Likelihood & Bayesian Inference in the Normal Model

(please report typos to gustavoc@msu.edu)

If y_i follows a Normal (aka Gaussian) distribution the pdf is

$$p(y_i) \sim \frac{e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}}{\sqrt{2\pi\sigma^2}} \qquad [1]$$

Or, in compact notation $y_i \sim N(\mu, \sigma^2)$ or, alternatively,$y_i \sim N(y_i|0, \sigma^2)$.

The distribution is indexed by two parameters μ and σ^2 $(\sigma^2 > 0)$ which are also the mean and variance of the RV, that is $E(y_i) = \int y_i p(y_i) dy_i = \mu$ and $Var(y_i) = \sigma^2$.

The R-functions `dnorm()`, `pnorm()` and `qnorm()` provide the density, cdf and quantiles of the normal distribution, respectively and `rnorm()` produces samples from the distribution.

Goal: to infer μ and σ^2 using data from a finite sample $Y = \{y_1, \dots, y_n\}$. We will focus on Bayesian inference but also briefly consider likelihood inference.

The Likelihood Function is the joint probability of the data viewed as a function of the parameters. If data is conditionally independent, then

$$p(y_1, \dots, y_n|\mu, \sigma^2) = p(y_1|\mu, \sigma^2)p(y_2|\mu, \sigma^2) \dots . p(y_n|\mu, \sigma^2)$$

Furthermore, if the data is IID Gaussian (each following [1]) we have

$$p(y_1, \dots, y_n|\mu, \sigma^2) = \frac{e^{-\frac{1}{2\sigma^2}(y_1-\mu)^2}}{\sqrt{2\pi\sigma^2}} \times \frac{e^{-\frac{1}{2\sigma^2}(y_2-\mu)^2}}{\sqrt{2\pi\sigma^2}} \times \dots \times \frac{e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2}}{\sqrt{2\pi\sigma^2}} = \prod_{i=1}^{n} \frac{e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}}{\sqrt{2\pi\sigma^2}}$$

Combining terms, we get

$$p(y_1, \dots, y_n|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{RSS(y,\mu)}{2\sigma^2}} \qquad [2]$$

where

$$RSS(y, \mu) = \sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu n\bar{y} \, . \qquad [3]$$

Maximum Likelihood (ML) Estimation

The steps we often follow to find analytic solutions to ML problems are: (i) take the log of [2] (because the log is a monotonic transformation, maximizing the likelihood or the log-likelihood renders the same solution; however, maximization of the log-likelihood is often easier), (ii) differentiate the log-likelihood with respect to each of the parameters of interest (μ and σ^2, in this problem), (iii) set the two derivatives equal to zero (FOC=first order conditions) and (iv) solve the FOC for the parameters. Technically we also. need to check the 2nd derivatives to be sure that the stationary point defined by the FCO corresponds to a maximum and not a minimum. However, for the likelihood problems that we will work in this 2nd order conditions will always be satisfied because the likelihoods we will considered are concave.

(i) $l = \log\{p(y_1, \dots, y_n|\mu, \sigma^2)\} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS(y, \mu)$

(ii) $\frac{dl}{d\mu} = -\frac{1}{2\sigma^2}[2n\mu - 2n\bar{y}]$ [4a]

and $\frac{dl}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{RSS(y,\mu)}{2(\sigma^2)^2}$ [4b]

(iii) Setting [4a] equal to zero renders: $\hat{\mu}_{ML} = \bar{y}$. Plugging this into [4b] and solving for the variance gives $\frac{n}{2\hat{\sigma}^2_{ML}} = \frac{RSS(y,\bar{y})}{2\left(\hat{\sigma}^2_{ML}\right)^2} \leftrightarrow n = \frac{RSS(y,\bar{y})}{\hat{\sigma}^2_{ML}} \leftrightarrow \hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}$

Thus: $\hat{\mu}_{ML} = \bar{y}$ and $\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}$ [5]

