STAT 431 — Applied Bayesian Analysis — Course Notes

One-Parameter Normal Models

Spring 2019

- Model:

$$Y_1, \dots Y_n \mid \mu, \sigma^2 \quad \sim \quad \text{i.i.d. } \mathrm{N}(\mu, \sigma^2)$$

Let

$$\begin{aligned}
\boldsymbol{Y} &= (Y_1, \dots Y_n) \\
\boldsymbol{y} &= (y_1, \dots y_n) \qquad \text{(observation of } \boldsymbol{Y}\text{)} \\
\bar{y} &= \frac{1}{n} \sum_i y_i \quad = \quad \text{usual estimate of } \mu
\end{aligned}$$

The **precision** is defined as

$$\tau^2 \;=\; 1/\sigma^2$$

which

- measures *concentration*, not spread
- can lead to less complicated derivations (later)
- is used in an alternative parameterization, especially in some Bayesian software

Known Variance

Assume σ^2 (or τ^2) is known.

- Likelihood

  Joint density of $\boldsymbol{Y}$:

$$
\begin{aligned}
p(\boldsymbol{y} \mid \mu) &= \prod_i p(y_i \mid \mu) \\
&= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\
&\propto e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2}
\end{aligned}
$$

  (where the proportionality is in μ)

Can show

$$\sum_i (y_i - \mu)^2 \quad = \quad \sum_i (y_i - \bar{y})^2 \ + \ n(\bar{y} - \mu)^2$$

so

$$\begin{aligned} L(\mu; \boldsymbol{y}) \quad &\propto \quad e^{-\frac{1}{2\sigma^2}\left(\sum_i (y_i - \bar{y})^2 \,+\, n(\bar{y}-\mu)^2\right)} \\ &= \quad e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\bar{y})^2} \ \cdot \ e^{-\frac{1}{2\sigma^2} n(\bar{y}-\mu)^2} \\ &\propto \quad e^{-\frac{1}{2\sigma^2} n(\mu-\bar{y})^2} \end{aligned}$$

(where the proportionality is in μ)

Note: $\bar{y}$ is a **sufficient statistic**. (How can you tell?)

[Draw likelihood ...]

- Conjugate Prior

$$\mu \quad \sim \quad \mathrm{N}(\mu_0, \sigma_0^2) \quad = \quad \mathrm{N}\big(\mu_0, 1/\tau_0^2\big)$$

Why is this conjugate? Let's derive the posterior ...

$$\begin{aligned} p(\mu \mid \boldsymbol{y}) &\propto p(\mu)\, L(\mu; \boldsymbol{y}) \\ &\propto e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \cdot e^{-\frac{1}{2\sigma^2}n(\mu-\bar{y})^2} \\ &= e^{-\frac{1}{2}\left(\tau_0^2(\mu-\mu_0)^2 + n\tau^2(\mu-\bar{y})^2\right)} \end{aligned}$$

The exponent is a concave quadratic function of μ, and thus the expression is the kernel of a normal distribution.

Next we identify the posterior mean and variance ...

$$\tau_0^2(\mu-\mu_0)^2 + n\tau^2(\mu-\bar{y})^2$$

$$= (\tau_0^2 + n\tau^2)\mu^2 - 2(\tau_0^2\mu_0 + n\tau^2\bar{y})\mu + \text{constant (without } \mu\text{)}$$

$$
\begin{aligned}
\tau_0^2(\mu-\mu_0)^2 + n\tau^2(\mu-\bar{y})^2 \\
&= (\tau_0^2 + n\tau^2)\mu^2 \;-\; 2(\tau_0^2\mu_0 + n\tau^2\bar{y})\mu \\
&\quad + \text{ constant (without } \mu) \\
&= \cdots \;\; \text{(complete the square)} \;\; \cdots \\
&= (\tau_0^2 + n\tau^2)\left(\mu \;-\; \frac{\tau_0^2\mu_0 + n\tau^2\bar{y}}{\tau_0^2 + n\tau^2}\right)^2 \\
&\quad + \text{ constant (without } \mu)
\end{aligned}
$$

$$
\begin{aligned}
\tau_0^2(\mu-\mu_0)^2 + n\tau^2(\mu-\bar{y})^2 & \\
&= (\tau_0^2+n\tau^2)\mu^2 \;-\; 2(\tau_0^2\mu_0+n\tau^2\bar{y})\mu \\
&\qquad + \text{ constant (without } \mu) \\
&= \cdots \;\; \text{(complete the square)} \;\; \cdots \\
&= (\tau_0^2+n\tau^2)\left(\mu \;-\; \frac{\tau_0^2\mu_0+n\tau^2\bar{y}}{\tau_0^2+n\tau^2}\right)^2 \\
&\qquad + \text{ constant (without } \mu) \\
&= \tau_1^2(\mu-\mu_1)^2 \quad + \text{ constant (without } \mu)
\end{aligned}
$$

where

$$
\tau_1^2 \;=\; \tau_0^2+n\tau^2 \qquad\qquad \mu_1 \;=\; \frac{\tau_0^2\mu_0+n\tau^2\bar{y}}{\tau_0^2+n\tau^2}
$$

