at the MLE:
> profLoglikMle <- profLoglik(thetaHat)
> relProfLoglik <- function(theta)
  {
      profLoglik(theta) - profLoglikMle
  }
> ## now compute the profile CI bounds:
> lower <- uniroot(function(theta){relProfLoglik(theta) + 1.92},
                   c(-0.99, thetaHat))
> upper <- uniroot(function(theta){relProfLoglik(theta) + 1.92},
                   c(thetaHat, 0.99))
> (profLogLikCi <- c(lower$root, upper$root))
[1] -0.47766117 -0.09330746
> ## compare with Wald interval
> thetaWald
[1] -0.48500548 -0.09920505
```

Tab. 5.1: Probability of offspring's blood group given allele frequencies in parental generation, and sample realisations.

Blood group	Probability	Observation
A={AA,A0}	$\pi_1 = p^2 + 2pr$	$x_1 = 182$
B={BB,B0}	$\pi_2 = q^2 + 2qr$	$x_2 = 60$
AB={AB}	$\pi_3 = 2pq$	$x_3 = 17$
0={00}	$\pi_4 = r^2$	$x_4 = 176$

The two 95% confidence intervals are quite close, and neither contains the reference value zero. Therefore, the P-value for testing the null hypothesis of no risk difference between the two groups against the two-sided alternative must be smaller than 5%.

7. The AB0 blood group system was described by Karl Landsteiner in 1901, who was awarded the Nobel Prize for this discovery in 1930. It is the most important blood type system in human blood transfusion, and comprises four different groups: A, B, AB and 0.

Blood groups are inherited from both parents. Blood groups A and B are dominant over 0 and codominant to each other. Therefore a phenotype blood group A may have the genotype AA or A0, for phenotype B the genotype is BB or B0, for phenotype AB there is only the genotype AB, and for phenotype 0 there is only the genotype 00.

Let p, q and r be the proportions of alleles A, B, and 0 in a population, so $p+q+r = 1$ and $p, q, r > 0$. Then the probabilities of the four blood groups for the offspring generation are given in Table 5.1. Moreover, the realisations in a sample of size $n = 435$ are reported.

a) Explain how the probabilities in Table 5.1 arise. What assumption is tacitly made?

▶ *The core assumption is random mating, i. e. there are no mating restrictions, neither genetic or behavioural, upon the population, and that therefore all recombination is possible. We assume that the alleles are independent, so the probability of the haplotype a_1/a_2 (i. e. the alleles in the order mother/father) is given by $\mathsf{Pr}(a_1)\,\mathsf{Pr}(a_2)$ when $\mathsf{Pr}(a_i)$ is the frequency of allele a_i in the population. Then we look at the haplotypes which produce the requested phenotype, and sum their probabilities to get the probability for the requested phenotype. For example, phenotype A is produced by the haplotypes A/A, A/0 and 0/A, having probabilities $p \cdot p$, $p \cdot r$ and $r \cdot p$, and summing up gives π_1.*

b) Write down the log-likelihood kernel of $\boldsymbol{\theta} = (p, q)^\top$. To this end, assume that $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\top$ is a realisation from a multinomial distribution with parameters

$n = 435$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^\top$.

▶ *We assume that*

$$\boldsymbol{X} = (X_1, X_2, X_3, X_4)^\top \sim \mathrm{M}_4\left\{n = 435, \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^\top\right\}.$$

The log-likelihood kernel of $\boldsymbol{\pi}$ *is*

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{4} x_i \log(\pi_i),$$

and by inserting the parametrisation from Table 5.1, we obtain the log-likelihood kernel of p *and* q *as*

$$\begin{aligned} l(p, q) = {} & x_1 \log\{p^2 + 2p(1-p-q)\} + x_2 \log\{q^2 + 2q(1-p-q)\} \\ & + x_3 \log(2pq) + x_4 \log\{(1-p-q)^2\}. \end{aligned}$$

Note that we have used here that $r = 1 - p - q$*, so there are only two parameters in this problem.*

c) Compute the MLEs of p and q numerically, using the R function `optim`. Use the option `hessian = TRUE` in `optim` and process the corresponding output to receive the standard errors of $\hat{p}_{\mathrm{ML}}$ and $\hat{q}_{\mathrm{ML}}$.

▶

```
> ## observed data:
> data <- c(182, 60, 17, 176)
> n <- sum(data)
> ## the loglikelihood function of theta = (p, q)
> loglik <- function(theta,data) {
      p <- theta[1]
      q <- theta[2]
      r <- 1-p-q

      ## check for valid parameters:
      if ((p>0) && (p<1) && (r>0) && (r<1) && (q>0) && (q<1)) {
          probs <- c(p^2+2*p*r,q^2+2*q*r, 2*p*q, r^2)
          return(dmultinom(data,prob=probs,size=sum(data),log=TRUE))
      } else {                            # if not valid, return NA
          return(NA)
      }
  }
> ## numerically optimise the log-likelihood
> optimResult <- optim(c(0.1,0.3), loglik,
                       control = list(fnscale=-1),
                       hessian = TRUE,    # also return the hessian!
                       data = data)
> ## check convergence:
> optimResult[["convergence"]] == 0
[1] TRUE
> ## and extract MLEs and standard errors
> (thetaMl <- optimResult$par)
[1] 0.26442773 0.09317313
```

```
> (thetaCov <- solve(- optimResult$hessian))
              [,1]          [,2]
[1,]  0.0002639946 -0.0000280230
[2,] -0.0000280230  0.0001023817
> (thetaSe <- sqrt(diag(thetaCov)))
[1] 0.01624791 0.01011839
```

So we have $\hat{p}_{\mathrm{ML}} = 0.264$ *and* $\hat{q}_{\mathrm{ML}} = 0.093$ *with corresponding standard errors* $\mathrm{se}(\hat{p}_{\mathrm{ML}}) = 0.016$ *and* $\mathrm{se}(\hat{q}_{\mathrm{ML}}) = 0.01$.

d) Finally compute $\hat{r}_{\mathrm{ML}}$ and $\mathrm{se}(\hat{r}_{\mathrm{ML}})$. Make use of Section 5.4.3.

▶ *By the invariance of the MLE we have* $\hat{r}_{\mathrm{ML}} = 1 - \hat{p}_{\mathrm{ML}} - \hat{q}_{\mathrm{ML}} = 0.642$*. Moreover,*

$$r = 1 - p - q = 1 + (-1, -1)\begin{pmatrix} p \\ q \end{pmatrix} = g(p, q)$$

we can apply the multivariate delta method in the special case of a linear transformation $g(\boldsymbol{\theta}) = \boldsymbol{a}^\top \cdot \boldsymbol{\theta} + b$*, and we have* $D(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \boldsymbol{a}^\top$*. Thus,*

$$\mathrm{se}\big\{g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\big\} = \sqrt{\boldsymbol{a}^\top \boldsymbol{I}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})^{-1} \boldsymbol{a}}.$$

In our case, $\boldsymbol{\theta} = (p, q)^\top$, $\boldsymbol{a}^\top = (-1, -1)$*, and* $b = 1$*, so*

$$\mathrm{se}(\hat{r}_{\mathrm{ML}}) = \mathrm{se}\big(g(p, q)\big) = \sqrt{(-1, -1)\boldsymbol{I}(\hat{p}_{\mathrm{ML}}, \hat{q}_{\mathrm{ML}})^{-1}\begin{pmatrix} -1 \\ -1 \end{pmatrix}} = \sqrt{\sum_{i,j=1}^{2} \{\boldsymbol{I}(\hat{p}_{\mathrm{ML}}, \hat{q}_{\mathrm{ML}})^{-1}\}_{ij}}.$$

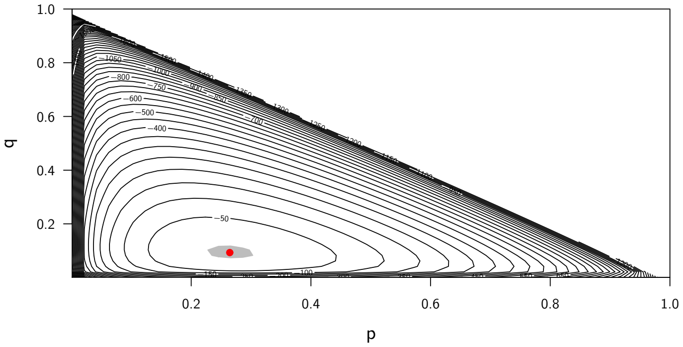
```
> (rMl <- 1 - sum(thetaMl))
[1] 0.6423991
> (rSe <- sqrt(sum(thetaCov)))
[1] 0.01761619
```

We thus have $\hat{r}_{\mathrm{ML}} = 0.642$ *and* $\mathrm{se}(\hat{r}_{\mathrm{ML}}) = 0.018$.

e) Create a contour plot of the relative log-likelihood and mark the 95% likelihood confidence region for $\boldsymbol{\theta}$. Use the R-functions `contourLines` and `polygon` for sketching the confidence region.

▶

```
> ## fill grid with relative log-likelihood values
> gridSize <- 50
> eps <- 1e-3
> loglikgrid <- matrix(NA, gridSize, gridSize)
> p <- seq(eps,1,length=gridSize)
> q <- seq(eps,1,length=gridSize)
> for (i in 1:length(p)) {
      for (j in 1:length(q)) {
          loglikgrid[i,j] <-
              loglik(c(p[i],q[j]),data=data) - loglik(thetaMl, data = data)
      }
  }
> ## plot
> contour(p,q,
          loglikgrid,
          nlevels = 50,
          xlab=expression (p),ylab= expression (q), xaxs = "i", yaxs = "i")
> ## add confidence region and MLE:
> region <- contourLines(x = p, y = q, z = loglikgrid, levels = log (0.05))[[1]]
> polygon(region$x, region$y, density = NA, col = "gray")
> points(thetaMl[1], thetaMl[2], pch = 19, col = 2)
```



f) Use the χ^2 and the G^2 test statistic to analyse the plausibility of the modeling assumptions.

▶ *The fitted ("expected") frequencies e_i in the restricted model are*

$$\begin{aligned} e_1 &= n(\hat{p}_{\text{ML}}^2 + 2\hat{p}_{\text{ML}}\hat{r}_{\text{ML}}) = 178.201, \\ e_2 &= n(\hat{q}_{\text{ML}}^2 + 2\hat{q}_{\text{ML}}\hat{r}_{\text{ML}}) = 55.85, \\ e_3 &= 2n\hat{p}_{\text{ML}}\hat{q}_{\text{ML}} = 21.435 \\ \textit{und} \quad e_4 &= n\hat{r}_{\text{ML}}^2 = 179.514. \end{aligned}$$

With the observed frequencies x_i, we can compute the two statistics as

$$G^2 = 2\sum_{i=1}^{4} x_i \log\left(\frac{x_i}{e_i}\right) \quad \textit{and} \quad \chi^2 = \sum_{i=1}^{4} \frac{(x_i - e_i)^2}{e_i}.$$

In R, we can compute the values as follows.

```
> ## values
> (x <- data)
[1] 182  60  17 176
> (e <- n * c(thetaMl[1]^2 + 2 * thetaMl[1] * rMl,
              thetaMl[2]^2 + 2 * thetaMl[2] * rMl,
              2 * prod(thetaMl),
              rMl^2
              ))
[1] 178.20137  55.84961  21.43468 179.51434
> ## Statistics
> (G2 <- 2 * sum(x * log(x / e)))
```

```
[1] 1.438987
> (Chi2 <- sum((x - e)^2 / x))
[1] 1.593398
```

We have $k = 4$ categories, i. e. 3 free probabilities, and $r = 2$ free parameters (p and q). Under the null hypothesis that the model is correct, both test statistics have asymptotically $\chi^2(1)$ distribution. The corresponding 95%-quantile is 3.84. Since neither test statistic exceeds this threshold, the null hypothesis cannot be rejected at the 5% significance level.

8. Let $T \sim \mathrm{t}(n-1)$ be a standard t random variable with $n-1$ degrees of freedom.

a) Derive the density function of the random variable

$$W = n \log\left(1 + \frac{T^2}{n-1}\right),$$

see Example 5.15, and compare it graphically with the density function of the $\chi^2(1)$ distribution for different values of n.

▶ *$W = g(T^2)$ is a one-to-one differentiable transformation of T^2, so we can apply the change-of-variables formula (Appendix A.2.3) to derive the density of W. We first need to derive the density of T^2 from that of T. This is not a one-to-one transformation and we will derive the density directly as follows. For $x \geq 0$, F_{T^2} the distribution function of T^2, and F_T the distribution function of T, we have*

$$\begin{aligned} F_{T^2}(x) &= \mathsf{Pr}(T^2 \leq x) \\ &= \mathsf{Pr}(|T| \leq \sqrt{x}) \\ &= \mathsf{Pr}(-\sqrt{x} \leq T \leq \sqrt{x}) \\ &= F_T(\sqrt{x}) - F_T(-\sqrt{x}) \\ &= 2F_T(\sqrt{x}) - 1. \end{aligned}$$

The last equality follows from the symmetry of the t distribution around 0. The density f_{T^2} of T^2 is thus

$$f_{T^2}(x) = \frac{d}{dx}F_{T^2}(x) = 2f_T(\sqrt{x}) \cdot \frac{1}{2}x^{-\frac{1}{2}} = \frac{f_T(\sqrt{x})}{\sqrt{x}},$$

where f_T is the density of T.
Now, the inverse transformation corresponding to g is

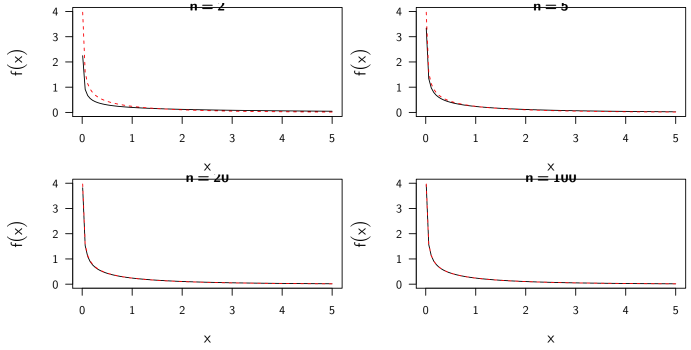
$$g^{-1}(w) = (n-1)\{\exp(w/n) - 1\}.$$

Using the change-of-variables formula, we obtain the following form for the density f_W of W:

$$\begin{aligned} f_W(x) &= f_{T^2}\{g^{-1}(x)\}\left|\frac{d}{dx}g^{-1}(x)\right| \\ &= \frac{f_T\left[\sqrt{(n-1)\{\exp(x/n)-1\}}\right]}{\sqrt{(n-1)\{\exp(x/n)-1\}}}\,\frac{n-1}{n}\exp(x/n). \end{aligned}$$

We can now draw the resulting density.

```
> ## implementation of the density
> densityW <- function(x, n)
  {
      y <- sqrt((n - 1) * (exp(x / n) - 1))
      dt(y, df = n - 1) / y * (n - 1) / n * exp(x / n)
  }
> ## testing the density through a comparison with the histogram
> ## set.seed(1234)
>
> ## n <- 10
> ## m <- 1e+5
> ## T2 <- rt(m, df = n - 1)^2
> ## W <- n * log(1 + T2 / (n - 1))
> ## hist(W[W < 5], breaks = 100, prob = TRUE)
> grid <- seq(from = 0.01, to = 5, length = 101)
> ## lines (grid, densityW(grid, n = n), col = 2)
>
> ## plot for different values of n
> par(mfrow = c(2, 2))
> for(n in c(2, 5, 20, 100)){
      ## density of W
      plot(grid, densityW(grid, n = n), type = "l",
           xlab = expression(x), ylab = expression(f(x)),
           ylim = c(0, 4), main = paste("n =", n)
           )
      ## density of chi^2(1)
      lines(grid, dchisq(grid, df = 1), lty = 2, col = 2)
  }
```



Indeed we can see that the differences between the densities are small already for $n = 20$.

b) Show that for $n \to \infty$, W follows indeed a $\chi^2(1)$ distribution.

▶ *Since $T \xrightarrow{D} \mathrm{N}(0,1)$ as $n \to \infty$, we have that $T^2 \xrightarrow{D} \chi^2(1)$ as $n \to \infty$. Further, the transformation g is for large n close to identity, as*

$$g(x) = \log\left\{\left(1 + \frac{x}{n-1}\right)^n\right\}$$

and $\{1 + x/(n-1)\}^n \to \exp(x)$ as $n \to \infty$. Altogether therefore $W = g(T^2) \xrightarrow{D} \chi^2(1)$ as $n \to \infty$.

9. Consider the χ^2 statistic given k categories with n observations. Let

$$D_n = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} \quad \text{and} \quad W_n = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{np_{i0}}\right).$$

Show that $W_n - D_n \xrightarrow{P} 0$ for $n \to \infty$.

▶ *In the notation of Section 5.5, we now have $n_i = x_i$ the observed frequencies and $np_{i0} = e_i$ the expected frequencies in a multinomial model with true probabilities π_i, and p_{i0} are maximum likelihood estimators under a certain model. If that model is true, then both $(n_i/n - \pi_i)$ and $(p_{i0} - \pi_i)$ converge to zero in probability and both $\sqrt{n}(n_i/n - \pi_i)$ and $\sqrt{n}(p_{i0} - \pi_i)$ are bounded in probability as $n \to \infty$, the first one by the central limit theorem and the second one by the standard likelihood theory.*

We can write

$$W_n = 2n\sum_{i=1}^{k} \frac{n_i}{n} \log\left(1 + \frac{n_i/n - p_{i0}}{p_{i0}}\right).$$

The Taylor expansion (cf. Appendix B.2.3) of $\log(1+x)$ around $x = 0$ yields

$$\log(1+x) = x - \frac{1}{2}x^2 + O(x^3),$$

cf. the Landau notation (Appendix B.2.6). By plugging this result into the equation above, we obtain that

$$\begin{aligned} W_n &= 2n\sum_{i=1}^{k}\left\{p_{i0} + \left(\frac{n_i}{n} - p_{i0}\right)\right\} \cdot \left[\frac{n_i/n - p_{i0}}{p_{i0}} - \frac{1}{2}\left(\frac{n_i/n - p_{i0}}{p_{i0}}\right)^2 + O\{(n_i/n - p_{i0})^3\}\right] \\ &= 2n\sum_{i=1}^{k}\left[(n_i/n - p_{i0}) + \frac{1}{2}\cdot\frac{(n_i/n - p_{i0})^2}{p_{i0}} + O\{(n_i/n - p_{i0})^3\}\right] \\ &= D_n + n\sum_{i=1}^{k} O\{(n_i/n - p_{i0})^3\}, \end{aligned}$$

where the last equality follows from the fact that both n_i/n and p_{i0} must sum up to one.

Since both $(n_i/n - \pi_i)$ and $(p_{i0} - \pi_i)$ converge to zero in probability and both $\sqrt{n}(n_i/n - \pi_i)$ and $\sqrt{n}(p_{i0} - \pi_i)$ are bounded in probability as $n \to \infty$, the last sum converges to zero in probability as $n \to \infty$, i. e. $W_n - D_n \xrightarrow{P} 0$ as $n \to \infty$.

10. In a psychological experiment the forgetfulness of probands is tested with the recognition of syllable triples. The proband has ten seconds to memorise the triple, afterwards it is covered. After a waiting time of t seconds it is checked whether the proband remembers the triple. For each waiting time t the experiment is repeated n times.

Let $y = (y_1, \dots, y_m)$ be the relative frequencies of correctly remembered syllable triples for the waiting times of $t = 1, \dots, m$ seconds. The *power model* now assumes, that

$$\pi(t; \boldsymbol{\theta}) = \theta_1 t^{-\theta_2}, \quad 0 \leq \theta_1 \leq 1, \theta_2 > 0,$$

is the probability to correctly remember a syllable triple after the waiting time $t \geq 1$.

a) Derive an expression for the log-likelihood $l(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

▶ *If the relative frequencies of correctly remembered triples are independent for different waiting times, then the likelihood kernel is*

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^{m} (\theta_1 t_i^{-\theta_2})^{ny_i} (1 - \theta_1 t_i^{-\theta_2})^{n - ny_i} \\ &= \theta_1^{n \sum_{i=1}^{m} y_i} \prod_{i=1}^{m} t_i^{-\theta_2 n y_i} (1 - \theta_1 t_i^{-\theta_2})^{n - ny_i}. \end{aligned}$$

The log-likelihood kernel is thus

$$l(\theta_1, \theta_2) = n \log(\theta_1) \sum_{i=1}^{m} y_i - n\theta_2 \sum_{i=1}^{m} y_i \log(t_i) + n \sum_{i=1}^{m} (1 - y_i) \log(1 - \theta_1 t_i^{-\theta_2}).$$

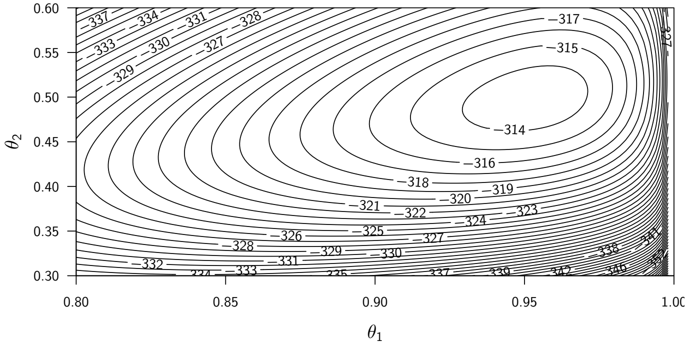
b) Create a contour plot of the log-likelihood in the parameter range $[0.8, 1] \times [0.3, 0.6]$ with $n = 100$ and

$$y = (0.94, 0.77, 0.40, 0.26, 0.24, 0.16), \; t = (1, 3, 6, 9, 12, 18).$$

▶

```
> loglik.fn <- function(theta1, theta2, n, y, t){
    return(
    n*log(theta1)*sum(y)
    - n*theta2*sum(y*log(t))
    + n*sum((1-y)*log(1-theta1*t^(-theta2)))
    )
  }
> y <- c(0.94,0.77,0.40,0.26,0.24,0.16)
> t <- c(1, 3, 6, 9, 12, 18)
```

```
> n <- 100
> gridSize <- 100
> theta1grid <- seq(0.8,1,length=gridSize)
> theta2grid <- seq(0.3,0.6,length=gridSize)
> loglik <- matrix(NA, nrow=length(theta1grid), ncol=length(theta2grid))
> for(i in 1:length(theta1grid)){
    for(j in 1:length(theta2grid)){
       loglik[i, j] <-
       loglik.fn(theta1=theta1grid[i], theta2=theta2grid[j], n=n, y=y, t=t)
       }
    }
> contour(theta1grid, theta2grid, loglik, nlevels=50, xlab=math (theta[1]),
  ylab=math (theta[2]), xaxs = "i", yaxs = "i", labcex=1)
```



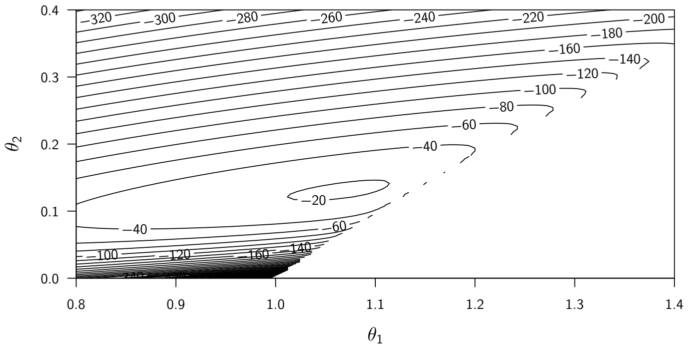
11. Often the *exponential model* is used instead of the power model (described in Exercise 10), assuming:

$$\pi(t; \boldsymbol{\theta}) = \min\{1, \theta_1 \exp(-\theta_2 t)\}, \quad t > 0, \theta_1 > 0 \text{ and } \theta_2 > 0.$$

a) Create a contour plot of the log-likelihood in the parameter range $[0.8, 1.4] \times [0, 0.4]$ for the same data as in Exercise 10.

▶

```
> pi.exp <- function(theta1, theta2, t){
    return(
    pmin( theta1*exp(-theta2*t), 1 )
    )
  }
> loglik.fn <- function(theta, n, y, t){
    return(
     sum( dbinom(x=n*y, size=n,
               prob=pi.exp(theta1=theta[1], theta2=theta[2], t=t), log=TRUE) )
    )
  }
> y <- c(0.94,0.77,0.40,0.26,0.24,0.16)
> t <- c(1, 3, 6, 9, 12, 18)
> n <- 100
> gridSize <- 100
> theta1grid <- seq(0.8,1.4,length=gridSize)
> theta2grid <- seq(0,0.4,length=gridSize)
> loglik <- matrix(NA, nrow=length(theta1grid), ncol=length(theta2grid))
> for(i in 1:length(theta1grid)){
    for(j in 1:length(theta2grid)){
       loglik[i, j] <-
       loglik.fn(theta=c(theta1grid[i], theta2grid[j]), n=n, y=y, t=t)
       }
    }
> contour(theta1grid, theta2grid, loglik, nlevels=50, xlab=math (theta[1]),
  ylab=math (theta[2]), xaxs = "i", yaxs = "i", labcex=1)
```

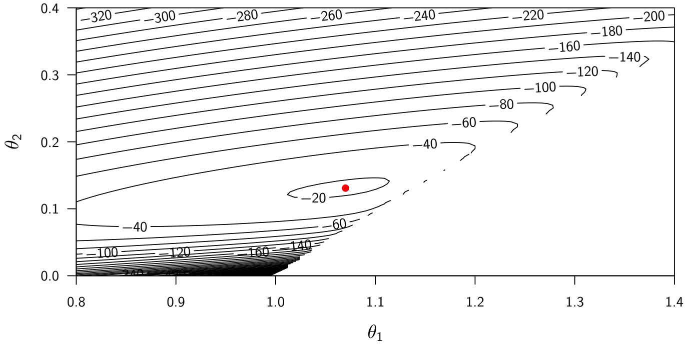


Note that we did not evaluate the likelihood values in the areas where $\pi(t;\boldsymbol{\theta}) = 1$ for some t. These are somewhat particular situations, because when $\pi = 1$, the corresponding likelihood contribution is 1, no matter what we observe as the corresponding y and no matter what the values of θ_1 and θ_2 are.

b) Use the R-function `optim` to numerically compute the MLE $\hat{\boldsymbol{\theta}}_{\text{ML}}$. Add the MLE to the contour plot from 11a).

▶

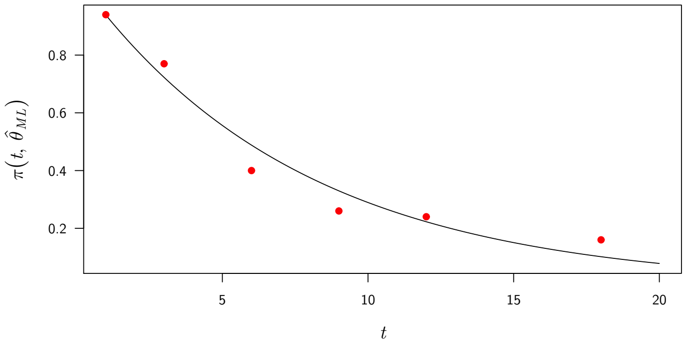
```
> ## numerically optimise the log-likelihood
> optimResult <- optim(c(1.0, 0.1), loglik.fn,
                       control = list(fnscale=-1),
                       n=n, y=y, t=t)
> ## check convergence:
> optimResult[["convergence"]] == 0
[1] TRUE
> ## extract the MLE
> thetaMl <- optimResult$par
> ## add the MLE to the plot
> contour(theta1grid, theta2grid, loglik, nlevels=50, xlab=math (theta[1]),
  ylab=math (theta[2]), xaxs = "i", yaxs = "i", labcex=1)
> points(thetaMl[1], thetaMl[2], pch = 19, col = 2)
```



c) For $0 \le t \le 20$ create a plot of $\pi(t; \hat{\boldsymbol{\theta}}_{\text{ML}})$ and add the observations y.

▶

```
> tt <- seq(1, 20, length=101)
> plot(tt, pi.exp(theta1=thetaMl[1], theta2=thetaMl[2], t=tt), type="l",
       xlab= math (t), ylab= math (pi(t, hat(theta)[ML])) )
> points(t, y, pch = 19, col = 2)
```



12. Let $X_{1:n}$ be a random sample from a log-normal $\text{LN}(\mu, \sigma^2)$ distribution, *cf.* Table A.2.

a) Derive the MLE of μ and σ^2. Use the connection between the densities of the normal distribution and the log-normal distribution. Also compute the corresponding standard errors.

▶ *We know that if X is normal, i. e. $X \sim \text{N}(\mu, \sigma^2)$, then $\exp(X) \sim \text{LN}(\mu, \sigma^2)$ (cf. Table A.2). Thus, if we have a random sample $X_{1:n}$ from log-normal distribution $\text{LN}(\mu, \sigma^2)$, then $Y_{1:n} = \{\log(X_1), \dots, log(X_n)\}$ is a random sample from normal distribution $\text{N}(\mu, \sigma^2)$. In Example 5.3 we computed the MLEs in the normal model:*

$$\hat{\mu}_{\text{ML}} = \bar{y} \quad \textit{and} \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

and in Section 5.2 we derived the corresponding standard errors $\operatorname{se}(\hat{\mu}_{\text{ML}}) = \hat{\sigma}_{\text{ML}}/\sqrt{n}$ *and* $\operatorname{se}(\hat{\sigma}^2_{\text{ML}}) = \hat{\sigma}^2_{\text{ML}}\sqrt{2/n}$.

b) Derive the profile log-likelihood functions of μ and σ^2 and plot them for the following data:

$$x = (225, 171, 198, 189, 189, 135, 162, 136, 117, 162).$$

Compare the profile log-likelihood functions with their quadratic approximations.

▶ *In Example 5.7, we derived the profile likelihood functions of* μ *and* σ^2 *for the normal model. Together with the relationship between the density of a log-normal and a normal random variable we obtain that, up to additive constants, the profile log-likelihood function of* μ *is*

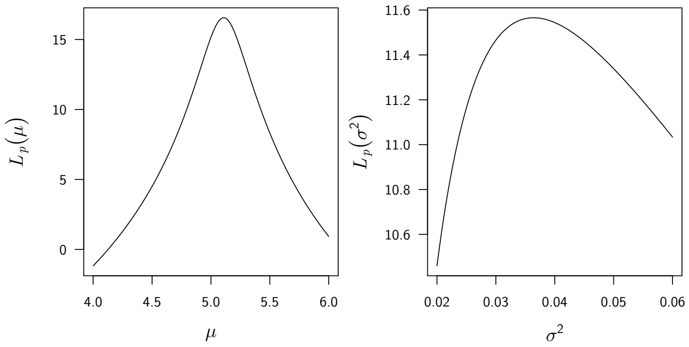
$$\begin{aligned} l_p(\mu) &= -\frac{n}{2}\log\left\{(\bar{y}-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} \\ &= -\frac{n}{2}\log\left\{(\hat{\mu}_{\text{ML}}-\mu)^2 + \hat{\sigma}^2_{\text{ML}}\right\}, \end{aligned}$$

whereas the profile log-likelihood function of σ^2 *is*

$$\begin{aligned} l_p(\sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= -\frac{n}{2}\left\{\log(\sigma^2) + \frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2}\right\}. \end{aligned}$$

For our data, they can be evaluated as follows.

```
> x <- c(225, 171, 198, 189, 189, 135, 162, 136, 117, 162)
> y <- log(x)
> loglik.mu <- function(y, mu){
      n <- length(y)
      return( -n/2*log( (mean(y)-mu)^2 + (n-1)/n*var(y) ) )
  }
> loglik.sigma2 <- function(y, sigma2){
      n <- length(y)
      return( -n/2*log(sigma2) - (n-1)*var(y)/2/sigma2 )
  }
> par(mfrow=c(1, 2))
> mu.x <- seq(4, 6, length=101)
> plot(mu.x, loglik.mu(y=y, mu=mu.x),
       type="l", xlab=math(mu), ylab=math(L[p](mu)))
> sigma2.x <- seq(0.02, 0.06, length=101)
> plot(sigma2.x, loglik.sigma2(y=y, sigma2=sigma2.x),
       type="l", xlab=math(sigma^2), ylab=math(L[p](sigma^2)))
```



Denoting $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, *and recalling that the inverse observed Fisher information that we derived in Section 5.2 for the normal model is*

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}_{\text{ML}})^{-1} = \begin{pmatrix} \frac{\hat{\sigma}^2_{\text{ML}}}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4_{\text{ML}}}{n} \end{pmatrix},$$

we obtain the quadratic approximations

$$\tilde{l}_p(\mu) \approx -\frac{n}{2}\frac{1}{\hat{\sigma}^2_{\text{ML}}}\cdot(\mu-\hat{\mu}_{\text{ML}})^2$$

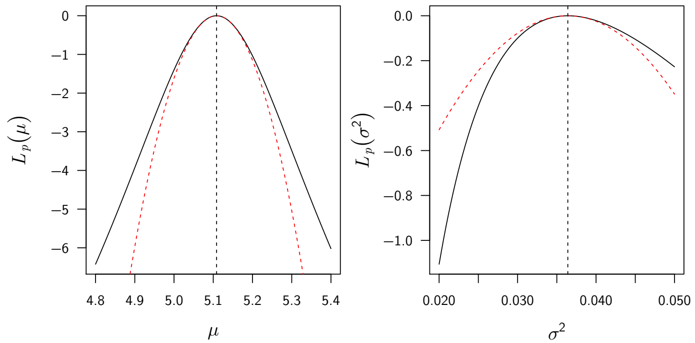
and

$$\tilde{l}_p(\sigma^2) \approx -\frac{n}{4}\frac{1}{\hat{\sigma}^4_{\text{ML}}}\cdot(\sigma^2-\hat{\sigma}^2_{\text{ML}})^2$$

to the corresponding relative profile log-likelihoods. We can compare the relative profile log-likelihoods to their quadratic approximations around the MLEs as follows.

```
> relloglik.mu <- function(y, mu){
      n <- length(y)
      return(
      -n/2*log( (mean(y)-mu)^2 + (n-1)/n*var(y) )
      +n/2*log( (n-1)/n*var(y) )
      )
  }
> relloglik.sigma2 <- function(y, sigma2){
      n <- length(y)
      hatsigma2 <- (n-1)/n*var(y)
      return(
      -n/2*log(sigma2) - (n-1)*var(y)/2/sigma2
      +n/2*log(hatsigma2) + n/2
      )
  }
> approxrelloglik.mu <- function(y, mu){
      n <- length(y)
      hatsigma2 <- (n-1)/n*var(y)
      return(
      -n/2/hatsigma2*(mean(y)-mu)^2
```

```
    )
  }
> approxrelloglik.sigma2 <- function(y, sigma2){
      n <- length(y)
      hatsigma2 <- (n-1)/n*var(y)
      return(
      -n/4/(hatsigma2)^2*(hatsigma2-sigma2)^2
      )
  }
> par(mfrow=c(1, 2))
> mu.x <- seq(4.8, 5.4, length=101)
> plot(mu.x, relloglik.mu(y=y, mu=mu.x),
       type="l", xlab=math(mu), ylab=math(L[p](mu)))
> lines(mu.x, approxrelloglik.mu(y=y, mu=mu.x),
        lty=2, col=2)
> abline(v=mean(y), lty=2)
> sigma2.x <- seq(0.02, 0.05, length=101)
> plot(sigma2.x, relloglik.sigma2(y=y, sigma2=sigma2.x),
       type="l", xlab=math(sigma^2), ylab=math(L[p](sigma^2)))
> lines(sigma2.x, approxrelloglik.sigma2(y=y, sigma2=sigma2.x),
        lty=2, col=2)
> abline(v=(length(y)-1)/length(y)*var(y), lty=2)
```



13. We assume an exponential model for the survival times in the randomised placebo-controlled trial of Azathioprine for primary biliary cirrhosis (PBC) from Section 1.1.8. The survival times (in days) of the $n = 90$ patients in the placebo group are denoted by x_i with censoring indicators γ_i $(i = 1, \dots, n)$, while the survival times of the $m = 94$ patients in the treatment group are denoted by y_i and have censoring indicators δ_i $(i = 1, \dots, m)$. The (partly unobserved) uncensored survival times follow exponential models with rates η and $\theta\eta$ in the placebo and treatment group, respectively $(\eta, \theta > 0)$.

a) Interpret η and θ. Show that their joint log-likelihood is

$$l(\eta,\theta) = (n\bar{\gamma} + m\bar{\delta})\log(\eta) + m\bar{\delta}\log(\theta) - \eta(n\bar{x} + \theta m\bar{y}),$$

where $\bar{\gamma}$, $\bar{\delta}$, $\bar{x}$, $\bar{y}$ are the averages of the γ_i, δ_i, x_i and y_i, respectively.

▶ *η is the rate of the exponential distribution in the placebo group, so the expected survival time is $1/\eta$. θ is the multiplicative change of the rate for the treatment group. The expected survival time changes to $1/(\eta\theta)$.*

The likelihood function is, by independence of all patients and the distributional assumptions (cf. Example 2.8):

$$\begin{aligned} L(\eta,\theta) &= \prod_{i=1}^{n} f(x_i;\eta)^{\gamma_i}\{1 - F(x_i;\eta)\}^{(1-\gamma_i)} \cdot \prod_{i=1}^{m} f(y_i;\eta,\theta)^{\delta_i}\{1 - F(y_i;\eta,\theta)\}^{(1-\delta_i)} \\ &= \prod_{i=1}^{n}\{\eta\exp(-\eta x_i)\}^{\gamma_i}\{\exp(-\eta x_i)\}^{(1-\gamma_i)} \cdot \prod_{i=1}^{m}\{\eta\theta\exp(-\eta\theta y_i)\}^{\delta_i}\{\exp(-\eta\theta y_i)\}^{(1-\delta_i)} \\ &= \eta^{n\bar{\gamma}+m\bar{\delta}}\theta^{m\bar{\delta}}\exp\left\{-\eta\left(n\bar{x} + \theta m\bar{y}\right)\right\}. \end{aligned}$$

Hence the log-likelihood is

$$\begin{aligned} l(\eta,\theta) &= \log\{L(\eta,\theta)\} \\ &= (n\bar{\gamma} + m\bar{\delta})\log(\eta) + m\bar{\delta}\log(\theta) - \eta(n\bar{x} + \theta m\bar{y}). \end{aligned}$$

b) Calculate the MLE $(\hat{\eta}_{\text{ML}}, \hat{\theta}_{\text{ML}})$ and the observed Fisher information matrix $\boldsymbol{I}(\hat{\eta}_{\text{ML}}, \hat{\theta}_{\text{ML}})$.

▶ *For calculating the MLE, we need to solve the score equations. The score function components are*

$$\frac{d}{d\eta}l(\eta,\theta) = \frac{n\bar{\gamma} + m\bar{\delta}}{\eta} - (n\bar{x} + \theta m\bar{y})$$

and

$$\frac{d}{d\theta}l(\eta,\theta) = \frac{m\bar{\delta}}{\theta} - \eta m\bar{y}.$$

From the second score equation $\frac{d}{d\theta}l(\eta,\theta) = 0$ we get $\hat{\theta}_{\text{ML}} = \bar{\delta}/(\hat{\eta}_{\text{ML}}\bar{y})$. Plugging this into the first score equation $\frac{d}{d\eta}l(\eta,\theta) = 0$ we get $\hat{\eta}_{\text{ML}} = \bar{\gamma}/\bar{x}$, and hence $\hat{\theta}_{\text{ML}} = (\bar{x}\bar{\delta})/(\bar{y}\bar{\gamma})$.

The ordinary Fisher information matrix is

$$\boldsymbol{I}(\eta,\theta) = \begin{pmatrix} \frac{n\bar{\gamma}+m\bar{\delta}}{\eta^2} & m\bar{y} \\ m\bar{y} & \frac{m\bar{\delta}}{\theta^2} \end{pmatrix},$$

so plugging in the MLEs gives the observed Fisher information matrix

$$\boldsymbol{I}(\hat{\eta}_{\text{ML}}, \hat{\theta}_{\text{ML}}) = \begin{pmatrix} (n\bar{\gamma} + m\bar{\delta}) \cdot (\bar{x}/\bar{\gamma})^2 & m\bar{y} \\ m\bar{y} & m\bar{\delta} \cdot \{(\bar{y}\bar{\gamma})/(\bar{x}\bar{\delta})\}^2 \end{pmatrix}.$$

c) Show that

$$\operatorname{se}(\hat{\theta}_{\text{ML}}) = \hat{\theta}_{\text{ML}} \cdot \sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}},$$

and derive a general formula for a $\gamma \cdot 100\%$ Wald confidence interval for θ. Use Appendix B.1.1 to compute the required entry of the inverse observed Fisher information.

▶ *From Section 5.2 we know that the standard error of $\hat{\theta}_{\text{ML}}$ is given by the square root of the corresponding diagonal element of the inverse observed Fisher information matrix. From Appendix B.1.1 we know how to compute the required entry:*

$$
\begin{aligned}
\left[\boldsymbol{I}(\hat{\eta}_{\text{ML}}, \hat{\theta}_{\text{ML}})^{-1}\right]_{22} &= \frac{(n\bar{\gamma} + m\bar{\delta}) \cdot (\bar{x}/\bar{\gamma})^2}{(n\bar{\gamma} + m\bar{\delta}) \cdot (\bar{x}/\bar{\gamma})^2 \cdot m\bar{\delta} \cdot (\bar{y}\bar{\gamma})^2/(\bar{x}\bar{\delta})^2 - (m\bar{y})^2} \\
&= \frac{(n\bar{\gamma} + m\bar{\delta}) \cdot (\bar{x}/\bar{\gamma})^2}{(n\bar{\gamma} + m\bar{\delta}) \cdot m \cdot \bar{y}^2/\bar{\delta} - (m\bar{y})^2} \\
&= \frac{(n\bar{\gamma} + m\bar{\delta}) \cdot (\bar{x}/\bar{\gamma})^2}{n\bar{\gamma} \cdot m \cdot \bar{y}^2/\bar{\delta}} \\
&= \left(\frac{\bar{x}\bar{\delta}}{\bar{y}\bar{\gamma}}\right)^2 \frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}} \\
&= \hat{\theta}_{\text{ML}}^2 \frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}.
\end{aligned}
$$

Hence the standard error is

$$
\operatorname{se}(\hat{\theta}_{\text{ML}}) = \hat{\theta}_{\text{ML}} \cdot \sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}},
$$

and the formula for a $\gamma \cdot 100\%$ Wald confidence interval for θ is

$$
\left[\frac{\bar{x}\bar{\delta}}{\bar{y}\bar{\gamma}} - z_{(1+\gamma)/2}\frac{\bar{x}\bar{\delta}}{\bar{y}\bar{\gamma}}\sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}}\ ,\ \frac{\bar{x}\bar{\delta}}{\bar{y}\bar{\gamma}} + z_{(1+\gamma)/2}\frac{\bar{x}\bar{\delta}}{\bar{y}\bar{\gamma}}\sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}},\right].
$$

d) Consider the transformation $\psi = \log(\theta)$. Derive a $\gamma \cdot 100\%$ Wald confidence interval for ψ using the delta method.

▶ *Due to the invariance of the MLE we have that*

$$
\hat{\psi}_{\text{ML}} = \log(\hat{\theta}_{\text{ML}}) = \log\{(\bar{x}\bar{\delta})/(\bar{y}\bar{\gamma})\}.
$$

For the application of the delta method, we need the derivative of the transformation, which is

$$
\frac{d}{d\theta}\log(\theta) = \frac{1}{\theta},
$$

giving the standard error for $\hat{\psi}_{\text{ML}}$ as

$$
\begin{aligned}
\operatorname{se}(\hat{\psi}_{\text{ML}}) &= \operatorname{se}(\hat{\theta}_{\text{ML}}) \left|\frac{d}{d\theta}\log(\hat{\theta}_{\text{ML}})\right| \\
&= \hat{\theta}_{\text{ML}}\sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}}\frac{1}{\hat{\theta}_{\text{ML}}} \\
&= \sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}}.
\end{aligned}
$$

Thus, a $\gamma \cdot 100\%$ Wald confidence interval for ψ is given by

$$
\left[\log\{(\bar{x}\bar{\delta})/(\bar{y}\bar{\gamma})\} - z_{(1+\gamma)/2}\sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}}\ ,\ \log\{(\bar{x}\bar{\delta})/(\bar{y}\bar{\gamma})\} + z_{(1+\gamma)/2}\sqrt{\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{\gamma} \cdot m\bar{\delta}}}\right].
$$

e) Derive the profile log-likelihood for θ. Implement an `R`-function which calculates a $\gamma \cdot 100\%$ profile likelihood confidence interval for θ.

▶ *Solving $\frac{d}{d\eta}l(\eta, \theta) = 0$ for η gives*

$$
\begin{aligned}
\frac{d}{d\eta}l(\eta, \theta) &= 0 \\
\frac{n\bar{\gamma} + m\bar{\delta}}{\eta} &= n\bar{x} + \theta m\bar{y} \\
\eta &= \frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{x} + \theta m\bar{y}},
\end{aligned}
$$

hence the profile log-likelihood of θ is

$$
\begin{aligned}
l_p(\theta) &= l(\hat{\eta}_{\text{ML}}(\theta), \theta) \\
&= (n\bar{\gamma} + m\bar{\delta})\log\left(\frac{n\bar{\gamma} + m\bar{\delta}}{n\bar{x} + \theta m\bar{y}}\right) + m\bar{\delta}\log(\theta) - (n\bar{\gamma} + m\bar{\delta}).
\end{aligned}
$$

Dropping the additive terms, we obtain the profile log-likelihood as

$$
l_p(\theta) = m\bar{\delta}\log(\theta) - (n\bar{\gamma} + m\bar{\delta})\log(n\bar{x} + \theta m\bar{y}).
$$

To obtain a $\gamma \cdot 100\%$ profile likelihood confidence interval, we need to find the solutions of

$$
2\{l_p(\hat{\theta}_{\text{ML}}) - l_p(\theta)\} = \chi^2_\gamma(1).
$$

In `R`:

