Ex. 7.2.4 Multinomial model

Sample $(s_1, s_2, \dots, s_n)$; Dirichlet prior $(\alpha_1, \alpha_2, \dots, \alpha_k)$ on $(\theta_1, \dots, \theta_{k-1})$ ⇒ the posterior distribution is

Dirichlet$(x_1+\alpha_1, x_2+\alpha_2, \dots, x_k+\alpha_k)$.

Can show mathematically that, if $(\theta_1, \dots, \theta_{k-1}) \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_k)$ then $\theta_i \sim$ Dirichlet(α_i, α_{-i}) = Beta(α_i, α_{-i}), where $\alpha_{-i} = \alpha_1+\alpha_2+\dots+\alpha_k-\alpha_i$ ⇒ the marginal posterior of θ_1 is Beta$(x_1+\alpha_1, x_2+\dots+x_k+\alpha_2+\dots+\alpha_k)$

Assuming that each $\alpha_i \geq 1$ & keeping in mind that $x_1+\dots+x_k = n$, the posterior mode is $\hat{\theta}_1 = \dfrac{x_1+\alpha_1+1}{n-2+\alpha_1+\dots+\alpha_k}$

For the uniform prior: $\alpha_1 = \dots = \alpha_k = 1$,

$$\hat{\theta}_1 = \frac{x_1}{n+k-2};$$

The posterior expectation is $E(\theta_1 | x) = \dfrac{x_1+\alpha_1}{n+\alpha_1+\dots+\alpha_k}$

The plug-in MLE of θ_1 is $\dfrac{x_1}{n}$ which is close to both Bayesian estimates for large n.

How to assess the accuracy of Bayesian estimates?
For posterior mean, ~~the~~ one possible approach is to compute Posterior variance. As an example, in Ex. 7.2.3, the posterior variance is $\left(\dfrac{1}{\tau_0^2}+\dfrac{n}{\sigma_0^2}\right)^{-1} \to 0$ as $\tau_0^2 \to 0$ and $\to \dfrac{\sigma_0^2}{n} = Var(\bar{x})$ as $\tau_0^2 \to \infty$.

7.2.2. Credible intervals

A credible interval for $\psi(\theta)$ is an interval $C(s) = [l(s), u(s)]$ that we believe will contain a true value of ψ. Typically we specify $\gamma \in (0,1)$ first. Then, we find an interval $C(s)$ such that $\Pi(\psi(\theta) \in C(s)/s) = \Pi(\{\theta : l(s) \le \psi(\theta) \le u(s)\}/s) \ge \gamma$ (*)

Then, $C(s)$ is a γ-credible interval for ψ.

Usually, the goal is to find a γ-credible interval $C(s)$ such that 1) $\Pi(\psi(\theta) \in C(s)/s)$ is as close to γ as possible and 2) $C(s)$ is as narrow as possible. One option to view the highest posterior density (HPD) intervals that are

$$C(s) = \{\psi : \omega(\psi/s) \ge c\}$$

where c is chosen as large as possible to satisfy (*)

Of course, $C(s)$ contains the mode whenever $c \le \max_{\psi} \omega(\psi/s)$. The width of such an interval is the measure of the accuracy of the mode of $\omega(\cdot/s)$ as an estimator of $\psi(\theta)$.

<u>Ex 7.2.7</u> Location normal model.

$$X_1, \ldots, X_n \sim N(\mu, \sigma_0^2); \quad \mu \sim N(\mu_0, \tau_0^2).$$

Here the posterior mean is $\hat{\mu}$ and the posterior variance is $\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}$, the distribution is symmetric around its mode.

$\hat{\mu}$, the shortest γ-HPD interval is

$$\hat{\mu} \pm \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1/2} c$$

where c is chosen in such a way that

$$\gamma = \Pi\left(\mu \in \left[\hat{\mu} \pm \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1/2} c\right] \Big/ x_1, \ldots, x_n\right) = \Pi\left(-c \le \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{1/2} (\mu - \hat{\mu}) \le c \,\Big|\, x_1, \ldots, x_n\right)$$

Because $\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{1/2}(\mu-\hat{\mu})\,/\,x_1,\ldots,x_n \sim N(0,1)$,
we find that $\gamma = \Phi(c) - \Phi(-c)$. Therefore,
$c = z_{\frac{1+\gamma}{2}}$, and the interval is

$$\hat{\mu} \pm \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1/2} z_{\frac{1+\gamma}{2}}.$$

Note that as $\tau_0^2 \to \infty$, this interval becomes $\bar{x} \pm \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}$

There are other possible types of credible intervals. One common method is to take $[\psi_e, \psi_r]$ where ψ_e is a $\frac{1-\gamma}{2}$ quantile of the posterior distribution and ψ_r is a $\frac{1+\gamma}{2}$ quantile for it. These tend to avoid some expensive computation that is needed for HPD intervals.

