Chapter 3: Normal Model

ST4234: Bayesian Statistics

Semester 2, AY 2019/2020

Department of Statistics and Applied Probability

National University of Singapore

LI Cheng

stalic@nus.edu.sg

Introduction

- The normal distribution is one of the most utilized probability model for data analysis.
- One of the reasons is due to the Central Limit Theorem (CLT) which states that under some general conditions, the sum (or mean) of a set of random variables is approximately normally distributed.
- It is a simple model with separate parameters for the population mean and variance – two quantities that are often of primary interest.
- In this chapter, we will discuss some properties of the normal distribution and Bayesian inference on the population mean and variance parameters.

Outline

The normal distribution

Inference for mean when variance is known

Inference when both mean and variance are unknown

Noninformative priors

- Improper priors
- Reference prior
- Jeffreys prior

Concluding Remarks

3.1 The normal distribution

Definition

- We say that a random variable Y is normally distributed with mean θ and variance σ^2, with the notation

  $$Y \sim \mathrm{N}(\theta, \sigma^2)$$

  if the pdf of Y is

  $$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}, \quad -\infty < y < +\infty.$$

- If $Y \sim \mathrm{N}(\theta, \sigma^2)$, then $Z = \dfrac{Y-\theta}{\sigma}$ has a standard normal distribution, i.e. $Z \sim \mathrm{N}(0, 1)$.

3.1 The normal distribution

Properties

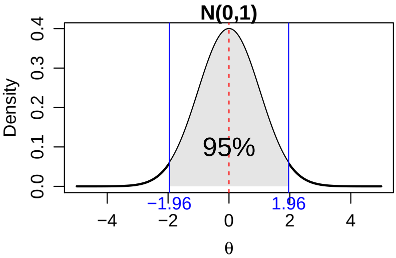
- The normal distribution is symmetric about θ and the mean, mode and median are all equal to θ.

- About 95% of the distribution lie within approximately two standard deviation of the mean.

- If $X \sim \mathrm{N}(\theta, \sigma^2)$ and $Y \sim \mathrm{N}(\mu, \tau^2)$ are independent, then

  $$aX+bY \sim \mathrm{N}(a\theta+b\mu, a^2\sigma^2+b^2\tau^2),$$

  where a and b are two arbitrary numbers.



R commands

- The commands `dnorm, pnorm, qnorm, rnorm` provide the density, distribution function, quantile function and random generation respectively for the normal distribution in `R`.

Example: Suppose $Y \sim \mathrm{N}(0, 4)$

- Find $p(1)$: `dnorm(x=1,mean=0,sd=2)` (Ans: 0.176)
- Find $\mathrm{P}(Y \leq 2)$: `pnorm(q=2,mean=0,sd=2)` (Ans: 0.841)
- Find α such that $P(Y \leq \alpha) = 0.9$: `qnorm(p=0.9,mean=0,sd=2)` (Ans: 2.56)
- Generate 50 random numbers from $\mathrm{N}(0, 4)$: `rnorm(n=50,mean=0,sd=2)`

3.1 The normal distribution

R commands

- Note that the commands `dnorm, pnorm, qnorm, rnorm` require specification of the mean and the standard deviation (not the variance!) of the normal distribution.

- For these commands, the default value for the mean is 0 and for the standard deviation is 1. Therefore, if we are using the standard normal, we do not need to specify the arguments mean and sd. For example, `rnorm(50)` generates 50 random variables from N(0,1).

Outline

The normal distribution

Inference for mean when variance is known

Inference when both mean and variance are unknown

Noninformative priors

- Improper priors
- Reference prior
- Jeffreys prior

Concluding Remarks

3.2 Inference for mean when variance is known

- We first assume that the variance σ^2 is **known** (e.g. $\sigma^2 = 4$), which means it is fixed and we don't need to assign a prior for σ^2.
- Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be an i.i.d. sample from $\mathrm{N}(\theta, \sigma^2)$.

$$y_i|\theta \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2) \quad \text{for } i = 1, \ldots, n.$$

- The likelihood function is

$$\begin{aligned} p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n} y_i^2 - 2\theta n\bar{y} + n\theta^2}{2\sigma^2}\right\}, \end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean.

3.2 Inference for mean when variance is known

Sufficient statistic

- We can rewrite the likelihood function as

$$p(\boldsymbol{y}|\theta) = \underbrace{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}}_{h(\boldsymbol{y})}\underbrace{\exp\left\{-\frac{n\theta^2}{2\sigma^2}\right\}}_{g(\theta)}\ \underbrace{\exp\left\{\frac{n\theta\bar{y}}{\sigma^2}\right\}}_{\exp\{\eta(\theta)t(\boldsymbol{y})\}}.$$

- This means that the normal model with a known variance belongs to the one-parameter exponential family, with

$$\eta(\theta) = n\theta/\sigma^2, \quad t(y) = \bar{y}.$$

- From the Factorization Theorem, $\bar{y}$ is the sufficient statistic for θ.

3.2 Inference for mean when variance is known

Conjugate prior

- A conjugate prior for θ is of the form

$$p(\theta) \propto \exp\left\{-c_1\theta^2 + c_2\theta\right\} \propto \exp\left\{-c_1\left(\theta - \frac{c_2}{2c_1}\right)^2\right\}. \qquad (1)$$

- Note that the range of the mean parameter θ is $\mathbb{R} = (-\infty, +\infty)$. The simplest class of probability densities on $\mathbb{R}$ with the form of (1) is the family of normal distributions.

- Since the normal model belongs to the exponential family, if we use a normal distribution as the prior for θ, then the posterior of θ is also a normal distribution.