Bayesian Inference

In Bayesian models we make inferences about the unknown parameters using the posterior distribution $p(\mu, \sigma^2|y_1, \dots, y_n)$. According to Bayes' theorem

$$p(\mu, \sigma^2|y_1, \dots, y_n) = \frac{p(\mu,\sigma^2,y_1,\dots,y_n)}{p(y_1,\dots,y_n)} = \frac{p(y_1,\dots,y_n|\mu,\sigma^2)p(\mu,\sigma^2)}{p(y_1,\dots,y_n)}$$

Since the marginal distribution of the data, $p(y_1, \dots, y_n)$, does not involve the unknown parameters, we have

$$p(\mu, \sigma^2|y_1, \dots, y_n) \propto p(y_1, \dots, y_n|\mu, \sigma^2)p(\mu, \sigma^2) \qquad [6]$$

The likelihood function, $p(y_1, \dots, y_n|\mu, \sigma^2)$, is given by expression [2].

Prior distribution: here we will assume independent priors that is $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$. We will choose conjugate priors that will give closed-form fully-conditional distributions. Specifically, we will assign a normal prior to the mean, $\mu \sim N(\mu_0, \sigma_0^2)$, and a scaled-inverse chi square to the variance parameter, $\sigma^2 \sim \chi^{-2}(S, df)$. That is:

$$p(\mu,\sigma^2) = \frac{e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}}{\sqrt{2\pi\sigma_0^2}} \times \frac{\left(\frac{S}{2}\right)^{\frac{df}{2}}}{\Gamma\left(\frac{df}{2}\right)}[\sigma^2]^{1-\frac{df}{2}} e^{-\frac{df\times S}{2\sigma^2}} \qquad [7]$$

Replacing the likelihood [2] and the prior [7] into the right-hand side of [6] we obtain the following expression for the joint-posterior distribution

$$p(\mu,\sigma^2|y_1,\dots,y_n) \propto (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{RSS(y,\mu)}{2\sigma^2}} \times \frac{e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}}{\sqrt{2\pi\sigma_0^2}} \times \frac{\left(\frac{S}{2}\right)^{\frac{df}{2}}}{\Gamma\left(\frac{df}{2}\right)}[\sigma^2]^{1-\frac{df}{2}} e^{-\frac{df\times S}{2\sigma^2}} \qquad [8]$$

The posterior distribution does not have a closed form. However, samples from the posterior distribution can be drawn using either composition sampling (this is the treatment made in the book) or using a Gibbs sampler; we discuss this approach next.

Gibbs Sampler

In a Gibbs sampler, we draw samples from a multivariate distribution by recursively sampling from fully-conditional distributions. Suppose we want to sample random variables W1 and W2 from the joint distribution $p(W_1, W_2)$. In a Gibbs sampler we achieve this by recursively sampling from $p(W_1|W_2)$ and $p(W_2|W_1)$.

For the Bayesian Gaussian model in [8] implementing a Gibbs sampler requires us to derive the fully-conditional distribution of the mean, $p(\mu|y_1,\dots,y_n,\sigma^2)$ and that of the variance, $p(\sigma^2|y_1,\dots,y_n,\mu)$. Both can be derived from the joint posterior, [8]. The strategy to derive fully-conditional distributions is as follows: (i) remove from the joint posterior all the terms that do not involve the random variable of interest, (ii) combine terms and, (iii) inspect whether the resulting object has the form of a known distribution.

Fully-conditional distribution for the mean. Removing from the right-hand-side of [8] all the terms that do not involve μ we get:

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto e^{-\frac{RSS(y,\mu)}{2\sigma^2}} \times e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \propto e^{-\frac{1}{2}\left[\frac{RSS(y,\mu)}{2\sigma^2}+\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]}$$

Furthermore, since $RSS(y,\mu) = \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu n\bar{y}$ and $(\mu-\mu_0)^2 = \mu^2 + \mu_0^2 - 2\mu\mu_0$

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto e^{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2+n\mu^2-2\mu n\bar{y}}{\sigma^2}+\frac{\mu^2+\mu_0^2-2\mu\mu_0}{\sigma_0^2}\right]}$$

