→ In practice, the first step in the analysis is the same as in frequentist framework
→ we have to decide/specify how we are best going to describe the data that we have
→ Once we have this, we have to choose a prior
→ Can do conjugate, but only one that goes w/ the data model we chose
⇒ This gives a posterior w/ a nice form, but it is not very flexible (limited # of shapes for a given prior)

→ We could also use a noninformative prior, but this will likely lead to an ugly posterior
⇒ Then the question becomes how we are going to sample from this ugly form??
⇒ Can use strategy in lab - Single-parameter-inference

Single parameter inference

STA 427/527, Fall 2019, Xin Wang

2.4 Choices of priors

2.4.1 Conjugate priors

- **Conjugacy of priors:** If $\mathcal{F}$ is a class of sampling distributions $p(\boldsymbol{y}|\theta)$, and $\mathcal{P}$ is a class of prior distributions $\pi(\theta)$, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if $p(\theta|\boldsymbol{y}) \in \mathcal{P}$.

  we want to choose a prior such that the prior & posterior distributions are in the same class of distributions
  ⇒ Keeps derivation of the posterior easier (this is the only way to guarantee a nice posterior dist)

- **Example 1:** Beta is conjugate for Binomial
  - $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Bin}(m, \theta)$, where m is known, $\theta \sim \text{Beta}(\alpha, \beta)$ $\quad \underset{\sim}{Y} = (Y_1, \ldots, Y_n)$
  - Posterior: $\theta|\boldsymbol{y} \sim \text{Beta}(\alpha + \sum_{i=1}^n y_i, \beta + mn - \sum_{i=1}^n y_i)$ ← end result

↳ show this → Data dist → $p(\underset{\sim}{y}|\theta) = p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n p(y_i|\theta) = \prod_{i=1}^n \binom{m}{y_i} \theta^{y_i}(1-\theta)^{m-y_i}$

$= \left(\prod_{i=1}^n \binom{m}{y_i}\right) \theta^{\sum_{i=1}^n y_i} (1-\theta)^{nm - \sum_{i=1}^n y_i}$

→ prior → $\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

→ Posterior → $p(\theta|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\theta)\pi(\theta)$

$\propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{nm - \sum_{i=1}^n y_i} \left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$

$= \theta^{(\sum_{i=1}^n y_i + \alpha) - 1} (1-\theta)^{(nm - \sum_{i=1}^n y_i + \beta) - 1}$

$\sim \text{Beta}\left(\sum_{i=1}^n y_i + \alpha,\ nm - \sum_{i=1}^n y_i + \beta\right)$ ⇒ posterior mean $E(\theta|\underset{\sim}{y}) = \frac{\sum_{i=1}^n y_i + \alpha}{nm + \beta + \alpha}$

- **Example 2:** Gamma is conjugate for Poisson
  - $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, prior of λ is $\lambda \sim \text{Gamma}(\alpha, \beta)$

→ Data dist → $p(\underset{\sim}{y}|\lambda) = p(y_1, \ldots, y_n|\lambda) = \prod_{i=1}^n p(y_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$

$= \left(\prod_{i=1}^n y_i!\right)^{-1} \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}$

→ prior → $\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$

→ Posterior → $p(\lambda|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\lambda)\pi(\lambda)$

$\propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda}$

$= \lambda^{(\sum_{i=1}^n y_i + \alpha) - 1} e^{-(n+\beta)\lambda}$

$\sim \text{Gamma}\left(\sum_{i=1}^n y_i + \alpha,\ n+\beta\right)$ ⇒ posterior mean $E(\lambda|\underset{\sim}{y}) = \frac{\sum_{i=1}^n y_i + \alpha}{n+\beta}$

↳ $\frac{\alpha}{\beta}$

now for new observations

– Question: What is prior predictive distribution and posterior predictive distribution? → Poisson example

→ we need to have an actual estimate of our model (b/c we don't have data yet)
⇒ first we make the prior predictive dist

① → prior predictive dist → $p(\tilde{y}) = \int p(\tilde{y}, \lambda)\, d\lambda = \int p(\tilde{y}|\lambda)\pi(\lambda)\, d\lambda$

$= \int \frac{\lambda^{\tilde{y}} e^{-\lambda}}{\tilde{y}!} \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \right) d\lambda$