```
> ## the profile log-likelihood of theta
> profLogLik <- function(theta,
                         n,
                         xbar,
                         gammabar,
                         m,
                         ybar,
                         deltabar)
  {
      m * deltabar * log(theta) -
          (n * gammabar + m * deltabar) * log(n * xbar + theta * m * ybar)
  }
> ## the generalised likelihood ratio statistic
> likRatioStat <- function(theta,
                           n,
                           xbar,
                           gammabar,
                           m,
                           ybar,
                           deltabar)
  {
      thetaMle <- (xbar * deltabar) / (ybar * gammabar)
      return(2 * (profLogLik(thetaMle, n, xbar, gammabar, m, ybar, deltabar) -
                  profLogLik(theta, n, xbar, gammabar, m, ybar, deltabar)))
  }
> ## compute a gamma profile likelihood confidence interval
> profLikCi <- function(gamma,
                        n,
                        xbar,
                        gammabar,
                        m,
                        ybar,
                        deltabar)
  {
      targetFun <- function(theta)
      {
          likRatioStat(theta, n, xbar, gammabar, m, ybar, deltabar) -
              qchisq(p=gamma, df=1)
      }
      thetaMle <- (xbar * deltabar) / (ybar * gammabar)
      lower <- uniroot(targetFun,
                       interval=c(1e-10, thetaMle))$root
      upper <- uniroot(targetFun,
                       interval=c(thetaMle, 1e10))$root
      return(c(lower, upper))
  }
```

f) Calculate 95% confidence intervals for θ based on 13c), 13d) and 13e). Also compute for each of the three confidence intervals 13c), 13d) and 13e) the corresponding P-value for the null hypothesis that the exponential distribution for the PBC survival times in the treatment group is not different from the placebo group.

▶ *First we read the data:*

```
> n <- sum(pbcFull$treat == 1)
> xbar <- mean(pbcFull$time[pbcFull$treat == 1])
> gammabar <- mean(pbcFull$d[pbcFull$treat == 1])
> m <- sum(pbcFull$treat == 2)
> ybar <- mean(pbcFull$time[pbcFull$treat == 2])
> deltabar <- mean(pbcFull$d[pbcFull$treat == 2])
```

Now we compute 95% confidence intervals using the three different methods:

```
> ## the MLE
> (thetaMle <- (xbar * deltabar) / (ybar * gammabar))
[1] 0.8685652
> ## the standard error:
> (thetaMleSe <- thetaMle * sqrt((n * gammabar + m * deltabar) /
                                 (n * gammabar * m * deltabar)))
[1] 0.1773336
> ## the Wald interval on the original scale:
> (waldCi <- thetaMle + c(-1, 1) * qnorm(0.975) * thetaMleSe)
[1] 0.5209977 1.2161327
> ## then the Wald interval derived from the log-scale:
> (transWaldCi <- exp(log(thetaMle) + c(-1, 1) * qnorm(0.975) *
                      sqrt((n * gammabar + m * deltabar) /
                           (n * gammabar * m * deltabar))))
[1] 0.5821219 1.2959581
> ## and finally the profile likelihood CI:
> (profLikCiRes <- profLikCi(0.95, n, xbar, gammabar, m, ybar, deltabar))
[1] 0.5810492 1.2969668
```

The Wald interval derived from the log-scale is much closer to the profile likelihood interval, which points to a better quadratic approximation of the relative profile log-likelihood for the transformed parameter ψ.
Now we compute two-sided P-values for testing $H_0 : \theta = 1$.

```
> ## the value to be tested:
> thetaNull <- 1
> ## first with the Wald statistic:
> waldStat <- (thetaMle - thetaNull) / thetaMleSe
> (pWald <- 2 * pnorm(abs(waldStat), lower.tail=FALSE))
[1] 0.458589
> ## second with the Wald statistic on the log-scale:
> transWaldStat <- (log(thetaMle) - log(thetaNull)) /
      sqrt((n * gammabar + m * deltabar) /
           (n * gammabar * m * deltabar))
> (pTransWald <- 2 * pnorm(abs(transWaldStat), lower.tail=FALSE))
[1] 0.4900822
> ## finally with the generalised likelihood ratio statistic:
> likStat <- likRatioStat(thetaNull, n, xbar, gammabar, m, ybar, deltabar)
> (pLikStat <- pchisq(likStat, df=1, lower.tail=FALSE))
[1] 0.4900494
```

All three test statistics say that the evidence against H_0 is not sufficient for the rejection.

14. Let $X_{1:n}$ be a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution.

a) First assume that σ^2 is known. Derive the likelihood ratio statistic for testing specific values of μ.

▶ *If σ^2 is known, the log-likelihood kernel is*

$$l(\mu; x) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

and the score function is

$$\begin{aligned} S(\mu; x) &= \frac{d}{d\mu} l(\mu; x) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i - \mu)\cdot(-1) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu). \end{aligned}$$

The root of the score equation is $\hat{\mu}_{\text{ML}} = \bar{x}$. Hence the likelihood ratio statistic is

$$\begin{aligned} W(\mu; x) &= 2\{l(\hat{\mu}_{\text{ML}}; x) - l(\mu; x)\} \\ &= 2\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2\right\} \\ &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu) + \sum_{i=1}^{n}(\bar{x} - \mu)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2\right\} \\ &= \frac{n}{\sigma^2}(\bar{x} - \mu)^2 \\ &= \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2. \end{aligned}$$

b) Show that, in this special case, the likelihood ratio statistic is an exact pivot and exactly has a $\chi^2(1)$ distribution.

▶ *From Example 3.5 we know that $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$ and $Z = \sqrt{n}/\sigma(\bar{X} - \mu) \sim \mathrm{N}(0, 1)$. Moreover, from Table A.2 in the Appendix we know that $Z^2 \sim \chi^2(1)$. It follows that $W(\mu; X) \sim \chi^2(1)$, and so the distribution of the likelihood ratio statistic does not depend on the unknown parameter μ. Therefore, the likelihood ratio statistic is an exact pivot. Moreover, the chi-squared distribution holds exactly for each sample size n, not only asymptotically for $n \to \infty$.*

c) Show that, in this special case, the corresponding likelihood ratio confidence interval equals the Wald confidence interval.

▶ *The Fisher information is*

$$I(\mu; x) = -\frac{d}{d\mu} S(\mu; x) = \frac{n}{\sigma^2},$$

implying the standard error of the MLE in the form

$$\operatorname{se}(\hat{\mu}_{\text{ML}}) = I(\mu)^{-1/2} = \sigma/\sqrt{n}.$$

Therefore the $\gamma \cdot 100\%$ Wald confidence interval for μ is

$$\left[\bar{x} - \sigma/\sqrt{n} z_{(1+\gamma)/2}\ ,\ \bar{x} + \sigma/\sqrt{n} z_{(1+\gamma)/2}\right]$$

and already found in Example 3.5. Now using the fact that the square root of the γ chi-squared quantile equals the $(1+\gamma)/2$ standard normal quantile, as mentioned in Section 4.4, we have

$$\begin{aligned} \mathsf{Pr}\left\{W(\mu; X) \leq \chi^2_\gamma(1)\right\} &= \mathsf{Pr}\left\{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \leq \chi^2_\gamma(1)\right\} \\ &= \mathsf{Pr}\left\{-\sqrt{\chi^2_\gamma(1)} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \sqrt{\chi^2_\gamma(1)}\right\} \\ &= \mathsf{Pr}\left\{-z_{(1+\gamma)/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{(1+\gamma)/2}\right\} \\ &= \mathsf{Pr}\left\{-\bar{X} - \sigma/\sqrt{n} z_{(1+\gamma)/2} \leq -\mu \leq -\bar{X} + \sigma/\sqrt{n} z_{(1+\gamma)/2}\right\} \\ &= \mathsf{Pr}\left\{\bar{X} + \sigma/\sqrt{n} z_{(1+\gamma)/2} \geq \mu \geq \bar{X} - \sigma/\sqrt{n} z_{(1+\gamma)/2}\right\} \\ &= \mathsf{Pr}\left\{\mu \in \left[\bar{X} - \sigma/\sqrt{n} z_{(1+\gamma)/2}\, , \bar{X} + \sigma/\sqrt{n} z_{(1+\gamma)/2}\right]\right\}. \end{aligned}$$

As the $\gamma \cdot 100\%$ likelihood ratio confidence interval is given by all values of μ with $W(\mu; X) \leq \chi^2_\gamma(1)$, we have shown that here it equals the $\gamma \cdot 100\%$ Wald confidence interval.

d) Now assume that μ is known. Derive the likelihood ratio statistic for testing specific values of σ^2.

▶ *If μ is known, the log-likelihood function for σ^2 is given by*

$$l(\sigma^2; x) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

and the score function is

$$\begin{aligned} S(\sigma^2; x) &= \frac{d}{d\sigma^2} l(\sigma^2; x) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned}$$

By solving the score equation $S(\sigma^2; x) = 0$ we obtain the MLE

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The likelihood ratio statistic thus is

$$\begin{aligned} W(\sigma^2; x) &= 2\{l(\hat{\sigma}^2_{\text{ML}}; x) - l(\sigma^2; x)\} \\ &= 2\left\{-\frac{n}{2}\log(\hat{\sigma}^2_{\text{ML}}) - \frac{n\hat{\sigma}^2_{\text{ML}}}{2\hat{\sigma}^2_{\text{ML}}} + \frac{n}{2}\log(\sigma^2) + \frac{n\hat{\sigma}^2_{\text{ML}}}{2\sigma^2}\right\} \\ &= -n\log\left(\frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2}\right) - n + n\frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2} \\ &= n\left\{\frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2} - \log\left(\frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2}\right) - 1\right\}. \end{aligned}$$

e) Compare the likelihood ratio statistic and its distribution with the exact pivot mentioned in Example 4.21. Derive a general formula for a confidence interval based on the exact pivot, analogously to Example 3.8.

▶ *In Example 4.21 we encountered the exact pivot*

$$V(\sigma^2; X) = n\frac{\hat{\sigma}^2_{\text{ML}}}{\sigma^2} \sim \chi^2(n) = \mathrm{G}(n/2, 1/2).$$

It follows that

$$V(\sigma^2; X)/n \sim \mathrm{G}(n/2, n/2).$$

The likelihood ratio statistic is a transformation of the latter:

$$W(\sigma^2; X) = n\left\{V(\sigma^2; X)/n - \log(V(\sigma^2; X)/n) - 1\right\}.$$

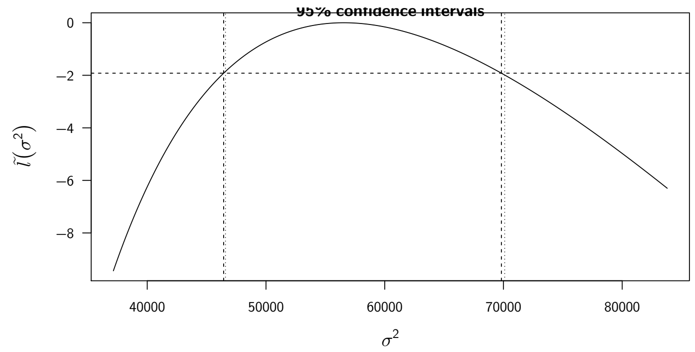
Analogously to Example 3.8 we can derive a $\gamma \cdot 100\%$ confidence interval for σ^2 based on the exact pivot $V(\sigma^2; X)$:

$$\begin{aligned} \gamma &= \Pr\left\{\chi^2_{(1-\gamma)/2}(n) \leq V(\sigma^2; X) \leq \chi^2_{(1+\gamma)/2}(n)\right\} \\ &= \Pr\left\{1/\chi^2_{(1-\gamma)/2}(n) \geq \frac{\sigma^2}{n\hat{\sigma}^2_{\text{ML}}} \geq 1/\chi^2_{(1+\gamma)/2}(n)\right\} \\ &= \Pr\left\{n\hat{\sigma}^2_{\text{ML}}/\chi^2_{(1+\gamma)/2}(n) \leq \sigma^2 \leq n\hat{\sigma}^2_{\text{ML}}/\chi^2_{(1-\gamma)/2}(n)\right\}. \end{aligned}$$

f) Consider the transformation factors from Table 1.3, and assume that the "mean" is the known μ and the "standard deviation" is $\hat{\sigma}_{\text{ML}}$. Compute both a 95% likelihood ratio confidence interval and a 95% confidence interval based on the exact pivot for σ^2. Illustrate the likelihood ratio confidence interval by plotting the relative log-likelihood and the cut-value, similar to Figure 4.8. In order to compute the likelihood ratio confidence interval, use the R-function `uniroot` (*cf*. Appendix C.1.1).

▶ *The data are $n = 185$, $\mu = 2449.2$, $\hat{\sigma}_{\text{ML}} = 237.8$.*

```
> ## define the data
> n <- 185
> mu <- 2449.2
> mlsigma2 <- (237.8)^2
> ## the log-likelihood
> loglik <- function(sigma2)
  {
      - n / 2 * log(sigma2) - n * mlsigma2 / (2 * sigma2)
  }
> ## the relative log-likelihood
> rel.loglik <- function(sigma2)
  {
      loglik(sigma2) - loglik(mlsigma2)
  }
> ## the likelihood ratio statistic
> likstat <- function(sigma2)
  {
      n * (mlsigma2 / sigma2 - log(mlsigma2 / sigma2) - 1)
  }
> ## find the bounds of the likelihood ratio CI
> lik.lower <- uniroot(function(sigma2){likstat(sigma2) - qchisq(0.95, df=1)},
                       interval=c(0.5*mlsigma2, mlsigma2))$root
> lik.upper <- uniroot(function(sigma2){likstat(sigma2) - qchisq(0.95, df=1)},
                       interval=c(mlsigma2, 2*mlsigma2))$root
> (lik.ci <- c(lik.lower, lik.upper))
[1] 46432.99 69829.40
> ## illustrate the likelihood ratio CI
> sigma2.grid <- seq(from=0.8 * lik.lower,
                     to=1.2 * lik.upper,
                     length=1000)
> plot(sigma2.grid,
       rel.loglik(sigma2.grid),
       type="l",
       xlab=math(sigma^2),
       ylab= math (tilde(l)(sigma^2)),
       main="95% confidence intervals",
       las=1)
> abline(h = - 1/2 * qchisq(0.95, df=1),
         v = c(lik.lower, lik.upper),
         lty=2)
> ## now compute the other 95% CI:
> (alt.ci <- c(n * mlsigma2 / qchisq(p=0.975, df=n),
               n * mlsigma2 / qchisq(p=0.025, df=n)))
[1] 46587.27 70104.44
> abline(v = c(alt.ci[1], alt.ci[2]),
         lty=3)
```



The two confidence intervals are very similar here.

15. Consider K independent groups of normally distributed observations with group-specific means and variances, *i. e.* let $X_{i,1:n_i}$ be a random sample from $\mathrm{N}(\mu_i, \sigma_i^2)$ for group $i = 1, \dots, K$. We want to test the null hypothesis that the variances are identical, *i. e.* $H_0 : \sigma_i^2 = \sigma^2$.

a) Write down the log-likelihood kernel for the parameter vector $\boldsymbol{\theta} = (\mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2)^\top$. Derive the MLE $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ by solving the score equations $S_{\mu_i}(\boldsymbol{\theta}) = 0$ and then $S_{\sigma_i^2}(\boldsymbol{\theta}) = 0$, for $i = 1, \dots, K$.

▶ *Since all random variables are independent, the log-likelihood is the sum of all individual log density contributions* $\log f(x_{ij}; \boldsymbol{\theta}_i)$*, where* $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2)^\top$*:*

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log f(x_{ij}; \boldsymbol{\theta}).$$

Hence the log-likelihood kernel is

$$\begin{aligned}
&\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left\{ -\frac{1}{2} \log(\sigma_i^2) - \frac{1}{2\sigma_i^2} (x_{ij} - \mu_i)^2 \right\} \\
&= \sum_{i=1}^{K} \left\{ -\frac{n_i}{2} \log(\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right\}.
\end{aligned}$$

Now the score function for the mean μ_i *of the* i*-th group is given by*

$$S_{\mu_i}(\boldsymbol{\theta}) = \frac{d}{d\mu_i} l(\boldsymbol{\theta}) = \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i),$$

and solving the corresponding score equation $S_{\mu_i}(\boldsymbol{\theta}) = 0$ *gives* $\hat{\mu}_i = \bar{x}_i$*, the average of the observations in the* i*-th group. The score function for the variance* σ_i^2 *of the* i*-th group is given by*

$$S_{\sigma_i^2}(\boldsymbol{\theta}) = \frac{d}{d\sigma_i^2} l(\boldsymbol{\theta}) = -\frac{n_i}{2\sigma_i^2} + \frac{1}{2(\sigma_i^2)^2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2,$$

which, after plugging in $\hat{\mu}_i = \bar{x}_i$ *for* μ_i*, is solved by* $\hat{\sigma}_i^2 = n_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$*. This can be done for any* $i = 1, \dots, K$*, giving the MLE* $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$*.*

b) Compute the MLE $\hat{\boldsymbol{\theta}}_0$ under the restriction $\sigma_i^2 = \sigma^2$ of the null hypothesis.

▶ *Under* H_0 *we have* $\sigma_i^2 = \sigma^2$*. Obviously the MLEs for the means* μ_i *do not change, so we have* $\hat{\mu}_{i,0} = \bar{x}_i$*. However, the score function for* σ^2 *now comprises all groups:*

$$S_{\sigma^2}(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{K} n_i + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2.$$

Plugging in the estimates $\hat{\mu}_{i,0} = \bar{x}_i$ *for* μ_i *and solving for* σ^2 *gives*

$$\hat{\sigma}_0^2 = \frac{1}{\sum_{i=1}^{K} n_i} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

which is a pooled variance estimate. So the MLE under the restriction $\sigma_i^2 = \sigma^2$ *of the null hypothesis is* $\hat{\boldsymbol{\theta}}_0 = (\bar{x}_1, \dots, \bar{x}_K, \hat{\sigma}_0^2, \dots, \hat{\sigma}_0^2)^\top$*.*

c) Show that the generalised likelihood ratio statistic for testing $H_0 : \sigma_i^2 = \sigma^2$ is

$$W = \sum_{i=1}^{K} n_i \log(\hat{\sigma}_0^2 / \hat{\sigma}_i^2)$$

where $\hat{\sigma}_0^2$ and $\hat{\sigma}_i^2$ are the ML variance estimates for the i-th group with and without the H_0 assumption, respectively. What is the approximate distribution of W under H_0?

▶ *The generalised likelihood ratio statistic is*

$$\begin{aligned}
W &= 2\{l(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - l(\hat{\boldsymbol{\theta}}_0)\} \\
&= 2 \sum_{i=1}^{K} \left\{ -\frac{n_i}{2} \log(\hat{\sigma}_i^2) - \frac{1}{2\hat{\sigma}_i^2} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)^2 + \frac{n_i}{2} \log(\hat{\sigma}_0^2) + \frac{1}{2\hat{\sigma}_0^2} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_{i,0})^2 \right\} \\
&= \sum_{i=1}^{K} n_i \log(\hat{\sigma}_0^2 / \hat{\sigma}_i^2) - \sum_{i=1}^{K} \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} + \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{\frac{1}{\sum_{i=1}^{K} n_i} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \\
&= \sum_{i=1}^{K} n_i \log(\hat{\sigma}_0^2 / \hat{\sigma}_i^2) - \sum_{i=1}^{K} n_i + \sum_{i=1}^{K} n_i \\
&= \sum_{i=1}^{K} n_i \log(\hat{\sigma}_0^2 / \hat{\sigma}_i^2).
\end{aligned}$$

Since there are $p = 2K$ *free parameters in the unconstrained model, and* $r = K+1$ *free parameters under the* H_0 *restriction, we have*

$$W \overset{a}{\sim} \chi^2(p - r) = \chi^2(K - 1)$$

under H_0*.*

d) Consider the special case with $K = 2$ groups having equal size $n_1 = n_2 = n$. Show that W is large when the ratio

$$R = \frac{\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2}{\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2}$$

is large or small. Show that W is minimal if $R = 1$. Which value has W for $R = 1$?

▶ *In this special case we have*

$$\hat{\sigma}_0^2/\hat{\sigma}_1^2 = \frac{\frac{1}{2n}\left\{\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2\right\}}{\frac{1}{n}\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2} = \frac{1}{2}(1 + 1/R)$$

and analogously

$$\hat{\sigma}_0^2/\hat{\sigma}_2^2 = \frac{1}{2}(1 + R).$$

Hence the likelihood ratio statistic is

$$\begin{aligned} W &= n\log(\hat{\sigma}_0^2/\hat{\sigma}_1^2) + n\log(\hat{\sigma}_0^2/\hat{\sigma}_2^2) \\ &= 2n\log(1/2) + n\left\{\log(1 + 1/R) + \log(1 + R)\right\}, \end{aligned}$$

which is large if $1/R$ or R is large, i. e. if R is small or large. We now consider the derivative of W with respect to R:

$$\begin{aligned} \frac{d}{dR}W(R) &= n\left\{\frac{1}{1 + 1/R}(-1)R^{-2} + \frac{1}{1 + R}\right\} \\ &= \frac{n}{1 + R}\left(-\frac{1}{R} + 1\right), \end{aligned}$$

which is equal to zero if and only if $R = 1$. Since we know that the function is increasing for small and large values of R, this is the minimum. The value of W for $R = 0$ is

$$W = 2n\log(1/2) + n\{\log(2) + \log(2)\} = -2n\log(2) + 2n\log(2) = 0.$$

e) Bartlett's modified likelihood ratio test statistic (Bartlett, 1937) is $B = T/C$ where in

$$T = \sum_{i=1}^{K}(n_i - 1)\log(s_0^2/s_i^2)$$

compared to W the numbers of observations n_i have been replaced by the degrees of freedom $n_i - 1$ and the sample variances $s_i^2 = (n_i - 1)^{-1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ define the pooled sample variance $s_0^2 = \{\sum_{i=1}^{K}(n_i - 1)\}^{-1}\sum_{i=1}^{K}(n_i - 1)s_i^2$. The correction factor

$$C = 1 + \frac{1}{3(K - 1)}\left\{\left(\sum_{i=1}^{K}\frac{1}{n_i - 1}\right) - \frac{1}{\sum_{i=1}^{K}(n_i - 1)}\right\}$$

is used because T/C converges more rapidly to the asymptotic $\chi^2(K - 1)$ distribution than T alone.

Write two R-functions which take the vector of the group sizes $(n_1, \ldots, n_K)$ and the sample variances $(s_1^2, \ldots, s_K^2)$ and return the values of the statistics W and B, respectively.

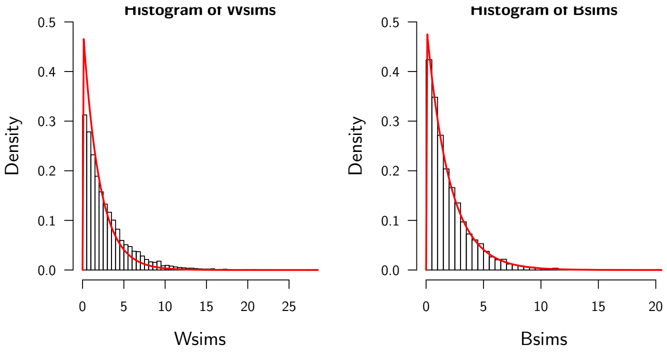
▶

```
> ## general function to compute the likelihood ratio statistic.
> W <- function(ni,               # the n_i's
                s2i)              # the s^2_i's
  {
      mleVars <- (ni - 1) * s2i / ni             # sigma^2_i estimate
      pooledVar <- sum((ni - 1) * s2i) / sum(ni) # sigma^2_0 estimate

      return(sum(ni * log(pooledVar / mleVars)))
  }
> ## general function to compute the Bartlett statistic.
> B <- function(ni,               # the n_i's
                s2i)              # the s^2_i's
  {
      pooledSampleVar <- sum((ni - 1) * s2i) / sum(ni - 1)

      T <- sum((ni - 1) * log(pooledSampleVar / s2i))
      C <- 1 + (sum(1 / (ni - 1)) - 1 / sum(ni - 1)) / (3 * (length(ni) - 1))

      return(T / C)
  }
```

f) In the H_0 setting with $K = 3$, $n_i = 5$, $\mu_i = i$, $\sigma^2 = 1/4$, simulate 10 000 data sets and compute the statistics W and B in each case. Compare the empirical distributions with the approximate $\chi^2(K - 1)$ distribution. Is B closer to $\chi^2(K - 1)$ than W in this case?

▶

```
> ## simulation setting under H0:
> K <- 3
> n <- rep(5, K)
> mu <- 1:K
> sigma <- 1/2
> ## now do the simulations
> nSim <- 1e4L
> Wsims <- Bsims <- numeric(nSim)
> s2i <- numeric(K)
> set.seed(93)
> for(i in seq_len(nSim))
  {
      ## simulate the sample variance for the i-th group
      for(j in seq_len(K))
      {
          s2i[j] <-  var(rnorm(n=n[j],
                               mean=mu[j],
                               sd=sigma))
      }

      ## compute test statistic results
      Wsims[i] <- W(ni=n,
                    s2i=s2i)
      Bsims[i] <- B(ni=n,
                    s2i=s2i)
  }
> ## now compare
> par(mfrow=c(1, 2))
> hist(Wsims,
       nclass=50,
       prob=TRUE,
       ylim=c(0, 0.5))
> curve(dchisq(x, df=K-1),
        col="red",
        add=TRUE,
        n=200,
        lwd=2)
> hist(Bsims,
       nclass=50,
       prob=TRUE,
       ylim=c(0, 0.5))
> curve(dchisq(x, df=K-1),
        col="red",
        add=TRUE,
        n=200,
        lwd=2)
```



We see that the empirical distribution of B is slightly closer to the $\chi^2(2)$ distribution than that of W, but the discrepancy is not very large.

g) Consider the alcohol concentration data from Section 1.1.7. Quantify the evidence against equal variances of the transformation factor between the genders using P-values based on the test statistics W and B.

▶

```
> ni <- c(33, 152)
> s2i <- c(220.1, 232.5)^2
> p.W <- pchisq(W(ni, s2i),
                df=3,
                lower.tail=FALSE)
> p.B <- pchisq(B(ni, s2i),
                df=3,
                lower.tail=FALSE)
> p.W
[1] 0.9716231
> p.B
[1] 0.9847605
```

According to both test statistics, there is very little evidence against equal variances.

16. In a 1:1 *matched case-control study*, one control (*i. e.* a disease-free individual) is matched to each case (*i. e.* a diseased individual) based on certain individual characteristics, *e. g.* age or gender. Exposure history to a potential risk factor is then determined for each individual in the study. If exposure E is binary (*e. g.* smoking history? yes/no) then it is common to display the data as frequencies of case-control pairs, depending on exposure history:

		History of control	
		Exposed	Unexposed
History of case	Exposed	a	b
	Unexposed	c	d

For example, a is the number of case-control pairs with positive exposure history of both the case and the control.
Let ω_1 and ω_0 denote the odds for a case and a control, respectively, to be exposed, such that

$$\Pr(E \,|\, \text{case}) = \frac{\omega_1}{1+\omega_1} \quad \text{and} \quad \Pr(E \,|\, \text{control}) = \frac{\omega_0}{1+\omega_0}.$$

To derive conditional likelihood estimates of the odds ratio $\psi = \omega_1/\omega_0$, we argue conditional on the number N_E of exposed individuals in a case-control pair. If $N_E = 2$ then both the case and the control are exposed so the corresponding a case-control pairs do not contribute to the conditional likelihood. This is also the case for the d case-control pairs where both the case and the control are unexposed ($N_E = 0$). In the following we therefore only consider the case $N_E = 1$, in which case either the case or the control is exposed, but not both.

a) Conditional on $N_E = 1$, show that the probability that the case rather than the control is exposed is $\omega_1/(\omega_0 + \omega_1)$. Show that the corresponding conditional odds are equal to the odds ratio ψ.

▶ *For the conditional probability we have*

$$\begin{aligned} \mathsf{Pr}(case\ E \,|\, N_E = 1) &= \frac{\mathsf{Pr}(case\ E,\ control\ not\ E)}{\mathsf{Pr}(case\ E,\ control\ not\ E) + \mathsf{Pr}(case\ not\ E,\ control\ E)} \\ &= \frac{\frac{\omega_1}{1+\omega_1}\cdot\frac{1}{1+\omega_0}}{\frac{\omega_1}{1+\omega_1}\cdot\frac{1}{1+\omega_0} + \frac{1}{1+\omega_1}\cdot\frac{\omega_0}{1+\omega_0}} \\ &= \frac{\omega_1}{\omega_1+\omega_0}, \end{aligned}$$

and for the conditional odds we have

$$\begin{aligned} \frac{\mathsf{Pr}(case\ E \,|\, N_E = 1)}{1 - \mathsf{Pr}(case\ E \,|\, N_E = 1)} &= \frac{\frac{\omega_1}{\omega_1+\omega_0}}{\frac{\omega_0}{\omega_1+\omega_0}} \\ &= \frac{\omega_1}{\omega_0}. \end{aligned}$$

b) Write down the binomial log-likelihood in terms of ψ and show that the MLE of the odds ratio ψ is $\hat{\psi}_{\text{ML}} = b/c$ with standard error $\operatorname{se}\{\log(\hat{\psi}_{\text{ML}})\} = \sqrt{1/b + 1/c}$. Derive the Wald test statistic for H_0: $\log(\psi) = 0$.

▶ *Note that* $\mathsf{Pr}(case\ E \,|\, N_E = 1) = \omega_1/(\omega_1 + \omega_0) = 1/(1 + 1/\psi)$ *and so* $\mathsf{Pr}(control\ E \,|\, N_E = 1) = \omega_0/(\omega_1 + \omega_0) = 1/(1 + \psi)$. *The conditional log-likelihood is*

$$\begin{aligned} l(\psi) &= b \log\left(\frac{1}{1 + 1/\psi}\right) + c \log\left(\frac{1}{1+\psi}\right) \\ &= -b \log\left(1 + \frac{1}{\psi}\right) - c \log(1 + \psi). \end{aligned}$$

Hence the score function is

$$\begin{aligned} S(\psi) &= \frac{d}{d\psi} l(\psi) \\ &= \frac{b}{1 + 1/\psi}\cdot\frac{1}{\psi^2} - \frac{c}{1+\psi} \\ &= \frac{b}{\psi(1+\psi)} - \frac{c}{1+\psi}, \end{aligned}$$

and the score equation $S(\psi) = 0$ *is solved by* $\hat{\psi}_{\text{ML}} = b/c$. *The Fisher information is*

$$\begin{aligned} I(\psi) &= -\frac{d}{d\psi} S(\psi) \\ &= \frac{b}{\psi^2(1+\psi)^2}\cdot\{(1+\psi)+\psi\} - \frac{c}{(1+\psi)^2} \\ &= \frac{1}{(1+\psi)^2}\cdot\left\{\frac{b}{\psi^2}\cdot(1+2\psi) - c\right\}, \end{aligned}$$

and the observed Fisher information is

$$\begin{aligned} I(\hat{\psi}_{\text{ML}}) &= \frac{1}{\{1 + (b/c)\}^2}\cdot\left\{\frac{b}{(b/c)^2}\cdot(1 + 2b/c) - c\right\} \\ &= \frac{c^2}{(c+b)^2}\cdot\frac{c\cdot(c+b)}{b} \\ &= \frac{c^3}{b(c+b)}. \end{aligned}$$

Note that, since the latter is positive, $\hat{\psi}_{\text{ML}} = b/c$ *indeed maximises the likelihood. By Result 2.1, the observed Fisher information corresponding to* $\log(\hat{\psi}_{\text{ML}})$ *is*

$$\begin{aligned} I\{\log(\hat{\psi}_{\text{ML}})\} &= \left\{\frac{d}{d\psi}\log(\hat{\psi}_{\text{ML}})\right\}^{-2}\cdot I(\hat{\psi}_{\text{ML}}) \\ &= \frac{b^2}{c^2}\cdot\frac{c^3}{b(c+b)} \\ &= \frac{b\cdot c}{c+b}. \end{aligned}$$

It follows that

$$\begin{aligned} \operatorname{se}\{\log(\hat{\psi}_{\text{ML}})\} &= [I\{\log(\hat{\psi}_{\text{ML}})\}]^{-1/2} \\ &= \sqrt{1/b + 1/c}. \end{aligned}$$

The Wald test statistic for H_0*:* $\log(\psi) = 0$ *is*

$$\frac{\log(\hat{\psi}_{\text{ML}}) - 0}{\operatorname{se}\{\log(\hat{\psi}_{\text{ML}})\}} = \frac{\log(b/c)}{\sqrt{1/b + 1/c}}.$$

c) Derive the standard error $\operatorname{se}(\hat{\psi}_{\text{ML}})$ of $\hat{\psi}_{\text{ML}}$ and derive the Wald test statistic for H_0: $\psi = 1$. Compare your result with the Wald test statistic for H_0: $\log(\psi) = 0$.

▶ *Using the observed Fisher information computed above, we obtain that*

$$\begin{aligned} \operatorname{se}(\hat{\psi}_{\text{ML}}) &= \{I(\hat{\psi}_{\text{ML}})\}^{-1/2} \\ &= \sqrt{\frac{b(c+b)}{c^3}} \\ &= \frac{b}{c}\cdot\sqrt{\frac{1}{b} + \frac{1}{c}}. \end{aligned}$$

The Wald test statistic for H_0*:* $\psi = 1$ *is*

$$\frac{\hat{\psi}_{\text{ML}} - 1}{\operatorname{se}(\hat{\psi}_{\text{ML}})} = \frac{b/c - 1}{b/c\cdot\sqrt{1/b + 1/c}} = \frac{b - c}{b\sqrt{1/b + 1/c}}.$$

d) Finally compute the score test statistic for H_0: $\psi = 1$ based on the expected Fisher information of the conditional likelihood.

▶ *We first compute the expected Fisher information.*

$$\begin{aligned}J(\psi) &= \mathsf{E}\{I(\psi)\}\\ &= \frac{1}{(1+\psi)^2}\cdot\left\{\frac{b+c}{1+1/\psi}\cdot\frac{1+2\psi}{\psi^2}-\frac{b+c}{1+\psi}\right\}\\ &= \frac{1}{(1+\psi)^2}\cdot\frac{b+c}{1+\psi}\cdot\frac{1+\psi}{\psi}\\ &= \frac{b+c}{\psi(1+\psi)^2}.\end{aligned}$$

The score test statistic is $S(\psi_0)/\sqrt{J(\psi_0)}$, where $\psi_0 = 1$. Using the results derived above, we obtain the statistic in the form

$$\frac{\frac{b}{1\cdot(1+1)}-\frac{c}{1+1}}{\sqrt{\frac{b+c}{1\cdot(1+1)^2}}} = \frac{b-c}{\sqrt{b+c}}.$$

17. Let $Y_i \overset{\text{ind}}{\sim} \text{Bin}(1, \pi_i)$, $i = 1, \dots, n$, be the binary response variables in a *logistic regression model*, where the probabilities $\pi_i = F(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ are parametrised via the inverse logit function

$$F(x) = \frac{\exp(x)}{1+\exp(x)}$$

by the regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$. The vector $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^\top$ contains the values of the p covariates for the i-th observation.

a) Show that F is indeed the inverse of the logit function $\text{logit}(x) = \log\{x/(1-x)\}$, and that $\frac{d}{dx}F(x) = F(x)\{1-F(x)\}$.

▶ *We have*

$$\begin{aligned} y = \text{logit}(x) &= \log\left(\frac{x}{1-x}\right)\\ \iff \exp(y) &= \frac{x}{1-x}\\ \iff \exp(y)(1-x) &= x\\ \iff \exp(y) &= x + x\exp(y) = x\{1+\exp(y)\}\\ \iff x &= \frac{\exp(y)}{1+\exp(y)} = F(y),\end{aligned}$$

which shows that $F = \text{logit}^{-1}$. For the derivative:

$$\begin{aligned}\frac{d}{dx}F(x) &= \frac{d}{dx}\left\{\frac{\exp(x)}{1+\exp(x)}\right\}\\ &= \frac{\exp(x)\{1+\exp(x)\}-\exp(x)\exp(x)}{\{1+\exp(x)\}^2}\\ &= \frac{\exp(x)}{1+\exp(x)}\frac{1}{1+\exp(x)}\\ &= F(x)\{1-F(x)\}.\end{aligned}$$

b) Use the results on multivariate derivatives outlined in Appendix B.2.2 to show that the log-likelihood, score vector and Fisher information matrix of $\boldsymbol{\beta}$, given the realisation $\boldsymbol{y} = (y_1, \dots, y_n)^\top$, are

$$\begin{aligned}l(\boldsymbol{\beta}) &= \sum_{i=1}^n y_i\log(\pi_i) + (1-y_i)\log(1-\pi_i),\\ \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^n (y_i-\pi_i)\boldsymbol{x}_i = \boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{\pi})\\ \text{and}\quad \boldsymbol{I}(\boldsymbol{\beta}) &= \sum_{i=1}^n \pi_i(1-\pi_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top = \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X},\end{aligned}$$

respectively, where $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top$ is the design matrix, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$ and $\boldsymbol{W} = \text{diag}\{\pi_i(1-\pi_i)\}_{i=1}^n$.

▶ *The probability mass function for Y_i is*

$$f(y_i; \boldsymbol{\beta}) = \pi_i^{y_i}(1-\pi_i)^{(1-y_i)},$$

where $\pi_i = F(\boldsymbol{x}_i^\top\boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$. Because of independence of the n random variables Y_i, $i = 1, \dots, n$, the log-likelihood of $\boldsymbol{\beta}$ is

$$\begin{aligned}l(\boldsymbol{\beta}) &= \sum_{i=1}^n \log\{f(y_i;\boldsymbol{\beta})\}\\ &= \sum_{i=1}^n y_i\log(\pi_i) + (1-y_i)\log(1-\pi_i).\end{aligned}$$

To derive the score vector, we have to take the derivative with respect to β_j, $j = 1, \dots, p$. Using the chain rule for the partial derivative of a function g of π_i, we obtain that

$$\frac{d}{d\beta_j}g(\pi_i) = \frac{d}{d\beta_j}g[F\{\eta_i(\boldsymbol{\beta})\}] = \frac{d}{d\pi_i}g(\pi_i)\cdot\frac{d}{d\eta_i}F(\eta_i)\cdot\frac{d}{d\beta_j}\eta_i(\boldsymbol{\beta}),$$

where $\eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \sum_{j=1}^p x_{ij}\beta_j$ *is the linear predictor for the* i*-th observation. In our case,*

$$\begin{aligned}
\frac{d}{d\beta_j} l(\boldsymbol{\beta}) &= \sum_{i=1}^n \left(\frac{y_i}{\pi_i} + \frac{y_i - 1}{1 - \pi_i}\right) \cdot \pi_i(1 - \pi_i) \cdot x_{ij} \\
&= \sum_{i=1}^n (y_i - \pi_i) x_{ij}.
\end{aligned}$$

Together we have obtained

$$\boldsymbol{S}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{d}{d\beta_1} l(\boldsymbol{\beta}) \\ \vdots \\ \frac{d}{d\beta_p} l(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n (y_i - \pi_i) x_{i1} \\ \vdots \\ \sum_{i=1}^n (y_i - \pi_i) x_{ip} \end{pmatrix} = \sum_{i=1}^n (y_i - \pi_i) \boldsymbol{x}_i.$$

We would have obtained the same result using vector differentiation:

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) \\
&= \sum_{i=1}^n \left(\frac{y_i}{\pi_i} + \frac{y_i - 1}{1 - \pi_i}\right) \cdot \pi_i(1 - \pi_i) \cdot \frac{\partial}{\partial \boldsymbol{\beta}} \eta_i(\boldsymbol{\beta}) \\
&= \sum_{i=1}^n (y_i - \pi_i) \boldsymbol{x}_i,
\end{aligned}$$

because the chain rule also works here, and $\frac{\partial}{\partial \boldsymbol{\beta}} \eta_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{x}_i^\top \boldsymbol{\beta} = \boldsymbol{x}_i$*. It is easily seen that we can also write this as* $\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\pi})$*.*

We now use vector differentiation to derive the Fisher information matrix:

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \boldsymbol{\beta}^\top} \boldsymbol{S}(\boldsymbol{\beta}) \\
&= -\sum_{i=1}^n (-\boldsymbol{x}_i) \cdot \pi_i(1 - \pi_i) \cdot \boldsymbol{x}_i^\top \\
&= \sum_{i=1}^n \pi_i(1 - \pi_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top.
\end{aligned}$$

We can rewrite this as $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ *if we define* $\boldsymbol{W} = \operatorname{diag}\{\pi_i(1 - \pi_i)\}_{i=1}^n$*.*

c) Show that the statistic $\boldsymbol{T}(y) = \sum_{i=1}^n y_i \boldsymbol{x}_i$ is minimal sufficient for $\boldsymbol{\beta}$.

▶ *We can rewrite the log-likelihood as follows:*

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^n y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \\
&= \sum_{i=1}^n y_i \{\log(\pi_i) - \log(1 - \pi_i)\} + \log(1 - \pi_i) \\
&= \sum_{i=1}^n y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sum_{i=1}^n \log\{1 - F(\boldsymbol{x}_i^\top \boldsymbol{\beta})\} \\
&= \boldsymbol{T}(y)^\top \boldsymbol{\beta} - A(\boldsymbol{\beta}), \qquad (5.4)
\end{aligned}$$

where $A(\boldsymbol{\beta}) = -\sum_{i=1}^n \log\{1 - F(\boldsymbol{x}_i^\top \boldsymbol{\beta})\}$*. Now, (5.4) is in the form of an exponential family of order* p *in canonical form; cf. Exercise 8 in Chapter 3. In that exercise, we showed that* $\boldsymbol{T}(y)$ *is minimal-sufficient for* $\boldsymbol{\beta}$*.*

d) Implement an R-function which maximises the log-likelihood using the Newton-Raphson algorithm (see Appendix C.1.2) by iterating

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{I}(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)}), \quad t = 1, 2, \ldots$$

until the new estimate $\boldsymbol{\beta}^{(t+1)}$ and the old one $\boldsymbol{\beta}^{(t)}$ are almost identical and $\hat{\boldsymbol{\beta}}_{\text{ML}} = \boldsymbol{\beta}^{(t+1)}$. Start with $\boldsymbol{\beta}^{(1)} = \boldsymbol{0}$.

▶ *Note that*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{I}(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)})$$

is equivalent to

$$\boldsymbol{I}(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) = \boldsymbol{S}(\boldsymbol{\beta}^{(t)}).$$

Since it is numerically more convenient to solve an equation directly instead of computing a matrix inverse and then multiplying it with the right-hand side, we will be solving

$$\boldsymbol{I}(\boldsymbol{\beta}^{(t)}) \boldsymbol{v} = \boldsymbol{S}(\boldsymbol{\beta}^{(t)})$$

for $\boldsymbol{v}$ *and then deriving the next iterate as* $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{v}$*.*