- Let us consider the prior $\theta \sim N(\mu_0, \tau_0^2)$. Then the prior density is

$$p(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right\}.$$

3.2 Inference for mean when variance is known

Posterior

- The posterior density of θ given $\boldsymbol{y}$ is

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &\propto \exp\left[-\frac{1}{2}\left\{\frac{n\theta^2 - 2\theta n\bar{y}}{\sigma^2} + \frac{(\theta^2 - 2\theta\mu_0)}{\tau_0^2}\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)\theta^2 - 2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right)\theta\right\}\right]. \end{aligned}$$

- We can make a whole square for θ in the exponent. Let

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \text{ and } \mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right).$$

3.2 Inference for mean when variance is known

Posterior

- Then

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto \exp\left[-\frac{1}{2}\left\{\frac{\theta^2 - 2\theta\mu_n}{\tau_n^2}\right\}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\frac{\theta^2 - 2\theta\mu_n + \mu_n^2}{\tau_n^2}\right\} + \frac{\mu_n^2}{2\tau_n^2}\right] \\ &\propto \exp\left\{-\frac{1}{2}\frac{(\theta - \mu_n)^2}{\tau_n^2}\right\}. \end{aligned}$$

- From the mathematical form, we recognize that the posterior density is

$$\theta \mid \boldsymbol{y} \ \sim \ \mathrm{N}(\mu_n, \tau_n^2).$$

- Hence, $p(\theta) = \mathrm{N}(\mu_0, \tau_0)$ is a conjugate prior, as it leads to a posterior distribution $p(\theta|\boldsymbol{y})$ that belongs to the same distributional family as the prior (both are normal distributions).

3.2 Inference for mean when variance is known

Prior and Posterior

- We have shown that when σ^2 is known,

$$\boxed{\begin{array}{ccl} \theta \sim \mathrm{N}(\mu_0, \tau_0^2) & & \theta|\boldsymbol{y} \sim \mathrm{N}(\mu_n, \tau_n^2), \text{ where} \\ Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2) & \Rightarrow & \dfrac{1}{\tau_n^2} = \dfrac{1}{\sigma^2/n} + \dfrac{1}{\tau_0^2}, \; \mu_n = \tau_n^2 \left(\dfrac{\bar{y}}{\sigma^2/n} + \dfrac{\mu_0}{\tau_0^2} \right). \end{array}}$$

- The sufficient statistic $\bar{y}$ has the distribution

$$\bar{y} \mid \theta \; \sim \; \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right).$$

Using $p(\theta|\bar{y}) \propto p(\bar{y}|\theta)p(\theta)$, we can check that $p(\theta|\boldsymbol{y}) = p(\theta|\bar{y})$ (left as an exercise).

3.2 Inference for mean when variance is known

Variance and precision

- Inverse variance is often referred to as the **precision**. Thus the relationship $\frac{1}{\tau^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$ can be thought of as

$$\text{Posterior precision} = \text{Data Precision} + \text{Prior Precision}$$

- The posterior variance is smaller than both the variance of the prior (τ_0^2 of $p(\theta)$), and the variance of the likelihood function (σ^2/n of $p(\bar{y}|\theta)$) (why?).

- Thus the posterior distribution offers a sensible compromise between our prior opinion and the observed data, and the combined strength of the two sources of information leads to increased precision in our understanding of θ.

3.2 Inference for mean when variance is known

Posterior mean

- We can write

$$\tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \text{ and } \mu_n = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \bar{y} + \frac{\frac{1}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \mu_0.$$

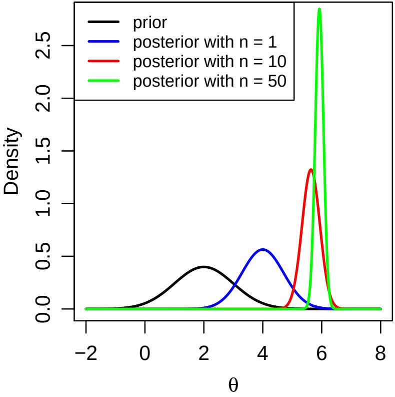
- The posterior mean is a weighted average of the prior mean and the sample mean.
- As $n \to \infty$, the posterior variance goes to zero (we become more certain about the value of θ), and the posterior mean approaches the sample mean.
- The data dominates the prior increasingly as the sample size increases.

3.2 Inference for mean when variance is known

Example: Posterior arising from different sample sizes

- Suppose that $\mu_0 = 2$, $\tau_0^2 = 1$, $\sigma^2 = 1$, $\bar{y} = 6$. The figure below plots the prior distribution, along with the posterior distributions arising from different sample sizes, $n = 1, 10, 50$.

- When $n = 1$, the prior and likelihood receive equal weight. Hence $\mu_n = 4$.
- When $n = 10$, the data dominate the prior, resulting in a posterior mean much closer to $\bar{y}$.
- The posterior variance shrinks as n increases, collapsing to a point mass at $\bar{y}$ when $n \to \infty$.



3.2 Inference for mean when variance is known

Prediction (with a known variance)

- Consider predicting a new observation $\tilde{y}$ from the population after observing $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- To find the predictive distribution, we will use the following fact

$$\tilde{y} \sim \mathrm{N}(\theta, \sigma^2) \;\Leftrightarrow\; \tilde{y} = \theta + \epsilon \;\text{ where }\; \epsilon \sim \mathrm{N}(0, \sigma^2).$$

  "$\tilde{y}$ is normal with mean θ" is the same as "$\tilde{y}$ is equal to θ plus some mean-zero normally distributed noise".

