

# Bayesian Theory

## Quick Notes

Ian S.W. Ma

Spring 2020

### Contents

1	Statistics Basics	1
2	Bayes Throrem	2
3	Conjugate Priors	3
4	Normal Distribution Priors	3
5	Jeffrey's Prior	4
6	Reference Prior	5
7	Point Estimates	5
8	Interval Estimates	6
9	Classic Hypothesis Testing	6
10	Bayesian Hypothesis Testing	6

# 1 Statistics Basics

## General Structure: Probability Space/Triple

Probability Space/Triple  $(\Omega, \mathcal{F}, \mathbb{P})$

- $\Omega$ : **Sample Space**, set of all possible outcomes.
- $\mathcal{F}$ : **Set of events**  $\sigma$ , each event is a subset of  $\Omega$
- $\mathbb{P}$  Probability of event  $\sigma \in \mathcal{F}$

## Conditional Probability

Conditional probability of  $A$  given  $B$ :  $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Leftrightarrow \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$

## General properties

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$   
Events are mutually exclusive  $\Rightarrow \mathbb{P}(A \cap B) = 0$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$   
Events are independent  $\Rightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$
- Law of Total Probability

If events  $A_1, A_2, A_N$  are mutually exclusive and exhaustive ( $\bigcup_{i=1}^N A_i = \Omega$ ) then:

$$\mathbb{P}(B) = \sum_{i=1}^N \mathbb{P}(A_i \cap B) = \sum_{i=1}^N \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

## Random Variables

$X \sim$ Distribution	$\mathbb{E}[X]$	$\text{Var}[X]$
Normal( $\mu, \sigma^2$ )	$\mu$	$\sigma^2$
Bernoulli( $\theta$ )	$\theta$	$\theta(1 - \theta)$
Binomial( $\theta$ )	$n\theta$	$n\theta(1 - \theta)$
Poisson( $\mu$ )	$\mu$	$\mu$
Uniform( $\alpha, \beta$ )	$\frac{\alpha+\beta}{2}$	$\frac{(\alpha+\beta)^2}{12}$
Beta( $\alpha, \beta$ )	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma( $\alpha, \beta$ )	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Inv. Gamma( $\alpha, \beta$ )	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Exponential( $\lambda$ )	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

## Random Vectors

$\mathbf{X} = X_1, \dots, X_k \sim$ Distribution	$\mathbb{E}[X_i]$	$\text{Var}[X_i]$	$\text{Cov}[X_i, X_j]$
Multivariate Normal( $\boldsymbol{\mu}, \Sigma$ )	$\mu_i$	$\Sigma_{i,i}$	$\Sigma_{i,j}$
Multinomial( $n, \theta_1, \dots, \theta_k$ )	$n\theta_i$	$n(1 - \theta_i)$	$-n\theta_i\theta_j$

## 2 Bayes Throrem

### Discrete Case

$$\begin{aligned}
 \mathbb{P}(X = x_i | Y = y_j) &= \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} \\
 &= \frac{\mathbb{P}(Y = y_j | X = x_i) \mathbb{P}(X = x_i)}{\mathbb{P}(Y = y_j)} \\
 &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}} \\
 &= \frac{\mathbb{P}(Y = y_j | X = x_i) \mathbb{P}(X = x_i)}{\sum_{k=1}^n \mathbb{P}(X = x_k, Y = y_j)} \\
 &= \frac{\mathbb{P}(Y = y_j | X = x_i) \mathbb{P}(X = x_i)}{\sum_{k=1}^n \mathbb{P}(Y = y_j | X = x_k) \mathbb{P}(X = x_k)} \\
 \Rightarrow \mathbb{P}((X|Y)) &= \frac{\mathbb{P}(Y|X) \mathbb{P}(X)}{\sum_x \mathbb{P}(X = x, Y)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}
 \end{aligned}$$

### Continous Case

$$f(X|Y) = \frac{f(X, Y)}{g(Y)} = \frac{g(Y|X)f(X)}{\int g(Y|X)f(X)dX}$$

### Bayesian Statistical Inference

- $L(\theta|y) = f(y|\theta)$ : Likelihood
- $\pi(\theta)$ : Prior
- $m(y) = p(y)$ : Marginal distribution/Normalizing constant

$$\text{Posterior Distribution: } p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int p(\theta, y)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta)$$

### Predictive Distrubutions

$$\text{Prior Predictive Distribution: } p(y^{new}) = \int_{\theta} f(y^{new}|\theta)\pi(\theta)d\theta$$

$$\text{Posterior Predictive Distribution: } p(y^{new}|y^{old}) = \int_{\theta} f(y^{new}|\theta, y^{old})p(\theta|y^{old})d\theta$$

