$$P(\theta \mid y) \propto \exp\left[\sum n_i(\theta)\left[\alpha_i + T_i(y)\right] - (1+\alpha_{k+1})A(\theta)\right]$$

Uncertainty Quantification

$$Y \sim B(n, \theta) \longrightarrow P(\theta|y):$$

$$\theta \sim Beta(\alpha, \beta)$$

$$Beta(y+\alpha, n-y+\beta)$$

$$E[\theta|y]: w\, y/n + (1-w)\frac{\alpha}{\alpha+\beta}$$

$$w = \frac{n}{n+\alpha+\beta}$$

$$Y_i \overset{iid}{\sim} Pois(v_i \lambda)$$

$$\lambda \sim Gam(a, b)$$

$$P(\lambda | y_1, \ldots y_n, v_1, \ldots v_n) \sim Gam\left(\sum y_i + a, \sum v_i + b\right)$$

$$E[\lambda | \ldots] = w \frac{\sum y_i}{\sum v_i} + (1-w)\frac{a}{b}$$

$$w = \frac{\sum v_i}{\sum v_i + b}$$

Posterior credible intervals

Random

- Frequentist interval: $Pr(l(Y) < \theta < u(Y) \mid \theta) = 0.95$
  - Probability that the interval covers the true value before the data are observed.

fixed.

  - Interval is random

Random.

- Bayesian interval: $Pr(l(y) < \theta < u(y) \mid Y = y) = 0.95$
  - Information about the parameter given $Y = y$.

const

  - θ is random (we can include a prior distribution) interval is random.

Posterior credible intervals (Quantile based method)

- The easiest way to find credible intervals is to use the quantiles of the posterior distribution

If we want a $100 \times (1-\alpha)$ interval, we find $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ such that

1. $p(\theta < \theta_{\alpha/2} \mid Y = y) = \alpha/2$
2. $p(\theta > \theta_{1-\alpha/2} \mid Y = y) = \alpha/2$

$$p(\theta \in [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y = y) = 1 - \alpha$$

e.g. [qbeta(.025, $\tilde{\alpha}$, $\tilde{\beta}$), qbeta(.975, $\tilde{\alpha}$, $\tilde{\beta}$)]

Interval for shooting skill in basketball

- The posterior distribution of Curry's shooting percentage is a

$$\text{Beta}(\underbrace{49}_{\alpha} + 478, \underbrace{50}_{\beta} + 873) = \text{Beta}(528, 924)$$

- For a 95% credible interval, $\alpha = 0.05$
  - Lower end point: `qbeta(0.025, 528, 9`
  - Upper end point: `qbeta(0.975, 528, 9`
  - $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$ = [0.34, 0.39]

Interval for shooting skill in basketball

- Bayesian credible interval $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$ = [0.34, 0
- Frequentist confidence interval: [0.39, 0.59
- End-of-season average: 0.37
- Credible interval is a confidence interval with different meanings!

Highest Posterior Density (HPD) Region

Highest Posterior Density Region (HPD)

Definition: A $100(1-\alpha)\%$ HPD region is a set of values of a parameter θ such that $R(y) \in \Theta$ such that

1. $\Pr(\theta \in R(y) | Y = y) = 1 - \alpha$
   - The probability that the parameter θ is within the HPD region is $1-\alpha$
2. If $\theta_a \in R(y)$, and $\theta_b \notin R(y)$, then $p(\theta_a | Y = y) > p(\theta_b | Y$
   - All points in the HPD region have higher posterior density than points outside the region

The HPD region is the smallest region with probability $(1-\alpha)$

Frequentist fallacy: Bayesian invalid

- Bayesian idealism is usually correct
- If our prior is well-calibrated and the sample correct, then the well-calibrated idealism
- And: as sample size grows, posterior is overwhelmed by the data. 95% of the posterior is empirical

$$\log P(\theta|y) \approx \log P(\hat{\theta}|y) + \overset{\text{linear term}}{0}$$

(expand around $\hat{\theta}$, posterior mode)

$$+ \frac{1}{2}(\hat{\theta}-\theta)' \frac{d^2}{d\theta} \log P(\hat{\theta}|y)(\hat{\theta}-\theta)$$

$$+ \ldots \text{ higher order terms}$$

$$-I(\hat{\theta})$$

$$\Rightarrow P(\theta|y_1,\ldots y_n) \xrightarrow{d} N(\hat{\theta}, I(\hat{\theta})^{-1})$$

Classic Freq: $\hat{\theta}_{MLE} \xrightarrow{d} N(\theta, I(\theta)^{-1})$

Bayes: $I(\hat{\theta})^{1/2}(\theta - \hat{\theta}_{mode}) \xrightarrow{n\to\infty} N(0, I)$

Freq: $I(\theta)^{1/2}(\theta - \hat{\theta}_{MLE}) \longrightarrow N(0, I)$

Conclusion: Freq/Bayes Agreement.