- The posterior mean and variance of $\tilde{y}$ are

$$\begin{aligned} \mathrm{E}(\tilde{y}|\boldsymbol{y}) &= \mathrm{E}(\theta + \epsilon|\boldsymbol{y}) = \mathrm{E}(\theta|\boldsymbol{y}) + \mathrm{E}(\epsilon|\boldsymbol{y}) = \mu_n. \\ \mathrm{Var}(\tilde{y}|\boldsymbol{y}) &= \mathrm{Var}(\theta + \epsilon|\boldsymbol{y}) = \mathrm{Var}(\theta|\boldsymbol{y}) + \mathrm{Var}(\epsilon) = \tau_n^2 + \sigma^2. \end{aligned}$$

  Note: ϵ is the noise for the new observation, which is independent of both θ and $\boldsymbol{y}$.

3.2 Inference for mean when variance is known

Prediction (with a known variance)

- Since both θ and ϵ conditioned on $\boldsymbol{y}$ are normally distributed, so is $\tilde{y} = \theta + \epsilon$.
- Therefore the predictive distribution is

$$\tilde{y} \mid \boldsymbol{y} \sim N(\mu_n, \tau_n^2 + \sigma^2).$$

- Uncertainty about a new sample $\tilde{y}$ stems from
  - (i) uncertainty about θ (with variance τ_n^2) and
  - (ii) uncertainty due to sampling variability or noise ϵ (with variance σ^2).
- As n increases, we become more certain about the value of θ, and $\tau_n^2 = \dfrac{1}{n/\sigma^2 + 1/\tau_0^2}$ goes to zero. But uncertainty due to the noise (sampling variability) remains. Hence our uncertainty about $\tilde{y}$ never goes below σ^2.
- We have explained a similar phenomenon before for the prediction of a Poisson model.

3.2 Inference for mean when variance is known

Example: Midge wing length

- Grogan and Wirth (1981) provide data on the wing length in millimeters of 9 members of a species of midge (small, two-winged flies). From these 9 measurements, we wish to make inference on the population mean θ. The observations are $\boldsymbol{y} = (1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)$, giving $\bar{y} = 1.804$.
- Studies suggest that wing lengths are typically around 1.9 mm, so we set the prior mean $\mu_0 = 1.9$.
- Lengths must be positive, implying that $\theta > 0$. Ideally, we should use a prior distribution $p(\theta)$ that has mass only on $\theta > 0$.
- Let us approximate this restriction by using the fact that for any normal distribution, most of the probability is within two standard deviations of the mean.
- Thus we choose τ_0^2 so that $\mu_0 - 2\tau_0 > 0 \Rightarrow \tau_0 < 0.95$. For now, we take $\tau_0 = 0.95$, which somewhat overstates our prior uncertainty about θ.

3.2 Inference for mean when variance is known

Example: Midge wing length

- Let us assume that the wing length of the *i*th midge,

$$y_i \sim \mathrm{N}(\theta, \sigma^2).$$

- Then $\theta \mid \boldsymbol{y} \sim \mathrm{N}(\mu_n, \tau_n^2)$ where

$$\tau_n^2 = 1/\left(\frac{9}{\sigma^2} + \frac{1}{0.95^2}\right), \quad \mu_n = \tau_n^2\left(\frac{9(1.804)}{\sigma^2} + \frac{1.9}{0.95^2}\right).$$

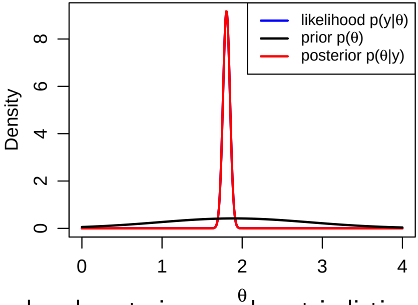
- If we estimate $\sigma^2 \approx s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.017$, then

$$\theta \mid \boldsymbol{y} \sim \mathrm{N}(1.805, 0.001871).$$

- Here we pretended that we know the value of σ^2. Later we will discuss how to address unknown σ^2.

3.2 Inference for mean when variance is known

Example: Midge wing length



Likelihood and posterior are almost indistinguishable.

- A 95% CI for θ based on this distribution is (1.72, 1.89). However, this interval assumes that we are certain $\sigma^2 = s^2$, when s^2 is only a rough estimate based on 9 observations. For a more accurate representation of our information we need to account for the fact that σ^2 is unknown.

3.2 Inference for mean when variance is known

R commands

- Now, let us look at how these computations can be performed in R. First we can input the data, compute the sample mean and variance, and specify the prior. Then we compute the posterior parameters using the formulas.

```
> y <- c(1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)
> n <- length(y)
> ybar <- mean(y)
> sigma2 <- var(y)
> mu0 <- 1.9
> tau02 <- 0.95^2
> (taun2 <- 1 / (n/sigma2 + 1/tau02))
[1] 0.00187142
> (mun <- taun * (n*ybar/sigma2 + mu0/tau02))
[1] 1.804643
> qnorm(c(0.025,0.975), mean=mun, sd=sqrt(taun2))
[1] 1.719855 1.889430
```

Outline

The normal distribution

Inference for mean when variance is known

Inference when both mean and variance are unknown

Noninformative priors

- Improper priors
- Reference prior
- Jeffreys prior

Concluding Remarks

3.3 Inference when both mean and variance are unknown

- Suppose we observe data $\boldsymbol{y} = (y_1, \ldots, y_n)$ where

$$y_i \mid \theta, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2), \quad \text{for} \quad i = 1, \ldots, n,$$

  where both θ and σ^2 are unknown.