We now combine the two quadratic forms, $\left[\frac{\sum_{i=1}^{n} y_i^2+n\mu^2-2\mu n\bar{y}}{\sigma^2} + \frac{\mu^2+\mu_0^2-2\mu\mu_0}{\sigma_0^2}\right]$

$$e^{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2+n\mu^2-2\mu n\bar{y}}{\sigma^2}+\frac{\mu^2+\mu_0^2-2\mu\mu_0}{\sigma_0^2}\right]} = e^{-\frac{1}{2}\left[\frac{\mu^2-2\mu n\bar{y}}{\sigma^2}+\frac{\mu^2-2\mu\mu_0}{\sigma_0^2}\right]} e^{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right]}$$
$$\propto e^{-\frac{1}{2}\left[\frac{n\mu^2-2\mu n\bar{y}}{\sigma^2}+\frac{\mu^2-2\mu\mu_0}{\sigma_0^2}\right]}$$

(Note: above we have removed $e^{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right]}$, this can be done because the expression does not involve μ.)

Now

$$\frac{n\mu^2-2\mu n\bar{y}}{\sigma^2}+\frac{\mu^2-2\mu\mu_0}{\sigma_0^2} = \mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)-2\mu\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\mu^2-2\mu\frac{\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right]$$

Thus,

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto e^{-\frac{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}{2}\left[\mu^2-2\mu\frac{\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right]}$$

Let $V_{\mu|y} = \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$ and mean $E_{\mu|y} = \frac{\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$, thus

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto e^{-\frac{1}{2V_{\mu|y}}\left[\mu^2-2\mu E_{\mu|y}\right]} \qquad [9]$$

Since neither $E_{\mu|y} = \frac{\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$ nor $V_{\mu|y} = \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$ involve μ we can multiply the right-hand-side of [9] by by $e^{-\frac{E_{\mu|y}^2}{2V_{\mu|y}}}$, thus,

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto e^{-\frac{1}{2V_{\mu|y}}\left[\mu^2-2\mu E_{\mu|y}\right]} e^{-\frac{E_{\mu|y}^2}{2V_{\mu|y}}} \propto e^{-\frac{1}{2V_{\mu|y}}\left[\mu^2-2\mu E_{\mu|y}+E_{\mu|y}^2\right]}$$
$$\propto e^{-\frac{1}{2V_{\mu|y}}\left[\mu-E_{\mu|y}\right]^2}$$

The right-hand-side expression can be recognized as the kernel of a normal distribution with variance $V_{\mu|y}$ and mean $E_{\mu|y}$; therefore, we conclude that the fully-conditional distribution of the mean is

$$p(\mu|y_1, \dots, y_n, \sigma^2) = N\left(E_{\mu|y}, V_{\mu|y}\right) \text{ where } E_{\mu|y} = \frac{\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)} \text{ and } V_{\mu|y} = \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)} \qquad [10]$$

Fully-conditional distribution for the variance. Removing from the right-hand-side of [8] all the terms that do not involve μ we get:

$$p(\sigma^2|y_1, \dots, y_n, \mu) \propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{RSS(y,\mu)}{2\sigma^2}} [\sigma^2]^{1-\frac{df}{2}} e^{-\frac{df \times S}{2\sigma^2}}$$

Combining terms, we get

$$p(\sigma^2|y_1, \dots, y_n, \mu) \propto (\sigma^2)^{1-\frac{df+n}{2}} e^{-\frac{RSS(y,\mu)+df \times S}{2\sigma^2}}$$

or

$$p(\sigma^2|y_1, \dots, y_n, \mu) \propto (\sigma^2)^{1-\frac{df+n}{2}} e^{-\frac{(n+df)\frac{(RSS(y,\mu)+df \times S)}{(n+df)}}{2\sigma^2}}$$

The righ-hand side can be recognized as the kernel of a scaled-inverse Chi-squared density with degree of freedom $df + n$ and scale parameter $\frac{(RSS(y,\mu)+df \times S)}{(n+df)}$; therefore

$$p(\sigma^2|y_1, \dots, y_n, \mu) = \chi^{-2}\left(df + n, \frac{(RSS(y,\mu)+df \times S)}{(n+df)}\right) \qquad [11]$$

A Gibbs sampler for the normal model will proceed recursively sampling from [10] and [11]. An example is provided in the GitHub repository

https://github.com/gdlc/STT465/blob/master/NormalModel_MeanAndVariance.md