### Bayes Theorem for Multiple Parameters $\Theta = \{\theta_1, \dots, \theta_q\}$

$$\mathbb{P}(\Theta|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\Theta)\mathbb{P}(\Theta)}{\mathbb{P}(\mathbf{y})} \propto \mathbb{P}(\mathbf{y}|\Theta)\mathbb{P}(\Theta)$$

### 3 Conjugate Priors

Sample Distribution	Parameter	Prior
Bernoulli( $\theta$ )	$\theta$	Beta( $\alpha, \beta$ )
Binomial( $n, \theta$ )	$\theta$	Beta( $\alpha, \beta$ )
Poisson( $\theta$ )	$\theta$	Gamma( $\alpha, \beta$ )
Exponential( $\theta$ )	$\theta$	Gamma( $\alpha, \beta$ )
Normal( $\mu, \sigma_0^2$ )	$\mu$	Normal( $\mu_h, \sigma_h^2$ )
Normal( $\mu_0, \sigma^2$ )	$\sigma^2$	Inverse Gamma( $\alpha, \beta$ )
Normal( $\mu_0, 1/\tau$ )	$\tau = 1/\sigma^2$	Gamma( $\alpha, \beta$ )

### 4 Normal Distribution Priors

#### Unknown $\mu$ , known $\sigma^2$

- Sample  $\mathbf{y} \in \mathbb{R}^n$
- Prior  $\mu \sim \text{Normal}(\mu_0, \sigma_0^2) = \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) \Rightarrow m = \sigma^2/\sigma_0^2$
- Posterior:

$$\mu|\mathbf{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{\sigma^2}{m+n}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{\sigma^2}{m+n}\right)$$

$$w = \frac{m}{m+n}$$

#### Unknown $\mu$ , known $\tau$

- Sample  $\mathbf{y} \in \mathbb{R}^n$
- Prior  $\mu \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) = \text{Normal}\left(\mu_0, \frac{1}{\tau m}\right)$
- Posterior:

$$\mu|\mathbf{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{1}{\tau(m+n)}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{1}{\tau(m+n)}\right)$$

$$w = \frac{m}{m+n}$$

#### Unknown $\tau$ , known $\mu$

- Sample  $\mathbf{y} \in \mathbb{R}^n$
- Prior  $\tau \sim \text{Gamma}(\alpha, \beta)$
- Posterior:

$$\tau|\mathbf{y}, \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{z^2}{2}\right)$$

$$z^2 = \sum_i (y_i - \mu)^2$$

## Unknown $\sigma^2$ , known $\mu$

- Sample  $\mathbf{y} \in \mathbb{R}^n$
- Prior  $\mu, \sigma^2 \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{\kappa}\right) \times \text{Inverse Gamma}(\alpha, \beta)$
- Posterior:

$$\mu, \sigma^2 | \mathbf{y} \sim \text{Normal}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{\sigma^2}{\kappa + n}\right) \times \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa n(\bar{y} - \mu_0)^2}{2(\kappa + n)}\right)$$

$$s^2 = \sum_i (y_i - \bar{y})^2 / (n-1)$$

$\kappa = \sigma^2(\alpha - 1)/\beta$  or given (seriously it's not written anywhere Ken, I'm just guessing here man)

## 5 Jeffrey's Prior

### Definition

Jeffery Prior:  $\pi_{JP}(\theta) \propto \sqrt{I(\theta|y)}$

Fisher Information:

$$I(\theta|y) = \mathbb{E} \left[ \left( \frac{d \log f(y|\theta)}{d\theta} \right)^2 \right] = -\mathbb{E} \left[ \frac{d^2 \log f(y|\theta)}{d\theta^2} \right]$$

### 1:1 transformation of Jeffrey's Prior

$$\pi(\phi) = \pi_{JP}(\theta) \left| \frac{d\theta}{d\phi} \right|$$

## Jeffrey's Prior for Multivariate Parameter Vector

Gradient of log likelihood:

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \partial_{\theta_1} \log f(x) \\ \vdots \\ \partial_{\theta_n} \log f(x) \end{bmatrix}$$

Hessian Matrix of log likelihood:

$$H(\boldsymbol{\theta}) = \text{Jacobian}[S(\boldsymbol{\theta})] = \begin{bmatrix} \partial_{\theta_1} \partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_1} \partial_{\theta_n} \ln f(x) \\ \vdots & \ddots & \vdots \\ \partial_{\theta_n} \partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_n} \partial_{\theta_n} \ln f(x) \end{bmatrix}$$

Fisher Information:  $I(\theta|x) = -\mathbb{E}[H(\theta)]$

Jeffrey's Prior:  $\pi_{JP} \propto \sqrt{\det(I(\theta|x))}$

## 6 Reference Prior

### Kullback-Leibler (KL) divergence

A measure of the difference between two pmfs/pdfs. When  $KL(f, g) = 0$ , the two distributions are identical.