```
> ## first implement score vector and Fisher information:
> scoreVec <- function(beta,
                       data) ## assume this is a list of vector y and matrix X
  {
      yMinusPiVec <- data$y - plogis(data$X %*% beta) # (y - pi)
      return(crossprod(data$X, yMinusPiVec)) # X^T * (y - pi)
  }
> fisherInfo <- function(beta,
                         data)
  {
      piVec <- as.vector(plogis(data$X %*% beta))
      W <- diag(piVec * (1 - piVec))
      return(t(data$X) %*% W %*% data$X)
  }
> ## here comes the Newton-Raphson algorithm:
> computeMle <- function(data)
  {
      ## start with the null vector
      p <- ncol(data$X)
      beta <- rep(0, p)
      names(beta) <- colnames(data$X)

      ## loop only to be left by returning the result
      while(TRUE)
      {
          ## compute increment vector v
          v <- solve(fisherInfo(beta, data),
                     scoreVec(beta, data))

          ## update the vector
          beta <- beta + v

          ## check if we have converged
          if(sum(v^2) < 1e-8)
          {
              return(beta)
          }
      }
  }
```

e) Consider the data set `amlxray` on the connection between X-ray usage and acute myeloid leukaemia in childhood, which is available in the R-package `faraway`. Here $y_i = 1$ if the disease was diagnosed for the i-th child and $y_i = 0$ otherwise (`disease`). We include an intercept in the regression model, *i. e.* we set $x_1 = 1$. We want to analyse the association of the diabetes status with the covariates x_2 (`age` in years), x_3 (1 if the child is male and 0 otherwise, `Sex`), x_4 (1 if the mother ever have an X-ray and 0 otherwise, `Mray`) and x_5 (1 if the father ever have an X-ray and 0 otherwise, `Fray`).

Interpret $\beta_2, \dots, \beta_5$ by means of odds ratios. Compute the MLE $\hat{\boldsymbol{\beta}}_{\text{ML}} = (\hat{\beta}_1, \dots, \hat{\beta}_5)^\top$ and standard errors $\text{se}(\hat{\beta}_i)$ for all coefficient estimates $\hat{\beta}_i$, and construct 95% Wald confidence intervals for β_i ($i = 1, \dots, 5$). Interpret the results, and compare them with those from the R-function `glm` (using the `binomial` family).

▶ *In order to interpret the coefficients, consider two covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The modelled odds ratio for having leukaemia ($y = 1$) is then*

$$\frac{\pi_i/(1-\pi_i)}{\pi_j/(1-\pi_j)} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})} = \exp\{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{\beta}\}.$$

If now $x_{ik} - x_{jk} = 1$ for one covariate k and $x_{il} = x_{jl}$ for all other covariates $l \neq k$, then this odds ratio is equal to $\exp(\beta_k)$. Thus, we can interpret β_2 as the log odds ratio for person i versus person j who is one year younger than person i; β_3 as the log odds ratio for a male versus a female; likewise, the mother's ever having had an X-ray gives odds ratio of $\exp(\beta_3)$, so the odds are $\exp(\beta_3)$ times higher; and the father's ever having had an X-ray changes the odds by factor $\exp(\beta_4)$.

We now consider the data set `amlxray` and first compute the ML estimates of the regression coefficients:

```
> library(faraway)
> # formatting the data
> data.subset <- amlxray[, c("disease", "age", "Sex", "Mray", "Fray")]
> data.subset$Sex <- as.numeric(data.subset$Sex)-1
> data.subset$Mray <- as.numeric(data.subset$Mray)-1
> data.subset$Fray <- as.numeric(data.subset$Fray)-1
> data <- list(y = data.subset$disease,
               X =
               cbind(intercept=1,
                     as.matrix(data.subset[, c("age", "Sex", "Mray", "Fray")])))
> (myMle <- computeMle(data))
                  [,1]
intercept -0.282982807
age       -0.002951869
Sex        0.124662103
Mray      -0.063985047
Fray       0.394158386
> (myOr <- exp(myMle))
               [,1]
intercept 0.7535327
age       0.9970525
Sex       1.1327656
Mray      0.9380190
Fray      1.4831354
```

We see a negative association of the leukaemia risk with age and the mother's ever having had an X-ray, and a positive association of the leukaemia risk with male gender and the father's ever having had an X-ray, and could now say the same as described above, e. g. being a male increases the odds for leukaemia by 13.3 %. However, these are only estimates and we need to look at the associated standard errors before we make conclusions.

We easily obtain the standard errors by inverting the observed Fisher information and taking the square root of the resulting diagonal. This leads to the (transformed) Wald confidence intervals:

```
> (obsFisher <- fisherInfo(myMle, data))
           intercept        age        Sex      Mray
intercept  58.698145  414.95349  29.721114  4.907156
age       414.953492 4572.35780 226.232641 38.464410
Sex        29.721114  226.23264  29.721114  1.973848
Mray        4.907156   38.46441   1.973848  4.907156
Fray       16.638090  128.19588   8.902213  1.497194
                Fray
intercept  16.638090
age       128.195880
Sex         8.902213
```

```
Mray        1.497194
Fray       16.638090
> (mySe <- sqrt(diag(solve(obsFisher))))
 intercept        age        Sex       Mray       Fray
0.25790020 0.02493209 0.26321083 0.47315533 0.29059384
> (myWaldCis <- as.vector(myMle) + qnorm(0.975) * t(sapply(mySe, "*", c(-1, +1))))
                 [,1]       [,2]
intercept -0.78845792 0.22249230
age       -0.05181788 0.04591414
Sex       -0.39122164 0.64054584
Mray      -0.99135245 0.86338235
Fray      -0.17539508 0.96371185
> (myWaldOrCis <- exp(myWaldCis))
               [,1]     [,2]
intercept 0.4545452 1.249186
age       0.9495018 1.046985
Sex       0.6762303 1.897516
Mray      0.3710745 2.371167
Fray      0.8391254 2.621409
```

Note that all the confidence intervals for interesting coefficients cover zero; thus, neither of the associations interpreted above is significant.
We can compare the results with those from the standard `glm` *function:*

```
> amlGlm <- glm(disease ~ age + Sex + Mray + Fray,
                family=binomial(link="logit"),
                data=amlxray)
> summary(amlGlm)
Call:
glm(formula = disease ~ age + Sex + Mray + Fray, family = binomial(link = "logit"),
    data = amlxray)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.279  -1.099  -1.043   1.253   1.340

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.282983   0.257896  -1.097    0.273
age         -0.002952   0.024932  -0.118    0.906
SexM         0.124662   0.263208   0.474    0.636
Mrayyes     -0.063985   0.473146  -0.135    0.892
Frayyes      0.394158   0.290592   1.356    0.175

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.86  on 237  degrees of freedom
Residual deviance: 326.72  on 233  degrees of freedom
AIC: 336.72

Number of Fisher Scoring iterations: 3
> myMle
                  [,1]
intercept -0.282982807
age       -0.002951869
Sex        0.124662103
```

```
Mray      -0.063985047
Fray       0.394158386
> mySe
 intercept        age        Sex       Mray       Fray
0.25790020 0.02493209 0.26321083 0.47315533 0.29059384
```

We now have a confirmation of our results.

f) Implement an R-function which returns the profile log-likelihood of one of the p parameters. Use it to construct 95% profile likelihood confidence intervals for them. Compare with the Wald confidence intervals from above, and with the results from the R-function `confint` applied to the `glm` model object.

▶ *In order to compute the profile log-likelihood of a coefficient β_j, we have to think about how to obtain $\hat{\boldsymbol{\beta}}_{-j}(\beta_j)$, the ML estimate of all other coefficients in the vector $\boldsymbol{\beta}$ except the j-th one, given a fixed value for β_j. We know that we have to solve all score equations except the j-th one. And we can do that again with the Newton-Raphson algorithm, by now iterating*

$$\boldsymbol{\beta}^{(t+1)}_{-j} = \boldsymbol{\beta}^{(t)}_{-j} + \boldsymbol{I}(\boldsymbol{\beta}^{(t)})^{-1}_{-j}\boldsymbol{S}(\boldsymbol{\beta}^{(t)})_{-j}, \quad t = 1, 2, \dots,$$

where $\boldsymbol{I}(\boldsymbol{\beta}^{(t)})_{-j}$ is the Fisher information matrix computed from the current estimate of $\boldsymbol{\beta}_{-j}$ and the fixed β_j, and then leaving out the j-th row and column. Likewise, $\boldsymbol{S}(\boldsymbol{\beta}^{(t)})_{-j}$ is the score vector without the j-th element.
The rest is then easy, we just plug in $\hat{\boldsymbol{\beta}}_{-j}(\beta_j)$ and β_j into the full log-likelihood:

```
> ## the full log-likelihood:
> fullLogLik <- function(beta,
                         data)
  {
      piVec <- as.vector(plogis(data$X %*% beta))
      ret <- sum(data$y * log(piVec) + (1 - data$y) * log(1 - piVec))
      return(ret)
  }
> ## compute the MLE holding the j-th coefficient fixed:
> computeConditionalMle <- function(data,
                                    value, # the fixed value
                                    index) # j
  {
      ## start with the MLE except for the value
      p <- ncol(data$X)
      beta <- computeMle(data)
      beta[index, ] <- value

      ## loop only to be left by returning the result
      while(TRUE)
      {
          ## compute increment vector v for non-fixed part
          v <- solve(fisherInfo(beta, data)[-index, -index, drop=FALSE],
                     scoreVec(beta, data)[-index, , drop=FALSE])

          ## update the non-fixed part of the beta vector
          beta[-index, ] <- beta[-index, ] + v
          ## check if we have converged
          if(sum(v^2) < 1e-8)
          {
              return(beta)
          }
      }
  }
> ## so the profile log-likelihood is:
> profileLogLik <- function(beta,         # this is scalar now!
                            data,
                            index)
  {
      fullLogLik(computeConditionalMle(data=data,
                                       value=beta,
                                       index=index),
                 data=data)
  }
> ## now for our data, compute 95% profile likelihood CIs:
> profLogLikCis <- matrix(nrow=ncol(data$X),
                          ncol=2,
                          dimnames=
                          list(colnames(data$X),
                               c("lower", "upper")))
> for(j in seq_len(ncol(data$X)))
  {
      ## this function must equal 0
      targetFun <- function(beta)
      {
          likRatioStat <-
              - 2 * (profileLogLik(beta, data=data, index=j) -
                     profileLogLik(myMle[j], data=data, index=j))
          likRatioStat - qchisq(0.95, df=1)
      }

      ## compute bounds
      ## note that we cannot use too large intervals because
      ## the numerics are not very stable... so we start with the Wald
      ## interval approximation first
      addSpace <- abs(diff(myWaldCis[j,])) / 10
      lower <- uniroot(targetFun,
                       interval=c(myWaldCis[j,1] - addSpace, myMle[j]))$root
      upper <- uniroot(targetFun,
                       interval=c(myMle[j], myWaldCis[j,2] + addSpace))$root

      ## save them correctly
      profLogLikCis[j, ] <- c(lower, upper)
  }
> ## our result:
> profLogLikCis
                lower      upper
intercept -0.79330733 0.22054433
age       -0.05203807 0.04595938
Sex       -0.39139772 0.64191554
Mray      -1.01577053 0.86400486
Fray      -0.17431963 0.96781339
```

```
> ## comparison with the R results:
> confint(amlGlm)
                  2.5 %     97.5 %
(Intercept) -0.79331983 0.22054713
age         -0.05202761 0.04594725
SexM        -0.39140156 0.64192116
Mrayyes     -1.01582557 0.86404508
Frayyes     -0.17432400 0.96782542
> ## so we are quite close to these!
>
> ## the Wald results:
> myWaldCis
                 [,1]       [,2]
intercept -0.78845792 0.22249230
age       -0.05181788 0.04591414
Sex       -0.39122164 0.64054584
Mray      -0.99135245 0.86338235
Fray      -0.17539508 0.96371185
```

Unfortunately, our algorithm is not very stable, which means that we must start with the full MLE configuration to search the constrained solution, and we must not choose parameter values too far away from the MLE. Here we pragmatically start with the Wald interval bounds and add an amount depending on the width of the Wald interval, which gives us a sensible parameter range. In practice one should always use the `confint` *routine to compute profile likelihood intervals in GLMs.*

In comparison to the Wald intervals, the profile likelihood intervals are a little shifted to the left for the coefficients β_2, β_3 and β_4, slightly wider for the intercept β_1 and slightly shorter for the coefficient β_5. However, all these differences are very small in size.

g) We want to test if the inclusion of the covariates x_2, x_3, x_4 and x_5 improves the fit of the model to the data. To this end, we consider the null hypothesis $H_0 : \beta_2 = \cdots = \beta_5 = 0$.

How can this be expressed in the form $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\delta}$, where $\boldsymbol{C}$ is a $q \times p$ contrast matrix (of rank $q \leq p$) and $\boldsymbol{\delta}$ is a vector of length q? Use a result from Appendix A.2.4 to show that under H_0,

$$(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta})^\top \left\{ \boldsymbol{C}\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{\text{ML}})^{-1}\boldsymbol{C}^\top \right\}^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta}) \overset{\text{a}}{\sim} \chi^2(q). \tag{5.5}$$

► *If we set*

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

then $\boldsymbol{C\beta} = (\beta_2, \beta_3, \beta_4, \beta_5)$, and with the right hand side $\boldsymbol{\delta} = \mathbf{0}$ we have expressed the null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ in the form of a so-called linear hypothesis: $H_0 : \boldsymbol{C\beta} = \boldsymbol{\delta}$.
From Section 5.4.2 we know that

$$\hat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathrm{N}_p(\boldsymbol{\beta}, \boldsymbol{I}(\hat{\boldsymbol{\beta}}_{\text{ML}})^{-1}),$$

hence

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathrm{N}_q(\boldsymbol{C\beta}, \boldsymbol{C}\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{\text{ML}})^{-1}\boldsymbol{C}^\top).$$

Now under H_0, $\boldsymbol{C\beta} = \boldsymbol{\delta}$, therefore

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathrm{N}_q(\boldsymbol{\delta}, \boldsymbol{C}\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{\text{ML}})^{-1}\boldsymbol{C}^\top)$$

and

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta}) \overset{a}{\sim} \mathrm{N}_q(\mathbf{0}, \boldsymbol{I}_q)$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{\text{ML}})^{-1}\boldsymbol{C}^\top$. Analogous to Section 5.4, we finally get

$$\begin{aligned}
&\left\{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta})\right\}^\top \left\{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta})\right\} \\
&= (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta})^\top \left\{\boldsymbol{\Sigma}^{-1/2}\right\}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta}) \\
&= (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{ML}} - \boldsymbol{\delta}) \quad \overset{a}{\sim} \quad \chi^2(q).
\end{aligned}$$

h) Compute two P-values quantifying the evidence against H_0, one based on the squared Wald statistic (5.5), the other based on the generalised likelihood ratio statistic.

▶

```
> ## First the Wald statistic:
> (Cmatrix <- cbind(0, diag(4)))
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    1    0    0    0
[2,]    0    0    1    0    0
[3,]    0    0    0    1    0
[4,]    0    0    0    0    1
> (waldStat <- t(Cmatrix %*% myMle) %*%
               solve(Cmatrix %*% solve(obsFisher) %*% t(Cmatrix)) %*%
               (Cmatrix %*% myMle))
       [,1]
[1,] 2.1276
> (p.Wald <- pchisq(q=waldStat, df=nrow(Cmatrix), lower.tail=FALSE))
          [,1]
[1,] 0.7123036
```

```
> ## Second the generalized likelihood ratio statistic:
>
> ## We have to fit the model under the H0 restriction, so
> ## only with intercept.
> h0data <- list(y=data$y,
                 X=data$X[, "intercept", drop=FALSE])
> myH0mle <- computeMle(data=h0data)
> ## then the statistic is
> (likRatioStat <- 2 * (fullLogLik(beta=myMle,
                                   data=data) -
                        fullLogLik(beta=myH0mle,
                                   data=h0data)))
[1] 2.141096
> (p.likRatio <- pchisq(q=likRatioStat, df=3, lower.tail=FALSE))
[1] 0.5436436
```

We see that neither of the two statistics finds much evidence against H_0, so the association of the oral cancer risk with the set of these four covariates is not significant.
Note that we can compute the generalised likelihood ratio statistic for the comparison of two models in R *using the* `anova` *function. For example, we can reproduce the null deviance from above using the following code:*

```
> anova(update(amlGlm, . ~ 1), amlGlm, test="LRT")
Analysis of Deviance Table

Model 1: disease ~ 1
Model 2: disease ~ age + Sex + Mray + Fray
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       237     328.86
2       233     326.72  4   2.1411   0.7098
```

i) Since the data is actually from a matched case-control study, where pairs of one case and one control have been matched (by age, race and county of residence; the variable `ID` denotes the matched pairs), it is more appropriate to apply conditional logistic regression. Compute the corresponding MLEs and 95% confidence intervals with the R-function `clogit` from the package `survival`, and compare the results.

▶ *Since the controls and cases are matched by age, we cannot evaluate the effect of age. We therefore fit a conditional logistic regression model with* `Sex`*,* `Mray` *and* `Fray` *as covariates. We do not look at the estimate of the intercept either, as each pair has its own intercept in the conditional logistic regression.*

```
> library(survival)
> # compute the results
> amlClogit <- clogit( disease ~ Sex + Mray + Fray + strata(ID), data=amlxray )
> amlClogitWaldCis <- amlClogit$coefficients +
                      qnorm(0.975) * t(sapply(sqrt(diag(amlClogit$var)), "*", c(-1, +1)))
> colnames(amlClogitWaldCis, do.NULL=TRUE)
NULL
> rownames(amlClogitWaldCis, do.NULL=TRUE)
NULL
```

```
> colnames(amlClogitWaldCis) <-c("2.5 %", "97.5 %")
> rownames(amlClogitWaldCis) <-c("Sex", "Mray", "Fray")
> # compare the results
> summary(amlClogit)
Call:
coxph(formula = Surv(rep(1, 238L), disease) ~ Sex + Mray + Fray +
    strata(ID), data = amlxray, method = "exact")

  n= 238, number of events= 111

            coef exp(coef) se(coef)      z Pr(>|z|)
SexM     0.10107   1.10635  0.34529  0.293    0.770
Mrayyes -0.01332   0.98677  0.46232 -0.029    0.977
Frayyes  0.44567   1.56154  0.30956  1.440    0.150

        exp(coef) exp(-coef) lower .95 upper .95
SexM       1.1064     0.9039    0.5623     2.177
Mrayyes    0.9868     1.0134    0.3987     2.442
Frayyes    1.5615     0.6404    0.8512     2.865

Rsquare= 0.01   (max possible= 0.501 )
Likelihood ratio test= 2.36  on 3 df,   p=0.5003
Wald test            = 2.31  on 3 df,   p=0.5108
Score (logrank) test = 2.35  on 3 df,   p=0.5029
> summary(amlGlm)
Call:
glm(formula = disease ~ age + Sex + Mray + Fray, family = binomial(link = "logit"),
    data = amlxray)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.279  -1.099  -1.043   1.253   1.340

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.282983   0.257896  -1.097    0.273
age         -0.002952   0.024932  -0.118    0.906
SexM         0.124662   0.263208   0.474    0.636
Mrayyes     -0.063985   0.473146  -0.135    0.892
Frayyes      0.394158   0.290592   1.356    0.175

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.86  on 237  degrees of freedom
Residual deviance: 326.72  on 233  degrees of freedom
AIC: 336.72

Number of Fisher Scoring iterations: 3
> amlClogitWaldCis
          2.5 %    97.5 %
Sex  -0.5756816 0.7778221
Mray -0.9194578 0.8928177
Fray -0.1610565 1.0523969
> confint(amlGlm)[3:5, ]
             2.5 %    97.5 %
SexM    -0.3914016 0.6419212
Mrayyes -1.0158256 0.8640451
Frayyes -0.1743240 0.9678254
```

The results from the conditional logistic regression are similar to those obtained above in the direction of the effects (negative or positive) and, more importantly, in that neither of the associations is significant. The actual numerical values differ, of course.

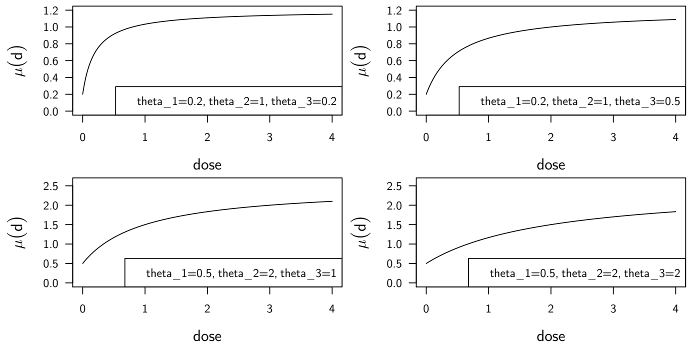
18. In clinical dose-finding studies, the relationship between the dose $d \geq 0$ of the medication and the average response $\mu(d)$ in a population is to be inferred. Considering a continuously measured response y, then a simple model for the individual measurements assumes $y_{ij} \overset{\text{ind}}{\sim} \mathrm{N}(\mu(d_{ij};\boldsymbol{\theta}),\sigma^2)$, $i = 1,\dots,K$, $j = 1,\dots,n_i$. Here n_i is the number of patients in the i-th dose group with dose d_i (placebo group has $d = 0$). The *Emax model* has the functional form

$$\mu(d;\boldsymbol{\theta}) = \theta_1 + \theta_2 \frac{d}{d+\theta_3}.$$

a) Plot the function $\mu(d;\boldsymbol{\theta})$ for different choices of the parameters $\theta_1, \theta_2, \theta_3 > 0$. Give reasons for the interpretation of θ_1 as the mean placebo response, θ_2 as the maximum treatment effect, and θ_3 as the dose giving 50% of the maximum treatment effect.

▶ *If $d = 0$ (the dose in the placebo group), then $\mu(0;\boldsymbol{\theta}) = \theta_1 + \theta_2 \cdot 0/(0+\theta_3) = \theta_1$. Thus, θ_1 is the mean response in the placebo group. Further, $\mu(d;\boldsymbol{\theta}) = \theta_1 + \theta_2 \cdot d/(d+\theta_3) = \theta_1 + \theta_2 \cdot 1/(1+\theta_3/d)$. Hence, the mean response is smallest for the placebo group ($\mu(0;\boldsymbol{\theta}) = \theta_1$) and increases for increasing d. As d tends to infinity, $\theta_2 \cdot 1/(1+\theta_3/d)$ approaches θ_2 from below. Thus, θ_2 is the maximum treatment effect. Finally, $\mu(\theta_3;\boldsymbol{\theta}) = \theta_1 + \theta_2 \cdot \theta_3/(2\theta_3) = \theta_1 + \theta_2/2$. This dose therefore gives 50% of the maximum treatment effect θ_2. We now draw $\mu(d;\boldsymbol{\theta})$ for various values of $\boldsymbol{\theta}$.*

```
> # define the Emax function
> Emax <- function(dose, theta1, theta2, theta3){
          return( theta1 + theta2*dose/(theta3+dose) )
    }
> # we look at doses between 0 and 4
> dose <- seq(from=0, to=4, by=0.001)
> # choice of the parameter values
> theta1 <- c(0.2, 0.2, 0.5, 0.5)
> theta2 <- c(1, 1, 2, 2)
> theta3 <- c(0.2, 0.5, 1, 2)
> ##
> # store the responses for the different parameter combinations in a list
> resp <- list()
> for(i in 1:4){
    resp[[i]] <-  Emax(dose, theta1[i], theta2[i], theta3[i])
    }
> # plot the different dose-mean-response relationships
> m.resp <- expression(mu(d))    # abbreviation for legend
> par(mfrow=c(2,2))
> plot(x=dose, y=resp[[1]], type="l", ylim=c(0,1.2), xlab="dose", ylab=m.resp)
> legend("bottomright", legend= "theta_1=0.2, theta_2=1, theta_3=0.2")
> plot(x=dose, y=resp[[2]], type="l", ylim=c(0,1.2), xlab="dose", ylab=m.resp)
> legend("bottomright", legend= "theta_1=0.2, theta_2=1, theta_3=0.5")
> plot(x=dose, y=resp[[3]], type="l", ylim=c(0,2.6), xlab="dose", ylab=m.resp)
> legend("bottomright", legend= "theta_1=0.5, theta_2=2, theta_3=1")
> plot(x=dose,resp[[4]], type="l", ylim=c(0,2.6), xlab="dose", ylab=m.resp)
> legend("bottomright", legend= "theta_1=0.5, theta_2=2, theta_3=2")
```



b) Compute the expected Fisher information for the parameter vector $\boldsymbol{\theta}$. Using this result, implement an R function which calculates the approximate covariance matrix of the MLE $\hat{\boldsymbol{\theta}}_{\text{ML}}$ for a given set of doses $d_1, \dots, d_K$, a total sample size $N = \sum_{i=1}^K n_i$, allocation weights $w_i = n_i/N$ and given error variance σ^2.

▶ *The log-likelihood kernel is*

$$l(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \{y_{ij} - \mu(d_{ij}; \boldsymbol{\theta})\}^2,$$

the score function is

$$\mathbf{S}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{d}{d\theta_1} l(\boldsymbol{\theta}) \\ \frac{d}{d\theta_2} l(\boldsymbol{\theta}) \\ \frac{d}{d\theta_3} l(\boldsymbol{\theta}) \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^K \tilde{y}_i \\ \sum_{i=1}^K \tilde{y}_i \cdot \frac{d_i}{d_i+\theta_3} \\ -\theta_2 \sum_{i=1}^K \tilde{y}_i \cdot \frac{d_i}{(d_i+\theta_3)^2} \end{pmatrix},$$

and the Fisher information matrix is

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^K n_i & \sum_{i=1}^K \frac{n_i d_i}{\delta_i} & -\theta_2 \sum_{i=1}^K \frac{n_i d_i}{\delta_i^2} \\ \sum_{i=1}^K \frac{n_i d_i}{\delta_i} & \sum_{i=1}^K \frac{n_i d_i^2}{\delta_i^2} & \sum_{i=1}^K \frac{d_i}{\delta_i^3}(\delta_i \tilde{y}_i - n_i d_i \theta_2) \\ -\theta_2 \sum_{i=1}^K \frac{n_i d_i}{\delta_i^2} & \sum_{i=1}^K \frac{d_i}{\delta_i^3}(\delta_i \tilde{y}_i - n_i d_i \theta_2) & \theta_2 \sum_{i=1}^K \frac{d_i}{\delta_i^4}(-2\delta_i \tilde{y}_i + n_i d_i \theta_2) \end{pmatrix},$$

where $\tilde{y}_i = \sum_{j=1}^{n_i} \{y_{ij} - \mu(d_i; \boldsymbol{\theta})\}$ and $\delta_i = d_i + \theta_3$. The expected Fisher information matrix is thus

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^K n_i & \sum_{i=1}^K \frac{n_i d_i}{\delta_i} & -\theta_2 \sum_{i=1}^K \frac{n_i d_i}{\delta_i^2} \\ \sum_{i=1}^K \frac{n_i d_i}{\delta_i} & \sum_{i=1}^K \frac{n_i d_i^2}{\delta_i^2} & -\theta_2 \sum_{i=1}^K \frac{n_i d_i^2}{\delta_i^3} \\ -\theta_2 \sum_{i=1}^K \frac{n_i d_i}{\delta_i^2} & -\theta_2 \sum_{i=1}^K \frac{n_i d_i^2}{\delta_i^3} & \theta_2^2 \sum_{i=1}^K \frac{n_i d_i^2}{\delta_i^4} \end{pmatrix}.$$

The following R function uses $\mathbf{J}(\boldsymbol{\theta})$ to obtain the approximate covariance matrix of the MLE $\hat{\boldsymbol{\theta}}_{\text{ML}}$ for a given set of doses $d_1, \dots, d_K$, a total sample size $N = \sum_{i=1}^K n_i$, allocation weights $w_i = n_i/N$, and given error variance σ^2.

```
> ApprVar <- function(doses, N, w, sigma, theta1, theta2, theta3){
    delta <- doses + theta3
    n <- w*N
    V <- matrix(NA, nrow=3, ncol=3)
    diag(V) <- c( N, sum(n*doses^2/delta^2), theta2^2*sum(n*doses^2/delta^4) )
    V[1, 2] <- V[2, 1] <- sum(n*doses/delta)
    V[1, 3] <- V[3, 1] <- -theta2*sum(n*doses/delta^2)
    V[2, 3] <- V[3, 2] <- -theta2*sum(n*doses^2/delta^3)
    return(sigma^2*solve(V))
  }
```

c) Assume $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = 0.5$ and $\sigma^2 = 1$. Calculate the approximate covariance matrix, first for $K = 5$ doses 0, 1, 2, 3, 4, and second for doses 0, 0.5, 1, 2, 4, both times with balanced allocations $w_i = 1/5$ and total sample size $N = 100$. Compare the approximate standard deviations of the MLEs of the parameters between the two designs, and also compare the determinants of the two calculated matrices.

▶

```
> # with the first set of doses:
> V1 <- ApprVar(doses=c(0:4), N=100, w=c(rep(1/5, 5)),
                sigma=1, theta1=0, theta2=1, theta3=0.5)
> # with the second set of doses:
> V2 <- ApprVar(doses=c(0, 0.5, 1, 2, 4), N=100, w=c(rep(1/5, 5)),
                sigma=1, theta1=0, theta2=1, theta3=0.5)
> # standard errors
> sqrt(diag(V1))
[1] 0.2235767 0.4081669 0.9185196
> sqrt(diag(V2))
[1] 0.2230535 0.3689985 0.6452322
> # determinant
> det(V1)
[1] 0.0007740515
> det(V2)
[1] 0.000421613
```

The second design attains achieves a lower variability of estimation by placing more doses on the increasing part of the dose-response curve.

d) Using the second design, determine the required total sample size N so that the standard deviation for estimation of θ_2 is 0.35 (so that the half-length of a 95% confidence interval is about 0.7).

▶

```
> sample.size <- function(N){
    V2 <- ApprVar(doses=c(0, 0.5, 1, 2, 4), N=N, w=c(rep(1/5, 5)),
                            sigma=1, theta1=0, theta2=1, theta3=0.5)
    return(sqrt(V2[2,2])-0.35)
  }
> # find the root of the function above
> uniroot(sample.size, c(50, 200))
£root
[1] 111.1509

£f.root
[1] -8.309147e-10

£iter
[1] 7

£estim.prec
[1] 6.103516e-05
```

We need a little more than 111 patients in total, that is a little more than 22 per dose. By taking 23 patients per dose, we obtain the desired length of the confidence interval:

```
> V2 <- ApprVar(doses=c(0, 0.5, 1, 2, 4), N=5*23, w=c(rep(1/5, 5)),
                            sigma=1, theta1=0, theta2=1, theta3=0.5)
> sqrt(V2[2,2])
[1] 0.3440929
```

6 Bayesian inference

1. In 1995, O. J. Simpson, a retired American football player and actor, was accused of the murder of his ex-wife Nicole Simpson and her friend Ronald Goldman. His lawyer, Alan M. Dershowitz stated on T.V. that only one-tenth of 1% of men who abuse their wives go on to murder them. He wanted his audience to interpret this to mean that the evidence of abuse by Simpson would only suggest a 1 in 1000 chance of being guilty of murdering her.
However, Merz and Caulkins (1995) and Good (1995) argue that a different probability needs to be considered: the probability that the husband is guilty of murdering his wife given both that he abused his wife *and* his wife was murdered. Both compute this probability using Bayes theorem, but in two different ways. Define the following events:

A : "The woman was abused by her husband."
M : "The woman was murdered by somebody."
G : "The husband is guilty of murdering his wife."

a) Merz and Caulkins (1995) write the desired probability in terms of the corresponding odds as:

$$\frac{\mathsf{Pr}(G \,|\, A, M)}{\mathsf{Pr}(G^c \,|\, A, M)} = \frac{\mathsf{Pr}(A \,|\, G, M)}{\mathsf{Pr}(A \,|\, G^c, M)} \cdot \frac{\mathsf{Pr}(G \,|\, M)}{\mathsf{Pr}(G^c \,|\, M)}. \tag{6.1}$$

They use the fact that, of the 4936 women who were murdered in 1992, about 1430 were killed by their husband. In a newspaper article, Dershowitz stated that "It is, of course, true that, among the small number of men who do kill their present or former mates, a considerable number did first assault them." Merz and Caulkins (1995) interpret "a considerable number" to be 1/2. Finally, they assume that the probability of a wife being abused by her husband, given that she was murdered by somebody else, is the same as the probability of a randomly chosen woman being abused, namely 0.05.
Calculate the odds (6.1) based on this information. What is the corresponding probability of O. J. Simpson being guilty, given that he has abused his wife and

she has been murdered?

▶ *Using the method of Merz and Caulkins (1995) we obtain*

$$\begin{aligned}\mathsf{Pr}(G\,|\,M) &= \frac{1430}{4936} \approx 0.29 \qquad \Rightarrow \mathsf{Pr}(G^c\,|\,M) = \frac{3506}{4936} \approx 0.71\\ \mathsf{Pr}(A\,|\,G, M) &= 0.5\\ \mathsf{Pr}(A\,|\,G^c, M) &= 0.05.\end{aligned}$$

The odds (6.1) *are therefore*

$$\begin{aligned}\frac{\mathsf{Pr}(G\,|\,A, M)}{\mathsf{Pr}(G^c\,|\,A, M)} &= \frac{\mathsf{Pr}(A\,|\,G, M)}{\mathsf{Pr}(A\,|\,G^c, M)} \cdot \frac{\mathsf{Pr}(G\,|\,M)}{\mathsf{Pr}(G^c\,|\,M)}\\ &= \frac{0.5}{0.05} \cdot \frac{\frac{1430}{4936}}{\frac{3506}{4936}}\\ &\approx 4.08,\end{aligned}$$

so that

$$\mathsf{Pr}(G\,|\,A, M) = \frac{4.08}{1+4.08} \approx 0.8.$$

That means the probability of O.J. Simpson being guilty, given that he has abused his wife and she has been murdered, is about 80%.

b) Good (1995) uses the alternative representation

$$\frac{\mathsf{Pr}(G\,|\,A, M)}{\mathsf{Pr}(G^c\,|\,A, M)} = \frac{\mathsf{Pr}(M\,|\,G, A)}{\mathsf{Pr}(M\,|\,G^c, A)} \cdot \frac{\mathsf{Pr}(G\,|\,A)}{\mathsf{Pr}(G^c\,|\,A)}. \tag{6.2}$$

He first needs to estimate $\mathsf{Pr}(G\,|\,A)$ and starts with Dershowitz's estimate of 1/1000 that the abuser will murder his wife. He assumes the probability is at least 1/10 that this will happen in the year in question. Thus $\mathsf{Pr}(G\,|\,A)$ is at least 1/10 000. Obviously $\mathsf{Pr}(M\,|\,G^c, A) = \mathsf{Pr}(M\,|\,A) \approx \mathsf{Pr}(M)$. Since there are about 25 000 murders a year in the U.S. population of 250 000 000, Good (1995) estimates $\mathsf{Pr}(M\,|\,G^c, A)$ to be 1/10 000.

Calculate the odds (6.2) based on this information. What is the corresponding probability of O. J. Simpson being guilty, given that he has abused his wife and she has been murdered?

▶ *Using the method of Good (1995) it follows:*

$$\begin{aligned}\mathsf{Pr}(G\,|\,A) &= \frac{1}{10000} \qquad \Rightarrow \mathsf{Pr}(G^c\,|\,A) = \frac{9999}{10000}\\ \mathsf{Pr}(M\,|\,G^c, A) &\approx \frac{1}{10000}\\ \mathsf{Pr}(M\,|\,G, A) &= 1.\end{aligned}$$

Now we obtain

$$\begin{aligned}\frac{\mathsf{Pr}(G\,|\,A, M)}{\mathsf{Pr}(G^c\,|\,A, M)} &= \frac{\mathsf{Pr}(M\,|\,G, A)}{\mathsf{Pr}(M\,|\,G^c, A)} \cdot \frac{\mathsf{Pr}(G\,|\,A)}{\mathsf{Pr}(G^c\,|\,A)}\\ &\approx \frac{1}{\frac{1}{10000}} \cdot \frac{\frac{1}{10000}}{\frac{9999}{10000}}\\ &\approx 1,\end{aligned}$$

so that $\mathsf{Pr}(G\,|\,A, M) = 0.5$. *That means the probability of O.J. Simpson being guilty, given that he has abused his wife and she has been murdered, is about* 50%.

c) Good (1996) revised this calculation, noting that approximately only a quarter of murdered victims are female, so $\mathsf{Pr}(M\,|\,G^c, A)$ reduces to 1/20 000. He also corrects $\mathsf{Pr}(G\,|\,A)$ to 1/2000, when he realised that Dershowitz's estimate was an annual and not a lifetime risk. Calculate the probability of O. J. Simpson being guilty based on this updated information.

▶ *The revised calculation is now*

$$\begin{aligned}\frac{\mathsf{Pr}(G\,|\,A, M)}{\mathsf{Pr}(G^c\,|\,A, M)} &= \frac{\mathsf{Pr}(M\,|\,G, A)}{\mathsf{Pr}(M\,|\,G^c, A)} \cdot \frac{\mathsf{Pr}(G\,|\,A)}{\mathsf{Pr}(G^c\,|\,A)}\\ &\approx \frac{1}{\frac{1}{20000}} \cdot \frac{\frac{1}{2000}}{\frac{1999}{2000}}\\ &\approx 10,\end{aligned}$$

so that $\mathsf{Pr}(G\,|\,A, M) \approx 0.91$. *Based on this updated information, the probability of O.J. Simpson being guilty, given that he has abused his wife and she has been murdered, is about* 90%.

2. Consider Example 6.4. Here we will derive the implied distribution of $\theta = \mathsf{Pr}(D+\,|\,T+)$ if the prevalence is $\pi \sim \mathrm{Be}(\tilde{\alpha}, \tilde{\beta})$.

a) Deduce with the help of Appendix A.5.2 that

$$\gamma = \frac{\tilde{\alpha}}{\tilde{\beta}} \cdot \frac{1-\pi}{\pi}$$

follows an F distribution with parameters $2\tilde{\beta}$ and $2\tilde{\alpha}$, denoted by $\mathrm{F}(2\tilde{\beta}, 2\tilde{\alpha})$.

▶ *Following the remark on page 336, we first show* $1-\pi \sim \mathrm{Be}(\tilde{\beta}, \tilde{\alpha})$ *and then we deduce* $\gamma = \frac{\tilde{\alpha}}{\tilde{\beta}} \cdot \frac{1-\pi}{\pi} \sim \mathrm{F}(2\tilde{\beta}, 2\tilde{\alpha})$.

Step 1: To obtain the density of $1-\pi$, *we apply the change-of-variables formula to the density of* π.

The transformation function is

$$g(\pi) = 1-\pi$$

and we have

$$\begin{aligned} g^{-1}(y) &= 1 - y, \\ \frac{dg^{-1}(y)}{dy} &= -1. \end{aligned}$$

This gives

$$\begin{aligned} f_{1-\pi}(y) &= \left|\frac{dg^{-1}(y)}{dy}\right| f_\pi\left(g^{-1}(y)\right) \\ &= f_\pi(1-y) \\ &= \frac{1}{B(\tilde{\alpha},\tilde{\beta})}(1-y)^{\tilde{\alpha}-1}(1-(1-y))^{\tilde{\beta}-1} \\ &= \frac{1}{B(\tilde{\beta},\tilde{\alpha})} y^{\tilde{\beta}-1}(1-y)^{\tilde{\alpha}-1}, \end{aligned}$$

where we have used that $B(\tilde{\alpha},\tilde{\beta}) = B(\tilde{\beta},\tilde{\alpha})$, which follows easily from the definition of the beta function (see Appendix B.2.1). Thus, $1 - \pi \sim \mathrm{Be}(\tilde{\beta},\tilde{\alpha})$.
Step 2: We apply the change-of-variables formula again to obtain the density of γ from the density of $1 - \pi$.
We have $\gamma = g(1 - \pi)$, where

$$\begin{aligned} g(x) &= \frac{\tilde{\alpha}}{\tilde{\beta}} \cdot \frac{x}{1-x}, \\ g^{-1}(y) &= \frac{y}{\tilde{\alpha}/\tilde{\beta} + y} = \frac{\tilde{\beta}y}{\tilde{\alpha} + \tilde{\beta}y} \quad \textit{and} \\ \frac{dg^{-1}(y)}{dy} &= \frac{\tilde{\beta}(\tilde{\alpha} + \tilde{\beta}y) - \tilde{\beta}^2 y}{(\tilde{\alpha} + \tilde{\beta}y)^2} = \frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha} + \tilde{\beta}y)^2}. \end{aligned}$$

Hence,

$$\begin{aligned} f_\gamma(y) &= f_{1-\pi}\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| \\ &= \frac{1}{B(\tilde{\beta},\tilde{\alpha})}\left(\frac{\tilde{\beta}y}{\tilde{\alpha}+\tilde{\beta}y}\right)^{\tilde{\beta}-1}\left(1 - \frac{\tilde{\beta}y}{\tilde{\alpha}+\tilde{\beta}y}\right)^{\tilde{\alpha}-1}\frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta}y)^2} \\ &= \frac{1}{B(\tilde{\beta},\tilde{\alpha})}\left(\frac{\tilde{\alpha}+\tilde{\beta}y}{\tilde{\beta}y}\right)^{1-\tilde{\beta}}\left(\frac{\tilde{\alpha}+\tilde{\beta}y}{\tilde{\alpha}}\right)^{1-\tilde{\alpha}}\frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta}y)^2} \\ &= \frac{1}{B(\tilde{\beta},\tilde{\alpha})y}\left(\frac{\tilde{\alpha}+\tilde{\beta}y}{\tilde{\beta}y}\right)^{-\tilde{\beta}}\left(\frac{\tilde{\alpha}+\tilde{\beta}y}{\tilde{\alpha}}\right)^{-\tilde{\alpha}} \\ &= \frac{1}{B(\tilde{\beta},\tilde{\alpha})y}\left(1 + \frac{\tilde{\alpha}}{\tilde{\beta}y}\right)^{-\tilde{\beta}}\left(1 + \frac{\tilde{\beta}y}{\tilde{\alpha}}\right)^{-\tilde{\alpha}}, \end{aligned}$$

which establishes $\gamma \sim \mathrm{F}(2\tilde{\beta}, 2\tilde{\alpha})$.

b) Show that as a function of γ, the transformation (6.12) reduces to

$$\theta = g(\gamma) = (1 + \gamma/c)^{-1}$$

where

$$c = \frac{\tilde{\alpha}\,\mathsf{Pr}(T+\,|\,D+)}{\tilde{\beta}\{1 - \mathsf{Pr}(T-\,|\,D-)\}}.$$

▶ *We first plug in the expression for ω given in Example 6.4 and then we express the term depending on π as a function of γ:*

$$\begin{aligned} \theta &= (1 + \omega^{-1})^{-1} \\ &= \left(1 + \frac{1 - \mathsf{Pr}(T-\,|\,D-)}{\mathsf{Pr}(T+\,|\,D+)} \cdot \frac{1-\pi}{\pi}\right)^{-1} \\ &= \left(1 + \frac{1 - \mathsf{Pr}(T-\,|\,D-)}{\mathsf{Pr}(T+\,|\,D+)} \cdot \frac{\tilde{\beta}\gamma}{\tilde{\alpha}}\right)^{-1} \\ &= \left(1 + \gamma\Big/\frac{\tilde{\alpha}\,\mathsf{Pr}(T+\,|\,D+)}{\tilde{\beta}\{1 - \mathsf{Pr}(T-\,|\,D-)\}}\right)^{-1} \\ &= (1 + \gamma/c)^{-1}. \end{aligned}$$

c) Show that

$$\frac{d}{d\gamma}g(\gamma) = -\frac{1}{c(1+\gamma/c)^2}$$

and that $g(\gamma)$ is a strictly monotonically decreasing function of γ.

▶ *Applying the chain rule to the function g gives*

$$\frac{d}{d\gamma}g(\gamma) = -(1+\gamma/c)^{-2} \cdot \frac{d}{d\gamma}(1+\gamma/c) = -\frac{1}{c(1+\gamma/c)^2}.$$

As $c > 0$, we have

$$\frac{d}{d\gamma}g(\gamma) < 0 \quad \textit{for all } \gamma \in [0,\infty),$$

which implies that $g(\gamma)$ is a strictly monotonically decreasing function of γ.

d) Use the change of variables formula (A.11) to derive the density of θ in (6.13).

▶ *We derive the density of $\theta = g(\gamma)$ from the density of γ obtained in 2a). Since g is strictly monotone by 2c) and hence one-to-one, we can apply the change-of-variables formula to this transformation. We have*

$$\begin{aligned} \theta &= (1 + \gamma/c)^{-1} \qquad \textit{by 2b) and} \\ \gamma &= g^{-1}(\theta) = \frac{c(1-\theta)}{\theta} = c(1/\theta - 1). \end{aligned}$$

Thus,

$$\begin{aligned} f_\theta(\theta) &= \left|\frac{dg(\gamma)}{d\gamma}\right|^{-1} \cdot f_\gamma(\gamma) \\ &= c \cdot (1+\gamma/c)^2 \cdot f_\gamma\left(c(1/\theta-1)\right) && (6.3) \\ &= c \cdot \theta^{-2} \cdot f_\mathrm{F}\left(c(1/\theta-1)\,;\,2\tilde{\beta},2\tilde{\alpha}\right), && (6.4) \end{aligned}$$

where we have used 2c) in (6.3) and 2a) in (6.4).

e) Analogously proceed with the negative predictive value $\tau = \mathsf{Pr}(D-\,|\,T-)$ to show that the density of τ is

$$f(\tau) = d\cdot\tau^{-2}\cdot f_\mathrm{F}\left(d(1/\tau-1); 2\tilde{\alpha}, 2\tilde{\beta}\right),$$

where

$$d = \frac{\tilde{\beta}\,\mathsf{Pr}(T-\,|\,D-)}{\tilde{\alpha}\{1-\mathsf{Pr}(T+\,|\,D+)\}}$$

and $f_\mathrm{F}(x\,;\,2\tilde{\alpha},2\tilde{\beta})$ is the density of the F distribution with parameters $2\tilde{\alpha}$ and $2\tilde{\beta}$.

▶ *In this case, the posterior odds can be expressed as*

$$\bar{\omega} := \frac{\mathsf{Pr}(D-\,|\,T-)}{\mathsf{Pr}(D+\,|\,T-)} = \frac{\mathsf{Pr}(T-\,|\,D-)}{1-\mathsf{Pr}(T+\,|\,D+)}\cdot\frac{1-\pi}{\pi},$$

so that

$$\tau = \mathsf{Pr}(D-\,|\,T-) = \frac{\bar{\omega}}{1+\bar{\omega}} = (1+\bar{\omega}^{-1})^{-1}.$$

Step a: We show that

$$\bar{\gamma} = \frac{\tilde{\beta}}{\tilde{\alpha}}\cdot\frac{\pi}{1-\pi} \sim \mathrm{F}(2\tilde{\alpha},2\tilde{\beta}).$$

We know that $\pi \sim \mathrm{Be}(\tilde{\alpha},\tilde{\beta})$ and we have $\bar{\gamma} = g(\pi)$ for

$$g(y) = \frac{\tilde{\beta}}{\tilde{\alpha}}\cdot\frac{y}{1-y}.$$

We are dealing with the same transformation function g as in step 2 of part (a), except that $\tilde{\alpha}$ and $\tilde{\beta}$ are interchanged. By analoguous arguments as in step 2 of part (a), we thus obtain $\bar{\gamma} \sim \mathrm{F}(2\tilde{\alpha},2\tilde{\beta})$.

Step b: Next, we express $\tau = \mathsf{Pr}(D-\,|\,T-)$ as a function of $\bar{\gamma}$.

$$\begin{aligned} \tau &= (1+\bar{\omega}^{-1})^{-1} \\ &= \left(1+\frac{1-\mathsf{Pr}(T+\,|\,D+)}{\mathsf{Pr}(T-\,|\,D-)}\cdot\frac{\pi}{1-\pi}\right)^{-1} \\ &= \left(1+\frac{1-\mathsf{Pr}(T+\,|\,D+)}{\mathsf{Pr}(T-\,|\,D-)}\cdot\frac{\tilde{\alpha}\bar{\gamma}}{\tilde{\beta}}\right)^{-1} \\ &= \left(1+\bar{\gamma}\Big/\frac{\tilde{\beta}\,\mathsf{Pr}(T+\,|\,D+)}{\tilde{\alpha}\{1-\mathsf{Pr}(T-\,|\,D-)\}}\right)^{-1} \\ &= (1+\bar{\gamma}/d)^{-1}. \end{aligned}$$

Step c: We show that the transformation $h(\bar{\gamma}) = (1+\bar{\gamma}/d)^{-1}$ is one-to-one by establishing strict monotonicity.
Applying the chain rule to the function h gives

$$\frac{d}{d\bar{\gamma}}h(\bar{\gamma}) = -\frac{1}{d(1+\bar{\gamma}/d)^2}.$$

As $d > 0$, we have

$$\frac{d}{d\bar{\gamma}}h(\bar{\gamma}) < 0 \quad \textit{for all } \bar{\gamma}\in[0,\infty),$$

which implies that $h(\bar{\gamma})$ is a strictly monotonically decreasing function of $\bar{\gamma}$.
Step d: We derive the density of $\tau = h(\bar{\gamma})$ from the density of $\bar{\gamma}$.
We have

$$\begin{aligned} \tau &= (1+\bar{\gamma}/d)^{-1} \qquad \textit{by Step b and} \\ \bar{\gamma} &= h^{-1}(\tau) = d(1/\tau-1). \end{aligned}$$

Thus,

$$\begin{aligned} f_\tau(\tau) &= \left|\frac{dg(\bar{\gamma})}{d\bar{\gamma}}\right|^{-1}\cdot f_{\bar{\gamma}}(\bar{\gamma}) \\ &= d\cdot(1+\bar{\gamma}/d)^2\cdot f_{\bar{\gamma}}\left(d(1/\tau-1)\right) && (6.5) \\ &= d\cdot\tau^{-2}\cdot f_\mathrm{F}\left(d(1/\tau-1)\,;\,2\tilde{\alpha},2\tilde{\beta}\right), && (6.6) \end{aligned}$$

where we have used Step c in (6.5) and Step a in (6.6).

3. Suppose the heights of male students are normally distributed with mean 180 and unknown variance σ^2. We believe that σ^2 is in the range $[22, 41]$ with approximately 95% probability. Thus we assign an inverse-gamma distribution $\mathrm{IG}(38, 1110)$ as prior distribution for σ^2.

a) Verify with R that the parameters of the inverse-gamma distribution lead to a prior probability of approximately 95% that $\sigma^2 \in [22, 41]$.

▶ *We use the fact that if $\sigma^2 \sim \mathrm{IG}(38, 1110)$, then $1/\sigma^2 \sim \mathrm{G}(38, 1110)$ (see Table A.2). We can thus work with the cumulative distribution function of the corresponding gamma distribution in R. We are interested in the probability*

$$\Pr\left(\frac{1}{\sigma^2} \in \left[\frac{1}{41}, \frac{1}{22}\right]\right) = \Pr\left(\frac{1}{\sigma^2} \leq \frac{1}{22}\right) - \Pr\left(\frac{1}{\sigma^2} < \frac{1}{41}\right).$$

```
> (prior.prob <- pgamma(1/22, shape=38, rate=1110)
                 - pgamma(1/41, shape=38, rate=1110))
[1] 0.9431584
```

b) Derive and plot the posterior density of σ^2 corresponding to the following data:

183, 173, 181, 170, 176, 180, 187, 176, 171, 190, 184, 173, 176, 179, 181, 186.

▶ *We assume that the observed heights $x_1 = 183, x_2 = 173, \ldots, x_{16} = 186$ are realisations of a random sample $X_{1:n}$ (in particular: $X_1, \ldots, X_n$ are independent) for $n = 16$. We know that*

$$X_i \mid \sigma^2 \sim \mathrm{N}(\mu = 180, \sigma^2), \quad i = 1, \ldots, n.$$

A priori we have $\sigma^2 \sim \mathrm{IG}(38, 1110)$ and we are interested in the posterior distribution of $\sigma^2 \mid x_{1:n}$. It can be easily verified, see also the last line in Table 6.2, that

$$\sigma^2 \mid x_i \sim \mathrm{IG}\left(38 + \frac{1}{2}, 1110 + \frac{1}{2}(x_i - \mu)^2\right).$$

As in Example 6.8, this result can be easily extended to a random sample $X_{1:n}$:

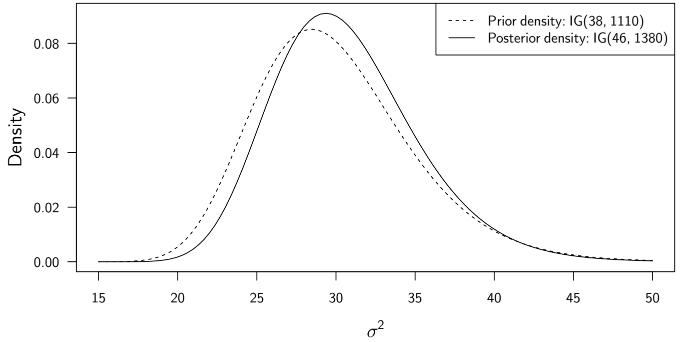
$$\sigma^2 \mid x_{1:n} \sim \mathrm{IG}\left(38 + \frac{n}{2}, 1110 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

```
>   # parameters of the inverse gamma prior
>   alpha <- 38
>   beta <- 1110
>   # prior mean of the normal distribution
>   mu <- 180
>   # data vector
>   heights <- c(183, 173, 181, 170, 176, 180, 187, 176,
                 171, 190, 184, 173, 176, 179, 181, 186)
>   # number of observations
>   n <- length(heights)
>   # compute the parameters of the inverse gamma posterior distribution
>   (alpha.post <- alpha + n/2)
[1] 46
```

```
>   (beta.post <- beta + 0.5*sum((heights - mu)^2))
[1] 1380
```

The posterior distribution of σ^2 is $\mathrm{IG}(46, 1380)$.

```
> library(MCMCpack)
> # plot the posterior distribution
> curve(dinvgamma(x, shape=alpha.post, scale=beta.post), from=15, to=50,
        xlab=expression(sigma^2), ylab="Density", col=1, lty=1)
> # plot the prior distribution
> curve(dinvgamma(x, shape=alpha, scale=beta), from=15, to=50,
        n=200, add=T, col=1, lty=2)
> legend("topright",
    c("Prior density: IG(38, 1110)",
       "Posterior density: IG(46, 1380)"), bg="white", lty=c(2,1), col=1)
```



c) Compute the posterior density of the standard deviation σ.

▶ *For notational convenience, let $Y = \sigma^2$. We know that $Y \sim \mathrm{IG}(\alpha = 46, \beta = 1380)$ and are interested in $Z = g(Y) = \sqrt{Y} = \sigma$, where $g(y) = \sqrt{y}$. Thus,*

$$g^{-1}(z) = z^2 \quad \textit{and}$$
$$\frac{dg^{-1}(z)}{dz} = 2z.$$

Using the change-of-variables formula, we obtain

$$\begin{aligned} f_Z(z) &= |2z| \frac{\beta^\alpha}{\Gamma(\alpha)} (z^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{z^2}\right) \\ &= \frac{2\beta^\alpha}{\Gamma(\alpha)} z^{-(2\alpha+1)} \exp\left(-\frac{\beta}{z^2}\right). \end{aligned}$$

This is the required posterior density of the standard deviation $Z = \sigma$ for $\alpha = 46$ and $\beta = 1380$.

4. Assume that n throat swabs have been tested for influenza. We denote by X the number of throat swabs which yield a positive result and assume that X is binomially distributed with parameters n and unknown probability π, so that $X \mid \pi \sim \mathrm{Bin}(n, \pi)$.

a) Determine the expected Fisher information and obtain Jeffreys' prior.

▶ *As* $X \mid \pi \sim \mathrm{Bin}(n, \pi)$*, the log-likelihood is given by*

$$l(\pi) = x \log \pi + (n - x) \log(1 - \pi),$$

so the score function is

$$S(\pi) = \frac{dl(\pi)}{d\pi} = \frac{x}{\pi} + \frac{n - x}{1 - \pi} \cdot (-1).$$

The Fisher information turns out to be

$$I(\pi) = -\frac{dS(\pi)}{d\pi} = \frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2}.$$

Thus, the expected Fisher information is

$$\begin{aligned} J(\pi) &= \mathsf{E}\{I(\pi; X)\} \\ &= \mathsf{E}\left\{\frac{X}{\pi^2}\right\} + \mathsf{E}\left\{\frac{n - X}{(1 - \pi)^2}\right\} \\ &= \frac{\mathsf{E}(X)}{\pi^2} + \frac{n - \mathsf{E}(X)}{(1 - \pi)^2} \\ &= \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1 - \pi)^2} \\ &= \frac{n}{\pi} + \frac{n}{1 - \pi} \\ &= \frac{n}{\pi(1 - \pi)}, \end{aligned}$$

as derived in Example 4.1. Jeffreys' prior therefore is

$$p(\pi) \propto \sqrt{J(\pi)} \propto \pi^{-1/2}(1 - \pi)^{-1/2},$$

which corresponds to the kernel of a $\mathrm{Be}(1/2, 1/2)$ *distribution.*

b) Reparametrise the binomial model using the log odds $\eta = \log\{\pi/(1-\pi)\}$, leading to

$$f(x \mid \eta) = \binom{n}{x} \exp(\eta x)\{1 + \exp(\eta)\}^{-n}.$$

Obtain Jeffreys' prior distribution directly for this likelihood and not with the change of variables formula.

▶ *We first deduce the reparametrisation given above:*

$$\begin{aligned} f(x \mid \pi) &= \binom{n}{x} \pi^x (1 - \pi)^{n - x} \\ &= \binom{n}{x} \left(\frac{\pi}{1 - \pi}\right)^x \left(\frac{1}{1 - \pi}\right)^{-n} \\ &= \binom{n}{x} \left(\frac{\pi}{1 - \pi}\right)^x \left(1 + \frac{\pi}{1 - \pi}\right)^{-n} \\ &= \binom{n}{x} \left(\exp\left(\log\left(\frac{\pi}{1 - \pi}\right) x\right)\right) \left(1 + \exp\left(\log\left(\frac{\pi}{1 - \pi}\right)\right)\right)^{-n}, \end{aligned}$$

so that for the log odds $\eta = \log\{\pi/(1 - \pi)\}$*, we have*

$$f(x \mid \eta) = \binom{n}{x} \exp(\eta x)\{1 + \exp(\eta)\}^{-n}.$$

The log-likelihood is therefore

$$l(\eta) = \eta x - n \log\{1 + \exp(\eta)\},$$

the score function is

$$S(\eta) = \frac{dl(\eta)}{d\eta} = x - \frac{n}{1 + \exp(\eta)} \cdot \exp(\eta),$$

and the Fisher information is

$$I(\eta) = -\frac{dS(\eta)}{d\eta} = \frac{\{1 + \exp(\eta)\} \cdot n \exp(\eta) - n\{\exp(\eta)\}^2}{\{1 + \exp(\eta)\}^2} = \frac{n \exp(\eta)}{\{1 + \exp(\eta)\}^2},$$

independent of x*. The expected Fisher information is therefore equal to the (observed) Fisher information,*

$$J(\eta) = \mathsf{E}\{I(\eta)\} = I(\eta),$$

so Jeffreys' prior is

$$f(\eta) \propto \sqrt{J(\eta)} \propto \sqrt{\frac{\exp(\eta)}{\{1 + \exp(\eta)\}^2}} = \frac{\exp(\eta)^{1/2}}{1 + \exp(\eta)}.$$

c) Take the prior distribution of 4a) and apply the change of variables formula to obtain the induced prior for η. Because of the invariance under reparameterization this prior density should be the same as in part 4b).

▶ *The transformation is*

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \eta$$

and we have

$$\begin{aligned} \frac{dg(\pi)}{d\pi} &= \frac{1 - \pi}{\pi} \cdot \frac{1}{(1 - \pi)^2} = \pi^{-1}(1 - \pi)^{-1}, \text{ and} \\ g^{-1}(\eta) &= \frac{\exp(\eta)}{1 + \exp(\eta)} = \pi. \end{aligned}$$

Applying the change-of-variables formula gives

$$\begin{aligned} f(\eta) &= \left|\frac{dg(\pi)}{d\pi}\right|^{-1} f_\pi(\pi) \\ &\propto \pi(1-\pi)\pi^{-1/2}(1-\pi)^{-1/2} \\ &= \pi^{1/2}(1-\pi)^{1/2} \\ &= \left(\frac{\exp(\eta)}{1+\exp(\eta)}\right)^{1/2}\left(\frac{1}{1+\exp(\eta)}\right)^{1/2} \\ &= \frac{\exp(\eta)^{1/2}}{1+\exp(\eta)}. \end{aligned}$$

This density is the same as the one we received in part 4b).

5. Suppose that the survival times $X_{1:n}$ form a random sample from an exponential distribution with parameter λ.

a) Derive Jeffreys' prior for λ und show that it is improper.

▶ *From Exercise 4a) in Chapter 4 we know the score function*

$$S(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Viewed as a random variable in $X_1, \dots, X_n$, its variance is

$$J(\lambda) = \operatorname{Var}\{S(\lambda)\} = n \cdot \operatorname{Var}(X_1) = \frac{n}{\lambda^2}.$$

Jeffreys' prior therefore is

$$f(\lambda) \propto \sqrt{J(\lambda)} \propto \lambda^{-1},$$

which cannot be normalised, since

$$\int_0^\infty \lambda^{-1}\, d\lambda = [\log(\lambda)]_0^\infty = \log(\infty) - \log(0) = \infty - (-\infty) = \infty,$$

so $f(\lambda)$ is improper.

b) Suppose that the survival times are only partially observed until the r-th death such that $n-r$ observations are actually censored. Write down the corresponding likelihood function and derive the posterior distribution under Jeffreys' prior.

▶ *Let $x_{(1)}, \dots, x_{(n)}$ denote the ordered survival times. Only $x_{(1)}, \dots, x_{(r)}$ are observed, the remaining survival times are censored. The corresponding likelihood function can be derived from Example 2.8 with $\delta_{(i)} = \mathsf{I}_{\{1,\dots,r\}}(i)$ and $\sum_{i=1}^n \delta_i = r$:*

$$L(\lambda) = \lambda^r \exp\left\{-\lambda\left(\sum_{i=1}^{r} x_{(i)} + (n-r)x_{(r)}\right)\right\},$$

because $x_{(r+1)}, \dots, x_{(n)}$ are assumed to be censored at time $x_{(r)}$.
The posterior distribution under Jeffreys' prior $f(\lambda) \propto \lambda^{-1}$ is thus

$$\begin{aligned} f(\lambda \mid x_{1:n}) &\propto f(\lambda) \cdot L(\lambda) \\ &= \lambda^{r-1} \exp\left\{-\lambda\left(\sum_{i=1}^{r} x_{(i)} + (n-r)x_{(r)}\right)\right\}. \end{aligned}$$

c) Show that the posterior is improper if all observations are censored.

▶ *If no death has occured prior to some time c, the likelihood of λ is*

$$L(\lambda) = \exp(-n\lambda c).$$

Using Jeffreys' prior $f(\lambda) \propto \lambda^{-1}$, we obtain the posterior

$$f(\lambda \mid x_{1:n}) \propto \frac{1}{\lambda}\exp(-nc\lambda).$$

This can be identified as the kernel of an improper $\mathrm{G}(0, nc)$ *distribution.*

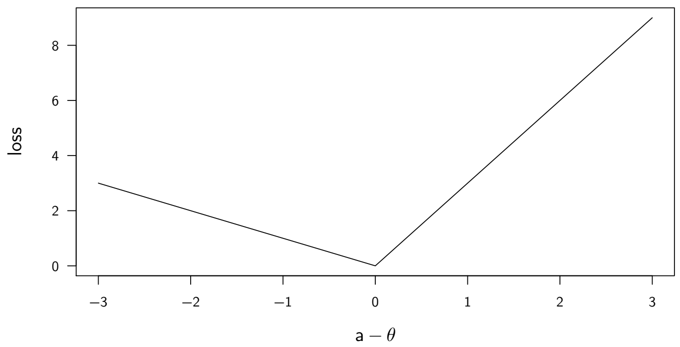
6. After observing a patient, his/her LDL cholesterol level θ is estimated by a. Due to the increased health risk of high cholesterol levels, the consequences of underestimating a patient's cholesterol level are considered more serious than those of overestimation. That is to say that $|a-\theta|$ should be penalised more when $a \leq \theta$ than when $a > \theta$. Consider the following loss function parameterised in terms of $c, d > 0$:

$$l(a, \theta) = \begin{cases} -c(a-\theta) & \text{if } a - \theta \leq 0 \\ d(a-\theta) & \text{if } a - \theta > 0 \end{cases}.$$

a) Plot $l(a, \theta)$ as a function of $a - \theta$ for $c = 1$ and $d = 3$.

▶

```
> # loss function with argument a-theta
> loss <- function(aMinusTheta, c, d)
  {
      ifelse(aMinusTheta <= 0, - c * aMinusTheta, d * aMinusTheta)
  }
> aMinusTheta <- seq(-3, 3, length = 101)
> plot(aMinusTheta, loss(aMinusTheta, c = 1, d = 3),
       type = "l", xlab = expression(a - theta), ylab = "loss")
```



b) Compute the Bayes estimate with respect to the loss function $l(a, \theta)$.

▶ *The expected posterior loss is*

$$\begin{aligned}\mathsf{E}\big(l(a,\theta)\,|\,x\big) &= \int l(a,\theta) f(\theta\,|\,x)\,d\theta \\ &= \int_{-\infty}^{a} d(a-\theta) f(\theta\,|\,x)\,d\theta + \int_{a}^{\infty} c(\theta - a) f(\theta\,|\,x)\,d\theta.\end{aligned}$$

To compute the Bayes estimate

$$\hat{a} = \arg\min_a \mathsf{E}\big(l(a,\theta)\,|\,x\big),$$

we take the derivative with respect to a, using Leibniz integral rule, see Appendix B.2.4. Using the convention $\infty \cdot 0 = 0$ *we obtain*

$$\begin{aligned}\frac{d}{da} E\big(l(a,\theta)\,|\,x\big) &= d \int_{-\infty}^{a} f(\theta\,|\,x)\,d\theta - c \int_{a}^{\infty} f(\theta\,|\,x)\,d\theta \\ &= dF(a\,|\,x) - c\big(1 - F(a\,|\,x)\big) \\ &= (c+d)F(a\,|\,x) - c.\end{aligned}$$

The root of this function in a is therefore

$$\hat{a} = F^{-1}\big(c/(c+d)\,|\,x\big),$$

i. e. the Bayes estimate $\hat{a}$ *is the* $c/(c+d) \cdot 100\%$ *quantile of the posterior distribution of* θ. *For* $c = d$ *we obtain as a special case the posterior median. For* $c = 1$ *and* $d = 3$ *the Bayes estimate is the 25%-quantile of the posterior distribution. Remark: If we choose* $c = 3$ *and* $d = 1$ *as mentioned in the Errata, then the Bayes estimate is the 75%-quantile of the posterior distribution.*

7. Our goal is to estimate the allele frequency at one bi-allelic marker, which has either allele A or B. DNA sequences for this location are provided for n individuals. We denote the observed number of allele A by X and the underlying (unknown) allele frequency with π. A formal model specification is then a binomial distribution $X\,|\,\pi \sim \text{Bin}(n, \pi)$ and we assume a beta prior distribution $\pi \sim \text{Be}(\alpha, \beta)$ where $\alpha, \beta > 0$.

a) Derive the posterior distribution of π and determine the posterior mean and mode.

▶ *We know*

$$X\,|\,\pi \sim \text{Bin}(n,\pi) \qquad \textit{and} \qquad \pi \sim \text{Be}(\alpha,\beta).$$

As in Example 6.3, we are interested in the posterior distribution $\pi\,|\,X$*:*

$$\begin{aligned}f(\pi\,|\,x) &\propto f(x\,|\,\pi) f(\pi) \\ &\propto \pi^x (1-\pi)^{n-x} \pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1},\end{aligned}$$

i.e. $\pi\,|\,x \sim \text{Be}(\alpha + x, \beta + n - x)$. *Hence, the posterior mean is given by* $(\alpha + x)/(\alpha + \beta + n)$ *and the posterior mode by* $(\alpha + x - 1)/(\alpha + \beta + n - 2)$.

b) For some genetic markers the assumption of a beta prior may be restrictive and a bimodal prior density, *e. g.*, might be more appropriate. For example, we can easily generate a bimodal shape by considering a mixture of two beta distributions:

$$f(\pi) = w f_{\text{Be}}(\pi; \alpha_1, \beta_1) + (1-w) f_{\text{Be}}(\pi; \alpha_2, \beta_2)$$

with mixing weight $w \in (0, 1)$.

i. Derive the posterior distribution of π.

▶

$$\begin{aligned}f(\pi\,|\,x) &\propto f(x\,|\,\pi) f(\pi) \\ &\propto \pi^x(1-\pi)^{n-x}\left\{\frac{w}{B(\alpha_1,\beta_1)}\pi^{\alpha_1-1}(1-\pi)^{\beta_1-1}\right. \\ &\quad \left. + \frac{1-w}{B(\alpha_2,\beta_2)}\pi^{\alpha_2-1}(1-\pi)^{\beta_2-1}\right\} \\ &= \frac{w}{B(\alpha_1,\beta_1)}\pi^{\alpha_1+x-1}(1-\pi)^{\beta_1+n-x-1} \\ &\quad + \frac{1-w}{B(\alpha_2,\beta_2)}\pi^{\alpha_2+x-1}(1-\pi)^{\beta_2+n-x-1}.\end{aligned}$$

ii. The posterior distribution is a mixture of two familiar distributions. Identify these distributions and the corresponding posterior weights.

▶ *We have*

$$
\begin{aligned}
f(\pi \mid x) \propto \; & \frac{w}{B(\alpha_1, \beta_1)} \frac{B(\alpha_1^\star, \beta_1^\star)}{B(\alpha_1^\star, \beta_1^\star)} \pi^{\alpha_1^\star - 1}(1-\pi)^{\beta_1^\star - 1} \\
& + \frac{1-w}{B(\alpha_2, \beta_2)} \frac{B(\alpha_2^\star, \beta_2^\star)}{B(\alpha_2^\star, \beta_2^\star)} \pi^{\alpha_2^\star - 1}(1-\pi)^{\beta_2^\star - 1}
\end{aligned}
$$

for

$$
\begin{aligned}
\alpha_1^\star &= \alpha_1 + x & \beta_1^\star &= \beta_1 + n - x, \\
\alpha_2^\star &= \alpha_2 + x & \beta_2^\star &= \beta_2 + n - x.
\end{aligned}
$$

Hence, the posterior distribution is a mixture of two beta distributions $\mathrm{Be}(\alpha_1^\star, \beta_1^\star)$ *and* $\mathrm{Be}(\alpha_2^\star, \beta_2^\star)$. *The mixture weights are proportional to*

$$
\gamma_1 = \frac{w \cdot B(\alpha_1^\star, \beta_1^\star)}{B(\alpha_1, \beta_1)} \qquad \text{and} \qquad \gamma_2 = \frac{(1-w) \cdot B(\alpha_2^\star, \beta_2^\star)}{B(\alpha_2, \beta_2)}.
$$

The normalized weights are $\gamma_1^\star = \gamma_1/(\gamma_1 + \gamma_2)$ *and* $\gamma_2^\star = \gamma_2/(\gamma_1 + \gamma_2)$.

iii. Determine the posterior mean of π.

▶ *The posterior distribution is a linear combination of two beta distributions:*

$$
f(\pi \mid x) = \gamma_1^\star \, \mathrm{Be}(\pi \mid \alpha_1^\star, \beta_1^\star) + (1 - \gamma_1^\star) \, \mathrm{Be}(\pi \mid \alpha_2^\star, \beta_2^\star),
$$

so the posterior mean is given by

$$
\begin{aligned}
\mathsf{E}(\pi \mid x) &= \gamma_1^\star \, \mathsf{E}(\pi \mid \alpha_1^\star, \beta_1^\star) + (1 - \gamma_1^\star) \, \mathsf{E}(\pi \mid \alpha_2^\star, \beta_2^\star) \\
&= \gamma_1^\star \frac{\alpha_1^\star}{\alpha_1^\star + \beta_1^\star} + (1 - \gamma_1^\star) \frac{\alpha_2^\star}{\alpha_2^\star + \beta_2^\star}.
\end{aligned}
$$

iv. Write an R-function that numerically computes the limits of an equi-tailed credible interval.

▶ *The posterior distribution function is*

$$
F(\pi \mid x) = \gamma_1^* F(\pi \mid \alpha_1^*, \beta_1^*) + (1 - \gamma_1^*) F(\pi \mid \alpha_2^*, \beta_2^*).
$$

The equi-tailed $(1-\alpha)$*-credible interval is therefore*

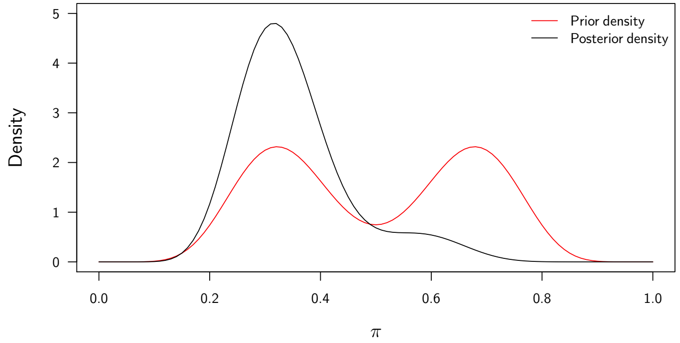
$$
[F^{-1}(\alpha/2 \mid x), F^{-1}(1 - \alpha/2 \mid x)],
$$

i. e. we are looking for arguments of π *where the distribution function has the values* $\alpha/2$ *and* $(1 - \alpha/2)$*, respectively.*

```
> # Distribution function of a mixture of beta distributions with
> # weight gamma1 of the first component and parameter vectors
> # alpha und beta
> pbetamix <- function(pi, gamma1, alpha, beta){
      gamma1 * pbeta(pi, alpha[1], beta[1]) +
          (1 - gamma1) * pbeta(pi, alpha[2], beta[2])
  }
> # corresponding quantile function
> qbetamix <- function(q, gamma1, alpha, beta){
      f <- function(pi){
          pbetamix(pi, gamma1, alpha, beta) - q
      }
      unirootResult <- uniroot(f, lower=0, upper=1)
      if(unirootResult$iter < 0)
          return(NA)
      else
          return(unirootResult$root)
  }
> # credibility interval with level level
> credBetamix <- function(level, gamma1, alpha, beta){
      halfa <- (1 - level)/2
      ret <- c(qbetamix(halfa, gamma1, alpha, beta),
               qbetamix(1 - halfa, gamma1, alpha, beta))
      return(ret)
  }
```

v. Let $n = 10$ and $x = 3$. Assume an even mixture ($w = 0.5$) of two beta distributions, $\mathrm{Be}(10, 20)$ and $\mathrm{Be}(20, 10)$. Plot the prior and posterior distributions in one figure.