<u>Additional example</u>. The same farmer wants to construct a 95% credible set for the average weight of the cows μ. Recall that the posterior mean $\mathbb{E}(\mu/x_1,\ldots,x_n) = 412.9\,kg$ while the posterior variance is $Var(\mu/x_1,\ldots,x_n) = \left(\frac{10}{400}+\frac{1}{100}\right)^{-1} = 28.6\,kg^2$. Thus, the posterior distribution is $\mu/x_1,\ldots,x_n \sim N(412.9, 28.6)$ and the 95% credible interval is

$$412.9 \pm z_{0.025}\sqrt{28.6} = (402.4,\ 423.4)$$

7.2.3. Hypothesis Testing and Bayesian Factors

Let $H_0 : \psi(\theta) = \psi_0$. Can compute the posterior probability $\Pi(\psi(\theta) = \psi_0 | s)$ (*). If it is small, we have the evidence against H_0

Ex. 7.2.9 - Let $H_0 : \theta \in A$. Let $\psi = I_A$, we can rewrite $H_0 : \psi(\theta) = 1$ and $\Pi(\psi(\theta) = 1 | s) = \Pi(A | s)$

The approach (*) has its problems. If the posterior distribution is a continuous one, then $\Pi(\psi(\theta) = \psi_0 | s) = 0$ for any s.

To avoid this problem, another approach is often used.

Realize that a region of low posterior probability occurs where the posterior density $\omega(\cdot | s)$ is relatively low.

This suggests computing a Bayesian P-value

$$\Pi(\{\theta : \omega(\psi(\theta)|s) \le \omega(\psi_0|s)\} | s) \quad (**)$$

If $\omega(\cdot|s)$ is unimodal, this is a tail probability. If (**) is small, then ψ_0 is surprising according to our posterior beliefs. If we decide to reject H_0 whenever P-val. $< 1-\gamma$, this is equivalent to computing a γ - HPD region for ψ and rejecting H_0 whenever ψ_0 is not in the region.

Ex. 7.2.10 Clearly, $\psi(\theta) = I_A(\theta)$ has the posterior $Ber(\Pi(A)/s)$

Thus, $\omega(\cdot/s)$ is defined as $\omega(0/s) = 1 - \Pi(A/s) = \Pi(A^c/s)$

$$\omega(1/s) = \Pi(A/s)$$

For $\psi_0 = 1$, so $\{\theta : \omega(\psi(\theta)|s) \le \omega(1|s)\} = \{\theta : \omega(I_A(\theta)|s) \le \Pi(A/s)\} =$

$$= \begin{cases} \Omega, & \Pi(A|s) \ge \Pi(A^c/s) \\ A, & \Pi(A|s) < \Pi(A^c/s) \end{cases}$$

Therefore, (**) becomes

$$\Pi(\{\theta : \omega(\psi(\theta)|s) \le \omega(1|s)\} | s) = \begin{cases} 1, & \Pi(A|s) \ge \Pi(A^c/s) \\ \Pi(A|s), & \Pi(A|s) < \Pi(A^c/s) \end{cases}$$

- again we have evidence against H_0 whenever $\Pi(A|s)$ is small -

Thus, we find that using (**) is equivalent to using (*) whenever the parameter $\psi(\theta)$ takes only two values. When the prior distribution is continuous, (*) doesn't work, but (**) has its own issues in this situation.

__Ex. 7.2.11__. Let the posterior distr. of θ be $\omega(\theta/s) = 2\theta$ when $0 \le \theta \le 1$; $H_0: \theta = \frac{3}{4}$. Then, $\omega(\theta/s) \le \omega(\frac{3}{4}/s)$ iff $\theta \le \frac{3}{4}$;

(**) becomes $\int_0^{3/4} 2\theta \, d\theta = \frac{9}{16}$.

On the other hand, 1-1 transformation $\rho = \theta^2$ brings $H_0: \rho = \frac{9}{16}$.