$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\tilde{y}!} \int \lambda^{(\tilde{y}+\alpha)-1} e^{-(\beta+1)\lambda}\, d\lambda$

$= \left(\frac{\Gamma(\tilde{y}+\alpha)}{(\beta+1)^{\tilde{y}+\alpha}} \right)$

$= \frac{(\tilde{y}+\alpha-1)!}{(\beta+1)^{\tilde{y}+\alpha}}$

$= \frac{(\tilde{y}+\alpha-1)!}{(\alpha-1)!\,\tilde{y}!} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{\tilde{y}}$

$\tilde{y} = 0, 1, 2, \ldots$

if α is an integer

$\sim$ Negative Binomial $(\alpha, \frac{\beta}{\beta+1})$

⇒ $\tilde{y}$ = # failures until α successes w/ prob $\frac{\beta}{\beta+1}$

→ If $X \sim \text{Gamma}(\alpha, \beta)$

⇒ $\int \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx = 1$

⇒ $\int x^{\alpha-1} e^{-\beta x} = \frac{\Gamma(\alpha)}{\beta^\alpha}$

→ If $X \sim NB(r, p)$

$f(x) = \binom{x-1}{r-1} p^r q^{x-r}$, x = # trials until r successes

OR

$f(y) = \binom{y+r-1}{y} p^r q^y$, y = # failures ...

② → Posterior predictive dist → $p(\tilde{y}|\underline{y}) = \int p(\tilde{y}, \lambda)\, d\lambda = \int \underbrace{p(\tilde{y}|\lambda)}_{\text{data dist}} \underbrace{p(\lambda|\underline{y})}_{\text{posterior}} d\lambda = \int \frac{\lambda^{\tilde{y}} e^{-\lambda}}{\tilde{y}!} \left(\frac{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}}{\Gamma(\sum_{i=1}^n y_i + \alpha)} \lambda^{(\sum_{i=1}^n y_i + \alpha)-1} e^{-(n+\beta)\lambda} \right) d\lambda$

$= \frac{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}}{\Gamma(\sum_{i=1}^n y_i + \alpha)} \frac{1}{\tilde{y}!} \int \lambda^{\left(\sum_{i=1}^n y_i + \alpha + \tilde{y}\right) - 1} e^{-(n+\beta+1)\lambda} d\lambda$

$= \downarrow$

$\frac{\Gamma(\sum_{i=1}^n y_i + \alpha + \tilde{y})}{(n+\beta+1)^{\sum_{i=1}^n y_i + \alpha + \tilde{y}}}$, $\tilde{y} = 0, 1, 2, \ldots$

$\alpha > 0$ integer $= \frac{(\tilde{y} + \sum_{i=1}^n y_i + \alpha - 1)!}{(\sum_{i=1}^n y_i + \alpha - 1)!\,\tilde{y}!} \left(\frac{n+\beta}{n+\beta+1} \right)^{\sum_{i=1}^n y_i + \alpha} \left(\frac{1}{n+\beta+1} \right)^{\tilde{y}}$

$\sim$ Negative Binomial $\left(\sum_{i=1}^n y_i + \alpha, \frac{n+\beta}{n+\beta+1} \right)$

⇒ posterior predictive mean

$E(\tilde{y}|\underline{y}) = \frac{\sum_{i=1}^n y_i + \alpha}{\frac{n+\beta}{n+\beta+1}} = \frac{\sum_{i=1}^n y_i + \alpha}{n+\beta}$

↳ $\frac{r(1-p)}{p}$

↳ This is a combo of info from the data ($\sum_{i=1}^n y_i + \alpha$) & the prior ($\alpha + \beta$)

→ In simulation → often we cannot this posterior predictive dist (or it would take lots of effort
⇒ So instead we can use simulation

→ 2 steps
→ 1) sample $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(M)}$ from posterior dist $\sim$ Gamma$(\sum y_i + \alpha, n+\beta)$
→ 2) Sample $\tilde{y}^{(m)}$, $m = 1, 2, \ldots, M$ from Poisson $(\lambda^{(m)})$

→ Posterior means as weighted average of data likelihood & prior mean

→ Beta – Binomial Example

$$E(\theta \mid \underset{\sim}{Y}) = \frac{\sum y_i + \alpha}{nm + \alpha + \beta} = \frac{\frac{\sum y_i}{nm}(nm) + \frac{\alpha}{\alpha+\beta}(\alpha+\beta)}{nm + \alpha + \beta}$$