```
> # data
> n <- 10
> x <- 3
> #
> # parameters for the beta components
> a1 <- 10
> b1 <- 20
> a2 <- 20
> b2 <- 10
> #
> # weight for the first mixture component
> w <- 0.5
> #
> # define a function that returns the density of a beta mixture
> # with two components
> mixbeta <- function(x, shape1a, shape2a, shape1b, shape2b, weight){
    y <- weight * dbeta(x, shape1=shape1a, shape2=shape2a) +
      (1-weight)* dbeta(x, shape1=shape1b, shape2=shape2b)
    return(y)
  }
> #
> # plot the prior density
> curve(mixbeta(x, shape1a=a1, shape2a=b1, shape1b=a2, shape2b=b2,weight=w),
    from=0, to=1, col=2, ylim=c(0,5), xlab=expression(pi), ylab="Density")
> #
> # parameters of the posterior distribution
> a1star <- a1 + x
> b1star <- b1 + n - x
> a2star <- a2 + x
> b2star <- b2 + n - x
> #
> # the posterior weights are proportional to
> gamma1 <- w*beta(a1star,b1star)/beta(a1,b1)
> gamma2 <- (1-w)*beta(a2star,b2star)/beta(a2,b2)
> #
> # calculate the posterior weight
> wstar <- gamma1/(gamma1 + gamma2)
> #
> # plot the posterior distribution
> curve(mixbeta(x, shape1a=a1star, shape2a=b1star,
                             shape1b=a2star, shape2b=b2star,weight=wstar),
    from=0, to=1, col=1, add=T)
> #
> legend("topright", c("Prior density", "Posterior density"),
         col=c(2,1), lty=1, bty="n")
```



8. The negative binomial distribution is used to represent the number of trials, x, needed to get r successes, with probability π of success in any one trial. Let X be negative binomial, $X \,|\, \pi \sim \mathrm{NBin}(r, \pi)$, so that

$$f(x \,|\, \pi) = \binom{x-1}{r-1} \pi^r (1-\pi)^{x-r},$$

with $0 < \pi < 1$, $r \in \mathbb{N}$ and support $\mathcal{T} = \{r, r+1, \dots\}$. As a prior distribution assume $\pi \sim \mathrm{Be}(\alpha, \beta)$,

$$f(\pi) = \mathrm{B}(\alpha, \beta)^{-1} \pi^{\alpha-1} (1-\pi)^{\beta-1},$$

with $\alpha, \beta > 0$.

a) Derive the posterior density $f(\pi \,|\, x)$. Which distribution is this and what are its parameters?

▶ *We have*

$$\begin{aligned} f(\pi \,|\, x) &\propto f(x \,|\, \pi) f(\pi) \\ &\propto \pi^r (1-\pi)^{x-r} \pi^{\alpha-1} (1-\pi)^{\beta-1} \\ &= \pi^{\alpha+r-1} (1-\pi)^{\beta+x-r-1}. \end{aligned}$$

This is the kernel of a beta distribution with parameters $\dot\alpha = \alpha + r$ and $\dot\beta = \beta + x - r$, that is $\pi \,|\, x \sim \mathrm{Be}(\alpha + r, \beta + x - r)$.

b) Define conjugacy and explain why, or why not, the beta prior is conjugate with respect to the negative binomial likelihood.

▶ *Definition of conjugacy (Def. 6.5): Let $L(\theta) = f(x \,|\, \theta)$ denote a likelihood function based on the observation $X = x$. A class $\mathcal{G}$ of distributions is called conjugate with respect to $L(\theta)$ if the posterior distribution $f(\theta \,|\, x)$ is in $\mathcal{G}$ for all x whenever the prior distribution $f(\theta)$ is in $\mathcal{G}$.*

The beta prior is conjugate with respect to the negative binomial likelihood since the resulting posterior distribution is also a beta distribution.

c) Show that the expected Fisher information is proportional to $\pi^{-2}(1-\pi)^{-1}$ and derive therefrom Jeffreys' prior and the resulting posterior distribution.

▶ *The log-likelihood is*

$$l(\pi) = r \log(\pi) + (x - r) \log(1 - \pi).$$

Hence,

$$\begin{aligned} S(\pi) &= \frac{dl(\pi)}{d\pi} = \frac{r}{\pi} - \frac{x-r}{1-\pi} \qquad \textit{and} \\ I(\pi) &= -\frac{d^2 l(\pi)}{d\pi^2} = \frac{r}{\pi^2} + \frac{x-r}{(1-\pi)^2}, \end{aligned}$$

which implies

$$\begin{aligned} J(\pi) &= \mathsf{E}(I(\pi; X)) \\ &= \frac{r}{\pi^2} + \frac{\mathsf{E}(X) - r}{(1-\pi)^2} \\ &= \frac{r}{\pi^2} + \frac{\frac{r}{\pi} - r}{(1-\pi)^2} \\ &= \frac{r(1-\pi)^2 + r\pi(1-\pi)}{\pi^2(1-\pi)^2} \\ &= \frac{r}{\pi^2(1-\pi)} \propto \pi^{-2}(1-\pi)^{-1}. \end{aligned}$$

Hence Jeffreys' prior is given by $\sqrt{J(\pi)} \propto \pi^{-1}(1-\pi)^{-1/2}$, which corresponds to a $\mathrm{Be}(0, 0.5)$ distribution and is improper.

By part (a), the posterior distribution is therefore $\pi \,|\, x \sim \mathrm{Be}(r, x - r + 0.5)$.

9. Let $X_{1:n}$ denote a random sample from a uniform distribution on the interval $[0, \theta]$ with unknown upper limit θ. Suppose we select a *Pareto distribution* $\mathrm{Par}(\alpha, \beta)$ with parameters $\alpha > 0$ and $\beta > 0$ as prior distribution for θ, *cf*. Table A.2 in Section A.5.2.

a) Show that $T(X_{1:n}) = \max\{X_1, \dots, X_n\}$ is sufficient for θ.

▶ *This was already shown in Exercise 6 of Chapter 2 (see the solution there).*

b) Derive the posterior distribution of θ and identify the distribution type.

▶ *The posterior distribution is also a Pareto distribution since for* $t = \max\{x_1, \dots, x_n\}$, *we have*

$$\begin{aligned} f(\theta \mid x_{1:n}) &\propto f(x_{1:n} \mid \theta) f(\theta) \\ &\propto \frac{1}{\theta^n} \mathsf{I}_{[0,\theta]}(t) \cdot \frac{1}{\theta^{\alpha+1}} \mathsf{I}_{[\beta,\infty)}(\theta) \\ &= \frac{1}{\theta^{(\alpha+n)+1}} \mathsf{I}_{[\max\{\beta,t\},\infty)}(\theta), \end{aligned}$$

that is $\theta \mid x_{1:n} \sim \mathrm{Par}(\alpha + n, \max\{\beta, t\})$.
Thus, the Pareto distribution is conjugate with respect to the uniform likelihood function.

c) Determine posterior mode $\mathrm{Mod}(\theta \mid x_{1:n})$, posterior mean $\mathsf{E}(\theta \mid x_{1:n})$, and the general form of the 95% HPD interval for θ.

▶ *The formulas for the mode and mean of the Pareto distribution are listed in Table A.2 in the Appendix. Here, we have*

$$\mathrm{Mod}(\theta \mid x_{1:n}) = \max\{\beta, t\} = \max\{\beta, x_1, \dots, x_n\}$$

and

$$\mathsf{E}(\theta \mid x_{1:n}) = \frac{(\alpha + n) \max\{\beta, t\}}{\alpha + n - 1},$$

where the condition $\alpha + n > 1$ *is satisfied for any* $n \geq 1$ *as* $\alpha > 0$.
Since the density $f(\theta \mid x_{1:n})$ *equals* 0 *for* $\theta < \max\{\beta, t\}$ *and is strictly monotonically decreasing for* $\theta \geq \max\{\beta, t\}$, *the 95% HPD interval for* θ *has the form*

$$[\max\{\beta, t\}, q],$$

where q *is the 95%-quantile of the* $\mathrm{Par}(\alpha + n, \max\{\beta, t\})$ *distribution.*

10. We continue Exercise 1 in Chapter 5, so we assume that the number of IHD cases is $D_i \mid \lambda_i \overset{\text{ind}}{\sim} \mathrm{Po}(\lambda_i Y_i), i = 1, 2$, where $\lambda_i > 0$ is the group-specific incidence rate. We use independent Jeffreys' priors for the rates λ_1 and λ_2.

a) Derive the posterior distribution of λ_1 and λ_2. Plot these in R for comparison.

▶ *We first derive Jeffreys' prior for* $\lambda_i, i = 1, 2$:

$$\begin{aligned} f(d_i \mid \lambda_i Y_i) &\propto (\lambda_i Y_i)^{d_i} \exp(-\lambda_i Y_i) \propto (\lambda_i)^{d_i} \exp(-\lambda_i Y_i), \\ l(\lambda_i) &\propto d_i \log(\lambda_i) - \lambda_i, \\ l'(\lambda_i) &\propto d_i / \lambda_i - 1, \\ I(\lambda_i) &= -l''(\lambda_i) \propto d_i \lambda_i^{-2}, \\ J(\lambda_i) &= \mathsf{E}(D_i) \lambda_i^{-2} \propto \lambda_i \lambda_i^{-2} = \lambda_i^{-1}. \end{aligned}$$

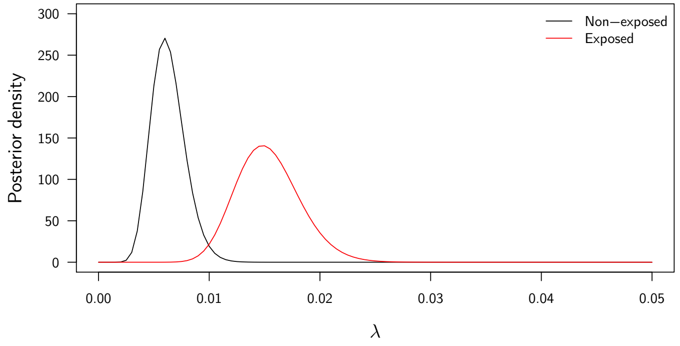
Thus, $f(\lambda_i) \propto \sqrt{J(\lambda_i)} \propto \lambda_i^{-1/2}$, *which corresponds to the improper* $\mathrm{G}(1/2, 0)$ *distribution (compare to Table 6.3 in the book). This implies*

$$\begin{aligned} f(\lambda_i \mid d_i) &\propto f(d_i \mid \lambda_i Y_i) f(\lambda_i) \\ &\propto (\lambda_i)^{d_i} \exp(-\lambda_i Y_i) \lambda_i^{-1/2} \\ &= (\lambda_i)^{d_i + 1/2 - 1} \exp(-\lambda_i Y_i), \end{aligned}$$

which is the density of the $\mathrm{G}(d_i + 1/2, Y_i)$ *distribution (compare to Table 6.2). Consequently,*

$$\begin{aligned} \lambda_1 \mid D_1 = 17 &\sim \mathrm{G}(17 + 1/2, Y_1) = \mathrm{G}(17.5, 2768.9), \\ \lambda_2 \mid D_2 = 28 &\sim \mathrm{G}(28 + 1/2, Y_2) = \mathrm{G}(28.5, 1857.5). \end{aligned}$$

```
> # the data is:
> # number of person years in the non-exposed and the exposed group
> y <- c(2768.9, 1857.5)
> # number of cases in the non-exposed and the exposed group
> d <- c(17, 28)
> #
> # plot the gamma densities for the two groups
> curve(dgamma(x, shape=d[1]+0.5, rate=y[1]),from=0,to=0.05,
        ylim=c(0,300), ylab="Posterior density", xlab=expression(lambda))
> curve(dgamma(x, shape=d[2]+0.5, rate=y[2]), col=2, from=0,to=0.05, add=T)
> legend("topright", c("Non-exposed", "Exposed"), col=c(1,2), lty=1, bty="n")
```



b) Derive the posterior distribution of the relative risk $\theta = \lambda_2/\lambda_1$ as follows:

i. Derive the posterior distributions of $\tau_1 = \lambda_1 Y_1$ and $\tau_2 = \lambda_2 Y_2$.

▶ *From Appendix A.5.2 we know that if* $X \sim \mathrm{G}(\alpha, \beta)$*, then* $c \cdot X \sim \mathrm{G}(\alpha, \beta/c)$*. Therefore, we have* $\tau_1 \sim \mathrm{G}(1/2 + d_1, 1)$ *and* $\tau_2 \sim \mathrm{G}(1/2 + d_2, 1)$*.*

ii. An appropriate multivariate transformation of $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ to work with is $\boldsymbol{g}(\boldsymbol{\tau}) = \boldsymbol{\eta} = (\eta_1, \eta_2)^\top$ with $\eta_1 = \tau_2/\tau_1$ and $\eta_2 = \tau_2 + \tau_1$ to obtain the joint density $f_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = f_{\boldsymbol{\tau}}\left(\boldsymbol{g}^{-1}(\boldsymbol{\eta})\right) \left|(\boldsymbol{g}^{-1})'(\boldsymbol{\eta})\right|$, *cf.* Appendix A.2.3.

▶ *We first determine the inverse transformation* $\boldsymbol{g}^{-1}(\boldsymbol{\eta})$ *by solving the two equations*

$$\eta_1 = \frac{\tau_2}{\tau_1} \qquad\qquad \eta_2 = \tau_2 + \tau_1;$$

for τ_1 *and* τ_2*, which gives*

$$\boldsymbol{g}^{-1}\left((\eta_1, \eta_2)^\top\right) = \left(\frac{\eta_2}{1+\eta_1}, \frac{\eta_1\eta_2}{1+\eta_1}\right)^\top.$$

The Jacobian matrix of $\boldsymbol{g}^{-1}$ *is*

$$\mathbf{J} := \left(\boldsymbol{g}^{-1}\right)'\left((\eta_1, \eta_2)^\top\right) = \begin{pmatrix} \frac{-\eta_2}{(1+\eta_1)^2} & \frac{1}{1+\eta_1} \\ \frac{\eta_2}{(1+\eta_1)^2} & \frac{\eta_1}{1+\eta_1} \end{pmatrix},$$

and its determinant is

$$\det(\mathbf{J}) = -\frac{\eta_2}{(\eta_1+1)^2}.$$

For ease of notation, let $\alpha_1 = 1/2 + d_1$ *and* $\alpha_2 = 1/2 + d_2$ *and note that* $f_{\boldsymbol{\tau}}$ *is a product of two gamma densities. Thus,*

$$\begin{aligned} f(\boldsymbol{\eta}) &= f_{\boldsymbol{\tau}}\left\{\boldsymbol{g}^{-1}\left((\eta_1, \eta_2)^\top\right)\right\} \frac{\eta_2}{(1+\eta_1)^2} \\ &= \frac{1}{\Gamma(\alpha_1)} \exp\left(-\frac{\eta_2}{1+\eta_1}\right)\left(\frac{\eta_2}{1+\eta_1}\right)^{\alpha_1 - 1} \\ &\cdot \frac{1}{\Gamma(\alpha_2)} \exp\left(-\frac{\eta_1\eta_2}{1+\eta_1}\right)\left(\frac{\eta_1\eta_2}{1+\eta_1}\right)^{\alpha_2 - 1} \frac{\eta_2}{(1+\eta_1)^2} \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \exp(-\eta_2)\eta_1^{\alpha_2 - 1}\eta_2^{\alpha_1+\alpha_2-1}(1+\eta_1)^{-\alpha_1-\alpha_2}. \end{aligned}$$

iii. Since $\eta_1 = \tau_2/\tau_1$ is the parameter of interest, integrate η_2 out of $f_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ and show that the marginal density is

$$f(\eta_1) = \frac{\eta_1^{\alpha_2 - 1}(1+\eta_1)^{-\alpha_1-\alpha_2}}{\mathrm{B}(\alpha_1, \alpha_2)},$$

which is a beta prime distribution with parameters α_2 and α_1.

▶

$$\begin{aligned} f(\eta_1) &= \int_0^\infty f(\eta_1, \eta_2) d\eta_2 \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \eta_1^{\alpha_2-1}(1+\eta_1)^{-\alpha_1-\alpha_2} \int_0^\infty \exp(-\eta_2)\eta_2^{\alpha_1+\alpha_2-1} d\eta_2 \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \eta_1^{\alpha_2-1}(1+\eta_1)^{-\alpha_1-\alpha_2}\Gamma(\alpha_1+\alpha_2) \\ &= \frac{\eta_1^{\alpha_2-1}(1+\eta_1)^{-\alpha_1-\alpha_2}}{\mathrm{B}(\alpha_1, \alpha_2)}, \end{aligned}$$

which is a beta prime distribution $\beta'(\alpha_2, \alpha_1)$ *with parameters* α_2 *and* α_1*. The beta prime distribution is a scaled F distribution. More preciseliy, if* $X \sim \beta'(\alpha_2, \alpha_1)$*, then* $\alpha_1/\alpha_2 \cdot X \sim \mathrm{F}(2\alpha_2, 2\alpha_1)$*. We write this as* $X \sim \alpha_2/\alpha_1 \times F(2\alpha_2, 2\alpha_1)$*.*

iv. From this distribution of τ_2/τ_1, the posterior distribution of λ_2/λ_1 is then easily found.

▶ *We know that*

$$\frac{\tau_2}{\tau_1} = \frac{\lambda_2 Y_2}{\lambda_1 Y_1} \sim \beta'(\alpha_2, \alpha_1),$$

so

$$\frac{\lambda_2}{\lambda_1} \sim \frac{Y_1}{Y_2} \times \beta'(\alpha_2, \alpha_1) = \frac{2768.9}{1857.5} \times \beta'(28.5, 17.5).$$

The relative risk $\theta = \lambda_2/\lambda_1$ *follows a* $2768.9/1857.5 \times \beta'(28.5, 17.5)$ *distribution, which corresponds to a*

$$\frac{Y_1(1/2+d_2)}{Y_2(1/2+d_1)} \times \mathrm{F}(1+2d_2, 1+2d_1) = \frac{78913.65}{32506.25} \times \mathrm{F}(57, 35) \qquad (6.7)$$

distribution.

c) For the given data, compute a 95% credible interval for θ and compare the results with those from Exercise 1 in Chapter 5.

▶ *We determine an equi-tailed credible interval. Since quantiles are invariant with respect to monotone one-to-one transformations, we can first determine the quantiles of the F distribution in (6.7) and then transform them accordingly:*

```
> # compute the 2.5 %- and the 97.5 %-quantile of the beta prime distribution
> # obtained in part (b)
> c <- y[1]*(0.5+ d[2])/(y[2]* (0.5+ d[1]))
> lower.limit <- c*qf(0.025, df1=2*(0.5+d[2]), df2=2*(0.5+d[1]) )
> upper.limit <- c*qf(0.975, df1=2*(0.5+d[2]), df2=2*(0.5+d[1]) )
> cred.int <- c(lower.limit, upper.limit)
> cred.int
```

```
[1] 1.357199 4.537326
```

An equi-tailed credible interval for θ is thus $[1.357, 4.537]$.

In Exercise 1 in Chapter 5, we have obtained the confidence interval $[0.296, 1.501]$ for $\log(\theta)$. Transforming the limits of this interval with the exponential function gives the confidence interval $[1.344, 4.485]$ for θ, which is quite similar to the credible interval obtained above. The credible interval is slightly wider and shifted towards slightly larger values than the confidence interval.

11. Consider Exercise 10 in Chapter 3. Our goal is now to perform Bayesian inference with an improper discrete uniform prior for the unknown number N of beds:

$$f(N) \propto 1 \quad \text{for} \quad N = 2, 3, \dots$$

a) Why is the posterior mode equal to the MLE?

▶ *This is due to Result 6.1: The posterior mode $\operatorname{Mod}(N \mid x_n)$ maximizes the posterior distribution, which is proportional to the likelihood function under a uniform prior:*

$$f(N \mid x_n) \propto f(x_n \mid N) f(N) \propto f(x_n \mid N).$$

Hence, the posterior mode must equal the value that maximizes the likelihood function, which is the MLE. In Exercise 10 in Chapter 3, we have obtained $\hat{N}_{\mathrm{ML}} = X_n$.

b) Show that for $n > 1$ the posterior probability mass function is

$$f(N \mid x_n) = \frac{n-1}{x_n}\binom{x_n}{n}\binom{N}{n}^{-1}, \quad \text{for } N \geq x_n.$$

▶ *We have*

$$f(N \mid x_n) = \frac{f(x_n \mid N) f(N)}{f(x_n)} \propto \frac{f(x_n \mid N)}{f(x_n)}.$$

From Exercise 10 in Chapter 3, we know that

$$f(x_n \mid N) = \binom{x_n - 1}{n-1}\binom{N}{n}^{-1} \quad \textit{for } N \geq x_n.$$

Next, we derive the marginal likelihood $f(x_n)$:

$$\begin{aligned} f(x_n) &= \sum_{N=1}^{\infty} f(x_n \mid N) \\ &= \sum_{N=x_n}^{\infty} \binom{x_n - 1}{n-1}\binom{N}{n}^{-1} \\ &= \binom{x_n - 1}{n-1} n! \sum_{N=x_n}^{\infty} \frac{(N-n)!}{N!}. \end{aligned}$$

To obtain to expression for $f(N \mid x_n)$ given in the exercise, we thus have to show that

$$\sum_{N=x_n}^{\infty} \frac{(N-n)!}{N!} = \left\{\frac{n-1}{x_n}\binom{x_n}{n} n!\right\}^{-1} = \frac{(x_n - n)!}{(n-1)(x_n-1)!}. \tag{6.8}$$

To this end, note that

$$\begin{aligned} \sum_{N=x_n}^{\infty} \frac{(N-n)!}{N!} &= \lim_{k \to \infty} \sum_{N=x_n}^{k} \frac{(N-n)!}{N!} \quad \textit{and} \\ \sum_{N=x_n}^{k} \frac{(N-n)!}{N!} &= \frac{(x_n-n)!}{(n-1)(x_n-1)!} - \frac{(k-(n-1))!}{k!(n-1)}, \end{aligned} \tag{6.9}$$

where (6.9) can be shown easily by induction on $k \geq x_n$ (and can be deduced by using the software `Maxima`*, for example). Now, the second term in on the right-hand side of (6.9) converges to 0 as $k \to \infty$ since*

$$\frac{(k-(n-1))!}{k!(n-1)} = \frac{1}{k(k-1)\cdots(k-(n-1)+1)} \cdot \frac{1}{n-1},$$

which impies (6.8) and completes the proof.

c) Show that the posterior expectation is

$$\mathsf{E}(N \mid x_n) = \frac{n-1}{n-2} \cdot (x_n - 1) \quad \text{for } n > 2.$$

▶ *We have*

$$\begin{aligned} \mathsf{E}(N \mid x_n) &= \sum_{N=0}^{\infty} N f(N \mid x_n) \\ &= \frac{n-1}{x_n}\binom{x_n}{n} n! \sum_{N=x_n}^{\infty} \frac{(N-n)!}{(N-1)!} \end{aligned} \tag{6.10}$$

and to determine the limit of the involved series, we can use (6.8) again:

$$\sum_{N=x_n}^{\infty} \frac{(N-n)!}{(N-1)!} = \sum_{N=x_n-1}^{\infty} \frac{(N-(n-1))!}{N!} = \frac{(x_n-n)!}{(n-2)(x_n-2)!}.$$

Plugging this result into expression (6.10) yields the claim.

d) Compare the frequentist estimates from Exercise 10 in Chapter 3 with the posterior mode and mean for $n = 48$ and $x_n = 1812$. Numerically compute the associated 95% HPD interval for N.

▶ *The unbiased estimator from Exercise 10 is*

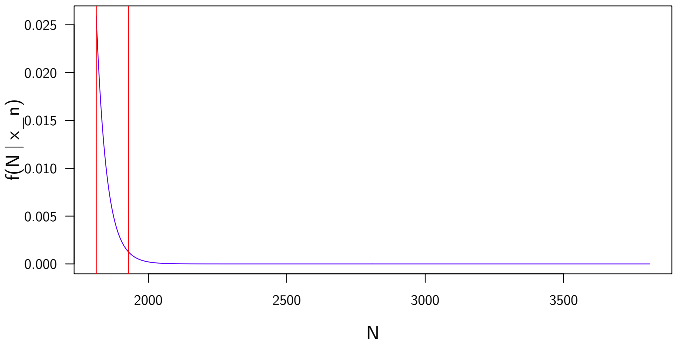
$$\hat{N} = \frac{n+1}{n} x_n - 1 = 1848.75,$$

which is considerably larger than the MLE and posterior mode $x_n = 1812$. The posterior mean

$$\mathsf{E}(N \mid x_n) = \frac{n-1}{n-2} \cdot (x_n - 1) = 1850.37$$

is even larger than $\hat{N}$. We compute the 95% HPD interval for N in R*:*

```
> # the data is
> n <- 48
> x_n <- 1812
> #
> # compute the posterior distribution for a large enough interval of N values
> N <- seq(from = x_n, length = 2000)
> posterior <- exp(log(n - 1) - log(x_n) + lchoose(x_n, n) - lchoose(N, n))
> plot(N, posterior, type = "l", col=4,
       ylab = "f(N | x_n)")
> # we see that this interval is large enough
> #
> # the posterior density is monotonically decreasing for values of N >= x_n
> # hence, the mode x_n is the lower limit of of the HPD interval
> level <- 0.95
> hpdLower <- x_n
> #
> # we next determine the upper limit
> # since the posterior density is monotonically decreasing
> # the upper limit is the smallest value of N for
> # which the cumulative posterior distribution function is larger or equal to 0.95
> cumulatedPosterior <- cumsum(posterior)
> hpdUpper <- min(N[cumulatedPosterior >= level])
> #
> # add the HPD interval to the figure
> abline(v = c(hpdLower, hpdUpper), col=2)
```



Thus, the HPD interval is $[1812, 1929]$.

12. Assume that $X_1, \ldots, X_n$ are independent samples from the binomial models $\mathrm{Bin}(m, \pi_i)$ and assume that $\pi_i \overset{\mathrm{iid}}{\sim} \mathrm{Be}(\alpha, \beta)$. Compute empirical Bayes estimates $\hat{\pi}_i$ of π_i as follows:

a) Show that the marginal distribution of X_i is beta-binomial, see Appendix A.5.1 for details. The first two moments of this distribution are

$$\mu_1 = \mathsf{E}(X_i) = m\frac{\alpha}{\alpha+\beta}, \tag{6.11}$$

$$\mu_2 = \mathsf{E}(X_i^2) = m\frac{\alpha\{m(1+\alpha)+\beta\}}{(\alpha+\beta)(1+\alpha+\beta)}. \tag{6.12}$$

Solve for α and β using the sample moments $\widehat{\mu}_1 = n^{-1}\sum_{i=1}^n x_i$, $\widehat{\mu}_2 = n^{-1}\sum_{i=1}^n x_i^2$ to obtain estimates of α and β.

▶ *The marginal likelihood of X_i can be found by integrating π_i out in the joint distribution of X_i and π_i:*

$$\begin{aligned} f(x_i) &= \int_0^1 \binom{m}{x_i} \pi_i^{x_i}(1-\pi_i)^{m-x_i} \cdot \frac{1}{B(\alpha,\beta)} \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1} d\pi_i \\ &= \binom{m}{x_i} \frac{B(\alpha+x_i, \beta+m-x_i)}{B(\alpha,\beta)} \\ &\quad \cdot \int_0^1 \frac{1}{B(\alpha+x_i,\beta+m-x_i)} \pi_i^{\alpha+x_i-1}(1-\pi_i)^{\beta+m-x_i-1} d\pi_i \\ &= \binom{m}{x_i} \frac{B(\alpha+x_i, \beta+m-x_i)}{B(\alpha,\beta)}, \end{aligned}$$

which is known as a beta-binomial distribution.

To derive estimates for α and β, we first solve the two given equations for α and β and then we replace μ_1 and μ_2 by the corresponding sample moments $\widehat{\mu_1}$ and $\widehat{\mu_2}$.

First, we combine the two equations by replacing part of the expression on the right-hand side of Equation (6.12) *by μ_1, which gives*

$$\mu_2 = \mu_1 \cdot \frac{m(1+\alpha)+\beta}{1+\alpha+\beta}.$$

We solve the above equation for β to obtain

$$\beta = \frac{(1+\alpha)(m\mu_1 - \mu_2)}{\mu_2 - \mu_1}. \tag{6.13}$$

Solving Equation (6.11) *for α yields*

$$\alpha = \beta \cdot \frac{\mu_1}{m - \mu_1}.$$

Next, we plug this expression for α into Equation (6.13) *and solve the resulting equation*

$$\beta = \frac{1}{\mu_2 - \mu_1}\left(\beta\left(\frac{m\mu_1^2 - \mu_1\mu_2}{m - \mu_1}\right) + m\mu_1 - \mu_2\right)$$

for β to obtain

$$\beta = \frac{(m\mu_1 - \mu_2)(m - \mu_1)}{m(\mu_2 - \mu_1(mu_1 + 1)) + \mu_1^2}$$

and consequently

$$\alpha = \beta \cdot \frac{\mu_1}{m - \mu_1} = \frac{(m\mu_1 - \mu_2)\mu_1}{m(\mu_2 - \mu_1(mu_1 + 1)) + \mu_1^2}.$$

The estimators for α and β are therefore

$$\begin{aligned}\widehat{\alpha} &= \frac{(m\widehat{\mu_1} - \widehat{\mu_2})\,\widehat{\mu_1}}{m\,(\widehat{\mu_2} - \widehat{\mu_1}\,(\widehat{\mu_1} + 1)) + \widehat{\mu_1}^2} \quad \textit{and}\\ \widehat{\beta} &= \frac{(m\widehat{\mu_1} - \widehat{\mu_2})\,(m - \widehat{\mu_1})}{m\,(\widehat{\mu_2} - \widehat{\mu_1}\,(\widehat{\mu_1} + 1)) + \widehat{\mu_1}^2}.\end{aligned}$$

b) Now derive the empirical Bayes estimates $\hat{\pi}_i$. Compare them with the corresponding MLEs.

▶ *By Example 6.3, the posterior $\pi_i|x_i$ has a Beta$(\widehat{\alpha}+x_i, \widehat{\beta}+m-x_i)$ distribution and the empirical Bayes estimate is thus the posterior mean*

$$\widehat{\pi}_i = \mathsf{E}[\pi_i|x_i] = \frac{\widehat{\alpha} + x_i}{\widehat{\alpha} + \widehat{\beta} + m}.$$

For comparison, the maximum likelihood estimate is $\hat{\pi}_{i_{\text{ML}}} = x_i/m$. Hence, the Bayes estimate is equal to the MLE if and only if $\widehat{\alpha} = \widehat{\beta} = 0$, which corresponds to an improper prior distribution. In general, the Bayes estimate

$$\frac{\widehat{\alpha} + x_i}{\widehat{\alpha} + \widehat{\beta} + m} = \frac{\widehat{\alpha} + \widehat{\beta}}{\widehat{\alpha} + \widehat{\beta} + m} \cdot \frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{\beta}} + \frac{m}{\widehat{\alpha} + \widehat{\beta} + m} \cdot \frac{x_i}{m}$$

is a weighted average of the prior mean $\widehat{\alpha}/(\widehat{\alpha} + \widehat{\beta})$ and the MLE x_i/m. The weights are proportional to the prior sample size $m_0 = \widehat{\alpha} + \widehat{\beta}$ and the data sample size m, respectively.

7 Model selection

1. Derive Equation (7.18).

▶ *Since the normal prior is conjugate to the normal likelihood with known variance (compare to Table 7.2), we can avoid integration and use Equation (7.16) instead to compute the marginal distribution:*

$$f(x \,|\, M_1) = \frac{f(x \,|\, \mu) f(\mu)}{f(\mu \,|\, x)}.$$

We know that

$$x \,|\, \mu \sim \mathrm{N}(\mu, \kappa^{-1}), \qquad \mu \sim \mathrm{N}(\nu, \delta^{-1})$$

and in Example 6.8, we have derived the posterior distribution of μ:

$$\mu \,|\, x \sim \mathrm{N}\left(\frac{n\kappa\bar{x} + \delta\nu}{n\kappa + \delta}, (n\kappa + \delta)^{-1}\right).$$

Consequently, we obtain

$$\begin{aligned}f(x \,|\, M_1) &= \frac{(2\pi\kappa^{-1})^{-n/2} \exp\left(-\kappa/2 \sum_{i=1}^n (x_i - \mu)^2\right) \cdot (2\pi\delta^{-1})^{-1/2} \exp\left(-\delta/2(\mu - \nu)^2\right)}{(2\pi(n\kappa + \delta)^{-1})^{-1/2} \exp\left(-(n\kappa + \delta)/2(\mu - \frac{n\kappa\bar{x}+\delta\nu}{n\kappa+\delta})^2\right)}\\
&= \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\delta}{n\kappa + \delta}\right)^{\frac{1}{2}}\\
&\cdot \exp\left[-\frac{1}{2}\left(\kappa\left(\sum_{i=1}^n x_i^2 - 2n\bar{x}\mu + m\mu^2\right) + \delta(\mu^2 - 2\mu\nu + \nu^2)\right.\right.\\
&\left.\left. - \left((n\kappa + \delta)\mu^2 - 2\mu(n\kappa\bar{x} + \delta\nu) + \frac{(n\kappa\bar{x} + \delta\nu)^2}{n\kappa + \delta}\right)\right)\right]\\
&= \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\delta}{n\kappa + \delta}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{\kappa}{2}\left(\sum_{i=1}^n x_i^2 + \frac{\delta\nu^2}{\kappa} - \frac{(n\kappa\bar{x} + \delta\nu)^2}{\kappa(n\kappa + \delta)}\right)\right]\\
&= \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\delta}{n\kappa + \delta}\right)^{\frac{1}{2}} \exp\left[-\frac{\kappa}{2}\left\{\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\delta}{n\kappa + \delta}(\bar{x} - \nu)^2\right\}\right].\end{aligned}$$

2. Let $Y_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \sigma^2)$, $i = 1, \dots, n$, be the response variables in a *normal regression model*, where the variance σ^2 is assumed known and the conditional means are $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. The design vectors $\boldsymbol{x}_i$ and the coefficient vector $\boldsymbol{\beta}$ have dimension p and are defined as for the logistic regression model (Exercise 17 in Chapter 5).

a) Derive AIC for this normal regression model.

▶ *Due to independence of $Y_i, i = 1, \dots, n$, the likelihood function is*

$$\begin{aligned}
L(y_{1:n}; \boldsymbol{\beta}) &= \prod_{i=1}^{n} L(y_i; \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2\right]
\end{aligned}$$

and thus the log-likelihood function is

$$l(y_{1:n}; \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2.$$

The coefficient vector $\boldsymbol{\beta}$ has dimension p, i. e. we have p parameters. Let $\hat{\boldsymbol{\beta}}_{\text{ML}}$ denote the ML estimate so that $\hat{\mu}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\text{ML}}$. Consequently, up to some constant, we obtain

$$\begin{aligned}
\text{AIC} &= -2\, l\left(y_{1:n}; \hat{\boldsymbol{\beta}}_{\text{ML}}\right) + 2\,p \\
&= n + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\text{ML}}\right)^2 + 2\,p \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + 2\,p + n.
\end{aligned}$$

Remark: As the ML estimate is the least squares estimate in the normal regression model, we have

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = (X^\top X)^{-1} X^\top y$$

for the $n \times p$ matrix $X = (x_1, \dots, x_n)$ and the n-dimensional vector $y = (y_1, \dots, y_n)^\top$.

b) *Mallow's C_p statistic*

$$C_p = \frac{\text{SS}}{s^2} + 2p - n$$

is often used to assess the fit of a regression model. Here $\text{SS} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ is the *residual sum of squares* and $\hat{\sigma}^2_{\text{ML}}$ is the MLE of the variance σ^2. How does AIC relate to C_p?

▶ *Up to some multiplicative constant, we have*

$$\text{AIC} = C_p + 2n,$$

which implies that rankings of different normal regression models with respect to AIC and C_p, respectively, will be the same.

c) Now assume that σ^2 is unknown as well. Show that AIC is given by

$$\text{AIC} = n\log(\hat{\sigma}^2_{\text{ML}}) + 2p + n + 2.$$

▶ *The log-likelihood function is the same as in part (a). Solving the score equation*

$$\frac{\partial\, l\left(y_{1:n}; (\boldsymbol{\beta}, \sigma^2)^T\right)}{\partial \sigma^2} = 0$$

gives

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\text{ML}})^2.$$

Plugging this estimate into the log-likelihood function yields

$$l\left(y_{1:n}; (\hat{\boldsymbol{\beta}}_{\text{ML}}, \hat{\sigma}^2_{\text{ML}})^T\right) = -n\log(\hat{\sigma}^2_{\text{ML}}) - \frac{n}{2}$$

up to some additive constant. In this model with unknown σ^2, there are $p + 1$ parameters. Consequently,

$$\begin{aligned}
\text{AIC} &= -2\, l\left(y_{1:n}; (\hat{\boldsymbol{\beta}}_{\text{ML}}, \hat{\sigma}^2_{\text{ML}})^T\right) + 2(p+1) \\
&= n\log(\hat{\sigma}^2_{\text{ML}}) + 2p + n + 2.
\end{aligned}$$

3. Repeat the analysis of Example 7.7 with unknown variance κ^{-1} using the conjugate normal-gamma distribution (see Example 6.21) as prior distribution for κ and μ.

a) First calculate the marginal likelihood of the model by using the rearrangement of Bayes' theorem in (7.16).

▶ *We know that for $\boldsymbol{\theta} = (\mu, \kappa)^T$, we have*

$$X \mid \boldsymbol{\theta} \sim \mathrm{N}(\mu, \kappa^{-1}) \quad \textit{and} \quad \boldsymbol{\theta} \sim \mathrm{NG}(\nu, \lambda, \alpha, \beta)$$

and in Example 6.21, we have seen that the posterior distribution is also a normal-gamma distribution:

$$\boldsymbol{\theta} \mid x_{1:n} \sim \mathrm{NG}(\nu^*, \lambda^*, \alpha^*, \beta^*),$$

where

$$\begin{aligned}
\nu^* &= (\lambda + n)^{-1}(\lambda\nu + n\bar{x}), \\
\lambda^* &= \lambda + n, \\
\alpha^* &= \alpha + n/2, \\
\beta^* &= \beta + \frac{n\hat{\sigma^2}_{\text{ML}} + (\lambda + n)^{-1} n\lambda(\nu - \bar{x})^2}{2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
f(x\,|\,\boldsymbol{\theta}) &= (2\pi/\kappa)^{-n/2}\exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right),\\
f(\boldsymbol{\theta}) &= \frac{\beta^{\alpha}}{\Gamma(\alpha)}(2\pi(\lambda\kappa)^{-1})^{-1/2}\kappa^{\alpha-1}\exp(-\beta\kappa)\cdot\exp\left\{-\frac{\lambda\kappa}{2}(\mu-\nu)^2\right\},\\
f(\boldsymbol{\theta}\,|\,x) &= \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)}(2\pi(\lambda^*\kappa)^{-1})^{-1/2}\kappa^{\alpha^*-1}\exp(-\beta^*\kappa)\cdot\exp\left\{-\frac{\lambda^*\kappa}{2}(\mu-\nu^*)^2\right\}.
\end{aligned}$$

Note that it is important to include the normalizing constants of the above densities in the following calculation to get

$$\begin{aligned}
f(x\,|\,M_1) =& \frac{f(x\,|\,\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{\theta}\,|\,x)}\\
=& \frac{(2\pi)^{-\frac{n}{2}}\frac{\beta^{\alpha}}{\Gamma(\alpha)}(2\pi)^{-\frac{1}{2}}\lambda^{1/2}}{\frac{(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)}(2\pi)^{-\frac{1}{2}}(\lambda^*)^{1/2}}\\
=& \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{\lambda}{\lambda+n}\right)^{1/2}\frac{\Gamma(\alpha+n/2)\beta^{\alpha}}{\Gamma(\alpha)}\cdot\\
& \left\{\beta+\frac{n\hat{\sigma}^2_{\text{ML}}+(\lambda+n)^{-1}n\lambda(\nu-\bar{x})^2}{2}\right\}^{-(\alpha+\frac{n}{2})}. \qquad (7.1)
\end{aligned}$$

b) Next, calculate explicitly the posterior probabilities of the four (*a priori* equally probable) models M_1 to M_3 using a $\text{NG}(2000, 5, 1, 50\,000)$ distribution as prior for κ and μ.

▶ *We work with model M_1 and M_2 only, as there is no model M_3 specified in Example 7.7.*

We first implement the marginal (log-)likelihood derived in part (a):

```
> ## marginal (log-)likelihood in the normal-normal-gamma model,
> ## for n realisations with mean "mean" and MLE estimate "var"
> ## for the variance
> marginalLikelihood <- function(n, mean, var,            # data
                                 nu, lambda, alpha, beta, # priori parameter
                                 log = FALSE      # should log(f(x))
                                                  # or f(x) be returned?
                                 )
  {
      betaStar <- beta +
          (n * var + n * lambda * (nu - mean)^2 / (lambda + n)) / 2

      logRet <- - n/2 * log(2 * pi) + 1/2 * log(lambda / (lambda + n)) +
          lgamma(alpha + n/2) - lgamma(alpha) + alpha * log(beta) -
              (alpha + n/2) * log(betaStar)
      if(log)
          return(logRet)
      else
          return(exp(logRet))
  }
```

Next, we store the alcohol concentration data given in Table 1.3 and use the given priori parameters to compute the marginal likelihoods of the the two models M_1 and M_2:

```
> ## store the data
> (alcoholdata <-
     data.frame(gender = c("Female", "Male", "Total"),
                n = c(33, 152, 185),
                mean = c(2318.5, 2477.5, 2449.2),
                sd = c(220.1, 232.5, 237.8))
   )
  gender   n   mean    sd
1 Female  33 2318.5 220.1
2   Male 152 2477.5 232.5
3  Total 185 2449.2 237.8
> attach(alcoholdata)
> ##
> ## priori parameter
> nu <- 2000
> lambda <- 5
> alpha <- 1
> beta <- 50000
> ##
> ## vector to store the marginal likelihood values
> logMargLik <- numeric(2)
> ## compute the marginal log-likelihood of model M_1
> ## use the accumulated data for both genders
>  logMargLik[1] <-
      marginalLikelihood(n = n[3], mean = mean[3], var = sd[3]^2,
                         nu, lambda, alpha, beta,
                         log = TRUE)
> ##
> ## compute the marginal log-likelihood of model M_2
> ## first compute the marginal log-likelihoods for the
> ## two groups (female and male)
> ## the marginal log-likelihood of model M_2 is the sum
> ## of these two marginal log-likelihoods
> logMargLikFem <-
         marginalLikelihood(n = n[1], mean = mean[1], var = sd[1]^2,
                         nu, lambda, alpha, beta,
                         log = TRUE)
> logMargLikMale <-
         marginalLikelihood(n = n[2], mean = mean[2], var = sd[2]^2,
                         nu, lambda, alpha, beta,
                         log = TRUE)
> logMargLik[2] <- logMargLikFem + logMargLikMale
> logMargLik
[1] -1287.209 -1288.870
```

Hence, the marginal likelihood of model M_1 is larger than the marginal likelihood of model M_2.
For equally probable models M_1 and M_2, we have

$$\Pr(M_i \mid x) = \frac{f(x \mid M_i)}{\sum_{j=1}^{2} f(x \mid M_j)} = \frac{f(x \mid M_i)/c}{\sum_{j=1}^{2} f(x \mid M_j)/c},$$

where the expansion with the constant c^{-1} ensures that applying the implemented exponential function to the marginal likelihood values in the range of -1290 does not return the value 0. Here we use $\log(c) = \min\{\log(f(x \mid M_1), \log(f(x \mid M_2))\}$:

```
> const <- min(logMargLik)
> posterioriProb <- exp(logMargLik - const)
> (posterioriProb <- posterioriProb / sum(posterioriProb))
[1] 0.8403929 0.1596071
```

Thus, given the data above, model M_1 is more likely than model M_2, i. e. the model using the same transformation factors for both genders is much more likely than the model using the different transformation factors for women and men, respectively.

c) Evaluate the behaviour of the posterior probabilities depending on varying parameters of the prior normal-gamma distribution.

▶ *The* R*-code from part (b) to compute the posterior probabilities can be used to define a function* `modelcomp`*, which takes the four priori parameters as arguments and returns the posterior probabilities of model M_1 and M_2 (rounded to three decimals). We will vary one parameter at a time in the following:*

```
> ## given parameters
> modelcomp(nu, lambda, alpha, beta)
[1] 0.84 0.16
> ## vary nu
> for(nuNew in c(1900, 1950, 2050, 2100))
      print(modelcomp(nuNew, lambda, alpha, beta))
[1] 0.987 0.013
[1] 0.95 0.05
[1] 0.614 0.386
[1] 0.351 0.649
> ## vary lambda
> for(lambdaNew in c(1, 3, 7, 9))
      print(modelcomp(nu, lambdaNew, alpha, beta))
[1] 0.166 0.834
[1] 0.519 0.481
[1] 0.954 0.046
[1] 0.985 0.015
> ## vary alpha
> for(alphaNew in c(0.2, 0.5, 2, 3))
      print(modelcomp(nu, lambda, alphaNew, beta))
```

```
[1] 0.952 0.048
[1] 0.892 0.108
[1] 0.864 0.136
[1] 0.936 0.064
> ## vary beta
> for(betaNew in c(10000, 30000, 70000, 90000))
      print(modelcomp(nu, lambda, alpha, betaNew))
[1] 0.931 0.069
[1] 0.863 0.137
[1] 0.839 0.161
[1] 0.848 0.152
```

The larger ν is chosen the smaller the posterior probability of the simpler model M_1 becomes. In contrast, larger values of λ favour the simpler model M_1. The choice of α and β only has a small influence on the posterior model probabilities.

4. Let $X_{1:n}$ be a random sample from a normal distribution with expected value μ and known variance κ^{-1}, for which we want to compare two models. In the first model (M_1) the parameter μ is fixed to $\mu = \mu_0$. In the second model (M_2) we suppose that the parameter μ is unknown with prior distribution $\mu \sim \mathrm{N}(\nu, \delta^{-1})$, where ν and δ are fixed.

a) Determine analytically the Bayes factor BF_{12} of model M_1 compared to model M_2.

▶ *Since there is no unknown parameter in model M_1, the marginal likelihood of this model equals the usual likelihood:*

$$f(x \mid M_1) = \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right).$$

The marginal likelihood of model M_2 is given in Equation (7.18) *in the book. The Bayes factor* BF_{12} *is thus*

$$\begin{aligned} B_{12} &= \frac{f(x \mid M_1)}{f(x \mid M_2)} \\ &= \left(\frac{n\kappa + \delta}{\delta}\right)^{\frac{1}{2}} \exp\left\{-\frac{\kappa}{2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n\delta}{n\kappa + \delta}(\bar{x} - \nu)^2\right)\right\}. \end{aligned} \tag{7.2}$$

b) As an example, calculate the Bayes factor for the centered alcohol concentration data using $\mu_0 = 0$, $\nu = 0$ and $\delta = 1/100$.

▶ *For $\mu_0 = \nu = 0$, the Bayes factor can be simplified to*

$$B_{12} = \left(\frac{n\kappa + \delta}{\delta}\right)^{\frac{1}{2}} \exp\left\{-\frac{n\kappa\bar{x}^2}{2}\left(1 + \frac{\delta}{n\kappa}\right)^{-1}\right\}.$$

We choose the parameter κ as the precision estimated from the data: $\kappa = 1/\hat{\sigma}^2$.