- We need to use an important relation (explain)

$$\sum_{i=1}^{n} (y_i - \theta)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2.$$

- The sample variance of $\boldsymbol{y}$ is defined to be

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

3.3 Inference when both mean and variance are unknown

- The likelihood function can be written as

$$
\begin{aligned}
p(\boldsymbol{y}|\theta,\sigma^2) &= \prod_{i=1}^{n} p(y_i|\theta,\sigma^2) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{2\sigma^2} - \frac{n(\bar{y}-\theta)^2}{2\sigma^2}\right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \exp\left\{-\frac{(\bar{y}-\theta)^2}{2\sigma^2/n}\right\},
\end{aligned}
$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ is the sample variance.

3.3 Inference when both mean and variance are unknown

Conjugate prior

- Since we have two parameters θ and σ^2, we must specify a **joint prior** on (θ, σ^2), i.e. $p(\theta, \sigma^2)$, for $\theta \in \mathbb{R}$ and $\sigma^2 > 0$.
- The form of the likelihood suggests a conjugate class of prior distributions which can be expressed as

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$$

and where

$$\theta|\sigma^2 \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n_0}\right), \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right).$$

- Here, $\mu_0 \in \mathbb{R}$, $n_0 > 0$, $\nu_0 > 0$, $\sigma_0^2 > 0$ are all constants (chosen by the user), called **hyperparameters**.
- Inv-Gamma stands for **inverse-gamma distribution**.

3.3 Inference when both mean and variance are unknown

Inverse-gamma distribution

- A random variable X is said to follow the **inverse-gamma distribution** with shape parameter a and scale parameter b, denoted by $X \sim \text{Inv-Gamma}(a, b)$, if X has the pdf

$$p(x) = \frac{b^a}{\Gamma(a)}(x)^{-a-1} \exp\left(-\frac{b}{x}\right) \quad \text{for} \quad x > 0,$$

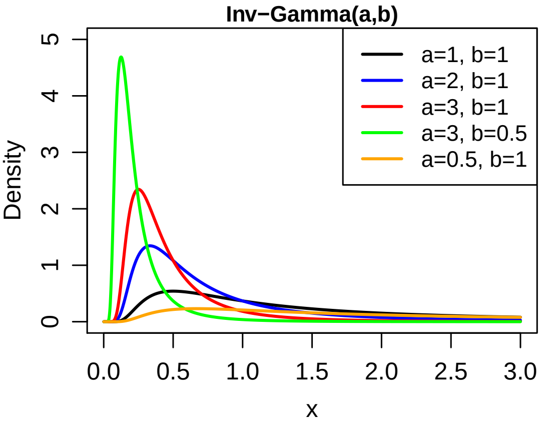
  where $\Gamma(\cdot)$ denotes the gamma function.

- $\text{E}(X) = \dfrac{b}{a-1}$ if $a > 1$, $\text{Var}(X) = \dfrac{b^2}{(a-1)^2(a-2)}$ if $a > 2$.
- Note that $X \sim \text{Gamma}(a, b) \Rightarrow Y = 1/X \sim \text{Inv-Gamma}(a, b)$.

3.3 Inference when both mean and variance are unknown

Inverse-gamma distribution

- The figure below shows some examples of inverse-gamma densities.



- Note that the inverse-gamma density is always zero when $x = 0$.

3.3 Inference when both mean and variance are unknown

Inverse-gamma distribution

- R base does not have direct functions for the inverse-gamma distribution. Several R packages include functions for the inverse-gamma distribution, such as `invgamma`, `MCMCpack`, `pscl`.
- Taking the `invgamma` package as an example. It has the functions `dinvgamma`, `pinvgamma`, `qinvgamma` and `rinvgamma`.
- For example, if $X \sim \text{Inv-Gamma}(3, 2)$, we can calculate the 20% quantile of X by

```
require(invgamma)
qinvgamma(0.5, shape=3, rate=2)
```

- However, usually we can directly use the gamma distribution from R base. For example, `1/qgamma(0.8,shape=3,rate=2)` gives the same 20% quantile for $X \sim \text{Inv-Gamma}(3, 2)$ (why?)
- To generate 100 random numbers from $\text{Inv-Gamma}(3, 2)$, we can simply use `1/rgamma(100,shape=3,rate=2)`.

3.3 Inference when both mean and variance are unknown

Posterior

- Using this prior $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$, the posterior is (check detailed derivation on board):

$$\begin{aligned}
p(\theta, \sigma^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta, \sigma^2)p(\theta, \sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \exp\left\{-\frac{(\bar{y}-\theta)^2}{2\sigma^2/n}\right\} \\
&\quad \times (\sigma^2)^{-1/2} \exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma^2/n_0}\right\} (\sigma^2)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\
&\propto (\sigma^2)^{-1/2} \exp\left\{-\frac{(\theta - \frac{n\bar{y}+n_0\mu_0}{n+n_0})^2}{2\sigma^2/(n+n_0)}\right\} \\
&\quad \times (\sigma^2)^{-\frac{\nu_0+n}{2}-1} \exp\left\{-\frac{(n-1)s^2 + \nu_0\sigma_0^2 + \frac{nn_0(\bar{y}-\mu_0)^2}{n+n_0}}{2\sigma^2}\right\}.
\end{aligned}$$

Posterior

- The posterior is of the form

$$p(\theta, \sigma^2|\boldsymbol{y}) = p(\theta|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y}),$$

where

$$\theta|\sigma^2, \boldsymbol{y} \sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \sigma^2|\boldsymbol{y} \sim \text{Inv-Gamma}\left(\frac{\nu_1}{2}, \frac{\nu_1\sigma_1^2}{2}\right),$$

$$\mu_1 = \frac{n\bar{y} + n_0\mu_0}{n + n_0}, \qquad \nu_1 = \nu_0 + n,$$

$$n_1 = n + n_0, \qquad \sigma_1^2 = \frac{1}{\nu_1}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{nn_0(\bar{y}-\mu_0)^2}{n+n_0}\right]$$

- Therefore, the posterior of (θ, σ^2) belongs to the same class of distributions as the prior. This conjugate prior $p(\theta, \sigma^2)$ in the normal model is called the **normal-inverse-gamma distribution**.