The posterior distr. of ρ is $\omega(\rho/s) = 1$, $0 \le \theta \le 1$;

thus it is trivially true that $\omega(\rho/s) \le \omega(\frac{9}{16}/s)$. $\Rightarrow$

$\Rightarrow$ (**) is equal to 1, no evidence against H_0 can be found using this parameterization. Thus, the dependence on parameterization seems inappropriate.

To avoid all of these difficulties, it seems advisable to assign a positive prior probability to the hypothesis H_0. Then Ex. 7.2.10 demonstrates that this is the same as using (*) to assess H_0. Usually, one assigns the prior

$$\Pi = p\Pi_1 + (1-p)\Pi_2,$$

where $\Pi_1(\psi(\theta) = \psi_0) = 1$, $\Pi_2(\psi(\theta) = \psi_0) = 0$, so that Π_1 is degenerate at ψ_0 and Π_2 is continuous at ψ_0.

Then, $\Pi(\psi(\theta) = \psi_0) = p\Pi_1(\psi(\theta) = \psi_0) + (1-p)\Pi_2(\psi(\theta) = \psi_0) = p > 0$ – this is the prior probability that H_0 is true.

The prior predictive (marginal) for the data s is

$$m(s) = p m_1(s) + (1-p) m_2(s)$$

where m_i is the prior predictive obtained via prior Π_i.

This implies that the posterior probability measure for θ, is

$$\Pi(A/s) = \frac{p m_1(s)}{p m_1(s) + (1-p) m_2(s)} \Pi_1(A/s) + \frac{(1-p) m_2(s)}{p m_1(s) + (1-p) m_2(s)} \Pi_2(A/s)$$

where $\Pi_i(\cdot/s)$ is the posterior measure obtained via the prior Π_i.

Now $\Pi_1(\cdot|s)$ is degenerate at θ_0 so

$$\Pi(\Psi(\theta)=\Psi_0|s)=\frac{pm_1(s)}{pm_1(s)+(1-p)m_2(s)}$$ and

we use this probability to assess H_0.

Ex. 7.2.17. Location Normal Model

Let $(x_1, \ldots, x_n) \sim N(\mu, \sigma_0^2)$; $H_0: \mu=\mu_0$.

Prior for μ is $N(\mu_0, \tau_0^2)$ .... this will be our $\Pi_2 \sim N(\mu_0, \tau_0^2)$

Consider now the prior $\Pi = p\Pi_1 + (1-p)\Pi_2$; take $\Pi_1(\{\mu_0\})=1$, and so $\Pi(\{\mu_0\})=p$. Recall that under Π_2 the posterior distribution of μ is $N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right), \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)$

and under Π_1 the posterior is the distribution degenerate at μ_0.

To be continued ...

Remark. The prior predictive probability measure for the data s with a mixture of Π_1 and Π_2 prior distribution is

$$m(s) = E_{\Pi} f_\theta(s) = \sum_\theta f_\theta(s)\Pi(\{\theta\}) = \sum_\theta f_\theta(s)\left(p\Pi_1(\theta) + (1-p)\Pi_2(\{\theta\})\right) =$$

$$= p\sum_\theta f_\theta(s)\Pi_1(\{\theta\}) + (1-p)\sum_\theta f_\theta(s)\,\Pi_2(\{\theta\}) =$$

$$= p f_{\theta_0}(s) + (1-p)\sum_\theta f_\theta(s)\Pi_2(\{\theta\}) = pm_1(s) + (1-p)m_2(s).$$

The posterior ~~predictive~~ probability is given by

$$\Pi(A|s) = \sum_{\theta\in A}\frac{f_\theta(s)\,\Pi(\{\theta\})}{m(s)} = \sum_{\theta\in A}\frac{f_\theta(s)\left[p\Pi_1(\{\theta\}) + (1-p)\Pi_2(\{\theta\})\right]}{pm_1(s)+(1-p)m_2(s)} =$$

$$= \frac{pm_1(s)}{pm_1(s)+(1-p)m_2(s)}\sum_{\theta\in A}\frac{f_\theta(s)\,\Pi_1(\{\theta\})}{m_1(s)} + \frac{(1-p)m_2(s)}{pm_1(s)+(1-p)m_2(s)}\sum_{\theta\in A}\frac{f_\theta(s)\,\Pi_2(\{\theta\})}{m_2(s)} =$$

$$= \frac{pm_1(s)}{pm_1(s)+(1-p)m_2(s)}\,\Pi_1(A|s) + \frac{(1-p)\,m_2(s)}{pm_1(s)+(1-p)m_2(s)}\,\Pi_2(A|s)$$

Bayes factors - another method of hypothesis assessment.