```
> # define the Bayes factor as a function of the data and delta
> bayesFactor <- function(n, mean, var, delta)
  {
      kappa <- 1 / var
      logbayesFactor <- 1/2 * (log(n * kappa + delta) - log(delta)) -
          n * kappa * mean^2 / 2 *
              (1 + delta / (n * kappa))^{-1}
      exp(logbayesFactor)
  }
> # centered alcohol concentration data
> n <- 185
> mean <- 0
> sd <- 237.8
> # compute the Bayes factor for the alcohol data
> bayesFactor(n, mean, sd^2, delta = 1/100)
[1] 1.15202
```

According to the Bayes factor, model M_1 is more likely than model M_2, i. e. the mean transformation factor does not differ from 0, as expected.

c) Show that the Bayes factor tends to ∞ for $\delta \to 0$ irrespective of the data and the sample size n.

▶ *The claim easily follows from Equation* (7.2) *since for $\delta \to 0$, the expression in the exponential converges to*

$$-\frac{\kappa}{2}\left(\sum_{i=1}^{n}(x_i-\mu_0)^2-\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

and the factor $\sqrt{(n\kappa+\delta)/\delta}$ diverges to ∞.
For the alcohol concentration data, we can obtain a large Bayes factor by using an extremely small δ:

```
> bayesFactor(n, mean, sd^2,
              delta = 10^{-30})
[1] 5.71971e+13
```

This is an example of Lindley's paradox.
(One can deduce in a similar way that for $\mu_0 = \nu = 0$, the Bayes factor converges to 1 as $\delta \to \infty$, i. e. the two models become equally likely.)

5. In order to compare the models

$$M_0 : X \sim \mathrm{N}(0, \sigma^2)$$
$$\text{and} \quad M_1 : X \sim \mathrm{N}(\mu, \sigma^2)$$

with known σ^2 we calculate the Bayes factor BF_{01}.

a) Show that

$$\mathrm{BF}_{01} \geq \exp\left\{-\frac{1}{2}z^2\right\},$$

for arbitrary prior distribution on μ, where $z = x/\sigma$ is standard normal under model M_0. The expression $\exp(-1/2z^2)$ is called the *minimum Bayes factor* (Goodman, 1999).

▶ *We denote the unknown prior distribution of μ by $f(\mu)$. Then, the Bayes factor BF_{01} can be expressed as*

$$\mathrm{BF}_{01} = \frac{f(x \mid M_0)}{f(x \mid M_1)} = \frac{f(x \mid M_0)}{\int f(x \mid \mu) f(\mu)\, d\mu}. \tag{7.3}$$

The model M_0 has no free parameters, so its marginal likelihood is the usual likelihood

$$f(x \mid M_0) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}z^2\right\},$$

for $z = x/\sigma$. To find a lower bound for BF_{01}, we have to maximise the integral in the denomiator in (7.3). Note that the density $f(\mu)$ averages over the values of the likelihood function $f(x \mid \mu)$. Hence, it is intuitively clear that the integral is maximized if we keep the density constant at its maximum value, which is reached at the MLE $\hat{\mu}_{\mathrm{ML}} = x$. We thus obtain

$$\begin{aligned} f(x \mid M_1) &\leq f(x \mid \hat{\mu}_{\mathrm{ML}}) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\hat{\mu}_{\mathrm{ML}}}{\sigma}\right)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{1}{2}}, \end{aligned}$$

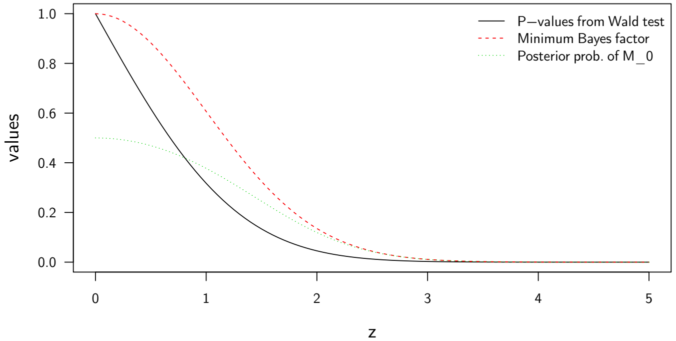
which implies

$$B_{01} = \frac{f(x \mid M_0)}{f(x \mid M_1)} \geq \exp\left\{-\frac{1}{2}z^2\right\}.$$

b) Calculate for selected values of z the two-sided P-value $2\{1-\Phi(|z|)\}$, the minimum Bayes factor and the corresponding posterior probability of M_0, assuming equal prior probabilities $\Pr(M_0) = \Pr(M_1) = 1/2$. Compare the results.

▶

```
> ## minimum Bayes factor:
> mbf <- function(z)
      exp(-1/2 * z^2)
> ## use these values for z:
> zgrid <- seq(0, 5, length = 101)
> ##
> ## compute the P-values, the values of the minimum Bayes factor and
> ## the corresponding posterrior probability of M_0
> ## note that under equal proir probabilities for the models,
> ## the posterior odds equals the Bayes factor
> pvalues <- 2 * (1 - pnorm(zgrid))
> mbfvalues <- mbf(zgrid)
> postprob.M_0 <- mbfvalues/(1 + mbfvalues)
> ##
> ## plot the obtained values
> matplot(zgrid, cbind(pvalues, mbfvalues, postprob.M_0), type = "l",
          xlab = expression(z), ylab = "values")
> legend("topright", legend = c("P-values from Wald test", "Minimum Bayes factor", "Poster
         col = 1:3, lty = 1:3, bty = "n")
> ## comparisons:
> all(pvalues <= mbfvalues)
[1] TRUE
> zgrid[pvalues == mbfvalues]
[1] 0
```



Thus, the P-values from the Wald test are smaller than or equal to the minimum Bayes factors for all considered values of z. Equality holds for $z = 0$ only and for $z > 3$, the P-values and the minimum Bayes factors are very similar.

6. Consider the models

$$
\begin{aligned}
& M_0 : p \sim \mathrm{U}(0,1) \\
\text{and} \quad & M_1 : p \sim \mathrm{Be}(\theta,1)
\end{aligned}
$$

where $0 < \theta < 1$. This scenario aims to reflect the distribution of a two-sided P-value p under the null hypothesis (M_0) and some alternative hypothesis (M_1), where smaller P-values are more likely (Sellke et al., 2001). This is captured by the decreasing density of the $\mathrm{Be}(\theta, 1)$ for $0 < \theta < 1$. Note that the data are now represented by the P-value.

a) Show that the Bayes factor for M_0 versus M_1 is

$$\mathrm{BF}(p) = \left\{ \int_0^1 \theta p^{\theta-1} f(\theta) d\theta \right\}^{-1}$$

for some prior density $f(\theta)$ for θ.

▶ *We have*

$$
\begin{aligned}
\mathrm{BF}(p) &= \frac{f(p \mid M_0)}{f(p \mid M_1)} \\
&= \frac{1}{\int_0^1 B(\theta,1)^{-1} p^{\theta-1} f(\theta)\, d\theta} \\
&= \left\{ \int_0^1 \theta p^{\theta-1} f(\theta) d\theta \right\}^{-1},
\end{aligned}
$$

since

$$B(\theta,1) = \frac{\Gamma(\theta)}{\Gamma(\theta+1)} = \frac{\Gamma(\theta)}{\theta\, \Gamma(\theta)} = 1/\theta.$$

To see that $\Gamma(\theta+1) = \theta\ \Gamma(\theta)$, we can use integration by parts:

$$
\begin{aligned}
\Gamma(\theta+1) &= \int_0^\infty t^\theta \exp(-t)\, dt \\
&= \left[-t^\theta \exp(-t)\right] \mid_0^\infty + \int_0^\infty \theta\ t^{\theta-1} \exp(-t)\, dt \\
&= \theta \int_0^\infty t^{\theta-1} \exp(-t)\, dt = \theta\ \Gamma(\theta).
\end{aligned}
$$

b) Show that the minimum Bayes factor mBF over all prior densities $f(\theta)$ has the form

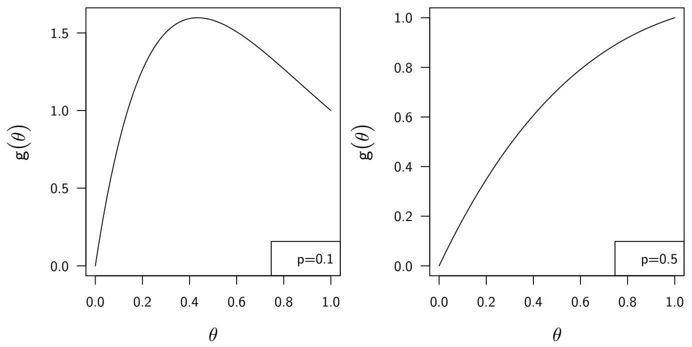
$$\mathrm{mBF}(p) = \begin{cases} -e\, p \log p & \text{for } p < e^{-1}, \\ 1 & \text{otherwise,} \end{cases}$$

where $e = \exp(1)$ is Euler's number.

▶ *We have*

$$\mathrm{mBF}(p) = \min_{f\, density} \left(\int_0^1 \theta p^{\theta-1} f(\theta)\, d\theta \right)^{-1} = \left(\max_{\theta \in [0,1]} \theta p^{\theta-1} \right)^{-1},$$

where the last equality is due to the fact that the above integral is maximum if the density $f(\theta)$ is chosen as a point mass at the value of θ which maximises $\theta p^{\theta-1}$.
We now consider the function $g(\theta) = \theta p^{\theta-1}$ to determine its maximum. For $p < 1/e$, the function g has a unique maximum in $(0, 1)$. For $p \geq 1/e$, the function g is strictly monotoically increasing on [0,1] and thus attains its maximum at $\theta = 1$ (compare to the figure below).



We now derive the maxima described above analytically:

i. *Case $p < 1/e$:*
We compute the maximum of the function $h(\theta) = log(g(\theta))$:

$$h(\theta) = log(\theta) + (\theta - 1)\log(p),$$
$$\frac{d}{d\theta}h(\theta) = \frac{1}{\theta} + \log(p) \qquad \textit{and hence}$$
$$\frac{d}{d\theta}h(\theta) = 0 \quad \Rightarrow \quad \theta = -\frac{1}{\log(p)}.$$

It is easy to see that $\frac{d}{d\theta}h(\theta) > 0$ for $\theta < -(\log p)^{-1}$ and $\frac{d}{d\theta}h(\theta) < 0$ for $\theta > -(\log p)^{-1}$, so that h and hence also g are strictly monotonically increasing for $\theta < -(\log p)^{-1}$ and strictly monotonically decreasing for $\theta > -(\log p)^{-1}$. Consequently, the maximum of g determined above is unique and we have

$$\begin{aligned}\sup_{\theta \in [0,1]} \theta p^{\theta - 1} &= g\left(-\frac{1}{\log p}\right)\\ &= -\frac{1}{\log p}\exp\left[\log p\left(-\frac{1}{\log p} - 1\right)\right]\\ &= -\frac{1}{\log(p) e\, p},\end{aligned}$$

which implies $\mathrm{mBF}(p) = -e\, p \log p$.

ii. *Case $p \geq 1/e$:*
We have

$$\frac{d}{d\theta}g(\theta) = p^{\theta-1}(1 + \theta\log p) \geq 0$$

for all $\theta \in [0,1]$ since $\log p \geq -1$ in this case. Thus, g is monotonically increasing on $[0,1]$ and

$$\mathrm{mBF}(p) = \left(\sup_{\theta \in [0,1]} \theta p^{\theta - 1}\right)^{-1} = (g(1))^{-1} = 1.$$

c) Compute and interpret the minimum Bayes factor for selected values of p (e.g. $p = 0.05$, $p = 0.01$, $p = 0.001$).

▶

```
> ## minimum Bayes factor:
> mbf <- function(p)
      { if(p < 1/exp(1))
           { - exp(1)*p*log(p) }
        else
           { 1 }
      }
> ## use these values for p:
> p <- c(0.05, 0.01, 0.001)
> ## compute the corresponding minimum Bayes factors
> minbf <- numeric(length =3)
> for (i in 1:3)
    {minbf[i] <- mbf(p[i])
    }
> minbf
[1] 0.40716223 0.12518150 0.01877723
> ratio <- p/minbf
> ratio
[1] 0.12280117 0.07988401 0.05325600
> ## note that the minimum Bayes factors are considerably larger
> ## than the corresponding p-values
```

For $p = 0.05$, we obtain a minimum Bayes factor of approximately 0.4. This means that given the data $p = 0.05$, model M_0 as at least 40% as likely as model M_1. If the prior odds of M_0 versus M_1 is 1, then the posterior odds of M_0 versus M_1 is at least 0.4. Hence, the data $p = 0.05$ does not correspond to strong evidence against model M_0. The other minimum Bayes factors have an analoguous interpretation.

7. Box (1980) suggested a method to investigate the compatibility of a prior with the observed data. The approach is based on computation of a P-value obtained from the prior predictive distribution $f(x)$ and the actually observed datum x_o. Small p-values indicate a *prior-data conflict* and can be used for *prior criticism.*
Box's p-value is defined as the probability of obtaining a result with prior predictive ordinate $f(X)$ equal to or lower than at the actual observation x_o:

$$\mathsf{Pr}\{f(X) \leq f(x_o)\},$$

here X is distributed according to the prior predictive distribution $f(x)$, so $f(X)$ is a random variable. Suppose both likelihood and prior are normal, *i. e.* $X \mid \mu \sim \mathrm{N}(\mu, \sigma^2)$ and $\mu \sim \mathrm{N}(\nu, \tau^2)$. Show that Box's p-value is the upper tail probability of a $\chi^2(1)$ distribution evaluated at

$$\frac{(x_o - \nu)^2}{\sigma^2 + \tau^2}.$$

▶ *We have already derived the prior predictive density for a normal likelihood with known variance and a normal prior for the mean μ in Exercise 1. By setting $\kappa = 1/\sigma^2$, $\delta = 1/\tau^2$ and $n = 1$ in Equation (7.18), we obtain*

$$f(x) = \left(\frac{1}{2\pi(\tau^2+\sigma^2)}\right) \exp\left(-\frac{1}{2(\tau^2+\sigma^2)}(x-\nu)^2\right),$$

that is $X \sim \mathrm{N}(\nu, \sigma^2+\tau^2)$. Consequently,

$$\frac{X-\nu}{\sqrt{\sigma^2+\tau^2}} \sim \mathrm{N}(0,1) \quad \textit{and} \quad \frac{(X-\nu)^2}{\sigma^2+\tau^2} \sim \chi^2(1) \tag{7.4}$$

(see Table A.2 for the latter fact).
Thus, Box's p-value is

$$\begin{aligned}
&\Pr\{f(X) \le f(x_o)\}\\
&= \Pr\left(\frac{1}{(2\pi(\sigma^2+\tau^2))^{1/2}} \exp\left(-\frac{(X-\nu)^2}{2(\sigma^2+\tau^2)}\right) \le \frac{1}{(2\pi(\sigma^2+\tau^2))^{1/2}} \exp\left(-\frac{(x_o-\nu)^2}{2(\sigma^2+\tau^2)}\right)\right)\\
&= \Pr\left(-\frac{(X-\nu)^2}{\sigma^2+\tau^2} \le -\frac{(x_o-\nu)^2}{\sigma^2+\tau^2}\right)\\
&= \Pr\left(\frac{(X-\nu)^2}{\sigma^2+\tau^2} \ge \frac{(x_o-\nu)^2}{\sigma^2+\tau^2}\right).
\end{aligned}$$

Due to (7.4), the latter probability equals the upper tail probability of a $\chi^2(1)$ distribution evaluated at $(x_o-\nu)^2/(\sigma^2+\tau^2)$.

8 Numerical methods for Bayesian inference

1. Let $X \sim \mathrm{Po}(e\lambda)$ with known e, and assume the prior $\lambda \sim \mathrm{G}(\alpha, \beta)$.

a) Compute the posterior expectation of λ.

▶ *The posterior density is*

$$\begin{aligned}
f(\lambda \mid x) &\propto f(x \mid \lambda) f(\lambda)\\
&\propto \lambda^x \exp(-e\lambda) \cdot \lambda^{\alpha-1} \exp(-\beta\lambda)\\
&= \lambda^{(\alpha+x)-1} \exp\bigl(-(\beta+e)\lambda\bigr),
\end{aligned}$$

which is the kernel of the $\mathrm{G}(\alpha+x, \beta+e)$ distribution (compare to Table 6.2). The posterior expectation is thus

$$\mathsf{E}(\lambda \mid x) = \frac{\alpha+x}{\beta+e}.$$

b) Compute the Laplace approximation of this posterior expectation.

▶ *We use approximation (8.6) with $g(\lambda) = \lambda$ und $n = 1$. We have*

$$\begin{aligned}
-k(\lambda) &= \log(f(x \mid \lambda)) + \log(f(\lambda))\\
&= (\alpha+x-1)\log(\lambda) - (\beta+e)\lambda\\
\textit{and} \quad -k_g(\lambda) &= \log(\lambda) - k(\lambda).
\end{aligned}$$

The derivatives of these functions are

$$\begin{aligned}
\frac{d(-k(\lambda))}{d\lambda} &= \frac{\alpha+x-1}{\lambda} - (\beta+e)\\
\textit{bzw.} \quad \frac{d(-k_g(\lambda))}{d\lambda} &= \frac{\alpha+x}{\lambda} - (\beta+e)
\end{aligned}$$

with roots

$$\hat{\lambda} = \frac{\alpha+x-1}{\beta+e} \quad \textit{and} \quad \hat{\lambda}_g = \frac{\alpha+x}{\beta+e}.$$

The negative curvatures of $-k(\lambda)$ and $-k_g(\lambda)$ at the above maxima turn out to be

$$\hat{\kappa} = \frac{d^2k(\hat{\lambda})}{d\lambda^2} = \frac{(\beta+e)^2}{\alpha+x-1} \quad \textit{and} \quad \hat{\kappa}_g = \frac{d^2k_g(\hat{\lambda}_g)}{d\lambda^2} = \frac{(\beta+e)^2}{\alpha+x},$$

which yields the following Laplace approximation of the posterior expectation:

$$
\begin{aligned}
\hat{\mathsf{E}}(\lambda \mid x) &= \sqrt{\frac{\alpha+x}{\alpha+x-1}}\exp\Big\{\log(\hat{\lambda}^*) + (\alpha+x-1)\log(\hat{\lambda}^*) - (\beta+e)\hat{\lambda}^* \\
&\qquad - (\alpha+x-1)\log(\hat{\lambda}) + (\beta+e)\hat{\lambda}\Big\} \\
&= \sqrt{\frac{\alpha+x}{\alpha+x-1}}\exp\left\{\log(\hat{\lambda}^*) + (\alpha+x-1)\log(\hat{\lambda}^*/\hat{\lambda}) + (\beta+e)(\hat{\lambda}-\hat{\lambda}^*)\right\} \\
&= \sqrt{\frac{\alpha+x}{\alpha+x-1}}\exp\left\{\log\left(\frac{\alpha+x}{e+\beta}\right) - (\alpha+x-1)\log\left(\frac{\alpha+x}{\alpha+x-1}\right) - 1\right\} \\
&= \exp\big\{(\alpha+x+0.5)\log(\alpha+x) - (\alpha+x-0.5)\log(\alpha+x-1) \\
&\qquad - \log(\beta+e) - 1\big\}.
\end{aligned}
$$

c) For $\alpha = 0.5$ and $\beta = 0$, compare the Laplace approximation with the exact value, given the observations $x = 11$ and $e = 3.04$, or $x = 110$ and $e = 30.4$. Also compute the relative error of the Laplace approximation.

▶ *We first implement the Laplace approximation and the exact formula:*

```
> ## Laplace approximation of the posterior expectation
> ## for data x, offset e und priori parameters alpha, beta:
> laplaceApprox1 <- function(x, e, alpha, beta)
  {
      logRet <- (alpha + x + 0.5) * log(alpha + x) -
          (alpha + x - 0.5) * log(alpha + x - 1) -
              log(beta + e) - 1
      exp(logRet)
  }
> ## exact calculation of the posterior expectation
> exact <- function(x, e, alpha, beta)
      (alpha + x) / (beta + e)
```

Using the values given above, we obtain

```
> (small <- c(exact = exact(11, 3.04, 0.5, 0),
              approx = laplaceApprox1(11, 3.04, 0.5, 0)))
   exact   approx
3.782895 3.785504
> (large <- c(exact = exact(110, 30.4, 0.5, 0),
              approx = laplaceApprox1(110, 30.4, 0.5, 0)))
   exact   approx
3.634868 3.634893
> ## relative errors:
> diff(small) / small["exact"]
      approx
0.0006897981
> diff(large) / large["exact"]
      approx
6.887162e-06
```

For a fixed ratio of observed value x and offset e, the Laplace approximation thus improves for larger values of x and e. If we consider the ratio of the Laplace approximation and the exact value

$$
\begin{aligned}
\frac{\hat{\mathsf{E}}(\lambda \mid x)}{\mathsf{E}(\lambda \mid x)} &= \exp\big\{(\alpha+x-0.5)\big(\log(\alpha+x) - \log(\alpha+x-1)\big) - 1\big\} \\
&= \left(1 + \frac{1}{\alpha+x-1}\right)^{\alpha+x-0.5} \Big/ \exp(1),
\end{aligned}
$$

it is not hard to see that this ratio converges to 1 as $x \to \infty$.

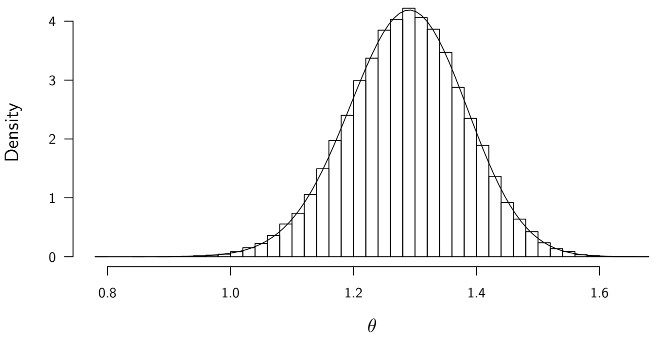
d) Now consider $\theta = \log(\lambda)$. First derive the posterior density function using the change of variables formula (A.11). Second, compute the Laplace approximation of the posterior expectation of θ and compare again with the exact value which you have obtained by numerical integration using the R-function `integrate`.

▶ *The posterior density is*

$$
\begin{aligned}
f_\theta(\theta \mid x) &= f_\lambda\big(g^{-1}(\theta) \mid x\big) \cdot \left|\frac{d}{d\theta} g^{-1}(\theta)\right| \\
&= \frac{(\beta+e)^{\alpha+x}}{\Gamma(\alpha+x)} \exp(\theta)^{\alpha+x-1} \exp\big(-(\beta+e)\exp(\theta)\big) \cdot \exp(\theta) \\
&= \frac{(\beta+e)^{\alpha+x}}{\Gamma(\alpha+x)} \exp\big\{(\alpha+x)\theta - (\beta+e)\exp(\theta)\big\},
\end{aligned}
$$

which does not correspond to any well-known distribution.

```
> ## posterior density of theta = log(lambda):
> thetaDens <- function(theta, x, e, alpha, beta, log = FALSE)
  {
      logRet <- (alpha + x) * (theta + log(beta + e)) -
          (beta + e) * exp(theta) - lgamma(alpha + x)
      if(log)
          return(logRet)
      else
          return(exp(logRet))
  }
> # check by simulation if the density is correct:
> x <- 110
> e <- 30.4
> alpha <- 0.5
> beta <- 0
> ## draw histogram of a sample from the distribution of log(theta)
> set.seed(59)
> thetaSamples <- log(rgamma(1e+5, alpha + x, beta + e))
> histResult <- hist(thetaSamples, prob= TRUE, breaks = 50,
                     xlab = expression(theta), main = "")
> ## plot the computed density
> thetaGrid <- seq(from = min(histResult$breaks),
                   to = max(histResult$breaks), length = 101)
> lines(thetaGrid, thetaDens(thetaGrid, x, e, alpha, beta))
> ## looks correct!
```



We first reparametrise the likelihood and the prior density. The transformation is $h(\lambda) = \log(\lambda)$. Since the likelihood is invariant with respect to one-to-one parameter transformations, we can just plug in $h^{-1}(\theta) = \exp(\theta)$ in place of λ:

$$f(x \mid h^{-1}(\theta)) = \frac{1}{x!}\exp(\theta)^{\alpha-1}\exp(-e\exp(\theta)).$$

To transform the proir density, we apply the change-of-variables formula to get

$$f(\theta) = \frac{\alpha^{\beta}}{\Gamma(\alpha)}\exp(\theta)^{\alpha}\exp(-\beta\exp(\theta)).$$

Thus,

$$\begin{aligned}
-k(\theta) &= \log\left\{f\left(x \mid h^{-1}(\theta)\right)\right\} + \log\{f(\theta)\} \\
&= (\alpha + x)\theta - (\beta + e)\exp(\theta) + \text{const} \qquad \textit{and} \\
-k_g(\theta) &= \log\left\{h^{-1}(\theta)\right\} - k(\theta) \\
&= \log(\theta) + (\alpha + x)\theta - (\beta + e)\exp(\theta) + \text{const}.
\end{aligned}$$

with $g = id$. The derivatives are

$$\begin{aligned}
-\frac{dk(\theta)}{d\theta} &= \alpha + x - (\beta + e)\exp(\theta) \qquad \textit{and} \\
-\frac{dk_g(\theta)}{d\theta} &= \frac{1}{\theta} + (\alpha + x) - (\beta + e)\exp(\theta).
\end{aligned}$$

The root of $k(\theta)$ is thus

$$\hat{\theta} = \log\left(\frac{\alpha + x}{\beta + e}\right) = \log(\alpha + x) - \log(\beta + e),$$

but the root of $k_g(\theta)$ cannot be computed analytically. We therefore proceed with numerical maximisation for $k_g(\theta)$ below. First, we complete the analytical calculations for $k(\theta)$:

$$-k(\hat{\theta}) = (\alpha + x)(\log(\alpha + x) - \log(\beta + e) - 1)$$

and the second-order derivative is

$$\frac{d^2k(\theta)}{d\theta^2} = (\beta + e)\exp(\theta),$$

which yields the following curvature of $k(\theta)$ at its minimum:

$$\hat{\kappa} = \frac{d^2k(\hat{\theta})}{d\theta^2} = \alpha + x.$$

```
> # numerical computation of Laplace approximation
> # for the posterior expectation of theta=log(lambda)
> numLaplace <- function(x, e, alpha, beta)
  {
    # first implement the formulas calculated above
    minus.k <- function(theta)
    {
      + (alpha+x)*theta - (beta +e)*exp(theta)
    }
    # location of maximum of -k(theta)
    theta.hat <- log(alpha+x) - log(beta+e)
    # curvature of -k(theta) at maximum
    kappa <- alpha+x

    # function to be maximised
    minus.kg <- function(theta)
    {
      log(theta) + (alpha+x)*theta - (beta +e)*exp(theta)
    }
    # numerical optimisation to find the maximum of -kg(theta)
    optimObj <- optim(par=1, fn=minus.kg, method="BFGS",
                      control = list(fnscale=-1), hessian=TRUE)
    # location of maximum of -kg(theta)
    thetag.hat <- optimObj$par
    # curvature at maximum
    kappa.g <- - optimObj$hessian
    # Laplace approximation
    approx <- sqrt(kappa/kappa.g) * exp(minus.kg(thetag.hat)
                                        - minus.k(theta.hat))
    return(approx)
  }
> # numerical integration
> numInt <- function(x, e, alpha, beta)
  {
    integrand <- function(theta, x, e, alpha, beta)
    {
      theta* thetaDens(theta, x, e, alpha, beta)
    }
    numInt <- integrate(integrand, x, e, alpha, beta, lower = -Inf,
                         upper = Inf, rel.tol = sqrt(.Machine$double.eps))
    return(numInt$value)
  }
> # comupte the exact and approximated values
> # use the data from part (c)
> (small <- c(exact = numInt(11, 3.04, 0.5, 0),
              approx = numLaplace(11, 3.04, 0.5, 0)))
```

```
   exact   approx
1.286382 1.293707
> (large <- c(exact = numInt(110, 30.4, 0.5, 0),
              approx = numLaplace(110, 30.4, 0.5, 0)))
   exact   approx
1.286041 1.286139
> ## relative errors:
> diff(small) / small["exact"]
     approx
0.005694469
> diff(large) / large["exact"]
     approx
7.602533e-05
```

The posterior expectations obtained by the Laplace approximation are close to the exact values. As observed before for λ, the Laplace approximation is more accurate for larger values of e and x if the ratio of e and x is kept fixed.

e) Additional exercise: Compute the Laplace approximation of the posterior expectation of $\exp(\theta)$ and compare it with the Laplace approximation in 1c) and with the exact value obtained by numerical integration.

▶ *The support of the parameter $\theta = g(\lambda) = \log(\lambda)$ is now the whole real line, which may lead to an improved Laplace approximation. To calculate the Laplace approximation of the posterior expectation of $\exp(\theta)$, we first reparametrise the likelihood and the prior density. The transformation is $h(\lambda) = \log(\lambda)$. Since the likelihood is invariant with respect to one-to-one parameter transformations, we can just plug in $h^{-1}(\theta) = \exp(\theta)$ in place of λ:*

$$f(x \mid h^{-1}(\theta)) = \frac{1}{x!}\exp(\theta)^{\alpha-1}\exp(-e\exp(\theta)).$$

To transform the proir density, we apply the change-of-variables formula to get

$$f(\theta) = \frac{\alpha^{\beta}}{\Gamma(\alpha)}\exp(\theta)^{\alpha}\exp(-\beta\exp(\theta)).$$

Hence, we have

$$\begin{aligned}
-k(\theta) &= \log\left\{f\left(x \mid h^{-1}(\theta)\right)\right\} + \log\left\{f(\theta)\right\} \\
&= (\alpha + x)\theta - (\beta + e)\exp(\theta) + \text{const} \qquad \textit{and} \\
-k_g(\theta) &= \log\left\{h^{-1}(\theta)\right\} - k(\theta) \\
&= (\alpha + x + 1)\theta - (\beta + e)\exp(\theta) + \text{const}.
\end{aligned}$$

with $g(\theta) = \exp(\theta)$. The derivatives are

$$\begin{aligned}
-\frac{dk(\theta)}{d\theta} &= \alpha + x - (\beta + e)\exp(\theta) \qquad \textit{and} \\
-\frac{dk_g(\theta)}{d\theta} &= \alpha + x + 1 - (\beta + e)\exp(\theta).
\end{aligned}$$

with roots

$$\begin{aligned}
\hat{\theta} &= \log\left(\frac{\alpha + x}{\beta + e}\right) = \log(\alpha + x) - \log(\beta + e), \qquad \textit{and} \\
\hat{\theta_g} &= \log\left(\frac{\alpha + x + 1}{\beta + e}\right).
\end{aligned}$$

The negative curvatures of $-k(\theta)$ and $-k_g(\theta)$ at the above maxima turn out to be

$$\hat{\kappa} = \frac{d^2k(\hat{\theta})}{d\theta^2} = \alpha + x \qquad \textit{and} \qquad \hat{\kappa_g} = \frac{d^2k_g(\hat{\theta}_g)}{d\theta^2} = \alpha + x + 1.$$

Combining these results yields the Laplace approximation

$$\hat{\mathsf{E}}\big(\exp(\theta) \mid x\big) = \exp\big\{(\alpha+x+0.5)\log(\alpha+x+1) - (\alpha+x-0.5)\log(\alpha+x) - \log(\beta+e) - 1\big\},$$

a very similar formula as in Exercise 1b). Note that these are two different approximations for the same posterior expectation since $\exp(\theta) = \lambda$. Here, the ratio of the approximated value and the true value is

$$\frac{\hat{\mathsf{E}}(\exp(\theta) \mid x)}{\mathsf{E}(\exp(\theta) \mid x)} = \left(1 + \frac{1}{\alpha + x}\right)^{\alpha + x + 0.5} \Big/ \exp(1),$$

that is one step in x closer to the limit 1 than the approximation in 1b). The approximation derived here is hence a bit more accurate than the one in 1b). We now compare the two Laplace approximations with the values obtained by numerical integration using the data from 1c):

```
> ## Laplace approximation of E(exp(theta))
> laplaceApprox2 <- function(x, e, alpha, beta)
  {
      logRet <- (alpha + x + 0.5) * log(alpha + x + 1) -
          (alpha + x - 0.5) * log(alpha + x) -
              log(beta + e) - 1
      exp(logRet)
  }
> ## numerical approximation
> numApprox <- function(x, e, alpha, beta)
  {
      integrand <- function(theta)
      ## important: add on log scale first and exponentiate afterwards
      ## to avoid numerical problems
          exp(theta + thetaDens(theta, x, e, alpha, beta, log = TRUE))

      intRes <- integrate(integrand, lower = -Inf, upper = Inf,
                          rel.tol = sqrt(.Machine$double.eps))
      if(intRes$message == "OK")
          return(intRes$value)
      else
          return(NA)
  }
> ## comparison of the three methods:
```

```
> (small <- c(exact = exact(11, 3.04, 0.5, 0),
              approx1 = laplaceApprox1(11, 3.04, 0.5, 0),
              approx2 = laplaceApprox2(11, 3.04, 0.5, 0),
              approx3 = numApprox(11, 3.04, 0.5, 0)))
   exact  approx1  approx2  approx3
3.782895 3.785504 3.785087 3.782895
> (large <- c(exact = exact(110, 30.4, 0.5, 0),
              approx = laplaceApprox1(110, 30.4, 0.5, 0),
              approx2 = laplaceApprox2(110, 30.4, 0.5, 0),
              approx3 = numApprox(110, 30.4, 0.5, 0)))
   exact   approx  approx2  approx3
3.634868 3.634893 3.634893 3.634868
> ## relative errors:
> (small[2:4] - small["exact"]) / small["exact"]
      approx1       approx2       approx3
 6.897981e-04  5.794751e-04 -1.643516e-15
> (large[2:4] - large["exact"]) / large["exact"]
       approx       approx2       approx3
 6.887162e-06  6.763625e-06 -4.105072e-14
```

Numerical integration using `integrate` *thus gives even more accurate results than the two Laplace approximations in this setting.*

2. In Example 8.3, derive the Laplace approximation (8.9) for the posterior expectation of π using the variance stabilising transformation.

▶ *As mentioned in Example 8.3, the variance stabilising transformation is* $\phi = h(\pi) = \arcsin(\sqrt{\pi})$ *and its inverse is* $h^{-1}(\phi) = \sin^2(\phi)$. *The relation* $\sin^2(\phi) + \cos^2(\phi) = 1$ *will be used several times in the following.*

We first reparametrise the likelihood and the prior density:

$$\begin{aligned} f(x \mid h^{-1}(\phi)) &= \binom{n}{x} \sin(\phi)^{2x}(1-\sin^2(\phi))^{n-x} \\ &= \binom{n}{x} \sin(\phi)^{2x}\cos(\phi)^{2(n-x)} \end{aligned}$$

and applying the change-of-variables formula gives

$$\begin{aligned} f(\phi) &= \mathrm{B}(0.5, 0.5)^{-1}\{\sin^2(\phi)(1-\sin^2(\phi))\}^{-\frac{1}{2}} 2\sin(\phi)\cos(\phi) \\ &= 2\,\mathrm{B}(0.5, 0.5)^{-1}, \end{aligned}$$

i. e. the transformed density is constant. Thus,

$$\begin{aligned} -k(\phi) &= \log\left\{f\left(x \mid h^{-1}(\phi)\right)\right\} + \log\{f(\phi)\} \\ &= 2x\log\{\sin(\phi)\} + 2(n-x)\log\{\cos(\phi)\} + \text{const} \qquad \textit{and} \\ -k_g(\phi) &= \log\left\{h^{-1}(\phi)\right\} - k(\phi) \\ &= \log\left\{\sin^2(\phi)\right\} + 2x\log\{\sin(\phi)\} + 2(n-x)\log\{\cos(\phi)\} + \text{const.} \end{aligned}$$

with $g = id$. *The derivatives are*

$$\begin{aligned} -\frac{dk(\phi)}{d\phi} &= \frac{2x\cos(\phi)}{\sin(\phi)} - \frac{2(n-x)\sin(\phi)}{\cos(\phi)} \\ &= 2\left(\frac{x}{\tan(\phi)} - (n-x)\tan(\phi)\right) = \frac{2(x-n\sin^2(\phi))}{\sin(\phi)\cos(\phi)} \qquad \textit{and} \\ -\frac{dk_g(\phi)}{d\phi} &= -\frac{dk(\pi)}{d\phi} + \frac{2\cos(\phi)}{\sin(\phi)} \\ &= 2\left(\frac{x+1}{\tan(\phi)} - (n-x)\tan(\phi)\right) \\ &= \frac{2\{x+1-(n+1)\sin^2(\phi)\}}{\sin(\phi)\cos(\phi)}. \end{aligned}$$

The different expressions for the derivatives will be useful for calculating the roots and the second-order derivatives, respectively. From the last expressions, we easily obtain the roots

$$\hat{\phi} = \arcsin\left(\sqrt{\frac{x}{n}}\right) \qquad \textit{and} \qquad \hat{\phi}_g = \arcsin\left(\sqrt{\frac{x+1}{n+1}}\right).$$

Exploiting the relation $\cos(\arcsin(x)) = (1-x^2)^{1/2}$ *gives*

$$\begin{aligned} -k(\hat{\phi}) &= 2\left\{x\log\left(\sqrt{\frac{x}{n}}\right) + (n-x)\log\left(\sqrt{\frac{n-x}{n}}\right)\right\} \qquad \textit{and} \\ -k_g(\hat{\phi}) &= \log\left(\frac{x+1}{n+1}\right) + 2\left\{x\log\left(\sqrt{\frac{x+1}{n+1}}\right) + (n-x)\log\left(\sqrt{\frac{n-x}{n+1}}\right)\right\}. \end{aligned}$$

By using for example

$$\frac{d\tan(\phi)}{d\phi} = \frac{1}{\cos^2(\phi)},$$

we obtain the second-order derivatives

$$\begin{aligned} \frac{d^2k(\phi)}{d\phi^2} &= 2\left(\frac{x}{\sin^2(\phi)} - \frac{n-x}{\cos^2(\phi)}\right) \qquad \textit{and} \\ \frac{d^2k_g(\phi)}{d\phi^2} &= 2\left(\frac{x+1}{\sin^2(\phi)} - \frac{n-x}{\cos^2(\phi)}\right), \end{aligned}$$

which yields the following curvatures of $k(\phi)$ *and* $k_g(\phi)$ *at their minima:*

$$\begin{aligned} \hat{\kappa} &= \frac{d^2k(\hat{\phi})}{d\phi^2} = 4n \qquad \textit{and} \\ \hat{\kappa_g} &= \frac{d^2k(\hat{\phi})}{d\phi^2} = 4(n+1). \end{aligned}$$

Combining the above results, we obtain

$$\begin{aligned}\hat{\mathsf{E}}_2(\pi\,|\,x) &= \sqrt{\frac{\hat{\kappa}}{\hat{\kappa}_g}}\exp\bigl[-\{k_g(\hat{\theta}_g)-k(\hat{\theta})\}\bigr]\\ &= \sqrt{\frac{n}{n+1}}\exp\bigl[(x+1)\log(x+1)-(n+1)\log(n+1)-x\log(x)+n\log(n)\bigr]\\ &= \frac{(x+1)^{x+1}n^{n+0.5}}{x^x(n+1)^{n+3/2}}.\end{aligned}$$

3. For estimating the odds ratio θ from Example 5.8 we will now use Bayesian inference. We assume independent $\mathrm{Be}(0.5, 0.5)$ distributions as priors for the probabilities π_1 and π_2.

a) Compute the posterior distributions of π_1 and π_2 for the data given in Table 3.1. Simulate samples from these posterior distributions and transform them into samples from the posterior distributions of θ and $\psi = \log(\theta)$. Use the samples to compute Monte Carlo estimates of the posterior expectations, medians, equi-tailed credible intervals and HPD intervals for θ and ψ. Compare with the results from likelihood inference in Example 5.8.

▶ *By Example 6.3, the posterior distributions are*

$$\begin{aligned}\pi_1\,|\,x_1 &\sim \mathrm{Be}(0.5+x_1, 0.5+n_1-x_1) = \mathrm{Be}(6.5, 102.5) \qquad \textit{and}\\ \pi_2\,|\,x_2 &\sim \mathrm{Be}(0.5+x_2, 0.5+n_2-x_2) = \mathrm{Be}(2.5, 101.5).\end{aligned}$$

We now generate random numbers from these two distributions and transform them to

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

and $\psi = \log(\theta)$, respectively, to obtain the desired Monte Carlo estimates:

```
> ## data from Table 3.1:
> x <- c(6, 2)
> n <- c(108, 103)
> ## simulate from the posterior distributions of pi1 and pi2
> size <- 1e+5
> set.seed(89)
> pi1 <- rbeta(size, 0.5 + x[1], 0.5 + n[1] - x[1])
> pi2 <- rbeta(size, 0.5 + x[2], 0.5 + n[2] - x[2])
> ## transform the samples
> theta <- (pi1 / (1 - pi1)) / (pi2 / (1 - pi2))
> psi <- log(theta)
> ##
> ## function for HPD interval calculation (see code in Example 8.9)
> ## approximate unimodality of the density is assumed!
> ## (HPD intervals consisting of several disjoint intervals
> ## cannot be found)
> hpd <- function(                          # returns a MC estimate
                  samples,                  # of the HPD interval
                  prob = 0.95               # based on samples
                  )                         # for the probability prob
  {
      ## sort the sample "samples"
      M <- length(samples)
      level <- 1 - prob
      samplesorder <- samples[order(samples)]