So we find

$$p(\mu \mid \boldsymbol{y}) \;\; \propto \;\; e^{-\frac{1}{2}\tau_1^2(\mu-\mu_1)^2}$$

which we recognize as the kernel of a $\mathrm{N}(\mu_1, 1/\tau_1^2)$:

$$\mu \mid \boldsymbol{y} \;\; \sim \;\; \mathrm{N}(\mu_1, 1/\tau_1^2)$$

So

$$\mathrm{E}(\mu \mid \boldsymbol{y}) \;\; = \;\; \mu_1 \qquad\qquad \mathrm{Var}(\mu \mid \boldsymbol{y}) \;\; = \;\; 1/\tau_1^2 \;\; \equiv \;\; \sigma_1^2$$

The posterior mean estimate of μ is thus μ_1, with a posterior standard deviation of σ_1.

(Notice: The posterior depends on the data only through the sufficient statistic $\bar{y}$.)

Notice that μ_1 is a weighted average of the prior mean μ_0 and sample average $\bar{y}$:

$$\mu_1 \quad = \quad \frac{\tau_0^2}{\tau_0^2 + n\tau^2}\,\mu_0 \;\; + \;\; \frac{n\tau^2}{\tau_0^2 + n\tau^2}\,\bar{y}$$

(What happens as $\tau_0^2 \to 0$? As $n \to \infty$?)

Notice that μ_1 is a weighted average of the prior mean μ_0 and sample average $\bar{y}$:

$$\mu_1 \quad = \quad \frac{\tau_0^2}{\tau_0^2 + n\tau^2}\,\mu_0 \;\; + \;\; \frac{n\tau^2}{\tau_0^2 + n\tau^2}\,\bar{y}$$

(What happens as $\tau_0^2 \to 0$? As $n \to \infty$?)

Note: μ_1 is generally biased as an estimator of μ. (Why?)

Letting $n_0 = \tau_0^2/\tau^2$,

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

so n_0 behaves like a "prior sample size" for "prior average" μ_0.

Also, $\tau_1^2 = (n_0+n)\,\tau^2$. (What if $n \to \infty$?)

Example: Jevons's Coin Data

- coins (gold sovereigns) collected in England ca. 1870
- legal standard weight: 7.9876 g
- min. legal weight: 7.9379 g

For $n = 24$ coins minted before 1830,

$$\begin{aligned} \bar{y} &= \text{avg. wt.} = 7.8730 \text{ g} \\ s &= \text{sample std. dev.} = 0.05353 \text{ g} \end{aligned}$$

For illustration, let's *assume*

$$\sigma^2 = s^2 = (0.05353)^2$$

Let's take a normal prior with

$$\begin{aligned} \mu_0 &= \text{standard weight} = 7.9876 \\ \sigma_0^2 &= (0.025)^2 \end{aligned}$$

(so that σ_0 is about half the difference between the standard and minimum legal weights)

How informative is this prior?

$$n_0 = \frac{\tau_0^2}{\tau^2} = \frac{1/(0.025)^2}{1/(0.05353)^2} \approx 4.6$$

so equivalent to 4 or 5 "prior observations."