3.3 Inference when both mean and variance are unknown

Interpretation

We first check the conditional posterior of θ given σ^2 and $\boldsymbol{y}$.

- For $\theta|\sigma^2 \sim \mathrm{N}\left(\mu_0, \sigma^2/n_0\right)$ and $\theta|\sigma^2, \boldsymbol{y} \sim \mathrm{N}\left(\mu_1, \sigma^2/n_1\right)$ where $\mu_1 = (n\bar{y} + n_0\mu_0)/(n + n_0)$ and $n_1 = n + n_0$.

- n_0 and μ_0 can be interpreted respectively as the sample size and mean of a prior set of observations with variance σ^2.

- n_1 is then the total number of (current and prior) observations and μ_1 is the sample mean of the current and prior observations.

3.3 Inference when both mean and variance are unknown

Interpretation

We then check the marginal posterior of σ^2 given $\boldsymbol{y}$.

- For $\sigma^2 \sim \text{Inv-Gamma}\left(\nu_0/2,\ \nu_0\sigma_0^2/2\right)$ and $\sigma^2 \mid \boldsymbol{y} \sim \text{Inv-Gamma}\left(\nu_1/2,\ \nu_1\sigma_1^2/2\right)$

  where $\nu_1 = \nu_0 + n$ and $\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \dfrac{nn_0(\bar{y}-\mu_0)^2}{n+n_0}$.

- ν_0 can be interpreted as a prior sample size.

- $(n-1)s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is often called "the sum of squares", and we can think of $\nu_0\sigma_0^2$ and $\nu_1\sigma_1^2$ as the prior and posterior sum of squares. Then we almost have "posterior sum of squares equals prior sum of squares plus data sum of squares".

- The third term in $\nu_1\sigma_1^2$ is a bit harder to understand: it says that a large value of $(\bar{y}-\mu_0)^2$ increases the posterior probability of a large σ^2 (explain).

3.3 Inference when both mean and variance are unknown

Marginal posterior of θ

- In many problems the main interest is the mean θ, while σ^2 is regarded as a nuisance parameter. The marginal posterior distribution of θ is

$$
\begin{aligned}
p(\theta|\boldsymbol{y}) &= \int_0^{+\infty} p(\theta, \sigma^2|\boldsymbol{y}) \, \mathrm{d}\sigma^2 \\
&\propto \int_0^{+\infty} (\sigma^2)^{-\frac{\nu_1+1}{2}-1} \exp\left\{-\frac{\nu_1\sigma_1^2 + n_1(\theta-\mu_1)^2}{2\sigma^2}\right\} \mathrm{d}\sigma^2 \\
&\propto \Gamma\left(\frac{\nu_1+1}{2}\right) \left[\frac{\nu_1\sigma_1^2 + n_1(\theta-\mu_1)^2}{2}\right]^{-\frac{\nu_1+1}{2}} . \\
&\propto \left[1 + \frac{1}{\nu_1}\left(\frac{\theta-\mu_1}{\sigma_1/\sqrt{n_1}}\right)^2\right]^{-\frac{\nu_1+1}{2}}
\end{aligned}
$$

This expression resembles the pdf of a Student's t-distribution.

3.3 Inference when both mean and variance are unknown

Marginal posterior of θ

- If t follows the Student's t-distribution with ν degrees of freedom, denoted by $t \sim t_\nu$, then the pdf of t is

$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad t \in \mathbb{R}.$$

- Define a new parameter $t = \dfrac{\theta - \mu_1}{\sigma_1/\sqrt{n_1}}$. t is a linear transformation of θ (recentered and then rescaled). So $\dfrac{\mathrm{d}\theta}{\mathrm{d}t} = \sqrt{\sigma_1^2/n_1}$. We apply the change-of-variable formula to obtain

$$p(t|\boldsymbol{y}) \propto \left(1 + \frac{t^2}{\nu_1}\right)^{-\frac{\nu_1+1}{2}} \times \left|\frac{\mathrm{d}\theta}{\mathrm{d}t}\right| \propto \left(1 + \frac{t^2}{\nu_1}\right)^{-\frac{\nu_1+1}{2}}.$$

Hence $t \sim t_{\nu_1}$.

3.3 Inference when both mean and variance are unknown

Change-of-variable formula

- For the univariate case, suppose the pdf of a r.v. X is $f_X(x)$ and $Y = g(X)$ where g is a bijective and differentiable function. Then the pdf of y is given by

$$f_Y(y) = f_X(x)|J|, \text{ where } J = \frac{dx}{dy}, \; x = g^{-1}(y).$$

- For the multivariate case, suppose $\mathbf{x} = (x_1, \ldots, x_p)$, $\boldsymbol{y} = (y_1, \ldots, y_p)$ and $\mathbf{Y} = g(\mathbf{X})$ where g is a bijective and differentiable function. Let $f_X(x)$ be the joint pdf of $\mathbf{X}$. Then the joint pdf of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(\mathbf{x}) \times |\det(J)|, \text{ where } J = \begin{bmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}y_1} & \cdots & \frac{\mathrm{d}x_1}{\mathrm{d}y_p} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}x_p}{\mathrm{d}y_1} & \cdots & \frac{\mathrm{d}x_p}{\mathrm{d}y_p} \end{bmatrix},$$

and $\det(J)$ is the determinant of J.

- Transformation of parameters is often carried out in Bayesian inference to reduce skewness of the posterior distribution or to make simulation or optimization procedures more straightforward.
- For example, it may be desirable to transform a parameter, which is constrained to be positive, to the real line.
- The change-of-variable formula is useful for finding the pdf of the transformed variables.