#### Properties

- $KL(f, g) \neq KL(g, f)$
- $KL(fmg) \geq 0$

#### Continuous parameter $x$

$$KL(f, g) = \int \ln \left[ \frac{f(x)}{g(x)} \right] dx = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

#### Discrete parameter $x$

$$KL(f, g) = \sum_{x \in X} \ln \left[ \frac{f(x)}{g(x)} \right] = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

### Reference Prior

Read week 4 notes (2.2)

## 7 Point Estimates

### Components of Decision Theory with emphasis on parameter estimation

- State Space  $\Theta$ , unknown true value  $\theta \in \Theta$
- Action Space  $\mathcal{A} \ni a$ , sometimes  $\mathcal{A} = \Theta$
- Sampling Distribution  $f(\mathbf{y}|\theta)$
- Loss Function  $\mathcal{L}(a|\theta)$
- Risk  $R_\theta(a|\mathbf{y}) = \mathbb{E}_\theta[L(a|\theta, \mathbf{y})] = \int_{\theta \in \Theta} \mathcal{L}(a|\theta) p(\theta|\mathbf{y}) d\theta$
- Bayes Estimator of a parameter  $\hat{\theta}_{BE} = \arg \min_{\hat{\theta} \in \Theta} R_\theta(\hat{\theta}|\mathbf{y})$

### Common Loss Functions and Corresponding Estimators

- **Squared Error Loss:**  $\mathcal{L}(\theta|\hat{\theta}) = (\theta - \hat{\theta})^2$ , Bayes Estimator  $\hat{\theta} = \mathbb{E}[\theta|\mathbf{y}]$
- **Absolute Error Loss:**  $\mathcal{L}(\theta|\hat{\theta}) = |\theta - \hat{\theta}|$ , Bayes Estimator  $\hat{\theta} = \theta_{0.5}$ ,  $(\mathbb{P}(\theta \leq \theta_{0.5}) = 0.5)$
- **0-1 Loss:**  $\mathcal{L}(\theta|\hat{\theta}) = I(\theta \neq \hat{\theta})$ , Bayes Estimator  $\hat{\theta}$  is the posterior mode

## 8 Interval Estimates

$$P \times 100\% \text{ Bayesian Credible interval} = [LB, UB] \text{ where } \int_{LB}^{UB} p(\theta, \mathbf{y}) d\theta = P$$

If looking for one side-confidence bounds then  $LB = -\infty$  or  $UB = \infty$

### Symmetric Credible Interval

$$[LB, UB] \text{ where } \mathbb{P}(\theta \leq LB) = \mathbb{P}(\theta \geq UB) = \alpha/2 = (1 - P)/2$$

### Highest Posterior Density Interval (HPDI)

Credible Interval where all values outside the interval has a density smaller than any of the values in the interval.

### Credible Regions

$$\iint p(\theta_1, \theta_2 | \mathbf{y}) d\theta_1 d\theta_2 = 1 - \alpha = P$$

## 9 Classic Hypothesis Testing

1. Assume hypothesis  $H_0$  is true
2. Calculate test statistic  $T(\mathbf{y}_{obs})$  based on observed sample data regarding to  $H_0$  and  $H_1$
3. Conditional on  $H_0$  being true,  $p\text{-value} = \mathbb{P}(T(\mathbf{y}) \text{ more extreme than } T(\mathbf{y}_{obs}) | \theta, H_0)$
4. Reject  $H_0$  and accept  $H_1$  for sufficiently small  $p\text{-values}$ , do not reject  $H_0$  otherwise

## 10 Bayesian Hypothesis Testing

Suppose there are two hypotheses about parameter  $\theta$ :

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$$

where  $\Theta_0 \cup \Theta_1 = \Theta$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ .

The Bayesian approach specifies prior probabilities on each hypotheses:

$$\begin{aligned} p_0 &= \mathbb{P}(H_0 \text{ is true}) = \mathbb{P}(\theta \in \Theta_0) \\ p_1 &= \mathbb{P}(H_1 \text{ is true}) = \mathbb{P}(\theta \in \Theta_1) \end{aligned}$$

Where the posteriors are as follows:

$$\begin{aligned} \mathbb{P}(H_0 | \mathbf{y}) &= \mathbb{P}(\theta \in \Theta_0 | \mathbf{y}) \\ \mathbb{P}(H_1 | \mathbf{y}) &= \mathbb{P}(\theta \in \Theta_1 | \mathbf{y}) \end{aligned}$$