↳ sample size

$$= \boxed{\frac{\sum y_i}{nm}}\left(\frac{nm}{nm+\alpha+\beta}\right) + \boxed{\frac{\alpha}{\alpha+\beta}}\left(\frac{\alpha+\beta}{nm+\alpha+\beta}\right)$$

{ frequentist point estimate = $\hat{p}_{MLE}$ for p of Binomial ; weight $\lim_{n\to\infty} = 1$; prior mean ; 1-weight $\lim_{n\to\infty} = 0$

⇒ as the sample size increases, the first weight increases

⇒ the first term dominates the posterior mean (data likelihood)

specifically → $\lim_{n\to\infty} E(\theta \mid \underset{\sim}{Y}) = \hat{\theta}_{MLE}$

e.g. $Y_1, Y_2, Y_3 \overset{iid}{\sim} Bin(m=5, p)$

↳ $\underset{\sim}{y} = (3,4,3) \Rightarrow \hat{p}_{MLE} = \frac{(\sum y_i)}{m} = \frac{10}{3} = \frac{10}{15} = \frac{2}{3}$

↳ $= \frac{\bar{y}}{m}$

→ Poisson – Gamma

$$E(\lambda \mid \underset{\sim}{Y}) = \frac{\sum y_i + \alpha}{n + \beta}$$

$$= \frac{\frac{\sum y_i}{n}(n) + \frac{\alpha}{\beta}(\beta)}{n+\beta}$$

$$= \boxed{\frac{\sum y_i}{n}}\left(\frac{n}{n+\beta}\right) + \boxed{\frac{\alpha}{\beta}}\left(\frac{\beta}{n+\beta}\right)$$

data MLE weight ; prior mean ; 1 – weight

- **Example 3: Conjugate priors for exponential family**
  A signle-parameter exponential family: The family consists of any distribution whose pmf/pdf can be written as

$$f(y|\theta) = \exp\{a(y)\,b(\theta) + c(\theta) + d(y)\}$$

$a(\cdot)$, $b(\cdot)$, $c(\cdot)$ & $d(\cdot)$ are known functions

where $a(y)$ and $d(y)$ do not depend on θ, $b(\theta)$ and $c(\theta)$ do not depend on y.

  - If we have n iid observations from the distribution above. The joint distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is

→ Data dist →

$$p(\underset{\sim}{y}|\theta) = \prod_{i=1}^{n} \exp\{a(y_i)\,b(\theta) + c(\theta) + d(y_i)\}$$

$$= \exp\{b(\theta)\sum_{i=1}^{n} a(y_i) + n\,c(\theta) + \sum_{i=1}^{n} d(y_i)\}$$

  - A conjugate prior for θ is

general form →

$$\pi(\theta) = g(k,\gamma)\exp[k\cdot\gamma\cdot b(\theta) + kc(\theta)]$$

  - Posterior distribution

$$p(\theta|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\theta)\,\pi(\theta)$$

$$= \exp\{b(\theta)\sum_{i=1}^{n} a(y_i) + n\,c(\theta) + \sum_{i=1}^{n} d(y_i)\} * [g(k,\gamma)\exp\{k\gamma\,b(\theta) + k\,c(\theta)\}]$$

$$\propto \exp\{b(\theta)(\sum_{i=1}^{n} a(y_i) + k\gamma) + (n+k)\,c(\theta)\}$$

- **Comments**
  - Easy to implement — once you specify the prior ⇒ you know what distribution the posterior will be & it will be relatively easy to sample from
  - Not flexible enough to allow other shapes of priors → not flexible enough to allow other shapes of priors; e.g. if have gamma prior ⇒ then can only have prior shapes like

  - Still possible to incorporate prior knowledge about θ

advantage / Disadvantage

- **Some conjugate priors**

Data likelihood	Prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Normal(σ^2 known)	Normal for μ
Normal(μ known)	Inverse-gamma for σ^2
Uniform$(0,\theta)$	Pareto for θ
Exponential	Gamma
Gamma(β is unknown, α is known)	Gamma