<u>Def. 7.2.1</u>. Prob. model with sample space S and probability measure P, the odds in favor of event $A \subset S$ is $\frac{P(A)}{P(A^c)}$.

Bayes factors compare posterior odds with prior odds.

<u>Def. 7.2.2</u>. The Bayes factor BF_{H_0} in favor of $H_0: \psi(\theta) = \psi_0$ is defined as

$$BF_{H_0} = \left\{ \frac{\Pi(\psi(\theta)=\psi_0 | s)}{1-\Pi(\psi(\theta)=\psi_0|s)} \right\} \Big/ \left\{ \frac{\Pi(\psi(\theta)=\psi_0)}{1-\Pi(\psi(\theta)=\psi_0)} \right\}$$

~~Small relationship~~ BF_{H_0} small provides evidence against H_0 and vice versa.

Note also that $\Pi(\psi(\theta)=\psi_0|s) = \frac{r BF_{H_0}}{1+r BF_{H_0}}$ where

$r = \frac{\Pi(\psi(\theta)=\psi_0)}{1-\Pi(\psi(\theta)=\psi_0)}$ is the prior odds in favor of H_0. So, when BF_{H_0} is small, then $\Pi(\psi(\theta)=\psi_0|s)$ is small and conversely.

The following result establishes connections with likelihood ratios.

<u>Th. 7.2.1</u>. If the prior Π is a mixture $\Pi = p\Pi_1 + (1-p)\Pi_2$, where $\Pi_1(A) = 1$, $\Pi_2(A^c) = 1$, and $H_0: \theta \in A$,

$$BF_{H_0} = \frac{m_1(s)}{m_2(s)}$$ where m_i is the prior predictive of the data under Π_i

<u>Proof</u>. Recall that, a prior that concentrates all prob. on the set, results in the same posterior

Then,
$$BF_{H_0} = \frac{\Pi(A|s)}{1-\Pi(A|s)} \Big/ \frac{\Pi(A)}{1-\Pi(A)} = \frac{p\, m_1(s)}{(1-p)\, m_2(s)} \Big/ \frac{p}{1-p} = \frac{m_1(s)}{m_2(s)}$$

Note that the Bayes factor is independent of p.

Remark 1. Note that BF_{H_0} does not depend on p. However, it is not immediately clear how to calibrate it. We can use e.g. $\Pi(\Psi(\theta)=\psi_0|s) = \frac{rBF_{H_0}}{1+rBF_{H_0}}$ (***) because the posterior probability is directly interpretable. This, however, requires the knowledge of p.

Ex. 7.2.13. Location - normal -

Continue: compute the prior predictive under Π_2.

Recall that the joint density of $(x_1,\dots,x_n)$ given μ is

$$(*)\quad (2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2\right)$$ thus

$$m_2(x_1,\dots,x_n) = \int_{-\infty}^{+\infty}(*)\times(2\pi\tau_0^2)^{-1/2}\exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right)d\mu =$$

$$=(2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\times\tau_0^{-1}(2\pi)^{-1/2}\int_{-\infty}^{+\infty}\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2\right)\exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right)d\mu$$

The part after (x)'' can be shown to be

$$(**)\quad \tau_0^{-1}\exp\left(\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)^2\right)\times\exp\left(-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{n\bar{x}^2}{\sigma_0^2}\right)\right)\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}\right)^{-1/2}$$

Because Π_1 is degenerate at μ_0 we have the prior predictive under Π_1 as $m_1(x_1,\dots,x_n) = (2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2\right)$

Therefore, BF_{H_0} is equal to $\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2\right)$ divided by $(**)$

For example, suppose that $\mu_0=0$, $\tau_0^2=2$, $\sigma_0^2=1$, $n=10$, $\bar{x}=0.2$. Then $\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2\right)=\exp\left(-\frac{10}{2}(0.2)^2\right)=0.81873$

and $(**)$ is equal to 0.21615:

$\Rightarrow BF_{H_0}$ is $=\frac{0.81873}{0.21615}=3.7878$ – some evidence in favor of $H_0: \mu=\mu_0$. If we take $p=\frac{1}{2}$ (complete indifference between H_0 being true and not being true), then $r=1$ and we find that from (***)

$$\Pi(\mu=0|x_1,\dots,x_n) = \frac{3.7878}{\ }=0.79114$$ – strongly supportive

Ex Trying to guess whether a man is from the UK or France. You offer him: beer, brandy/cognac, whisky, or wine and try to figure out who he is by the choice of the drink.

