

Background

Frequentist: Parameters are FIXED.

Bayesian: Parameters are RANDOM (i.e parameters have probability distributions)

Recall:

- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)}$

- Conditional Distribution:

Discrete r.v.'s $\rightarrow P(y_1 | y_2)$, Continuous r.v.'s $\rightarrow f(y_1 | y_2)$

$$f(y_1 | y_2) = \frac{f(y_1, y_2)}{f(y_2)} = \frac{f(y_2, y_1)}{f(y_2)}$$

- Joint Distribution:

Discrete r.v.'s $\rightarrow P(y_1, y_2) = P(y_2, y_1)$, Continuous r.v.'s $\rightarrow f(y_1, y_2) = f(y_2, y_1)$

$$f(y_1, y_2) = f(y_2, y_1) = f(y_2)f(y_1 | y_2)$$

- Marginal Distribution:

If y_1, y_2 are discrete r.v.'s: $P_1(y_1) = \sum_{y_2} P(y_1, y_2) = \sum_{i=1}^k P_2(y_2)P(y_1 | y_2)$

If y_1, y_2 are continuous r.v.'s $\Rightarrow f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 = \int_{-\infty}^{\infty} f_2(y_2) f(y_1 | y_2) dy_2$

- Expectation: $E(X) = \int x \cdot f(x) dx$, $E[g(X)] = \int g(x) \cdot f(x) dx$

- Variance: $Var(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$

- Precision: $\frac{1}{Var(X)}$

Bayes Rule

Assume sample space S is partitioned into $\{B_1, B_2, \dots, B_k\}$, such that $P(B_i) > 0$, then for any event A:

$$P(A) = \sum_{i=1}^k P(A | B_i)P(B_i) \quad \text{(Law of total probability)}$$

$$P(B_j | A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(B_j)P(A | B_j)}{\sum_{i=1}^k P(B_i)P(A | B_i)} \quad \text{(Bayes' Rule)}$$

$P(B_j)$ is called the "Prior", $P(A | B_j)$ is called the "Likelihood", and $P(B_j | A)$ is called the "Posterior".

In the continuous case:

$$f(y_2 | y_1) = \frac{f(y_1, y_2)}{f(y_1)} = \frac{f(y_2)f(y_1 | y_2)}{\int f(y_2)f(y_1 | y_2)dy_2}$$

Bayesians assign probability distributions to parameters:

$$f(\theta | Y) = \frac{f(Y, \theta)}{f(Y)} = \frac{f(\theta)f(Y | \theta)}{\int f(\theta)f(Y | \theta)d\theta}$$

Thus:

$f(\theta | Y) \propto f(\theta)f(Y | \theta)$ where:

- $f(\theta | Y)$ is the posterior
- $f(\theta)$ is the prior
- $f(Y | \theta)$ is the Likelihood
- The denominator is a constant with respect to θ

Bayesian Inference

Parameters are considered random variables.

Example: (Beta-Binomial)

Flip a coin n times, then estimate the probability $\pi = P(H)$.

Let X be the number of Heads. Thus $X | \pi \sim \text{Binomial}(n, \pi)$

Notation: Likelihood = $P(X | \pi) = \ell(\pi | X)$.

Likelihood:

$$\ell(\pi | X) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

Prior:

Since $\pi = P(H)$ is between 0 and 1, then a good prior distribution for π is the Beta distribution since the support for it is between 0 and 1. Plus, the *pdf* of the Beta distribution is quite flexible and can take an infinite number of shapes between 0 and 1.

$\pi \sim \text{Beta}(\alpha, \beta)$

$$P(\pi) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \pi^{\alpha-1} (1 - \pi)^{\beta-1}$$

Posterior:

$$P(\pi | X) \propto (\text{Prior}) \times (\text{Likelihood})$$

$$P(\pi | X) \propto \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1} \times \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

$$P(\pi | X) \propto \pi^{\alpha-1} (1-\pi)^{\beta-1} \times \pi^x (1-\pi)^{n-x}$$

$$P(\pi | X) \propto \pi^{\alpha+x-1} (1-\pi)^{\beta+n-x-1} \leftarrow \text{Kernel for Beta Distribution}$$

Thus: $\pi | X \sim \text{Beta}(\alpha^*, \beta^*)$ where $\alpha^* = \alpha + x$, and $\beta^* = \beta + n - x$