      ## determine and return the smallest interval
      ## containing at least 95% of the values in "samples"
      max.size <- round(M * level)
      size <- rep(NA, max.size)
      for(i in 1:max.size){
          lower <- samplesorder[i]
          upper <- samplesorder[M-max.size+i]
          size[i] <- upper - lower
      }
      size.min <- which.min(size)
      HPD.lower <- samplesorder[size.min]
      HPD.upper <- samplesorder[M-max.size+size.min]
      return(c(lower = HPD.lower, upper = HPD.upper))
  }
> ## compute the Monte Carlo estimates:
> ## estimates for theta
> quantile(theta, prob = c(0.025, 0.5, 0.975)) # equi-tailed CI with median
     2.5%        50%      97.5%
 0.6597731  2.8296129 16.8312942
> mean(theta)                                  # mean
[1] 4.338513
> (thetaHpd <- hpd(theta))                     # HPD interval
    lower      upper
 0.2461746 12.2797998
> ##
> ## estimates for psi
> quantile(psi, prob = c(0.025, 0.5, 0.975)) # equi-tailed CI with median
      2.5%        50%      97.5%
-0.4158593  1.0401399  2.8232399
> mean(psi)                                  # mean
[1] 1.082216
> (psiHpd <- hpd(psi))                       # HPD interval
     lower      upper
-0.5111745  2.7162842
```

The Bayesian point estimate $\mathsf{E}(\psi\,|\,x) \approx 1.082$ *is slightly smaller than the MLE* $\hat{\psi}_{\mathrm{ML}} \approx 1.089$. *The HPD interval* $[-0.511, 2.716]$ *is similar to the Wald conficence interval* $[-0.54, 2.71]$, *whereas to equi-tailed credible interval* $[-0.416, 2.823]$ *is more similar to the profile likelihood confidence interval* $[-0.41, 3.02]$. *(Also compare to Exercise 4 in Chapter 5.)*

b) Try to compute the posterior densities of θ and ψ analytically. Use the density functions to numerically compute the posterior expectations and HPD intervals. Compare with the Monte Carlo estimates from 3a).

▶ *To find the density of the odds ratio*

$$\theta = \frac{\pi_1}{1-\pi_1} \cdot \frac{1-\pi_2}{\pi_2},$$

we first derive the density of $\gamma = \pi/(1-\pi)$ *for* $\pi \sim \mathrm{Be}(a,b)$*: Similarly to Exercise 2 in Chapter 6, we apply the change-of-variables formula with transformation* $g(\pi) = \pi/(1-\pi)$*, inverse function* $g^{-1}(\gamma) = \gamma/(1+\gamma)$ *and derivative*

$$\frac{d}{d\gamma} g^{-1}(\gamma) = \frac{1}{(1+\gamma)^2}$$

to get

$$\begin{aligned} f_\gamma(\gamma) &= \frac{1}{B(a,b)} \left(\frac{\gamma}{\gamma+1}\right)^{a-1} \left(1 - \frac{\gamma}{\gamma+1}\right)^{b-1} \frac{1}{(\gamma+1)^2} \\ &= \frac{1}{B(a,b)} \left(\frac{\gamma}{\gamma+1}\right)^{a-1} \left(\frac{1}{\gamma+1}\right)^{b+1}. \end{aligned} \tag{8.1}$$

Since $\pi \sim \mathrm{Be}(a,b)$ *implies* $1-\pi \sim \mathrm{Be}(b,a)$, $(1-\pi)/\pi$ *also has a density of the form* (8.1) *with the roles of* a *and* b *interchanged. To obtain the density of* θ*, we use the following result:*

If X *and* Y *are two independent random variables with density* f_X *and* f_Y*, respectively, then the density of* $Z = X \cdot Y$ *is given by*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y\left(\frac{z}{x}\right) \frac{1}{|x|}\, dx.$$

Setting $\gamma_1 = \pi_1/(1-\pi_1)$ *and* $\gamma_2^* = (1-\pi_2)/\pi_2$*, we get*

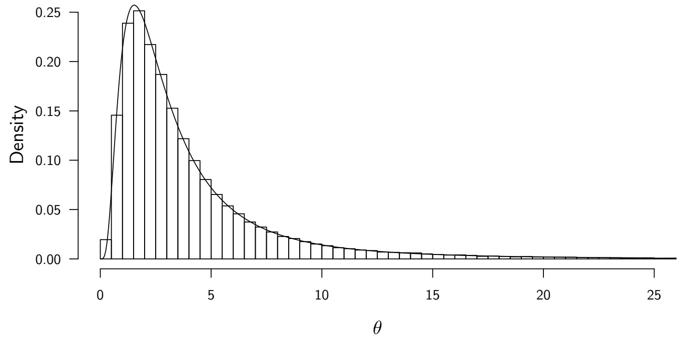
$$\begin{aligned} f_\theta(\theta) &= \int_{-\infty}^{\infty} f_{\gamma_1}(\gamma) f_{\gamma_2^*}(\gamma) \frac{1}{|\gamma|}\, d\gamma \\ &= \frac{1}{B(a_1,b_1)B(a_2,b_2)} \int_{-\infty}^{\infty} \left(\frac{\gamma}{\gamma+1}\right)^{a_1-1} \left(\frac{1}{\gamma+1}\right)^{b_1+1} \\ &\quad \cdot \left(\frac{\theta}{\gamma+\theta}\right)^{b_2-1} \left(\frac{\gamma}{\gamma+\theta}\right)^{a_2+1} \frac{1}{|\gamma|}\, d\gamma. \end{aligned}$$

where

$$\begin{aligned} a_1 &= x_1 + 0.5, & b_1 &= n_1 - x_1 + 0.5, \\ a_2 &= x_2 + 0.5, & b_2 &= n_2 - x_2 + 0.5. \end{aligned}$$

We use numerical integration to compute the above integral:

```
> ## density of theta=  pi_1/(1-pi_1)* (1-pi_2)/pi_2
> ## for pij ~ Be(a[j], b[j])
> thetaDens <- function(theta, a, b, log = FALSE)
  {
      logRet <- theta
      ## compute the value of the density function
      integrand <- function(gamma, theta)
      {
          ## use built-in distributions if possible
          logRet <-
            lbeta(a[1],b[1]+2) + lbeta(b[2],a[2]+2) -
            lbeta(a[1],b[1]) - lbeta(a[2],b[2]) +
            dbeta(gamma/(gamma+1), a[1], b[1]+2, log=TRUE) +
            dbeta(theta/(gamma+theta), b[2], a[2]+2, log=TRUE) -
            log(abs(gamma))
          exp(logRet)
      }
      for(i in seq_along(theta)){
          ## if the integration worked, save the result
          intRes <- integrate(integrand, lower = 0, upper = 1,
                              theta = theta[i],
                              stop.on.error = FALSE, rel.tol = 1e-6,
                              subdivisions = 200)
          if(intRes$message == "OK")
              logRet[i] <- log(intRes$value)
          else
              logRet[i] <- NA
      }
      ## return the vector of results
      if(log)
          return(logRet)
      else
          return(exp(logRet))
  }
> ## test the function using the simulated data:
> histRes <- hist(theta, prob = TRUE, xlim=c(0,25), breaks=1000,
                  main = "", xlab = expression(theta))
> thetaGrid <- seq(from = 0, to = 25,
                   length = 501)
> lines(thetaGrid, thetaDens(thetaGrid, 0.5 + x, 0.5 + n - x))
```



The log odds ratio ψ is the difference of the two independent log odds ϕ_i, $i = 1, 2$:

$$\psi = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = \operatorname{logit}(\pi_1) - \operatorname{logit}(\pi_2) = \phi_1 - \phi_2.$$

To compute the density of ψ, we therefore first compute the density of $\phi = g(\pi) = \operatorname{logit}(\pi)$ assuming $\pi \sim \operatorname{Be}(a, b)$. Since

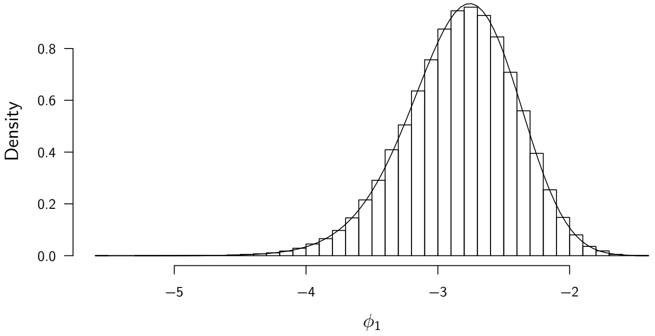
$$g^{-1}(\phi) = \frac{\exp(\phi)}{1+\exp(\phi)} \quad \text{and} \quad \frac{d}{d\phi} g^{-1}(\phi) = g^{-1}(\phi)\bigl(1 - g^{-1}(\phi)\bigr),$$

applying the change-of-variables formula gives

$$f(\phi) = \frac{1}{B(a,b)} \frac{\exp(\phi)^a}{\bigl(1+\exp(\phi)\bigr)^{a+b}}.$$

We can verify this result for ϕ_1 by using the simulated random numbers from part (a):

```
> ## density of phi = logit(pi) for pi ~ Be(a, b)
> phiDens <- function(phi, a, b)
  {
      pi <- plogis(phi)
      logRet <- a * log(pi) + b * log(1 - pi) - lbeta(a, b)
      return(exp(logRet))
  }
> ## simulated histogram
> histRes <- hist(qlogis(pi1), prob = TRUE, breaks = 50,
                  xlab = expression(phi[1]), main = "")
> ## analytic density function
> phiGrid <- seq(from = min(histRes$breaks), to = max(histRes$breaks),
                 length = 101)
> lines(phiGrid, phiDens(phiGrid, 0.5 + x[1], 0.5 + n[1] - x[1]))
```



Now let $\phi_i = \operatorname{logit}(\pi_i)$ for $\pi_i \sim \operatorname{Be}(a_i, b_i)$. As ϕ_1 and ϕ_2 are independent, the density of $\psi = \phi_1 - \phi_2$ can be calculated by applying the convolution theorem:

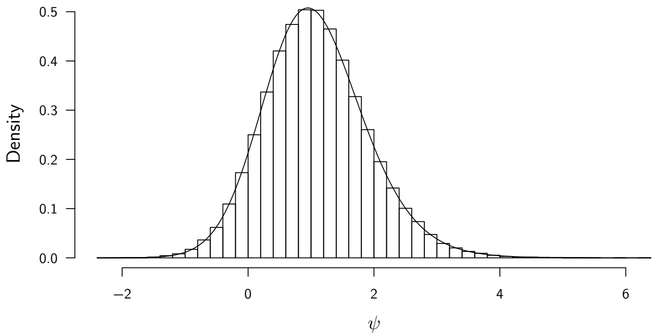
$$\begin{aligned} f_\psi(\psi) &= \int f_1(\psi + \phi_2) f_2(\phi_2)\, d\phi_2 \\ &= \frac{1}{B(a_1,b_1)B(a_2,b_2)} \int_{-\infty}^{\infty} \frac{\exp(\psi+\phi_2)^{a_1} \exp(\phi_2)^{a_2}}{\bigl(1+\exp(\psi+\phi_2)\bigr)^{a_1+b_1}\bigl(1+\exp(\phi_2)\bigr)^{a_2+b_2}}\, d\phi_2. \end{aligned}$$

By substituting $\pi = g^{-1}(\phi_2)$, the above integral can be expressed as

$$\int_0^1 g^{-1}\bigl(g(\pi)+\psi\bigr)^{a_1} \left\{1 - g^{-1}\bigl(g(\pi)+\psi\bigr)\right\}^{b_1} \pi^{a_2-1}(1-\pi)^{b_2-1}\, d\pi,$$

which cannot be simplified further. Thus, the density of ψ has to be computed by numerical integration:

```
> ## density of psi = logit(pi1) - logit(pi2) for pij ~ Be(a[j], b[j])
> psiDens <- function(psi, a, b, log = FALSE)
  {
      logRet <- psi
      ## compute the value of the density function
      integrand <- function(pi, psi)
      {
          ## use built-in distributions if possible
          logRet <-
            a[1] * plogis(qlogis(pi) + psi, log.p = TRUE) +
            b[1] * plogis(qlogis(pi) + psi, lower = FALSE, log.p = TRUE) -
            lbeta(a[1], b[1]) +
            dbeta(pi, a[2], b[2], log = TRUE)
          exp(logRet)
      }
      for(i in seq_along(psi)){
          ## if the integration worked, save the result
          intRes <- integrate(integrand, lower = 0, upper = 1, psi = psi[i],
                              stop.on.error = FALSE, rel.tol = 1e-6,
                              subdivisions = 200)
          if(intRes$message == "OK")
              logRet[i] <- log(intRes$value)
          else
              logRet[i] <- NA
      }
      ## return the vector of results
      if(log)
          return(logRet)
      else
          return(exp(logRet))
  }
> ## test the function using the simulated data:
> histRes <- hist(psi, prob = TRUE, breaks = 50,
                  main = "", xlab = expression(psi))
> psiGrid <- seq(from = min(histRes$breaks), to = max(histRes$breaks),
                 length = 201)
> lines(psiGrid, psiDens(psiGrid, 0.5 + x, 0.5 + n - x))
```



The problem here was at first that the integration routine did not converge in the range of $\psi \approx 4.7$ for the default settings. This problem could be solved by changing the relative tolerance (`rel.tol`) and the number of subdivisions (`subdivisions`) to different values. To compute the posterior expectation of ψ, we use numerical integration again. Note that we need to be able to calculate the density $f(\psi \mid x)$ for each value of ψ to do so!

```
> ## numerical integration
> integrate(function(theta) theta * thetaDens(theta, 0.5 + x, 0.5 + n - x),
            lower = -Inf, upper = Inf)
4.333333 with absolute error < 1.1e-06
> integrate(function(psi) psi * psiDens(psi, 0.5 + x, 0.5 + n - x),
            lower = -Inf, upper = Inf)
1.079639 with absolute error < 0.00012
> ## Monte-Carlo estimate was:
> mean(theta)
[1] 4.338513
> mean(psi)
[1] 1.082216
```

Thus, the Monte Carlo estimate is close to the value obtained by numerical integration for θ and relatively close for ψ. We now turn to the numerical calculation of HPD intervals. We write separate functions for θ and ϕ, which are tailored to the corresponding histograms:

```
> ## HPD interval computation for theta:
> ## for given h, the function outerdensTheta returns
> ## the probability mass of all theta values having density smaller than h
> ## (a, b are as for the function thetaDens)
> outerdensTheta <- function(h)
  {
      ## only valid for this data!
      mode <- 2                           # estimated from graph
      a <- 0.5 + x
      b <- 0.5 + n - x
      ## find the points of intersection of h and the density
      intersec.left <- uniroot(function(x) thetaDens(x, a, b) - h,
                               interval = c(-2, mode))$root
      intersec.right <- uniroot(function(x) thetaDens(x, a, b) - h,
                                interval = c(mode, 100))$root
      ## probability mass outside of the points of intersection
      p.left <- integrate(thetaDens, lower = -Inf, upper = intersec.left,
                          a = a, b = b)$value
      p.right <- integrate(thetaDens, lower = intersec.right, upper = Inf,
                           a = a, b = b)$value
      ## return this probability and the points of intersection
      return(c(prob = p.left + p.right,
               lower = intersec.left, upper = intersec.right))
  }
> ## determine the optimal h: want to have 5% of probability mass outside
> result <- uniroot(function(h) outerdensTheta(h)[1] - 0.05,
                    interval = c(0.001, 0.2))
> height <- result[["root"]]
> ## this numerical optimisation gives
> (thetaHpdNum <- outerdensTheta(height)[c(2,3)])
    lower      upper
0.2448303 12.2413591
> ## the Monte Carlo estimate was:
> thetaHpd
    lower      upper
0.2461746 12.2797998
```

```
> ## HPD interval computation for psi:
> ## for given h, the function outerdensPsi returns
> ## the probability mass of all psi values having density smaller than h
> ## (a, b are as for the function psiDens)
> outerdensPsi <- function(h)
  {
      ## only valid for this data!
      mode <- 1                            # estimated from graph
      a <- 0.5 + x
      b <- 0.5 + n - x
      ## find the points of intersection of h and the density
      intersec.left <- uniroot(function(x) psiDens(x, a, b) - h,
                               interval = c(-2, mode))$root
      intersec.right <- uniroot(function(x) psiDens(x, a, b) - h,
                                interval = c(mode, 100))$root
      ## probability mass outside of the points of intersection
      p.left <- integrate(psiDens, lower = -Inf, upper = intersec.left,
                          a = a, b = b)$value
      p.right <- integrate(psiDens, lower = intersec.right, upper = Inf,
                           a = a, b = b)$value
      ## return this probability and the points of intersection
      return(c(prob = p.left + p.right,
               lower = intersec.left, upper = intersec.right))
  }
> ## determine the optimal h: want to have 5% of probability mass outside
> result <- uniroot(function(h) outerdensPsi(h)[1] - 0.05,
                    interval = c(0.001, 0.4))
> height <- result[["root"]]
> ## this numerical optimisation gives
> (psiHpdNum <- outerdensPsi(height)[c(2,3)])
     lower      upper
-0.4898317  2.7333078
```

```
> ## the Monte Carlo estimate was:
> psiHpd
     lower      upper
-0.5111745  2.7162842
```

The Monte Carlo estimates of the HPD intervals are also close to the HPD intervals obtained by numerical methods. The above calculations illustrate that Monte Carlo estimation is considerably easier (e. g. we do not have to calculate any densities!) than the corresponding numerical methods and does not require any tuning of integration routines as the numerical methods do.

4. In this exercise we will estimate a Bayesian hierarchical model with MCMC methods. Consider Example 6.31, where we had the following model:

$$\begin{aligned} \hat{\psi}_i \,|\, \psi_i &\sim \mathrm{N}\left(\psi_i,\, \sigma_i^2\right), \\ \psi_i \,|\, \nu, \tau &\sim \mathrm{N}\left(\nu,\, \tau^2\right), \end{aligned}$$

where we assume that the empirical log odds ratios $\hat{\psi}_i$ and corresponding variances $\sigma_i^2 := 1/a_i + 1/b_i + 1/c_i + 1/d_i$ are known for all studies $i = 1, \dots, n$. Instead of empirical Bayes estimation of the hyper-parameters ν and τ^2, we here proceed in a fully Bayesian way by assuming hyper-priors for them. We choose $\nu \sim \mathrm{N}(0, 10)$ and $\tau^2 \sim \mathrm{IG}(1, 1)$.

a) Derive the full conditional distributions of the unknown parameters $\psi_1, \dots, \psi_n$, ν and τ^2.

▶ *Let $i \in \{1, \dots, n = 9\}$. The conditional density of ψ_i given all other parameters and the data $\mathcal{D}$ (which can be reduced to the empirical log odds ratios $\{\hat{\psi}_i\}$ and the corresponding variances $\{\sigma_i^2\}$) is*

$$\begin{aligned} f(\psi_i \,|\, \{\psi_j\}_{j \neq i}, \nu, \tau^2, \mathcal{D}) &\propto f(\boldsymbol{\psi}, \nu, \tau^2, \mathcal{D}) \\ &\propto f(\hat{\psi}_i \,|\, \psi_i) f(\psi_i \,|\, \nu, \tau^2). \end{aligned}$$

This corresponds to the normal model in Example 6.8: μ is replaced by ψ_i here, σ^2 by σ_i^2 and x by $\hat{\psi}_i$. We can thus use Equation (6.16) to obtain the full conditional distribution

$$\psi_i \,|\, \{\psi_j\}_{j \neq i}, \nu, \tau^2, \mathcal{D} \sim \mathrm{N}\left(\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} \left(\frac{\hat{\psi}_i}{\sigma_i^2} + \frac{\nu}{\tau^2}\right), \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

For the population mean ν, we have

$$f(\nu \,|\, \boldsymbol{\psi}, \tau^2, \mathcal{D}) \propto \prod_{i=1}^{n} f(\psi_i \,|\, \nu, \tau^2) \cdot f(\nu),$$

which corresponds to the normal model for a random sample. Using the last equation in Example 6.8 on page 182 with x_i replaced by ψ_i and σ^2 by τ^2 yields the full conditional distribution

$$\nu \,|\, \boldsymbol{\psi}, \tau^2, \mathcal{D} \sim \mathrm{N}\left(\left(\frac{n}{\tau^2} + \frac{1}{10}\right)^{-1} \left(\frac{n\bar{\psi}}{\tau^2} + \frac{0}{10}\right), \left(\frac{n}{\tau^2} + \frac{1}{10}\right)^{-1}\right).$$

We further have

$$\begin{aligned} f(\tau^2 \,|\, \boldsymbol{\psi}, \nu, \mathcal{D}) &\propto \prod_{i=1}^{n} f(\psi_i \,|\, \nu, \tau^2) \cdot f(\tau^2) \\ &\propto \prod_{i=1}^{n} (\tau^2)^{-\frac{1}{2}} \exp\left(-\frac{(\psi_i - \nu)^2}{2} \Big/ \tau^2\right) \cdot (\tau^2)^{-(1+1)} \exp(-1/\tau^2) \\ &= (\tau^2)^{-(\frac{n+2}{2}+1)} \exp\left(-\frac{\sum_{i=1}^{n} (\psi_i - \nu)^2 + 2}{2} \Big/ \tau^2\right), \end{aligned}$$

that is

$$\tau^2 \,|\, \boldsymbol{\psi}, \nu, \mathcal{D} \sim \mathrm{IG}\left(\frac{n+2}{2}, \frac{\sum_{i=1}^{n} (\psi_i - \nu)^2 + 2}{2}\right).$$

b) Implement a Gibbs sampler to simulate from the corresponding posterior distributions.

▶ *In the following R code, we iteratively simulate from the full conditional distributions of ν, τ^2, $\psi_1, \dots, \psi_n$:*

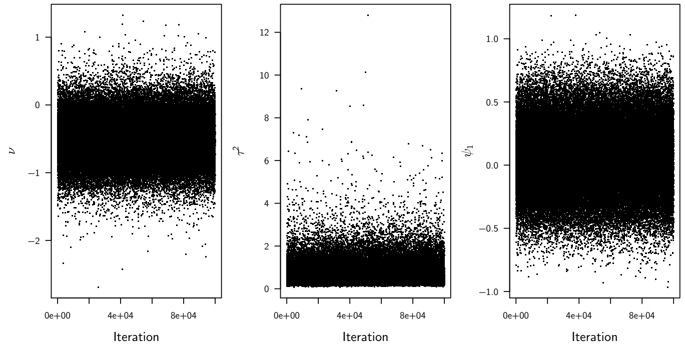
```
> ## the data is
> preeclampsia <- read.table ("../Daten/preeclampsia.txt", header = TRUE)
> preeclampsia
        Trial Diuretic Control Preeclampsia
1     Weseley       14      14          yes
2     Weseley      117     122           no
3     Flowers       21      17          yes
4     Flowers      364     117           no
5     Menzies       14      24          yes
6     Menzies       43      24           no
7      Fallis        6      18          yes
8      Fallis       32      22           no
9     Cuadros       12      35          yes
10    Cuadros      999     725           no
11 Landesman      138     175          yes
12 Landesman     1232    1161           no
13      Krans       15      20          yes
14      Krans      491     504           no
15    Tervila        6       2          yes
16    Tervila      102     101           no
17   Campbell       65      40          yes
18   Campbell       88      62           no
```

```
> ## functions for calculation of odds ratio and variance
> ## for a 2x2 table square
> oddsRatio <- function (square)
      (square[1,1] * square[2,2]) / (square[1,2] * square[2,1])
> variance <- function (square)
      sum (1 / square)
> ## extract the data we need
> groups <- split (subset (preeclampsia, select = c (Diuretic, Control)),
                   preeclampsia[["Trial"]])           # list of 2x2 tables
> (logOddsRatios <- log (sapply (groups, oddsRatio))) # psihat vector
   Campbell     Cuadros      Fallis     Flowers       Krans 
 0.13530539 -1.39102454 -1.47330574 -0.92367084 -0.26154993 
  Landesman     Menzies     Tervila     Weseley 
-0.29688945 -1.12214279  1.08875999  0.04184711 
> (variances <- sapply (groups, variance))            # sigma^2 vector
  Campbell    Cuadros     Fallis    Flowers      Krans 
0.06787728 0.11428507 0.29892677 0.11773684 0.12068745 
 Landesman    Menzies    Tervila    Weseley 
0.01463368 0.17801772 0.68637158 0.15960087 
> n <- length(groups)                                 # number of studies
```

```
> ## Gibbs sampler for inference in the fully Bayesian model:
> niter <- 1e+5
> s <- matrix(nrow = 2 + n, ncol = niter)
> rownames(s) <- c("nu", "tau2", paste("psi", 1:n, sep = ""))
> psiIndices <- 3:(n + 2)
> ## set initial values (other values in the domains are also possible)
> s[, 1] <- c(nu = mean(logOddsRatios),
              tau2 = var(logOddsRatios), logOddsRatios)
> set.seed(59)
> ## iteratively update the values
> for(j in 2:niter){
      ## nu first
      nuPrecision <- n / s["tau2",j-1] + 1 / 10

      psiSum <- sum(s[psiIndices,j-1])
      nuMean <- (psiSum / s["tau2",j-1]) / nuPrecision

      s["nu",j] <- rnorm(1, mean = nuMean, sd = 1 / sqrt(nuPrecision))

      ## then tau^2
      sumPsiNuSquared <- sum( (s[psiIndices,j-1] - s["nu",j])^2 )

      tau2a <- (n + 2) / 2
      tau2b <- (sumPsiNuSquared + 2) / 2

      s["tau2",j] <- 1 / rgamma(1, shape = tau2a, rate = tau2b)

      ## finally psi1, ..., psin
      for(i in 1:n){
          psiiPrecision <- 1 / variances[i] + 1 / s["tau2",j]
          psiiMean <- (logOddsRatios[i] / variances[i] + s["nu",j] /
                       s["tau2",j])/psiiPrecision

          s[psiIndices[i],j] <- rnorm(1, mean = psiiMean,
                                      sd = 1 / sqrt(psiiPrecision))
      }
  }
```

Having generated the Markov chain, one should check for convergence of the random numbers, e. g. in trace trace plots. For illustration, we generate such plots for the random variables ν, τ^2 und ψ_1:

```
> par(mfrow = c(1, 3))
> for(j in 1:3){
      plot(s[j, ], pch = ".",
           xlab = "Iteration", ylab = expression(nu, tau^2, psi[1])[j])
  }
```



The generated Markov chain seems to converge quickly so that a burn-in of 1000 *iterations seems sufficient.*

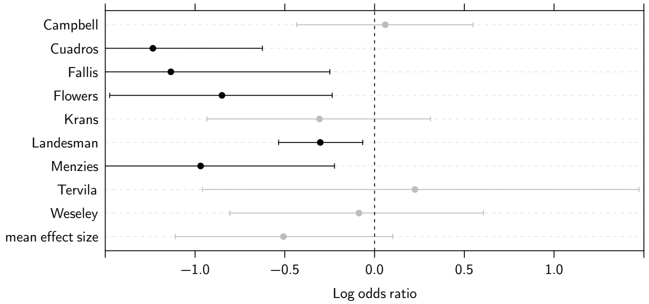
c) For the data given in Table 1.1, compute 95% credible intervals for $\psi_1, \ldots, \psi_n$ and ν. Produce a plot similar to Figure 6.15 and compare with the results from the empirical Bayes estimation.

▶ *We use the function* hpd *written in Exercise 3a) to calculate Monte Carlo estimates of 95% HPD intervals based on samples from the posterior distributions and produce a plot similar to Figure 6.15 as follows:*

```
> ## remove burn-in
> s <- s[, -(1:1000)]
> ## estimate the 95 % HPD credible intervals and the posterior expectations
> (mcmcHpds <- apply(s, 1, hpd))
              nu      tau2       psi1       psi2       psi3
lower -1.1110064 0.1420143 -0.4336897 -1.8715310 -2.0651776
upper  0.1010381 1.5214695  0.5484652 -0.6248588 -0.2488816
            psi4       psi5       psi6       psi7
lower -1.4753737 -0.9334114 -0.5345147 -1.7058202
upper -0.2362761  0.3118408 -0.0660448 -0.2232895
            psi8       psi9
lower -0.9592916 -0.8061968
upper  1.4732138  0.6063260
> (mcmcExpectations <- rowMeans(s))
```

```
         nu        tau2        psi1        psi2        psi3
-0.50788898  0.68177858  0.05943317 -1.23486599 -1.13470903
       psi4        psi5        psi6        psi7        psi8
-0.85002638 -0.30665961 -0.30249620 -0.96901680  0.22490430
       psi9
-0.08703254
> ## produce the plot
> library (lattice)
> ## define a panel function
> panel.ci <- function(
                       x,              # point estimate
                       y,              # height of the "bar"
                       lx,             # lower, lower limit of the interval
                       ux,             # upper, upper limit of the interval
                       subscripts,     # vector of indices
                       ...             # further graphics arguments
                       )
  {
      x <- as.numeric(x)
      y <- as.numeric(y)

      lx <- as.numeric(lx[subscripts])
      ux <- as.numeric(ux[subscripts])

      # normal dotplot for point estimates
      panel.dotplot(x, y, lty = 2, ...)
      # draw intervals
      panel.arrows(lx, y, ux, y,
                   length = 0.1, unit = "native",
                   angle = 90,             # deviation from line
                   code = 3,               # left and right whisker
                   ...)
      # reference line
      panel.abline (v = 0, lty = 2)
  }
> ## labels:
> studyNames <- c (names(groups), "mean effect size")
> studyNames <- ordered (studyNames, levels = rev(studyNames))
>     # levels important for order!
> indices <- c(psiIndices, nu = 1)
> ## collect the data in a dataframe
> ciData <- data.frame (low = mcmcHpds["lower", indices],
                        up = mcmcHpds["upper", indices],
                        mid = mcmcExpectations[indices],
                        names = studyNames
                        )
> ciData[["signif"]] <- with (ciData,
                              up < 0 | low > 0)
> ciData[["color"]] <- with (ciData,
                             ifelse (signif, "black", "gray"))
> randomEffectsCiPlot <- with (ciData,
                               dotplot (names ~ mid,
                                        panel = panel.ci,
                                        lx = low, ux = up,
                                        pch = 19, col = color,
                                        xlim = c (-1.5, 1.5),
                                        xlab = "Log odds ratio",
```

```
                                        scales = list (cex = 1)
                                        )
                               )
> print (randomEffectsCiPlot)
```



Compared to the results of the empirical Bayes analysis in Example 6.31, the credible intervals are wider in this fully Bayesian analysis. The point estimate of the mean effect ν is $\mathsf{E}(\nu \,|\, \mathcal{D}) = -0.508$ here, which is similar to the result $\hat{\nu}_{\text{ML}} = -0.52$ obtained with empirical Bayes. However, the credible interval for ν includes zero here, which is not the case for the empirical Bayes result. In addition, shrinkage of the Bayesian point estimates for the single studies towards the mean effect ν is less pronounced here, see for example the Tervila study.

5. Let X_i, $i = 1, \dots, n$ denote a random sample from a $\text{Po}(\lambda)$ distribution with gamma prior $\lambda \sim \text{G}(\alpha, \beta)$ for the mean λ.

a) Derive closed forms of $\mathsf{E}(\lambda \,|\, x_{1:n})$ and $\text{Var}(\lambda \,|\, x_{1:n})$ by computing the posterior distribution of $\lambda \,|\, x_{1:n}$.

▶ *We have*

$$\begin{aligned} f(\lambda \,|\, x_{1:n}) &= f(x_{1:n} \,|\, \lambda) f(\lambda) \\ &\propto \prod_{i=1}^{n} \left(\lambda^{x_i} \exp(-\lambda)\right) \lambda^{\alpha-1} \exp(-\beta\lambda) \\ &= \lambda^{\alpha + \sum_{i=1}^{n} x_i - 1} \exp(-(\beta + n)\lambda), \end{aligned}$$

that is $\lambda \,|\, x_{1:n} \sim \text{G}(\alpha + n\bar{x}, \beta + n)$. *Consequently,*

$$\begin{aligned} \mathsf{E}(\lambda \,|\, x_{1:n}) &= \frac{\alpha + n\bar{x}}{\beta + n} \qquad \textit{and} \\ \text{Var}(\lambda \,|\, x_{1:n}) &= \frac{\alpha + n\bar{x}}{(\beta + n)^2}. \end{aligned}$$

b) Approximate $\mathsf{E}(\lambda \mid x_{1:n})$ and $\operatorname{Var}(\lambda \mid x_{1:n})$ by exploiting the asymptotic normality of the posterior (*cf*. Section 6.6.2).

▶ *We use the following result from Section 6.6.2:*

$$\lambda \mid x_{1:n} \overset{a}{\sim} \mathrm{N}\big(\hat{\lambda}_n, I(\hat{\lambda}_n)^{-1}\big),$$

where $\hat{\lambda}_n$ denotes the MLE and $I(\hat{\lambda}_n)^{-1}$ the inverse observed Fisher information. We now determine these two quantities for the Poisson likelihood: The log-likelihood is

$$l(x_{1:n} \mid \lambda) = \sum_{i=1}^{n} x_i \log(\lambda) - n\lambda,$$

which yields the MLE $\hat{\lambda}_n = (\sum_{i=1}^n x_i)/n = \bar{x}$. We further have

$$I(\lambda) = -\frac{d^2 l(x_{1:n} \mid \lambda)}{d\lambda^2} = \frac{\sum_{i=1}^n x_i}{\lambda^2} = \frac{n\bar{x}}{\lambda^2}$$

and thus

$$I(\hat{\lambda}_n)^{-1} = \frac{\bar{x}^2}{n\bar{x}} = \frac{\bar{x}}{n}.$$

Consequently,

$$\mathsf{E}(\lambda \mid x_{1:n}) \approx \hat{\lambda}_n = \bar{x} \qquad \textit{and}$$
$$\operatorname{Var}(\lambda \mid x_{1:n}) \approx I(\hat{\lambda}_n)^{-1} = \frac{\bar{x}}{n}.$$

c) Consider now the log mean $\theta = \log(\lambda)$. Use the change of variables formula (A.11) to compute the posterior density $f(\theta \mid x_{1:n})$.

▶ *From part (a), we know that the posterior density of λ is*

$$f(\lambda \mid x_{1:n}) = \frac{(\beta+n)^{\alpha+n\bar{x}}}{\Gamma(\alpha+n\bar{x})} \lambda^{\alpha+n\bar{x}-1} \exp(-(\beta+n)\lambda).$$

Applying the change of variables formula with transformation function $g(y) = \log(y)$ gives

$$\begin{aligned} f(\theta \mid x_{1:n}) &= \frac{(\beta+n)^{\alpha+n\bar{x}}}{\Gamma(\alpha+n\bar{x})} \exp(\theta)^{\alpha+n\bar{x}-1} \exp(-(\beta+n)\exp(\theta))\exp(\theta) \\ &= \frac{(\beta+n)^{\alpha+n\bar{x}}}{\Gamma(\alpha+n\bar{x})} \exp(\theta)^{\alpha+n\bar{x}} \exp(-(\beta+n)\exp(\theta)). \end{aligned}$$

d) Let $\alpha = 1, \beta = 1$ and assume that $\bar{x} = 9.9$ has been obtained for $n = 10$ observations from the model. Compute approximate values of $\mathsf{E}(\theta \mid x_{1:n})$ and $\operatorname{Var}(\theta \mid x_{1:n})$ via:

i. the asymptotic normality of the posterior,

ii. numerical integration (*cf*. Appendix C.2.1),

iii. and Monte Carlo integration.

▶

i. *Analogously to part (b), we have*

$$\theta \mid x_{1:n} \overset{a}{\sim} \mathrm{N}\big(\hat{\theta}_n, I(\hat{\theta}_n)^{-1}\big),$$

where $\hat{\theta}_n$ denotes the MLE and $I(\hat{\theta}_n)^{-1}$ the inverse observed Fisher information. We exploit the invariance of the MLE with respect to one-to-one transformations to obtain

$$\mathsf{E}(\theta \mid x_{1:n}) \approx \hat{\theta}_n = \log(\hat{\lambda}_n) = \log(\bar{x}) = 2.2925.$$

To transform the observed Fisher information obtained in part (b), we apply Result 2.1:

$$\operatorname{Var}(\theta \mid x_{1:n}) \approx I(\hat{\theta}_n)^{-1} = I(\hat{\lambda}_n)^{-1}\left(\frac{d\exp(\hat{\theta}_n)}{d\theta}\right)^{-2} = \frac{1}{n\bar{x}} = 0.0101.$$

ii. *For the numerical integration, we work with the density of $\lambda = \exp(\theta)$ instead of θ to avoid numerical problems. We thus compute the posterior expectation and variance of $\log(\lambda)$. We use the* `R` *function* `integrate` *to compute the integrals numerically:*

```
> ## given data
> alpha <- beta <- 1   ## parameters of the prior distribution
> n <- 10   ## number of observed values
> xbar <- 9.9   ## mean of observed values
> ##
> ## function for numerical computation of
> ## posterior expectation and variance of theta=log(lambda)
> numInt <- function(alpha, beta, n, xbar)
  {
  ## posterior density of lambda
  lambdaDens <- function(lambda, alpha, beta, n, xbar, log=FALSE)
  {
    ## parameters of the posterior gamma density
    alphapost <- alpha + n*xbar
    betapost <- beta + n
    logRet <- dgamma(lambda, alphapost, betapost, log=TRUE)
    if(log)
          return(logRet)
     else
          return(exp(logRet))
  }
     # integrand for computation of posterior expectation
     integrand.mean <- function(lambda)
      {
           log(lambda) *
           lambdaDens(lambda, alpha, beta, n, xbar)
      }
```

```
        # numerical integration to get posterior expectation
        res.mean <- integrate(integrand.mean, lower = 0, upper = Inf,
                              stop.on.error = FALSE,
                              rel.tol = sqrt(.Machine$double.eps))
        if(res.mean$message == "OK")
              mean <- res.mean$value
        else
              mean <- NA

        # numerical computation of variance
        integrand.square <- function(lambda)
        {   (log(lambda))^2 *
                 lambdaDens(lambda, alpha, beta, n, xbar)
        }

        res.square <- integrate(integrand.square, lower = 0, upper = Inf,
                                stop.on.error = FALSE,
                                rel.tol = sqrt(.Machine$double.eps))
        if(res.square$message == "OK")
            var <- res.square$value - mean^2
        else
            var <- NA
        return(c(mean=mean,var=var))
    }
> # numerical approximation for posterior mean and variance of theta
> numInt(alpha, beta, n, xbar)
      mean        var
2.20226658 0.01005017
```

iii. *To obtain a random sample from the distribution of $\theta = \log(\lambda)$, we first generate a random sample from the distribution of λ - which is a Gamma distribution - and then transform this sample:*

```
> ## Monte-Carlo integration
> M <- 10000
> ## parameters of posterior distribution
> alphapost <- alpha + n*xbar
> betapost <- beta + n
> ## sample for lambda
> lambdaSam <- rgamma(M,alphapost,betapost)
> ## sample for theta
> thetaSam <- log(lambdaSam)
> # conditional expectation of theta
> (Etheta <- mean(thetaSam))
[1] 2.201973
> ## Monte Carlo standard error
> (se.Etheta <- sqrt(var(thetaSam)/M))
[1] 0.001001739
> EthetaSq <- mean(thetaSam^2)
> # conditional variance of theta
> (VarTheta <- EthetaSq - Etheta^2)
[1] 0.0100338
```

The estimates obtained by numerical integration and Monte Carlo integration are similar. The estimate of the posterior mean obtained from asymtotic normality is larger than the other two estimates. This difference is not surprising as the sample size $n = 10$ is quite small so that an asymtotic approximation may be inaccurate.

6. Consider the genetic linkage model from Exercise 5 in Chapter 2. Here we assume a uniform prior on the proportion ϕ, *i. e.* $\phi \sim \mathrm{U}(0, 1)$. We would like to compute the posterior mean $\mathsf{E}(\phi \mid \boldsymbol{x})$.

a) Construct a rejection sampling algorithm to simulate from $f(\phi \mid \boldsymbol{x})$ using the prior density as the proposal density.

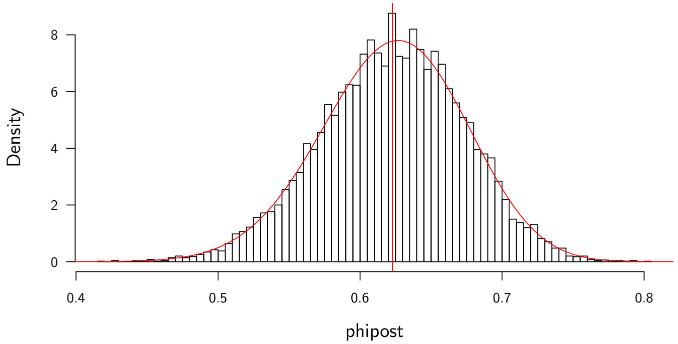
▶

```
> ## define the log-likelihood function (up to multiplicative constants)
> ## Comment: on log-scale the numerical calculations are more robust
> loglik <- function(phi, x)
  {
    loglik <- x[1]*log(2+phi)+(x[2]+x[3])*log(1-phi)+x[4]*log(phi)
    return(loglik)
  }
> ## rejection sampler (M: number of samples, x: data vector)
> rej <- function(M, x)
  {
    post.mode <- optimize(loglik, x=x, lower = 0, upper = 1,
                          maximum = TRUE)

    ## determine constant a to be ordinate at the mode
    ## a represents the number of trials up to the first success
    (a <- post.mode$objective)

    ## empty vector of length M
    phi <- double(M)

    ## counter to get M samples
    N <- 1
    while(N <=M)
    {
      while (TRUE)
      {
        ## value from uniform distribution
        u <- runif(1)
        ## proposal for phi
        z <- runif(1)
        ## check for acceptance
        ## exit the loop after acceptance
        if (u <= exp(loglik(phi=z, x=x)-a))
              break
      }
      ## save the proposed value
      phi[N] <- z
      ## go for the next one
      N <- N+1
    }
    return(phi)
  }
```

b) Estimate the posterior mean of ϕ by Monte Carlo integration using $M = 10\,000$ samples from $f(\phi \mid \boldsymbol{x})$. Calculate also the Monte Carlo standard error.

▶

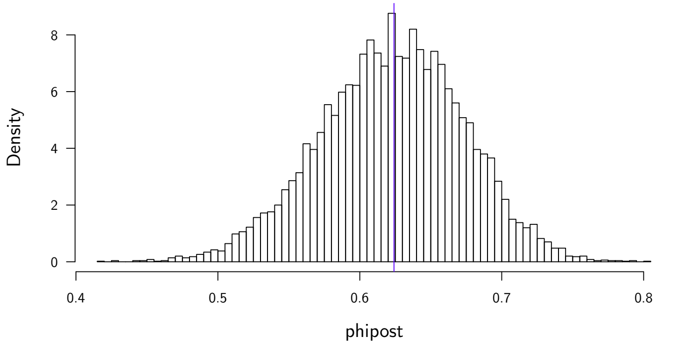
```
> ## load the data
> x <- c(125, 18, 20, 34)
> ## draw values from the posterior
> set.seed(2012)
> M <- 10000
> phipost <- rej(M=M, x=x)
> ## posterior mean by Monte Carlo integration
> (Epost <- mean(phipost))
[1] 0.6227486
> ## compute the Monte Carlo standard error
> (se.Epost <- sqrt(var(phipost)/M))
[1] 0.0005057188
> ## check the posterior mean using numerical integration:
> numerator   <- integrate(function(phi) phi * exp(loglik(phi, x)),
          lower=0,upper=1)$val
> denominator <- integrate(function(phi) exp(loglik(phi, x)),
          lower=0,upper=1)$val
> numerator/denominator
[1] 0.6228061
> ## draw histogram of sampled values
> hist(phipost, prob=TRUE, nclass=100, main=NULL)
> ## compare with density
> phi.grid <- seq(0,1,length=1000)
> dpost <- function(phi) exp(loglik(phi, x)) / denominator
> lines(phi.grid,dpost(phi.grid),col=2)
> abline(v=Epost, col="red")
```



c) In 6b) we obtained samples of the posterior distribution assuming a uniform prior on ϕ. Suppose we now assume a Be(0.5, 0.5) prior instead of the previous U(0, 1) = Be(1, 1). Use the importance sampling weights to estimate the posterior mean and Monte Carlo standard error under the new prior based on the old samples from 6b).

▶