Choosing Prior

→ Cannot mix & match priors & data dists just b/c they are in exponential family ⇒ choices have to follow the table

→ So best practice is to choose a prior that meets your conditions so that the posterior will have a nice form

→ Conjugate priors guarantee the posterior has the same parametric form, which is obviously good

→ BUT, this also locks in the shape of the posterior (even if we choose a prior w/ a large variance)

⇒ not flexible again

→ Ultimately, choice of prior is subjective

⇒ this is one of the main criticisms of Bayesian methods

2.5 Noninformative priors

1. Uniform prior → This is one example of an uninformative prior b/c we think all possible values of the parameter are equally likely

- **Example:** If we toss a coin, $Y \sim Bernoulli\,(p); p \sim Uniform\,(0,1)$

(reparameterizing (to illustrate a problem)) → Let $\pi(p) = 1 \quad 0 < p < 1$

→ If considering transformation $\alpha = \frac{p}{1-p} \quad (\alpha > 0)$

⇒ $\pi(\alpha) = ??$ (goal is to find this)

⇒ $p = \frac{\alpha}{\alpha+1} \rightarrow \frac{dp}{d\alpha} = \frac{1}{(\alpha+1)^2}$

→ need to calculate $|J|$ (Jacobian) $= \left|\frac{dp}{d\alpha}\right| = \frac{1}{(\alpha+1)^2}$

$\pi(\alpha) = \pi(p)\,|J| = 1 \cdot \frac{1}{(\alpha+1)^2} = \frac{1}{(\alpha+1)^2}$

↳ prior for α

↳ Can be parameterized with p (prob of success) or $\alpha = \frac{p}{1-p}$ (odds ratio)

⇒ want a prior that gives the same beliefs regardless of which parameterization is used

Problem → After doing the transformation, not all α values are equally likely (i.e properties of the prior changed) ⇒ generally true after a transformation unless ↓

2. Jeffrey's prior: Fisher information → solves problem

$$I(\theta) = J(\theta) = E\left[\left(\frac{\partial \ln L(\boldsymbol{y}|\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln L(\boldsymbol{y}|\theta)}{\partial\theta^2}\right] = -E\left[\frac{\partial^2 \ell(\theta)}{\partial\theta^2}\right]$$

(log-likelihood)

Then the corresponding Jeffrey's prior is $p(\theta) \propto \sqrt{I(\theta)}$

→ Consider $\varphi = h(\theta)$ (greek "phi") → new parameterization (some function of original parameter)

→ It turns out → $\pi_1(\varphi) \propto \sqrt{I(\varphi)}$ (new prior) & $\pi(\theta) \propto \sqrt{I(\theta)}$

⇒ "invariant" to the parameterization ⇔ These priors are related by the usual change of variables theorem

$p(\varphi) = p(\theta)\left|\frac{d\theta}{d\varphi}\right|$

- **Example:** Binomial data $Y|\theta \sim Binomial\,(n,\theta)$

→ $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

→ $\ln(p(y|\theta)) = \ln\binom{n}{y} + y\ln(\theta) + (n-y)\ln(1-\theta)$

→ $\frac{\partial \ln(p(y|\theta))}{\partial\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}$

→ $\frac{\partial^2 \ln(p(y|\theta))}{\partial\theta^2} = \frac{-y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$

→ $I(\theta) = -E\left[\frac{\partial^2 \ln(p(y|\theta))}{\partial\theta^2}\right] = -E\left[\frac{-y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right]$

$= \frac{1}{\theta^2}E(y) + \frac{1}{(1-\theta)^2}E(n-y)$ → $E(y) = n\theta$

$E(n-y) = n - E(y) = n\theta$

$= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2}$

lots of algebra ↓

$= \frac{n}{\theta} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)} + \frac{n}{1-\theta}\left(\frac{\theta}{\theta}\right)$

$= \frac{n(1-\theta) + n\theta}{\theta(1-\theta)}$

$= \frac{n - n\theta + n\theta}{\theta(1-\theta)}$

$= \frac{n}{\theta(1-\theta)}$

⇒ The Jeffrey's prior is $\pi(\theta) \propto \sqrt{I(\theta)}$

$= \sqrt{\frac{n}{\theta(1-\theta)}}$

$\propto \theta^{-1/2}(1-\theta)^{-1/2}$

$\sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$