Since in this case a Beta prior resulted in a Beta posterior, we call it a “conjugate prior.”

$$E(\pi | X) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n} \right) \left(\frac{\alpha}{\alpha + \beta} \right) + \left(\frac{n}{\alpha + \beta + n} \right) \left(\frac{x}{n} \right)$$

Posterior Mean = weighted average of prior mean and the sample mean.

$$= w \left(\frac{\alpha}{\alpha + \beta} \right) + (1-w) \bar{x}$$

95% Bayesian Credible Interval (BCI)/Highest (posterior Density Region (HDR):

In R: $qbeta(0.025, \alpha^*, \beta^*) < \pi < qbeta(0.975, \alpha^*, \beta^*)$

ALWAYS: In the case of binomial likelihood, if you assume a beta prior, you will get a beta posterior (different parameters). This is referred to as “conjugacy,” where the prior and posterior have the same distribution family. Conjugacy is desired because it results in simplification of calculations.

Example: Normal-Normal

Likelihood:

Data is distributed normally, where $\theta | X \sim \text{Normal}(\theta, \phi)$, where θ is the UNKNOWN mean and ϕ is the known variance.

Prior:

$\theta \sim \text{Normal}(\theta_0, \phi_0)$, where both θ_0 and ϕ_0 are known,

Posterior:

$$P(\theta | X) \propto (\text{Prior}) \times (\text{Likelihood})$$

$$P(\theta | X) \propto \exp\left(-\frac{\theta - \theta_0}{2\phi_0}\right) \times \exp\left(-\frac{x - \theta}{2\phi}\right).$$

After some messy Algebra (involves completing a square) we get:

$$P(\theta | X) \propto \exp\left(-\frac{\theta - \theta_1}{2\phi_1}\right) \times \exp(k), \text{ where } k \text{ is a constant}$$

$$\Rightarrow P(\theta | X) \propto \exp\left(-\frac{\theta - \theta_1}{2\phi_1}\right) \leftarrow \text{Kernel for Normal Dist}$$

where: $E(\theta | x) = \theta_1$ and $Var(\theta | x) = \phi_1$. These parameters have the following expressions:

$$\phi_1 = \frac{1}{\phi_0^{-1} + (\phi/n)^{-1}}$$

Recall: Precision = 1/Variance. Thus:

$$\text{Precision}(\theta | x) = \phi_1^{-1} = \phi_0^{-1} + \left(\frac{\phi}{n}\right)^{-1}$$

Posterior Precision = Prior Precision + Data Precision

$$\theta_1 = \phi_1 \left(\frac{\theta_0}{\phi_0} + \frac{\bar{x}}{(\phi/n)} \right) = \left(\frac{\phi_1}{\phi_0} \right) \theta_0 + \left(\frac{\phi_1}{(\phi/n)} \right) \bar{x} = (w)\theta_0 + (1-w)\bar{x}$$

Posterior Mean = weighted average of prior mean and the sample mean.

95% Bayesian Credible Interval (BCI)/Highest posterior Density Region (HDR):

$$\theta_1 \pm 1.96\sqrt{\phi_1}$$

ALWAYS: In the case of normal likelihood, if you assume a normal prior, you will get a normal posterior (different parameters).

Reference Prior: This is an uninformative prior. So for a population proportion π , an uninformative prior would be Uniform(0,1) or equivalently Beta(1,1). Both are flat between 0 and 1.

In the case of the normal distribution, we can get a “near” flat distribution if we assume that the variance is large, that is choose Normal(θ_0, n), where n is large.

With a “flat” prior, the posterior is not affected by the prior. In fact, it is completely driven by the data. In this case, Bayesian analysis agrees with frequentist analysis. We can always pick near flat priors to be conjugate which results in simplification of calculations.

Robustness

Statistical inference is a conclusion drawn from data. In Bayesian analysis we use “posterior inference,” which takes into account information provided by the data and the

prior distribution. An inference is robust if it is not seriously affected by changes in the assumptions it is based on. To check robustness of Bayesian inference, we compare the effects of different priors on the posterior.

Disclaimer: The information presented earlier is based in part on my own class notes for 6860 by Dr. Eric Suess (Summer 06) and *Bayesian Statistics: An Introduction* (third edition) by Peter M. Lee.