```
> ## Importance sampling -- try a new prior, the beta(0.5,0.5)
> ## posterior density ratio = prior density ratio!
> weights <- dbeta(phipost, .5, .5)/dbeta(phipost, 1, 1)
> (Epost2 <- sum(phipost*weights)/sum(weights))
[1] 0.6240971
> ## or simpler
> (Epost2 <- weighted.mean(phipost, w=weights))
[1] 0.6240971
> ##
> (se.Epost2 <- 1/sum(weights)*sum((phipost - Epost2)^2*weights^2))
[1] 0.001721748
> hist(phipost, prob=TRUE, nclass=100, main=NULL)
> abline(v=Epost2, col="blue")
```



7. As in Exercise 6, we consider the genetic linkage model from Exercise 5 in Chapter 2. Now, we would like to sample from the posterior distribution of ϕ using MCMC. Using the Metropolis-Hastings algorithm, an arbitrary proposal distribution can be used and the algorithm will always converge to the target distribution. However, the time until convergence and the degree of dependence between the samples depends on the chosen proposal distribution.

a) To sample from the posterior distribution, construct an MCMC sampler based on the following normal independence proposal (*cf*. approximation 6.6.2 in Section 6.6.2):

$$\phi^* \sim \mathrm{N}\big(\mathrm{Mod}(\phi \,|\, \boldsymbol{x}), F^2 \cdot C^{-1}\big),$$

where $\mathrm{Mod}(\phi \,|\, \boldsymbol{x})$ denotes the posterior mode, C the negative curvature of the log-posterior at the mode and F a factor to blow up the variance.

▶

```
> # define the log-likelihood function
> log.lik <- function(phi, x)
  {
    if((phi<1)&(phi>0)) {
      loglik <- x[1]*log(2+phi)+(x[2]+x[3])*log(1-phi)+x[4]*log(phi)
    }
    else {
      # if phi is not in the defined range return NA
      loglik <- NA
    }
    return(loglik)
  }
> # MCMC function with independence proposal
> # M: number of samples, x: data vector, factor: factor to blow up
> # the variance
> mcmc_indep <- function(M, x, factor)
  {
    # store samples here
    xsamples <- rep(NA, M)

    # Idea: Normal Independence proposal with mean equal to the posterior
    # mode and standard deviation equal to the standard error or to
    # a multiple of the standard error.
    mymean <- optimize(log.lik, x=x, lower = 0, upper = 1,
                       maximum = TRUE)$maximum

    # negative curvature of the log-posterior at the mode
    a <- -1*(-x[1]/(2+mymean)^2 - (x[2]+x[3])/(1-mymean)^2 - x[4]/mymean^2)
    mystd <- sqrt(1/a)

    ##################################################################
    ## Alternative using optim instead of optimize.
    ## Optim returns directly Hessian.
    #   eps <- 1E-9
    #   # if fnscale < 0 the maximum is comupted
    #   mycontrol <- list(fnscale=-1, maxit=100)
    #
    #   ml <- optim(0.5, log.lik, x=data, control=mycontrol, hessian=T,
    #               method="L-BFGS-B", lower=0+eps, upper=1-eps)
    #   mymean <- ml$par
    #   mystd <- sqrt(-1/ml$hessian)
    ##################################################################

    # count number of accepted and rejected values
    yes <- 0
    no <- 0

    # use as initial starting value mymean
    xsamples[1] <- mymean

    # Metropolis-Hastings iteration
    for(k in 2:M){
      # value of the past iteration
      old <- xsamples[k-1]

      # propose new value
      # factor fac blows up standard deviation
      proposal <- rnorm(1, mean=mymean, sd=mystd*factor)

      # compute acceptance ratio
      # under uniform proir: posterior ratio = likelihood ratio
      posterior.ratio <-  exp(log.lik(proposal, data)-log.lik(old, data))
      if(is.na(posterior.ratio)){
        # happens when the proposal is not between 0 and 1
        # => acceptance probability will be 0
        posterior.ratio <- 0
      }
      proposal.ratio <- exp(dnorm(old, mymean, mystd*factor, log=T) -
                            dnorm(proposal, mymean, mystd*factor, log=T))

      # get the acceptance probability
      alpha <- posterior.ratio*proposal.ratio

      # accept-reject step
      if(runif(1) <= alpha){
        # accept the proposed value
        xsamples[k] <- proposal
        # increase counter of accepted values
        yes <- yes + 1
      }
      else{
        # stay with the old value
        xsamples[k] <- old
        no <- no + 1
      }
    }

    # acceptance rate
    cat("The acceptance rate is: ", round(yes/(yes+no)*100,2), "%\n", sep="")
    return(xsamples)
  }
```

b) Construct an MCMC sampler based on the following random walk proposal:

$$\phi^* \sim \mathrm{U}(\phi^{(m)} - d, \phi^{(m)} + d),$$

where $\phi^{(m)}$ denotes the current state of the Markov chain and d is a constant.

▶

```
> # MCMC function with random walk proposal
> # M: number of samples, x: data vector, fac: factor to blow up
> # the variance
> mcmc_rw <- function(M, x, d){

    # store samples here
    xsamples <- rep(NA, M)

    # count number of accepted and rejected values
    yes <- 0
    no <- 0

    # specify a starting value
    xsamples[1] <- 0.5

    # Metropolis-Hastings iteration
    for(k in 2:M){
      # value of the past iteration
      old <- xsamples[k-1]

      # propose new value
      # use a random walk proposal: U(phi^(k) - d, phi^(k) + d)
      proposal <- runif(1, old-d, old+d)

      # compute acceptance ratio
      posterior.ratio <-  exp(log.lik(proposal, data)-log.lik(old, data))
      if(is.na(posterior.ratio)){
        # happens when the proposal is not between 0 and 1
        # => acceptance probability will be 0
        posterior.ratio <- 0
      }
      # the proposal ratio is equal to 1
      # as we have a symmetric proposal distribution
```

```
      proposal.ratio <- 1

      # get the acceptance probability
      alpha <- posterior.ratio*proposal.ratio

      # accept-reject step
      if(runif(1) <= alpha){
        # accept the proposed value
        xsamples[k] <- proposal
        # increase counter of accepted values
        yes <- yes + 1
      }
      else{
        # stay with the old value
        xsamples[k] <- old
        no <- no + 1
      }
    }

    # acceptance rate
    cat("The acceptance rate is: ", round(yes/(yes+no)*100,2),
        "%\n", sep="")
    return(xsamples)
  }
```

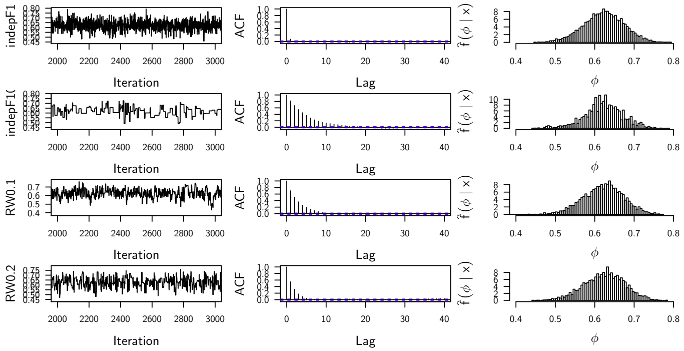
c) Generate $M = 10\,000$ samples from algorithm 7a), setting $F = 1$ and $F = 10$, and from algorithm 7b) with $d = 0.1$ and $d = 0.2$. To check the convergence of the Markov chain:

i. plot the generated samples to visually check the traces,
ii. plot the autocorrelation function using the R-function `acf`,
iii. generate a histogram of the samples,
iv. compare the acceptance rates.

What do you observe?

▶

```
> # data
> data <- c(125, 18, 20, 34)
> # number of iterations
> M <- 10000
> # get samples using Metropolis-Hastings with independent proposal
> indepF1 <- mcmc_indep(M=M, x=data, factor=1)
The acceptance rate is: 96.42%
> indepF10 <- mcmc_indep(M=M, x=data, factor=10)
The acceptance rate is: 12.22%
> # get samples using Metropolis-Hastings with random walk proposal
> RW0.1 <- mcmc_rw(M=M, x=data, d=0.1)
The acceptance rate is: 63.27%
> RW0.2 <- mcmc_rw(M=M, x=data, d=0.2)
The acceptance rate is: 40.33%
> ## some plots
> # independence proposal with F=1
> par(mfrow=c(4,3))
> # traceplot
> plot(indepF1, type="l", xlim=c(2000,3000), xlab="Iteration")
> # autocorrelation plot
> acf(indepF1)
> # histogram
> hist(indepF1, nclass=100, xlim=c(0.4,0.8), prob=T, xlab=expression(phi),
       ylab=expression(hat(f)*(phi ~"|"~ x)), main="")
> # ylab=expression(phi^{(k)})
> # independence proposal with F=10
> plot(indepF10, type="l", xlim=c(2000,3000), xlab="Iteration")
> acf(indepF10)
> hist(indepF10, nclass=100, xlim=c(0.4,0.8), prob=T, xlab=expression(phi),
       ylab=expression(hat(f)*(phi ~"|"~ x)), main="")
> # random walk proposal with d=0.1
> plot(RW0.1, type="l", xlim=c(2000,3000), xlab="Iteration")
> acf(RW0.1)
> hist(RW0.1, nclass=100, xlim=c(0.4,0.8), prob=T, xlab=expression(phi),
       ylab=expression(hat(f)*(phi ~"|"~ x)), main="")
> # random walk proposal with d=0.2
> plot(RW0.2, type="l", xlim=c(2000,3000), xlab="Iteration")
> acf(RW0.2)
> hist(RW0.2, nclass=100, xlim=c(0.4,0.8), prob=T, xlab=expression(phi),
       ylab=expression(hat(f)*(phi ~"|"~ x)), main="")
```



All four Markov chains converge quickly after a few hundred iterations. The independence proposal with the original variance (F=1) performs best: It produces uncorrelated samples and has a high acceptance rate. In contrast, the independence proposal with blown-up variance ($F = 10$) performs worst. It has a low acceptance rate and thus the Markov chain often gets stuck in the same value for several iterations, which leads to correlated samples. Regarding the random walk proposals, the one with the wider proposal distribution ($d = 0.2$) performs better since it yields less correlated samples and has a preferrable acceptance rate. (For random walk proposals, acceptance rates between 30% *and* 50% *are recommended.)*

8. Cole et al. (2012) describe a rejection sampling approach to sample from a posterior distribution as a simple and efficient alternative to MCMC. They summarise their approach as:

I. Define model with likelihood function $L(\theta; y)$ and prior $f(\theta)$.

II. Obtain the maximum likelihood estimate $\hat{\theta}_{\text{ML}}$.

III. To obtain a sample from the posterior:

i. Draw θ^* from the prior distribution (note: this must cover the range of the posterior).

ii. Compute the ratio $p = L(\theta^*; y)/L(\hat{\theta}_{\text{ML}}; y)$.

iii. Draw u from $\text{U}(0, 1)$.

iv. If $u \leq p$, then accept θ^*. Otherwise reject θ^* and repeat.

a) Using Bayes' rule, write out the posterior density of $f(\theta \,|\, y)$. In the notation of Section 8.3.3, what are the functions $f_X(\theta)$, $f_Z(\theta)$ and $L(\theta; x)$ in the Bayesian formulation?

▶ *By Bayes' rule, we have*

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

where $f(y) = \int f(y \,|\, \theta) f(\theta)\, d\theta$ is the marginal likelihood. The posterior density is the target, so $f(\theta \,|\, y) = f_Z(\theta)$, and the prior density is the proposal, so that $f(\theta) = f_X(\theta)$. As usual, the likelihood is $f(y \,|\, \theta) = L(\theta; y)$. Thus, we can rewrite the above equation as

$$f_X(\theta) = \frac{L(\theta; y)}{c'} f_Z(\theta) \tag{8.2}$$

with constant $c' = f(y)$.

b) Show that the acceptance probability $f_X(\theta^*)/\{a f_Z(\theta^*)\}$ is equal to $L(\theta^*; y)/L(\hat{\theta}_{\text{ML}}; y)$. What is a?

▶ *Let U denote a random variable with $U \sim \text{U}(0, 1)$. Then, the acceptance probability is*

$$\Pr(U \leq p) = p = L(\theta^*; y)/L(\hat{\theta}_{\text{ML}}; y)$$

as $p \in [0, 1]$. *Solving the equation*

$$\frac{f_X(\theta^*)}{a f_Z(\theta^*)} = \frac{f(\theta^* \,|\, y)}{a f(\theta^*)} = \frac{L(\theta^*; y)}{L(\hat{\theta}_{\mathrm{ML}}; y)}$$

for a *and applying Bayes' rule yields*

$$a = L(\hat{\theta}_{\mathrm{ML}}; y) \frac{f(\theta^* \,|\, y)}{L(\theta^*; y) f(\theta^*)} = \frac{L(\hat{\theta}_{\mathrm{ML}}; y)}{c'}. \tag{8.3}$$

Note that the constant $c' = f(y)$ *is not explicitly known.*

c) Explain why the inequality $f_X(\theta) \leq a f_Z(\theta)$ is guaranteed by the approach of Cole et al. (2012).

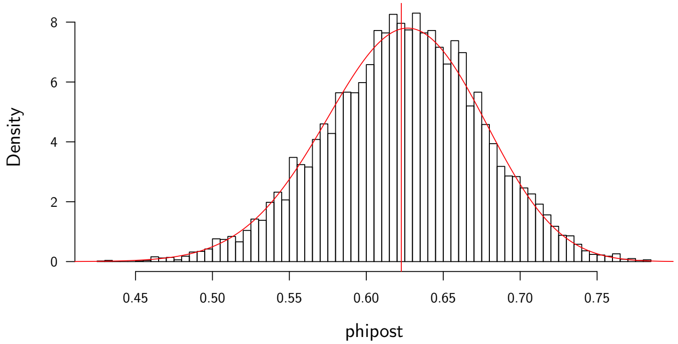
▶ *Since* $L(\theta; y) \leq L(\hat{\theta}_{\mathrm{ML}}; y)$ *for all* θ *by construction, the inequality follows easily by combining* (8.2) *and* (8.3). *Note that the expression for* a *given in* (8.3) *is the smallest constant* a *such that the sampling criterion* $f_X(\theta) \leq a f_Z(\theta)$ *is satisfied. This choice of* a *will result in more samples being accepted than for a larger* a.

d) In the model of Exercise 6c), use the proposed rejection sampling scheme to generate samples from the posterior of ϕ.

▶ *In Exercise 6c), we have produced a histogram of the posterior distribution of* ϕ. *We therefore know that the* $\mathrm{Be}(0.5, 0.5)$ *distribution, which has range* $[0, 1]$, *covers the range of the posterior distribution so that the condition in Cole et al.'s rejection sampling algorithm is satisfied.*

```
> # data
> x <- c(125,18,20,34)
> n <- sum(x)
> ## define the log-likelihood function (up to multiplicative constants)
> loglik <- function(phi, x)
  {
    loglik <- x[1]*log(2+phi)+(x[2]+x[3])*log(1-phi)+x[4]*log(phi)
    return(loglik)
  }
> ## rejection sampler (M: number of samples, x: data vector);
> ## approach by Cole et al.
> rejCole <- function(M, x)
  {
    # determine the MLE for phi
    mle <- optimize(loglik, x=x, lower = 0, upper = 1,
                    maximum = TRUE)$maximum

    ## empty vector of length M
    phi <- double(M)

    ## counter to get M samples
    N <- 1
    while(N <=M)
    {
      while (TRUE)
      {
```

```
        ## proposal from prior distribution
        z <- rbeta(1,0.5,0.5)
        ## compute relative likelihood of the proposal
        p <- exp(loglik(z,x) - loglik(mle,x))
        ## value from uniform distribution
        u <- runif(1)
        ## check for acceptance
        ## exit the loop after acceptance
        if (u <= p)
        break
      }
      ## save the proposed value
      phi[N] <- z
      ## go for the next one
      N <- N+1
    }
    return(phi)
  }
> ## draw histogram of sampled values
> phipost <- rejCole(10000,x)
> hist(phipost, prob=TRUE, nclass=100, main=NULL)
> ## compare with density from Ex. 6
> phi.grid <- seq(0,1,length=1000)
> dpost <- function(phi) exp(loglik(phi, x)) / denominator
> lines(phi.grid,dpost(phi.grid),col=2)
> abline(v=Epost, col="red")
```



9 Prediction

1. Five physicians participate in a study to evaluate the effect of a medication for migraine. Physician $i = 1, \dots, 5$ treats n_i patients with the new medication and it shows positive effects for y_i of the patients. Let π be the probability that an arbitrary migraine patient reacts positively to the medication. Given that

$$n = (3, 2, 4, 4, 3) \quad \text{and} \quad y = (2, 1, 4, 3, 3)$$

a) Provide an expression for the likelihood $L(\pi)$ for this study.

▶ *We make two assumptions:*

i. *The outcomes for different patients treated by the same physician are independent.*

ii. *The study results of different physicians are independent.*

By assumption (i), the y_i can be modelled as realisations of a binomial distribution:

$$Y_i \overset{iid}{\sim} \operatorname{Bin}(n_i, \pi), \quad i = 1, \dots, 5.$$

By assumption (ii), the likelihood of π is then

$$\begin{aligned} L(\pi) &= \prod_{i=1}^{5} \binom{n_i}{y_i} \pi^{y_i} (1-\pi)^{n_i - y_i} \\ &\propto \pi^{5\bar{y}} (1-\pi)^{5\bar{n} - 5\bar{y}}, \end{aligned}$$

where $\bar{y} = 1/5 \sum_{i=1}^{5} y_i$ is the mean number of successful treatments per physicians and $\bar{n} = 1/5 \sum_{i=1}^{5} n_i$ the mean number of patients treated per study.

b) Specify a conjugate prior distribution $f(\pi)$ for π and choose appropriate values for its parameters. Using these parameters derive the posterior distribution $f(\pi \mid n, y)$.

▶ *It is easy to see that the beta distribution* $\operatorname{Be}(\alpha, \beta)$ *with kernel*

$$f(\pi) \propto \pi^{\alpha - 1} (1-\pi)^{\beta - 1}$$

is conjugate with respect to the above likelihood (or see Example 6.7). We choose the non-informative Jeffreys' prior as prior for π, i. e. we choose $\alpha = \beta = 1/2$ (see Table 6.3). This gives the following posterior distribution for π:

$$\pi \mid n, y \sim \operatorname{Be}(5\bar{y} + 1/2, 5\bar{n} - 5\bar{y} + 1/2).$$

c) A sixth physician wants to participate in the study with $n_6 = 5$ patients. Determine the posterior predictive distribution for y_6 (the number of patients out of the five for which the medication will have a positive effect).

▶ *The density of the posterior predictive distribution is*

$$\begin{aligned} f(y_6 \mid n_6, y, n) &= \int_0^1 f(y_6 \mid \pi, n_6) f(\pi \mid y, n)\, d\pi \\ &= \int_0^1 \binom{n_6}{y_6} \pi^{y_6}(1-\pi)^{n_6-y_6} \\ &\quad \cdot \frac{1}{B(5\bar{y}+1/2, 5\bar{n}-5\bar{y}+1/2)} \pi^{5\bar{y}-1/2}(1-\pi)^{5\bar{n}-5\bar{y}-1/2}\, d\pi \\ &= \binom{n_6}{y_6} B(5\bar{y}+1/2, 5\bar{n}-5\bar{y}+1/2)^{-1} \\ &\quad \cdot \int_0^1 \pi^{5\bar{y}+y_6-1/2}(1-\pi)^{5\bar{n}+n_6-5\bar{y}-y_6-1/2}\, d\pi \qquad (9.1) \\ &= \binom{n_6}{y_6} \frac{B(5\bar{y}+y_6+1/2, 5\bar{n}+n_6-5\bar{y}-y_6+1/2)}{B(5\bar{y}+1/2, 5\bar{n}-5\bar{y}+1/2)}, \end{aligned}$$

where in (9.1)*, we have used that the integrand is the kernel of a* $\text{Be}(5\bar{y}+y_6+1/2, 5\bar{n}+n_6-5\bar{y}-y_6+1/2)$ *density. The obtained density is the density of a beta-binomial distribution (see Table A.1), more precisely*

$$y_6 \mid n_6, y, n \sim \text{BeB}(n_6, 5\bar{y}+1/2, 5\bar{n}-5\bar{y}+1/2).$$

Addition: Based on this posterior predictive distribution, we now compute a point prediction and a prognostic interval for the given data:

```
> ## given observations
> n <- c(3, 2, 4, 4, 3)
> y <- c(2, 1, 4, 3, 3)
> ## parameters of the beta-binomial posterior predictive distribution
> ## (under Jeffreys' prior)
> alphaStar <- sum(y) + 0.5
> betaStar <- sum(n - y) + 0.5
> nNew <- 5  ## number of patients treated by the additional physician
> ## point prediction: expectation of the post. pred. distr.
> (expectation <- nNew * alphaStar / (alphaStar + betaStar))
[1] 3.970588
> ## compute cumulative distribution function to get a prediction interval
> library(VGAM, warn.conflicts = FALSE)
> rbind(0:5,
        pbetabinom.ab(0:5, size = nNew, alphaStar, betaStar))
```

```
            [,1]       [,2]       [,3]      [,4]      [,5]
[1,] 0.000000000 1.00000000 2.00000000 3.0000000 4.0000000
[2,] 0.001729392 0.01729392 0.08673568 0.2824352 0.6412176
     [,6]
[1,]    5
[2,]    1
```

Thus, the 2.5% *quantile is* 2 *and the* 97.5% *quantile is* 5 *so that the* 95% *prediction interval is* [2, 5]*. Clearly, this interval does not contain exactly* 95% *of the probability mass of the predictive distribution since the distribution of* Y_6 *is discrete. In fact, the predictive probability for* Y_6 *to fall into* [2, 5] *is larger:*

$$1 - \Pr(Y_6 \leq 1) = 0.9827.$$

d) Calculate the likelihood prediction as well.

▶ *The extended likelihood function is*

$$L(\pi, y_6) = \binom{n_6}{y_6} \pi^{5\bar{y}+y_6}(1-\pi)^{5\bar{n}+n_6-5\bar{y}-y_6}.$$

If y_6 *had been observed, then the ML estimate of* π *would be*

$$\hat{\pi}(y_6) = \frac{5\bar{y}+y_6}{5\bar{n}+n_6},$$

which yields the predictive likelihood

$$\begin{aligned} L_p(y_6) &= L(\hat{\pi}(y_6), y_6) \\ &= \binom{n_6}{y_6} \left(\frac{5\bar{y}+y_6}{5\bar{n}+n_6}\right)^{5\bar{y}+y_6} \left(\frac{5\bar{n}+n_6-5\bar{y}-y_6}{5\bar{n}+n_6}\right)^{5\bar{n}+n_6-5\bar{y}-y_6}. \end{aligned}$$

The likelihood prediction

$$f_p(y_6) = \frac{L_p(y_6)}{\sum_{y=0}^{n_6} L_p(y)}$$

can now be calculated numerically:

```
> ## predictive likelihood
> predLik <- function(yNew, nNew)
  {
      sumY <- sum(y) + yNew
      sumN <- sum(n) + nNew
      pi <- sumY / sumN

      logRet <- lchoose(nNew, yNew) + sumY * log(pi)
               + (sumN - sumY) * log(1 - pi)
      return(exp(logRet))
  }
> ## calculate values of the discrete likelihood prediction:
> predictiveProb <- predLik(0:5, 5)
> (predictiveProb <- predictiveProb / sum(predictiveProb))
```

```
[1] 0.004754798 0.041534762 0.155883020 0.312701779
[5] 0.333881997 0.151243644
> ## distribution function
> cumsum(predictiveProb)
[1] 0.004754798 0.046289560 0.202172580 0.514874359
[5] 0.848756356 1.000000000
```

The values of the discrete distribution function are similar to ones obtained from the Bayes prediction. The 95% prediction interval is also $[2, 5]$ *here. The point estimate from the likehood prediction turns out to be:*

```
> sum((0:5) * predictiveProb)
[1] 3.383152
```

This estimate is close to 3.9706 *from the Bayes prediction.*

2. Let $X_{1:n}$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution from which a further observation $Y = X_{n+1}$ is to be predicted. Both the expectation μ and the variance σ^2 are unknown.

a) Start by determining the plug-in predictive distribution.

▶ *Note that in contrast to Example 9.2, the variance* σ^2 *is unknown here. By Example 5.3, the ML estimates are*

$$\hat{\mu}_{\mathrm{ML}} = \bar{x} \quad \textit{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{9.2}$$

The plug-in predictive distribution is thus

$$Y \sim \mathrm{N}\left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right).$$

b) Calculate the likelihood and the bootstrap predictive distributions.

▶ *The extended likelihood is*

$$\begin{aligned} L(\mu, \sigma^2, y) &= f(y \mid \mu, \sigma^2) \cdot L(\mu, \sigma^2) \\ &\propto \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \cdot (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)\right\} \\ &= (\sigma^2)^{-\frac{n+1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left((y-\mu)^2 + n(\bar{x}-\mu)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right\} \\ &\propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left((y-\mu)^2 + n(\bar{x}-\mu)^2\right)\right\} \end{aligned}$$

and the ML estimates of the parameters based on the extended data set are

$$\hat{\mu}(y) = \frac{n\bar{x} + y}{n+1} \quad \textit{and} \quad \hat{\sigma}^2(y) = \frac{1}{n+1}\left(\sum_{i=1}^{n}(x_i - \hat{\mu}(y))^2 + (y - \hat{\mu}(y))^2\right).$$

This yields the predictive likelihood

$$L_p(y) = L(\hat{\mu}(y), \hat{\sigma}^2(y), y),$$

which can only be normalised numerically for a given data set to obtain the likelihood prediction $f(y) = L_p(y) / \int L_p(u)\, du$.

To determine the bootstrap predictive distribution, we need the distribution of the ML estimators in (9.2). *In Example 3.5 and Example 3.8, respectively, we have seen that*

$$\hat{\mu}_{\mathrm{ML}} \mid \mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2/n) \quad \textit{and} \quad \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \mid \sigma^2 \sim \chi^2(n-1).$$

In addition, the two above random variables are independent. Since $\chi^2(d) = \mathrm{G}(d/2, 1/2)$, *we can deduce*

$$\hat{\sigma}^2_{\mathrm{ML}} \mid \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \mid \sigma^2 \sim \mathrm{G}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right)$$

by using the fact that the second parameter of the Gamma distribution is an inverse scale parameter (see Appendix A.5.2).

The bootstrap predictive distribution of y *given* $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ *has density*

$$\begin{aligned} g(y; \boldsymbol{\theta}) &= \int_{0}^{\infty}\int_{-\infty}^{\infty} f(y \mid \hat{\mu}_{\mathrm{ML}}, \hat{\sigma}^2_{\mathrm{ML}}) f(\hat{\mu}_{\mathrm{ML}} \mid \mu, \sigma^2) f(\hat{\sigma}^2_{\mathrm{ML}} \mid \mu, \sigma^2)\, d\hat{\mu}_{\mathrm{ML}}\, d\hat{\sigma}^2_{\mathrm{ML}} \\ &= \int_{0}^{\infty}\int_{-\infty}^{\infty} f(y \mid \hat{\mu}_{\mathrm{ML}}, \hat{\sigma}^2_{\mathrm{ML}}) f(\hat{\mu}_{\mathrm{ML}} \mid \mu, \sigma^2)\, d\hat{\mu}_{\mathrm{ML}} f(\hat{\sigma}^2_{\mathrm{ML}} \mid \mu, \sigma^2)\, d\hat{\sigma}^2_{\mathrm{ML}}. \end{aligned} \tag{9.3}$$

The inner integral in (9.3) *corresponds to the marginal likelihood in the normal-normal model. From* (7.18) *we thus obtain*

$$\begin{aligned} &\int_{-\infty}^{\infty} f(y \mid \hat{\mu}_{\mathrm{ML}}, \hat{\sigma}^2_{\mathrm{ML}}) f(\hat{\mu}_{\mathrm{ML}} \mid \mu, \sigma^2)\, d\hat{\mu}_{\mathrm{ML}} \\ &= (2\pi\hat{\sigma}^2_{\mathrm{ML}})^{-\frac{1}{2}} \left(\frac{1/\sigma^2}{1/\hat{\sigma}^2_{\mathrm{ML}} + 1/\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2_{\mathrm{ML}}}\left(\frac{1/\sigma^2}{1/\hat{\sigma}^2_{\mathrm{ML}} + 1/\sigma^2}(y-\mu)^2\right)\right\}. \end{aligned} \tag{9.4}$$

Analytical computation of the outer integral in (9.3) *is however difficult. To calculate* $g(y; \boldsymbol{\theta})$, *we can use Monte Carlo integration instead: We draw a large number of random numbers* $(\hat{\sigma}^2_{\mathrm{ML}})^{(i)}$ *from the* $\mathrm{G}((n-1)/2, n/(2\hat{\sigma}^2_{\mathrm{ML}}))$ *distribution, plug them into* (9.4) *and compute the mean for the desired values of* y. *Of course, this only works for a given data set* $x_1, \ldots, x_n$.

c) Derive the Bayesian predictive distribution under the assumption of the reference prior $f(\mu, \sigma^2) \propto \sigma^{-2}$.

▶ *As in Example 6.24, it is convenient to work with the precision $\kappa = (\sigma^2)^{-1}$ and the corresponding reference prior $f(\mu, \kappa) \propto \kappa^{-1}$, which formally corresponds to the normal-gamma distribution $\mathrm{NG}(0, 0, -1/2, 0)$. By (6.26), the posterior distribution is again a normal-gamma distribution:*

$$(\mu, \kappa) \,|\, x_{1:n} \sim \mathrm{NG}\left(\mu^*, \lambda^*, \alpha^*, \beta^*\right),$$

where

$$\mu^* = \bar{x}, \qquad \lambda^* = n,$$
$$\alpha^* = \frac{1}{2}(n-1) \qquad \text{and} \qquad \beta^* = \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

In Exercise 3 in Chapter 7, we have calculated the prior predictive distribution for this model. Since we have a conjugate prior distribution, we can infer the posterior predictive distribution from the prior predictive distribution by replacing the prior parameters by the posterior parameters and adjusting the likelihood appropriately. For the latter task, we set $n = 1$, $\hat{\sigma}^2_{\text{ML}} = 0$ and $\bar{x} = y$ in formula (7.1) in the solutions since we want to predict one observation only. In addidtion, we replace the prior parameters by the posterior parameters in (7.1) to obtain

$$\begin{aligned}
f(y \,|\, x) &= \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{\lambda^*}{\lambda^*+1}\right)^{1/2} \frac{\Gamma(\alpha^*+1/2)(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)} \cdot \\
&\quad \left\{\beta^* + \frac{(\lambda^*+1)^{-1}\lambda^*(\nu^*-y)^2}{2}\right\}^{-(\alpha^*+\frac{1}{2})} \\
&\propto \left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{(n+1)^{-1}n(\bar{x}-y)^2}{2}\right)^{-(\frac{n-1}{2}+\frac{1}{2})} \\
&= \left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{-\frac{n}{2}} \left(1 + \frac{2n(y-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2(n+1)2}\right)^{-\frac{n}{2}} \\
&\propto \left(1 + \frac{(y-\bar{x})^2}{(n-1)\left(1+\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{-\frac{(n-1)+1}{2}}.
\end{aligned}$$

This is the kernel of the t distribution mentioned at the end of Example 9.7, i.e. the posterior predictive distribution is

$$Y \,|\, x_{1:n} \sim \mathrm{t}\left(\bar{x}, \left(1+\frac{1}{n}\right) \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}, n-1\right).$$

3. Derive Equation (9.11).

▶ *We proceed analoguously as in Example 9.7. By Example 6.8, the posterior distribution of μ is*

$$\mu \,|\, x_{1:n} \sim \mathrm{N}\left(\bar{\mu}, \frac{\sigma^2}{n+\delta\sigma^2}\right),$$

where

$$\bar{\mu} = \mathsf{E}(\mu \,|\, x_{1:n}) = \frac{\sigma^2}{\delta\sigma^2+n}\left(\frac{n\bar{x}}{\sigma^2} + \delta\nu\right).$$

The posterior predictive distribution of $Y \,|\, x_{1:n}$ has thus density

$$\begin{aligned}
f(y \,|\, x_{1:n}) &= \int f(y \,|\, \mu) f(\mu \,|\, x_{1:n})\, d\mu \\
&\propto \int \exp\left[-\frac{1}{2}\left\{\frac{(\mu-y)^2}{\sigma^2} + \frac{(\delta\sigma^2+n)(\mu-\bar{\mu})^2}{\sigma^2}\right\}\right] d\mu.
\end{aligned}$$

We now use Appendix B.1.5 to combine these two quadratic forms:

$$\begin{aligned}
&\frac{(\mu-y)^2}{\sigma^2} + \frac{(\delta\sigma^2+n)(\mu-\bar{\mu})^2}{\sigma^2} \\
=&\frac{\delta\sigma^2+n+1}{\sigma^2}(\mu-c)^2 + \frac{\delta\sigma^2+n}{\sigma^2(\delta\sigma^2+n+1)}(y-\bar{\mu})^2,
\end{aligned}$$

for

$$c = \frac{y + (\delta\sigma^2+n)\bar{\mu}}{\delta\sigma^2+n+1}.$$

Since the second term does not depend on μ, this implies

$$\begin{aligned}
f(y \,|\, x_{1:n}) &\propto \exp\left\{-\frac{\delta\sigma^2+n}{2\sigma^2(\delta\sigma^2+n+1)}(y-\bar{\mu})^2\right\} \underbrace{\int \exp\left\{-\frac{\delta\sigma^2+n+1}{2\sigma^2}(\mu-c)^2\right\} d\mu}_{=\sqrt{2\pi}\sigma/\sqrt{\delta\sigma^2+n+1}} \\
&\propto \exp\left\{-\frac{\delta\sigma^2+n}{2\sigma^2(\delta\sigma^2+n+1)}(y-\bar{\mu})^2\right\}, \qquad (9.5)
\end{aligned}$$

where we have used that the above integrand is the kernel of a normal density. Now, (9.5) is the kernel of a normal density with mean $\bar{\mu}$ and variance $\sigma^2(\delta\sigma^2+n+1)/(\delta\sigma^2+n)$, so the posterior predictive distribution is

$$Y \,|\, x_{1:n} \sim \mathrm{N}\left(\bar{\mu}, \sigma^2\left(\frac{\delta\sigma^2+n+1}{\delta\sigma^2+n}\right)\right).$$

4. Prove Murphy's decomposition (9.16) of the Brier score.

▶ *We first establish a useful identity: Since the observations are binary, the mean of the squared observations $(\overline{y^2})$ equals the mean of the original observations $(\bar{y})$. First, we have*

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \overline{y^2} - \bar{y}^2 = \bar{y}(1-\bar{y}). \qquad (9.6)$$

Let the observations be assigned to J groups and denoted by y_{ji}, $i = 1, \dots, n_j$, $j = 1, \dots, J$ with group means

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}$$

and respresentative prediction probabilities π_j. In total, there are $N = \sum_{j=1}^{J} n_j$ observed values with overall mean (or relative frequency) $\bar{y}$.

We now calculate the right-hand side of Murphy's decomposition (9.16) *and use* (9.6)*:*

$$\begin{aligned}&\bar{y}(1-\bar{y}) + \mathrm{SC} - \mathrm{MR}\\ =&\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2 + \frac{1}{N}\sum_{j=1}^{J} n_j(\bar{y}_j-\pi_j)^2 - \frac{1}{N}\sum_{j=1}^{J} n_j(\bar{y}_j-\bar{y})^2\\ =&\frac{1}{N}\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2 + n_j(\bar{y}_j-\pi_j)^2 - n_j(\bar{y}_j-\bar{y})^2\right).\end{aligned}$$

The aim is to obtain the mean Brier score

$$\overline{\mathrm{BS}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji}-\pi_j)^2,$$

which we can isolate from the first term above since

$$\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2 = \sum_{i=1}^{n_j}(y_{ji}-\pi_j)^2 + 2(\pi_j-\bar{y})\sum_{i=1}^{n_j}(y_{ji}-\pi_j) + n_j(\pi_j-\bar{y})^2.$$

Consequently,

$$\begin{aligned}&\bar{y}(1-\bar{y}) + \mathrm{SC} - \mathrm{MR}\\ =&\overline{\mathrm{BS}} + \frac{1}{N}\sum_{j=1}^{J}\left(2(\pi_j-\bar{y})n_j(\bar{y}_j-\pi_j) + n_j(\pi_j-\bar{y})^2 + n_j(\bar{y}_j-\pi_j)^2 - n_j(\bar{y}_j-\bar{y})^2\right)\end{aligned}$$

and by expanding the quadratic terms on the right-hand side of the equation, we see that they all cancel. This completes the proof of Murphy's decomposition.

5. Investigate if the scoring rule

$$S(f(y), y_o) = -f(y_o)$$

is proper for a binary observation Y.

▶ *Let $\mathrm{B}(\pi_0)$ denote the true distribution of the observation Y_o and f the probability mass of the predictive distribution $Y \sim \mathrm{B}(\pi)$ as introduced in Definition 9.9. The expected score under the true distribution is then*

$$\begin{aligned}\mathsf{E}[S(f(y), Y_o)] = -\,\mathsf{E}[f(Y_o)] &= -f(0)\cdot(1-\pi_0) - f(1)\cdot\pi_0\\ &= -(1-\pi)(1-\pi_0) - \pi\cdot\pi_0\\ &= (1-2\pi_0)\pi + \pi_0 - 1.\end{aligned}$$

As a function of π, the expected score is thus a line with slope $1-2\pi_0$. If this slope is positive or negative, respectively, then the score is minimised by $\pi = 0$ or $\pi = 1$, respectively (compare to the proof of Result 9.2 for the absolute score). Hence, the score is in general not minimised by $\pi = \pi_0$, i. e. this scoring rule is not proper.

6. For a normally distributed prediction show that it is possible to write the CRPS as in (9.17) using the formula for the expectation of the folded normal distribution in Appendix A.5.2.

▶ *The predictive distribution here is the normal distribution $\mathrm{N}(\mu, \sigma^2)$. Let Y_1 und Y_2 be independent random variables with $\mathrm{N}(\mu, \sigma^2)$ distribution. From this, we deduce*

$$Y_1 - y_o \sim \mathrm{N}(\mu - y_o, \sigma^2) \quad \textit{and} \quad Y_1 - Y_2 \sim \mathrm{N}(0, 2\sigma^2),$$

where for the latter result, we have used $\mathrm{Var}(Y_1 + Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$ due to independence (see Appendix A.3.5). This implies (see Appendix A.5.2)

$$|Y_1 - y_o| \sim \mathrm{FN}(\mu - y_o, \sigma^2) \quad \textit{and} \quad |Y_1 - Y_2| \sim \mathrm{FN}(0, 2\sigma^2).$$

The CRPS is therefore

$$\begin{aligned}CRPS(f(y), y_o) &= \mathsf{E}\{|Y_1 - y_o|\} - \frac{1}{2}\,\mathsf{E}\{|Y_1 - Y_2|\}\\ &= 2\sigma\varphi\left(\frac{\mu - y_o}{\sigma}\right) + (\mu - y_o)\left\{2\Phi\left(\frac{\mu - y_o}{\sigma}\right) - 1\right\} - \frac{1}{2}\left\{2\sqrt{2}\sigma\varphi(0) + 0\right\}\\ &= 2\sigma\varphi\left(\frac{y_o - \mu}{\sigma}\right) + (\mu - y_o)\left[2\left\{1 - \Phi\left(\frac{y_o - \mu}{\sigma}\right)\right\} - 1\right] - \frac{\sqrt{2}\sigma}{\sqrt{2\pi}}\\ &= 2\sigma\varphi(\tilde{y}_o) + \left(\frac{\mu - y_o}{\sigma}\right)\sigma\left\{1 - 2\Phi(\tilde{y}_o)\right\} - \frac{\sigma}{\sqrt{\pi}}\\ &= \sigma\left[\tilde{y}_o\{2\Phi(\tilde{y}_o) - 1\} + 2\varphi(\tilde{y}_o) - \frac{1}{\sqrt{\pi}}\right].\end{aligned}$$

Bibliography

Bartlett M. S. (1937) Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 160(901):268–282.

Box G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). Journal of the Royal Statistical Society, Series A, 143:383–430.

Cole S. R., Chu H., Greenland S., Hamra G. and Richardson D. B. (2012) Bayesian posterior distributions without Markov chains. American Journal of Epidemiology, 175(5):368–375.

Davison A. C. (2003) Statistical Models. Cambridge University Press, Cambridge.

Dempster A. P., Laird N. M. and Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

Good I. J. (1995) When batterer turns murderer. Nature, 375(6532):541.

Good I. J. (1996) When batterer becomes murderer. Nature, 381(6532):481.

Goodman S. N. (1999) Towards evidence-based medical statistics. 2.: The Bayes factor. Annals of Internal Medicine, 130:1005–1013.

Merz J. F. and Caulkins J. P. (1995) Propensity to abuse - Propensity to murder? Chance, 8(2):14.

Rao C. R. (1973) Linear Statistical Inference and Its Applications. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York.

Sellke T., Bayarri M. J. and Berger J. O. (2001) Calibration of p values for testing precise null hypotheses. The American Statistician, 55:62–71.

Index

A
arithmetic mean 14

B
beta distribution 195
beta-binomial distribution 196
binomial distribution 195
bootstrap predictive distribution 199
burn-in 177

C
case-control study
  matched 95
change-of-variables formula 170
convolution theorem 171

E
Emax model 111
examples
  analysis of survival times 80
  blood alcohol concentration 94, 145
  capture-recapture method 7
  prevention of preeclampsia 64, 174
exponential model 75

F
Fisher information 57

H
HPD interval 172

I
inverse gamma distribution 175

J
Jeffreys' Priori 195

L
likelihood function
  extended 197
likelihood]extended 198
Lindley's paradox 150

M
Mallow's C_p statistic 144
marginal likelihood 199
minimum Bayes factor 151
Monte Carlo estimate 172, 199

N
normal distribution
  folded 203
normal-gamma distribution 200
normal-normal model 199
numerical integration 163, 168, 171, 172

P
P-value 152
Pareto distribution 134
point prediction 196
power model 74
prediction interval 198
predictive likelihood 199
prior
  -data conflict 155
  criticism 155
prior distribution
  non-informative 195
prior predictive distribution *see* marginal likelihood
profile likelihood confidence interval 61
prognostic interval 196

R
regression model
  logistic 98
  normal 144
risk
  log relative 59
  relative 59

S
score equations 56
significance test 71

T
Taylor approximation 73