Where  $\mathbb{P}(H_0 | \mathbf{y}) + \mathbb{P}(H_1 | \mathbf{y}) = 1$

## Simple

- Single parameter values  $H_0 : \theta = \theta_0$   $H_1 : \theta = \theta_1$
- Posteriors:

$$\begin{aligned}\mathbb{P}(H_0|\mathbf{y}) &= \mathbb{P}(\theta = \theta_0|\mathbf{y}) = \frac{f(\mathbf{y}|\theta_0)p_0}{m(\mathbf{y})} = \frac{f(\mathbf{y}|\theta_0)p_0}{f(\mathbf{y}|\theta_0)p_0 + f(\mathbf{y}|\theta_1)p_1} \\ \mathbb{P}(H_1|\mathbf{y}) &= \mathbb{P}(\theta = \theta_1|\mathbf{y}) = \frac{f(\mathbf{y}|\theta_1)p_1}{m(\mathbf{y})} = \frac{f(\mathbf{y}|\theta_1)p_1}{f(\mathbf{y}|\theta_0)p_0 + f(\mathbf{y}|\theta_1)p_1} \\ \mathbb{P}(H_0|\mathbf{y}) &= 1 - \mathbb{P}(H_1|\mathbf{y})\end{aligned}$$

- posterior odds of  $H_0$  against  $H_1$ :

$$\frac{\mathbb{P}(H_0|\mathbf{y})}{\mathbb{P}(H_1|\mathbf{y})} = \frac{f(\mathbf{y}|\theta_0)p_0}{f(\mathbf{y}|\theta_1)p_1}$$

## Composite

- Single parameter values  $H_0 : \theta \in \Theta_0$   $H_1 : \theta \in \Theta_1$
- Prior:  $p_i = \int_{\theta \in \Theta_i} \pi(\theta) d\theta$
- Posteriors:

$$\begin{aligned}\mathbb{P}(H_i|\mathbf{y}) &= \frac{p(\mathbf{y}, H_i)}{m(\mathbf{y})} = \frac{p_i p(\mathbf{y}|H_i)}{m(\mathbf{y})} = \frac{p_i \int p(\mathbf{y}, \theta|H_i) d\theta}{m(\mathbf{y})} = \frac{p_i \int f(\mathbf{y}|\theta) \pi(\theta|H_i) d\theta}{m(\mathbf{y})} \\ &= \frac{p_i \int_{\theta \in \Theta_i} f(\mathbf{y}|\theta) \frac{\pi(\theta)}{p_i} d\theta}{m(\mathbf{y})} = \frac{\int_{\theta \in \Theta_i} f(\mathbf{y}|\theta) \pi(\theta) d\theta}{m(\mathbf{y})} = \int_{\theta \in \Theta_i} p(\theta|\mathbf{y}) d\theta = \mathbb{P}(\theta \in \Theta_i|\mathbf{y})\end{aligned}$$

- posterior odds of  $H_0$  against  $H_1$ :

$$\frac{\mathbb{P}(H_0|\mathbf{y})}{\mathbb{P}(H_1|\mathbf{y})} = \frac{\int_{\theta \in \Theta_0} f(\mathbf{y}|\theta) \pi(\theta) d\theta}{\int_{\theta \in \Theta_1} f(\mathbf{y}|\theta) \pi(\theta) d\theta} = \frac{\mathbb{P}(\theta \in \Theta_0|\mathbf{y})}{\mathbb{P}(\theta \in \Theta_1|\mathbf{y})} = \frac{\mathbb{P}(\theta \in \Theta_0|\mathbf{y})}{1 - \mathbb{P}(\theta \in \Theta_0|\mathbf{y})}$$

## Multiple models

- Set of models  $M_1, \dots, M_K$
- $H_i$ : The correct model is  $M_i$

$$\mathbb{P}(H_i|\mathbf{y}) = \frac{\mathbb{P}(H_i, \mathbf{y})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(H_i, \mathbf{y})}{\sum_{j=1}^K \mathbb{P}(H_j, \mathbf{y})}$$



## Bayes Factor

- Prior odds for  $H_0$  against  $H_1 = p_0/p_1$
- Posterior odds for  $H_0$  against  $H_1 = \mathbb{P}(H_0|\mathbf{y})/\mathbb{P}(H_1|\mathbf{y})$
- Bayes Factor for  $H_0$  against  $H_1$ :

$$BF_{01} = \frac{\text{Posterior odds for } H_0 \text{ against } H_1}{\text{Prior odds for } H_0 \text{ against } H_1} = \frac{\mathbb{P}(H_0|\mathbf{y})/\mathbb{P}(H_1|\mathbf{y})}{p_0/p_1}$$

$$BF_{10} = \frac{1}{BF_{01}}$$

$BF_{ij}$	Interpretation
$< 3$	No edvidence for $H_i$ over $H_j$
$> 3$	Positive edvidence for $H_i$
$> 20$	Strong edvidence for $H_i$
$> 150$	Very strong edvidence for $H_i$