3.3 Inference when both mean and variance are unknown

Example: Midge wing length

- We revisit the midge wing length example. This time, we assume that both θ and σ^2 are unknown.
- Studies of other populations suggest that the true mean and standard deviation of our population under study should not be too far from 1.9 mm and 0.1 mm respectively. So we set $\mu_0 = 1.9$, $\sigma_0^2 = 0.01$.
- Since the observed data may differ from other populations, we choose $n_0 = 1$ and $\nu_0 = 1$, such that the prior of θ and σ^2 is only weakly centered around 1.9 and 0.01.
- The prior $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$ is described by

$$\theta|\sigma^2 \sim \mathrm{N}(1.9, \sigma^2), \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{1}{2}, \frac{0.01}{2}\right).$$

- The sample mean is $\bar{y} = 1.804$. The sample variance is $s^2 = 0.0169$.

Example: Midge wing length

- We can calculate the following quantities

$$
\begin{aligned}
\mu_1 &= \frac{n\bar{y} + n_0\mu_0}{n + n_0} = \frac{9 \times 1.804 + 1.9}{9 + 1} = 1.814, \\
n_1 &= n + n_0 = 9 + 1 = 10, \quad \nu_1 = \nu_0 + n = 1 + 9 = 10, \\
\sigma_1^2 &= \frac{1}{\nu_1}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{nn_0(\bar{y} - \mu_0)^2}{n + n_0}\right], \\
&= \frac{1}{10}\left[0.01 + 8 \times 0.0169 + \frac{9 \times (1.804 - 1.9)^2}{9 + 1}\right] = 0.0153.
\end{aligned}
$$

- The posterior $p(\theta, \sigma^2|\boldsymbol{y}) = p(\theta|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$ is given by

$$
\theta|\sigma^2, \boldsymbol{y} \sim \mathrm{N}\left(1.814, \frac{\sigma^2}{10}\right), \quad \sigma^2|\boldsymbol{y} \sim \text{Inv-Gamma}(5, 0.077).
$$

3.3 Inference when both mean and variance are unknown

R commands

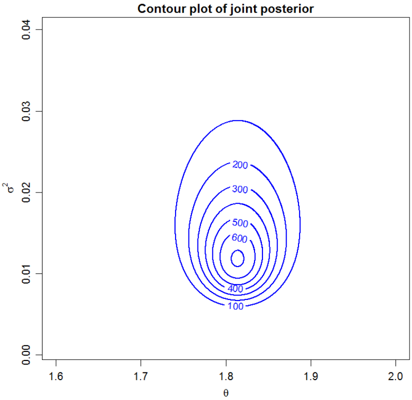
Calculate the posterior parameters.

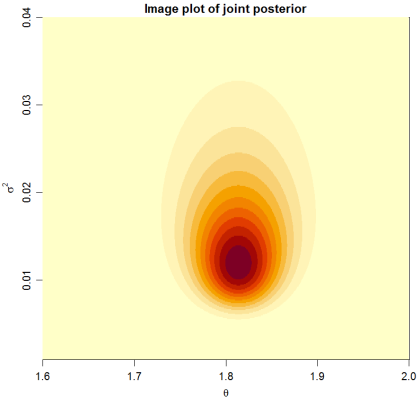
```
> y <- c(1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)
> n <- length(y)
> ybar <- mean(y)
> s2 <- var(y)
> mu0 <- 1.9
> sigma02 <- 0.01
> n0 <- 1
> nu0 <- 1
> (n1 <- n + n0)
[1] 10
> (mu1 <- (n*ybar+n0*mu0)/n1)
[1] 1.814
> (nu1 <- n + nu0)
[1] 10
> (sigma12 <- (nu0*sigma02 + (n-1)*s2 + n*n0*(ybar-mu0)^2/n1)/nu1)
[1] 0.015324
```

3.3 Inference when both mean and variance are unknown

Example: Midge wing length

- The contour and image plots of the joint posterior $p(\theta, \sigma^2 | \boldsymbol{y})$ are as follows. `R` codes to generate the plots can be found in the file `Chapter3Rcodes.R`.





Outline

The normal distribution

Inference for mean when variance is known

Inference when both mean and variance are unknown

Noninformative priors

- Improper priors
- Reference prior
- Jeffreys prior

Concluding Remarks

3.4 Noninformative priors

- In Bayesian inference, unknown parameters are regarded as random variables. Whatever beliefs we may have about these parameters before examining the data is quantified using a prior distribution.
- Determining an appropriate form of the prior can be a difficult task. Typically, these distributions are specified based on information from past studies or expert opinions. To simplify the computation, the prior is often limited to some familiar distributional family.
- In some cases, no prior information exists or inference based dominantly on the data is desired. Suppose we could find a distribution $p(\theta)$ that contained "no information" about θ in the sense that it did not favor one θ value over another. We might refer to such a distribution as a noninformative prior for θ.

3.4 Noninformative priors

- For example, suppose the parameter space is discrete and finite, i.e. $\Theta = \{\theta_1, \ldots, \theta_n\}$. Then the distribution

  $$p(\theta_i) = 1/n, \quad i = 1, \ldots, n,$$

  does not favor any one candidate θ value over any other, and, as such, is noninformative for θ.

- If instead we have a bounded continuous parameter space, $\Theta = [a, b]$, $-\infty < a < b < \infty$, then the uniform distribution $p(\theta) = 1/(b - a)$ for $a \leq \theta \leq b$ is arguably noninformative for θ (though this conclusion can be questioned as we will see later).