Posterior is normal with

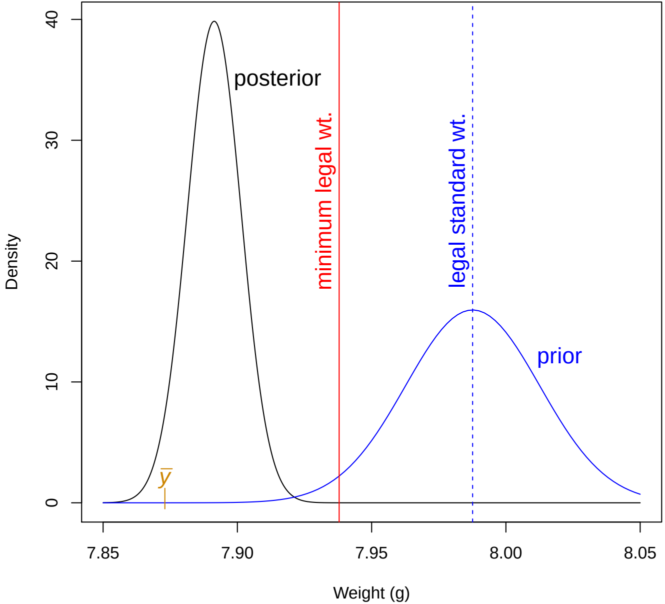
$$\mu_1 = \frac{\tau_0^2\mu_0 + n\tau^2\bar{y}}{\tau_0^2 + n\tau^2} \approx 7.891381$$

$$\sigma_1^2 = \frac{1}{\tau_0^2 + n\tau^2} \approx 0.0001002444$$

$$(\sigma_1 \approx 0.01001221)$$

So $\bar{y} = 7.8730$ is barely within 2 posterior standard deviations of the posterior mean.

Perhaps our prior is a bit too informative (too much bias)?



posterior
minimum legal wt.
legal standard wt.
prior
$\bar{y}$
Density
Weight (g)
0
10
20
30
40
7.85
7.90
7.95
8.00
8.05

- Jeffreys Prior

Notice:

$$\begin{aligned} \ln p(\boldsymbol{y} \mid \mu) &= -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 + \text{constant (no } \mu) \\ &= \text{a strictly concave quadratic in } \mu \end{aligned}$$

and the coefficient of μ^2 doesn't depend on $\boldsymbol{y}$.

Thus

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2} \ln p(\boldsymbol{y} \mid \mu) &= \text{twice the coefficient of } \mu^2 \\ &= \text{negative constant (no } \mu \text{ or } \boldsymbol{y}) \end{aligned}$$

So the Fisher information is

$$
\begin{aligned}
I(\mu) &= -\,\mathrm{E}\left(\frac{\partial^2}{\partial \mu^2} \ln p(\boldsymbol{Y} \mid \mu) \;\middle|\; \mu \right) \\
&= \text{some positive constant (no } \mu\text{)}
\end{aligned}
$$

and the Jeffreys prior is

$$
p(\mu) \;\propto\; \sqrt{I(\mu)} \;\propto\; 1 \quad (-\infty < \mu < \infty)
$$

That is, the Jeffreys prior for μ is "flat."

(Is this prior *proper* or *improper*?)

We must check that the posterior is proper ...

$$\begin{aligned} p(\mu \mid \boldsymbol{y}) \;&\propto\; \underbrace{p(\mu)}_{\propto 1} p(\boldsymbol{y} \mid \mu) \\ &\propto\; L(\mu;\, \boldsymbol{y}) \;\propto\; e^{-\frac{n}{2\sigma^2}\,(\mu-\bar{y})^2} \end{aligned}$$

Recognize as the kernel of $\mathrm{N}(\bar{y},\, \sigma^2/n)$ (why?), so the posterior is indeed proper:

$$\mu \mid \boldsymbol{y} \;\sim\; \mathrm{N}(\bar{y},\, \sigma^2/n)$$

Note: The posterior mean is $\bar{y}$ and the posterior standard deviation is the (frequentist) standard error of $\bar{y}$.

Can show that, under this Jeffreys prior,

- credible intervals for μ are the same as confidence intervals
- the posterior probability of a *one-sided* H_0 is the same as a p-value

  (Not true for the two-sided case.)
- this posterior is the limit as $\tau_0^2 \to 0$ (equiv. $n_0 \to 0$) in the conjugate case

(Note typo in Cowles, formula (6.6), p. 92.)

Notation: For a flat prior, write, e.g.

$$\mu \;\sim\; 1\, d\mu$$

Eg: Jevons's Coin Data

Posterior under Jeffreys prior:

$$\mu \mid \boldsymbol{y} \quad \sim \quad \mathrm{N}(\mu_1 = \bar{y} = 7.8730,\ \sigma_1^2 = \sigma^2/n \approx 0.0001194)$$

(so $\sigma_1 \approx 0.01093$)

[Draw density w/ probability limits ...]

So an approx. 95% credible interval is

$$\mu_1 \;\pm\; 1.96\,\sigma_1 \quad \approx \quad (7.8516,\; 7.8944)$$

It excludes all values meeting the min. legal weight (7.9379).