We have the following table:

State	beer	brandy	whisky	wine
France	10%	20%	10%	60%
UK	50%	10%	20%	20%

H_0: France H_1: UK

Assume some prior probabilities $\pi_0 = P(\text{France})$ and $\pi_1 = P(UK)$.

-- perhaps a non-informative prior $\pi_0 = \pi_1 = 0.5$

We can apply the simplest Bayes rule for this case and reject then the null if the likelihood ratio $\lambda_y = \frac{P_{UK}(Y)}{P_{France}(Y)} > 1$.

We decide that the man is from UK if beer or whisky and France if wine or brandy. The values of λ are: $5 \quad 1/2 \quad 2 \quad 1/3$

7.2.4. Prediction.

We have an unobserved response value t in a sample space T and observed response $s \in S$. The statistical model is $\{P_\theta : \theta \in \Omega\}$ for s and the conditional statistical model $\{Q_\theta(\cdot|s) : \theta \in \Omega\}$ for t given s. Assume that both models have the same true value of $\theta \in \Omega$. The objective is to construct a prediction $\tilde{t}(s) \in T$ of the unobserved t based on s. The value of t is unknown because it is e.g. a future outcome.

Use $q_\theta(\cdot|s)$ to obtain the joint distribution of (θ, s, t):

$$q_\theta(\cdot|s) f_\theta(s) \pi(\theta).$$

Once we have observed s, the conditional density of (t,θ) given s is

$$\frac{q_\theta(t|s) f_\theta(s)\pi(\theta)}{\int_\Omega \int_T q_\theta(t|s) f_\theta(s)\pi(\theta)\,dt\,d\theta} = \frac{q_\theta(t|s) f_\theta(s)\pi(\theta)}{\int_\Omega f_\theta(s)\pi(\theta)\,d\theta} = \frac{q_\theta(t|s) f_\theta(s)\pi(\theta)}{m(s)}$$

The marginal posterior distribution of t, known as the posterior predictive of t, is

$$q(t|s) = \int_\Omega \frac{q_\theta(t|s) f_\theta(s)\,\pi(\theta)}{m(s)}\, d\theta = \int_\Omega q_\theta(t|s)\,\pi(\theta|s)\, d\theta.$$

Remark. The posterior predictive of t is obtained by averaging $q_\theta(t|s)$ with respect to the posterior distribution of θ.

Ex. 7-2.14 Given $(x_1, \dots, x_n) \sim Ber(\theta)$, $\theta \sim Beta(\alpha, \beta)$, predict X_{n+1}. The posterior probability f n of X_{n+1} is (at t) given by

$$q(t|x_1, \dots, x_n) = \int_0^1 \theta^t (1-\theta)^{1-t} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \times \theta^{n\bar{x}+\alpha-1} (1-\theta)^{n(1-\bar{x})+\beta-1}$$

(Keep in mind that X_{n+1} is independent of $x_1, \dots, x_n$)

$$= \cancel{\frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)}\,\Gamma(n\bar{x}+\alpha+t)\,\Gamma(n(1-\bar{x})+\beta+1-t)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \int_0^1 \theta^{n\bar{x}+\alpha+t-1} (1-\theta)^{n(1-\bar{x})+\beta+(1-t)-1}\, d\theta =$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \cdot \frac{\Gamma(n\bar{x}+\alpha+t)\,\Gamma(n(1-\bar{x})+\beta+1-t)}{\Gamma(n+\alpha+\beta+1)} =$$

$$= \begin{cases} \dfrac{n\bar{x}+\alpha}{n+\alpha+\beta}, & t=1 \\[2ex] \dfrac{n(1-\bar{x})+\beta}{n+\alpha+\beta}, & t=0 \end{cases} \sim Ber\left(\frac{n\bar{x}+\alpha}{n+\alpha+\beta}\right)$$

Using the posterior mode as the predictor we maximize $q(t|x_1, \dots, x_n)$ w.r.t. t; $\Rightarrow \hat{t} = \begin{cases} 1, & \text{if } \dfrac{n\bar{x}+\alpha}{n+\alpha+\beta} \geq \dfrac{n(1-\bar{x})+\beta}{n+\alpha+\beta} \\ 0 & \text{otherwise} \end{cases}$

At the same time, $\mathbb{E}(t|x_1, \dots, x_n) = \dfrac{n\bar{x}+\alpha}{n+\alpha+\beta}$ – can be anywhere in $[0, 1)$