3.4.1 Improper priors

- When the parameter space Θ is unbounded, the situation is even less clear. Suppose $\Theta = (-\infty, +\infty)$.
- In the case where we observe $y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$ with unknown mean θ and known variance σ^2, we considered a prior $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$.
- The resulting posterior is $\theta | \boldsymbol{y} \sim \mathrm{N}(\mu_n, \tau_n^2)$ where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \text{ and } \mu_n = \tau_n^2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right). \tag{2}$$

- As $\tau_0^2 \to \infty$, the posterior distribution approaches $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ and the likelihood becomes dominant in determining the posterior.

3.4.1 Improper priors

- As for the prior, we can imagine it becoming flatter as the variance τ_0^2 increases so that when $\tau_0^2 \to \infty$, it is essentially flat or uniform over the whole real line and the location of μ_0 no longer matters.
- This notion of being uniform over the whole real line can be described as
$$p(\theta) = c, \quad -\infty < \theta < \infty,$$
for some constant $c > 0$. In application, we simply use the notation $p(\theta) \propto 1$.
- This distribution is **improper** as $\int_{-\infty}^{\infty} p(\theta) d\theta = +\infty$ for any $c > 0$ (Any proper pdf must integrate to 1).
- In particular, substituting $\tau_0^2 = \infty$ into the normal density yields $p(\theta) = 0$ which cannot be a proper density.

3.4.1 Improper priors

- Even though improper priors may appear inappropriate, sometimes, an improper prior can be combined with the likelihood to give a proper posterior.

- If $p(\theta) = c$ for some $c > 0$, $-\infty < \theta < +\infty$,

$$p(\theta|y) = \frac{p(y|\theta) \cdot c}{\int p(y|\theta) \cdot c \, \mathrm{d}\theta} = \frac{p(y|\theta)}{\int p(y|\theta) \, \mathrm{d}\theta}.$$

  Thus the posterior is proper provided $\displaystyle\int p(y|\theta) \, \mathrm{d}\theta$ is finite.

- Note: $\displaystyle\int p(y|\theta) \, \mathrm{d}y = 1$ for any $\theta \in \Theta$ but $\displaystyle\int p(y|\theta) \, \mathrm{d}\theta$ is not necessarily finite.

3.4.1 Improper priors

- In the problem of inferring θ when σ^2 is known, if we assume $p(\theta) \propto 1$,

  $$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta) \propto \exp\left\{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2\right\}.$$

  The posterior is $\theta|\boldsymbol{y} \sim \mathrm{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$, which is a proper density.

- We might say that $p(\theta) \propto 1$ is a "noninformative prior" for θ and argue that all of the information resulting in the posterior arose from the data.

- Care must be taken when using improper priors, since proper posteriors will not always result.

- In some models, the information in the data may be insufficient to identify all the parameters and, consequently, at least some of the prior distributions on the individual parameters must be informative.

3.4.2 Reference prior

- Noninformative priors are closely related to the notion of a reference prior. There are some subtle differences among authors on the precise meaning of this term. We will follow Box and Tiao (1973), who suggest that all that is important is that the data should dominate whatever information is contained in the reference prior, since as long as this happens, the precise form of the prior is unimportant.

- Previously, we have discussed the uniform prior and how it is arguably noninformative in the sense that it does not favor any value. Then, is the uniform prior always a good "reference"?

3.4.2 Reference prior

- Unfortunately, the answer is no, as it is not always invariant under reparameterization.
- As an example, suppose we claim ignorance concerning a univariate parameter θ defined on $[0,1]$ and adopt the prior, $p(\theta) = 1$, $\theta \in [0,1]$.
- A sensible reparameterization is $\gamma = \log(\theta)$, since this converts the support of the parameter to the real line.
- The prior on γ is given by $p_\gamma(\gamma) = p(\theta)|J| = |J|$, where $J = \dfrac{\mathrm{d}\theta}{\mathrm{d}\gamma}$, the Jacobian of the inverse transformation. Here, $\theta = e^\gamma$ so $J = e^\gamma$ and

$$p_\gamma(\gamma) = e^\gamma, \;\; -\infty < \gamma < 0,$$

  a prior that is clearly not uniform.

3.4.2 Reference prior

- Hence, using uniformity as a universal definition of prior ignorance, it is possible that "ignorance about θ" does not imply "ignorance about γ".
- A possible remedy to this problem is to rely on the particular modeling context to provide the most reasonable parameterization and subsequently, apply the uniform prior on this scale.

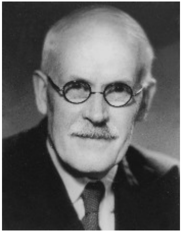
3.4.3 Jeffreys prior

The Jeffreys prior offers an easy-to-compute alternative that is invariant to transformation, named after Sir Harold Jeffreys (1891-1989).

- If the model only has a univariate parameter θ, this prior is given by

  $$p(\theta) \propto \sqrt{I(\theta)},$$

  where $I(\theta)$ is the expected Fisher information in the model, namely,



$$I(\theta) = -\mathsf{E}_{\mathbf{Y}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{y}|\theta)\right]$$

Note that the form of the likelihood helps to determine the prior but not the actual observed data since we are averaging over Y.

3.4.3 Jeffreys prior

- It can be shown that the Jeffreys prior is invariant to 1-1 transformation. Thus if $\gamma = g(\theta)$ is a 1-1 transformation, then

  $$\sqrt{I(\gamma)} = \sqrt{I(\theta)} \, \left| \frac{\mathrm{d}\theta}{\mathrm{d}\gamma} \right|$$

  so that computing the Jeffreys prior for γ directly produces the same answer as computing the Jeffreys prior for θ and then performing the usual Jacobian transformation to the γ scale.