Indeed, for $\qquad H_0 :\; \mu \geq 7.9379 \qquad$ we find

$$\mathrm{P}(H_0 \mid \boldsymbol{y}) \;\;=\;\; 1 \;-\; \Phi\!\left(\frac{7.9379 - \mu_1}{\sigma_1}\right) \;\;\approx\;\; 10^{-9}$$

(same as a p-value, in this case)

- Posterior Predictive Distribution

  Let Y^* be a hypothetical new observation sampled independently of the data (conditional on μ).

  Then

$$Y^* \mid \mu \;\; = \;\; Y^* \mid \mu, \boldsymbol{y} \;\; \sim \;\; \mathrm{N}(\mu, \sigma^2)$$

  and we can write

$$Y^* \;\; = \;\; \mu \, + \, \varepsilon^* \qquad\qquad \varepsilon^* \; \sim \; \mathrm{N}(0, \sigma^2)$$

  where ε^* is independent of μ and $\boldsymbol{Y}$ (why?).

So

$$\begin{aligned} \mu \mid \boldsymbol{y} &\sim \mathrm{N}(\mu_1, \sigma_1^2) \\ \varepsilon^* \mid \boldsymbol{y} &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

and μ and ε^* are conditionally independent given $\boldsymbol{Y}$.

This makes it easy to find the posterior predictive distribution:

$$Y^* \mid \boldsymbol{y} \;=\; \mu + \varepsilon^* \mid \boldsymbol{y} \;\sim\; \mathrm{N}(\mu_1, \sigma_1^2 + \sigma^2)$$

(Why?)

Note: This distribution always has variance at least σ^2, no matter how small σ_1^2 is.

Eg: Jevons's Coin Data

Consider randomly selecting another coin of the same kind (minted before 1830). Its (random) weight will be Y^*.

Using the Jeffreys prior,

$$Y^* \mid \boldsymbol{y} \quad \sim \quad \mathrm{N}\big(7.8730,\, 0.0001194 + (0.05353)^2\big)$$

The posterior predictive standard deviation works out to be about 0.05463.

The posterior predictive prob. that *this coin* is of legal weight:

$$\begin{aligned} \mathrm{P}(Y^* \geq 7.9379 \mid \boldsymbol{y}) \quad &\approx \quad 1 \;-\; \Phi\left(\frac{7.9379 - 7.8730}{0.05463}\right) \\ &\approx \quad 0.1174 \end{aligned}$$

Known Mean

Assume μ is known, but not σ^2.

- Likelihood

$$
\begin{aligned}
p(\boldsymbol{y} \mid \sigma^2) &= \prod_i p(y_i \mid \sigma^2) \\
&= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2} \quad \propto \quad \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{n\nu}{2\sigma^2}}
\end{aligned}
$$

in terms of sufficient statistic

$$
\nu = \frac{1}{n}\sum_i (y_i - \mu)^2
$$

So

$$L(\sigma^2;\, \boldsymbol{y}) \quad \propto \quad \frac{1}{(\sigma^2)^{n/2}}\, e^{-\frac{n\nu}{2\sigma^2}} \qquad\qquad \sigma^2 > 0$$

[Draw likelihood ...]

(Can show ν is the MLE.)

- Conjugate Prior

We say X has an **inverse gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if it has (continuous) density

$$p(x) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)}\,\frac{1}{x^{\alpha+1}}\,e^{-\beta/x} \qquad x > 0$$

and write

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

If

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

it can be shown that

- $$1/X \sim \mathrm{gamma}(\alpha, \beta)$$

  (in the parameterization of Cowles, Table A.2)

- if $\alpha > 1$,

  $$\mathrm{E}(X) = \frac{\beta}{\alpha - 1}$$

- if $\alpha > 2$,

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

The inverse gamma distribution is a conjugate prior for σ^2:

Suppose

$$\sigma^2 \quad \sim \quad \mathrm{IG}(\alpha, \beta)$$

Then (for $\sigma^2 > 0$)

$$\begin{aligned} p(\sigma^2 \mid \boldsymbol{y}) \quad &\propto \quad \frac{1}{(\sigma^2)^{\alpha+1}}\, e^{-\beta/\sigma^2} \quad \cdot \quad \frac{1}{(\sigma^2)^{n/2}}\, e^{-\frac{n\nu}{2\sigma^2}} \\ &= \quad \frac{1}{(\sigma^2)^{\alpha + n/2 + 1}}\, e^{-(\beta + n\nu/2)/\sigma^2} \end{aligned}$$

which is the kernel of $\mathrm{IG}(\alpha + n/2,\, \beta + n\nu/2)$:

$$\sigma^2 \mid \boldsymbol{y} \quad \sim \quad \mathrm{IG}(\alpha + n/2,\, \beta + n\nu/2)$$

We could alternatively consider the *reparameterization*

$$\tau^2 \;=\; 1/\sigma^2$$

It follows that the prior

$$\tau^2 \;\sim\; \mathrm{gamma}(\alpha, \beta)$$

produces posterior

$$\tau^2 \mid \boldsymbol{y} \;\sim\; \mathrm{gamma}(\alpha + n/2,\, \beta + n\nu/2)$$

so the gamma distribution is conjugate for this situation.

- Jeffreys Prior

To derive the Fisher information ...

$$\ln p(\boldsymbol{y} \mid \sigma^2) \;=\; -\frac{n}{2}\ln\sigma^2 \;-\; \frac{n\nu}{2\sigma^2} \;+\; \text{constant}$$

so

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ln p(\boldsymbol{y} \mid \sigma^2) \;=\; \frac{n}{2(\sigma^2)^2} \;-\; \frac{n\nu}{(\sigma^2)^3}$$

and

$$\begin{aligned} I(\sigma^2) \;&=\; -\,\mathrm{E}\left(\frac{\partial^2}{\partial(\sigma^2)^2}\ln p(\boldsymbol{Y} \mid \sigma^2) \;\Big|\; \sigma^2\right) \\ &=\; -\,\frac{n}{2(\sigma^2)^2} \;+\; \frac{n\,\mathrm{E}(\mathcal{V} \mid \sigma^2)}{(\sigma^2)^3} \end{aligned}$$

Since

$$\begin{aligned}
\mathrm{E}(\mathcal{V} \mid \sigma^2) &= \mathrm{E}\left(\frac{1}{n}\sum_i (Y_i - \mu)^2 \;\middle|\; \sigma^2\right) \\
&= \frac{1}{n}\sum_i \mathrm{E}\big((Y_i - \mu)^2 \mid \sigma^2\big) \\
&= \frac{1}{n}\sum_i \sigma^2 \;=\; \sigma^2
\end{aligned}$$

we obtain

$$I(\sigma^2) \;=\; -\frac{n}{2(\sigma^2)^2} + \frac{n\sigma^2}{(\sigma^2)^3} \;=\; \frac{n}{2(\sigma^2)^2}$$

so the Jeffreys prior is

$$p(\sigma^2) \;\propto\; \sqrt{I(\sigma^2)} \;\propto\; \frac{1}{\sigma^2} \qquad \sigma^2 > 0$$

This Jeffreys prior is improper:

$$\int_0^\infty \frac{1}{\sigma^2} \, d\sigma^2 \;\; = \;\; \infty$$

[Draw prior curve area ...]

Therefore, one must verify that the posterior will be proper — see Cowles, Sec. 6.3.3.

Recall that the inverse gamma prior density for σ^2 has kernel

$$\frac{1}{(\sigma^2)^{\alpha+1}}\, e^{-\beta/\sigma^2}$$

We obtain the Jeffreys prior by setting "$\alpha = 0$" and "$\beta = 0$."

We conclude that the inverse gamma prior becomes less informative (more "vague") as we let α and β approach zero.

Since the Jeffreys prior is parameterization-invariant, we can use the Jeffreys prior for σ^2 to derive the Jeffreys prior for $\tau^2 = 1/\sigma^2$.

We use the transformation-of-variables formula (which also works for improper densities):

$$\frac{d\sigma^2}{d\tau^2} \;=\; \frac{d}{d\tau^2}\left(\frac{1}{\tau^2}\right) \;=\; -\frac{1}{(\tau^2)^2}$$

so

$$p(\tau^2) \;=\; p(\sigma^2)\left|\frac{d\sigma^2}{d\tau^2}\right| \;=\; \frac{1}{\sigma^2}\cdot\frac{1}{(\tau^2)^2} \;=\; \frac{1}{\tau^2}$$

(Note: This is like the conjugate gamma prior density for τ^2 with "$\alpha = 0$" and "$\beta = 0$.")

Notation:

$$\sigma^2 \sim \frac{1}{\sigma^2}\, d\sigma^2$$

$$\tau^2 \sim \frac{1}{\tau^2}\, d\tau^2$$

Writing the differential is important! It emphasizes that the parameter is σ^2 (or τ^2), rather than σ (or τ).