3.4.3 Jeffreys prior

- If $\boldsymbol{\theta}$ is multi-dimensional, the Jeffreys prior is given by

  $$p(\boldsymbol{\theta}) \propto \sqrt{\det\{I(\boldsymbol{\theta})\}},$$

  where $\det(\cdot)$ denotes the determinant, and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix, whose (i, j)th entry is

  $$I_{ij}(\boldsymbol{\theta}) = -\mathrm{E}_{\mathbf{Y}|\theta}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\boldsymbol{y}|\boldsymbol{\theta})\right].$$

- This approach provides a recipe for obtaining noninformative priors, but can be cumbersome to use if the dimension of $\boldsymbol{\theta}$ is high. A more common approach is to obtain a noninformative prior for each parameter individually and then form the joint prior as the product of these individual priors.

3.4.3 Jeffreys prior

Example: Normal model

Suppose that $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$. We can show the following:

- If θ is unknown and σ is known, then the Jeffreys prior for θ is given by $p(\theta) \propto 1$, for $\theta \in \mathbb{R}$.
- If θ is known and σ is unknown, then the Jeffreys prior for σ is given by $p(\sigma) \propto 1/\sigma$, for $\sigma > 0$.
- If θ and σ are both unknown, then the bivariate Jeffreys prior is given by $p(\theta, \sigma) \propto 1/\sigma^2$, for $\theta \in \mathbb{R}$, $\sigma > 0$. Note that this is slightly different from the joint prior obtained by simply multiplying the two individual Jeffreys priors for θ and σ^2.

Note that here the Jeffreys priors here are described for σ, not σ^2.

3.4.3 Jeffreys prior

Example: Normal model

Derivation of Jeffreys priors:

Define the log-likelihood function

$$\ell(\theta, \sigma) = \log p(\boldsymbol{y}|\theta, \sigma) = -\frac{1}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}.$$

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^n (y_i - \theta)}{\sigma^2}, \qquad \frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (y_i - \theta)^2}{\sigma^3},$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{n}{\sigma^2}, \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3\sum_{i=1}^n (y_i - \theta)^2}{\sigma^4}, \qquad \frac{\partial^2 \ell}{\partial \theta \partial \sigma} = -\frac{2\sum_{i=1}^n (y_i - \theta)}{\sigma^3}.$$

3.4.3 Jeffreys prior

Example: Normal model

Derivation of Jeffreys priors (cont'd):

- If θ is unknown and σ is known,

$$p(\theta) \propto \sqrt{I(\theta)} = \sqrt{-\mathsf{E}_{\boldsymbol{y}|\theta}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right]} = \sqrt{-\mathsf{E}_{\boldsymbol{y}|\theta}[-\frac{1}{\sigma^2}]} = \frac{1}{\sigma} \propto 1.$$

- If θ is known and σ is unknown, then we use the fact $\mathsf{E}_{y_i|\sigma}(y_i - \theta)^2 = \sigma^2$ and obtain that

$$\begin{aligned} p(\sigma) \propto \sqrt{I(\sigma)} &= \sqrt{-\mathsf{E}_{\boldsymbol{y}|\sigma}\left[\frac{\partial^2 \ell}{\partial \sigma^2}\right]} \\ &= \sqrt{-\mathsf{E}_{\boldsymbol{y}|\sigma}\left[\frac{n}{\sigma^2} - \frac{3\sum_{i=1}^{n}(y_i - \theta)^2}{\sigma^4}\right]} \\ &= \frac{\sqrt{2n}}{\sigma} \propto \frac{1}{\sigma}. \end{aligned}$$

3.4.3 Jeffreys prior

Example: Normal model

Derivation of Jeffreys priors (cont'd):

- If both θ and σ are unknown, then

$$\begin{aligned} I(\theta,\sigma) &= -\mathsf{E}_{\boldsymbol{y}|\theta,\sigma}\begin{bmatrix} \frac{\partial^2\ell}{\partial\theta^2} & \frac{\partial^2\ell}{\partial\theta\partial\sigma} \\ \frac{\partial^2\ell}{\partial\sigma\partial\theta} & \frac{\partial^2\ell}{\partial\sigma^2} \end{bmatrix} \\ &= \mathsf{E}_{\boldsymbol{y}|\theta,\sigma}\begin{bmatrix} \frac{n}{\sigma^2} & \frac{2\sum_{i=1}^n(y_i-\theta)}{\sigma^3} \\ \frac{2\sum_{i=1}^n(y_i-\theta)}{\sigma^3} & -\frac{n}{\sigma^2}+\frac{3\sum_{i=1}^n(y_i-\theta)^2}{\sigma^4} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}. \end{aligned}$$

Therefore, $p(\theta,\sigma) \propto \sqrt{\det\{I(\theta,\sigma)\}} = \sqrt{\frac{2n^2}{\sigma^4}} \propto \frac{1}{\sigma^2}$. Thus the Jeffreys prior $p(\theta,\sigma)$ is of a different form compared to the product of two independent Jeffreys priors $p(\theta)p(\sigma) \propto \frac{1}{\sigma}$.

Outline

The normal distribution

Inference for mean when variance is known

Inference when both mean and variance are unknown

Noninformative priors

- Improper priors
- Reference prior
- Jeffreys prior

Concluding Remarks

3.5 Concluding Remarks

- The normal distribution with both mean and variance unknown is an example of a multi-parameter model.
- While we have considered mainly conjugate prior distributions so far in order to make posterior computations simple, sometimes non-conjugate priors are preferred if they are able to better represent the available prior information.
- In some situations, the posterior distribution may no longer be a distribution from any known families. Monte Carlo or more advanced computational methods are required to derive and summarize the posterior.
- We will consider more examples of multi-parameter models when we cover Bayesian computational methods.