Frequentist Bayesian Uncertainty Quantification

Under regularity conditions:

$$p(\theta \mid y_1, \ldots, y_n) \approx \mathrm{N}\left(\hat{\theta}, [I(\hat{\theta})]^{-1}\right)$$

Classical result

$$\hat{\theta} \approx N(\theta, [I(\theta)]^{-1})$$

Sequential Bayesian Updating

$$\underbrace{p(\theta)}_{\text{prior}} \to \underbrace{p\left(\theta \mid y_1\right)}_{\text{reveal the first observation}} \to \underbrace{p\left(\theta \mid y_1, y_2\right)}_{\text{reveal the second observation}} \cdots$$

$$\to \cdots \to \underbrace{p\left(\theta \mid y_1, \ldots, y_n\right)}_{\text{reveal all } n \text{ observations}}$$

When a data point is obtained, the posterior becomes the prior for the next data point. This is called sequential updating.

Beta-Bin:

Post: $\boxed{\text{Beta}(y+\alpha,\ n-y+\beta)}$

What are valid α, β?

$y+\alpha > 0$ $\quad$ $n-y+\beta > 0$

Improper prior distributions

- For the Beta distribution we assumed a uniform prior ($p(\theta) \propto \text{const}$). This is so that we c
  - $\int_0^1 p(\theta) d\theta = \text{const} < \infty$
  - We say this prior is a proper distribution integrable
- For the Poisson distribution, the uniform prior $p(\lambda) \propto \text{const}$
  - $\int_0^\infty p(\lambda) d\lambda = \infty$
  - In this case we say $p(\lambda)$ is an improper prior

$\mathrm{Gam}(\Sigma y_i, \Sigma v_i)$ when $P(\lambda) \propto$ const.

Improper prior dist

- Sometimes we have an objective prior
- The prior distribution doesn't matter as long as the posterior does!
  - A proper distribution guarantees convergence
  - If you use a proper improper prior, you need to check that the posterior is proper

Objective Bias

- Also called "reference", "non-informative" prior distributions
- Laplace's principle of insufficient reason
- Principle of maximum entropy (MEMT)
- Matching prior distributions
  - Find prior distribution that corresponds to a statistical model with optimal properties
- Invariant prior

$-\sum P(x) \log P(x)$

info. theory

Laplace's principle of insufficient reason

Uniform distribution

```
y    E oE      ey t tgt
5        e    gVE E/      e gnE
.  E E              e    e    gnE   . tgEm
   E E          e          5 g
```

$y \sim Bin(n, p)$

Laplace's approximation to the posterior

- Assume $Y \sim \text{Binomial}(n, \theta)$ but we want to work with the log odds:

$$\gamma = \text{logodds}(\theta) = \log \frac{\theta}{1-\theta}$$

- What prior should we have to be "non-informative"?

Diiculties owrimtaat ipvo in e i n

logistic density.

Method of transformations

Assume the $p(\theta)$ is uniform. What is the density of the transformed variable, $\gamma = g(\theta)$?

$p(\theta) \sim \text{Unif}$

1. Find the inverse, $\theta = g^{-1}(\gamma)$

$\gamma = \log\left(\frac{\theta}{1-\theta}\right) \sim ?$

2. Compute $\frac{dg^{-1}(\gamma)}{d\gamma}$

$\frac{\theta}{1-\theta} = e^{\gamma}, \quad \theta = \frac{e^{\gamma}}{1+e^{\gamma}}$

$\text{"}g^{-1}\text{"}$

3. Find $p_\gamma(\gamma) = \left|\frac{dg^{-1}(\gamma)}{d\gamma}\right| \times p_\theta(g^{-1}(\gamma))$

$\frac{dg^{-1}}{d\gamma} = \frac{e^{\gamma}}{(1+e^{\gamma})^2}$

$\underbrace{\text{Unif}[\,] \; \frac{e^{\gamma}}{1+e^{\gamma}}}_{\text{const}} \rightarrow \boxed{P_\gamma(\gamma) = \frac{e^{\gamma}}{(1+e^{\gamma})^2}}$

Jeffreys prior $\sqrt{I(\theta)}$

- Idea: find a prior that is invariant under transformations $\phi = g(\theta)$
- Derivation:

$$I_\phi(\phi) = I_\theta(\theta)\left(\frac{d\theta}{d\phi}\right)^2 \quad \text{identity}$$

$$P_\theta(\theta) = \sqrt{I(\theta)}$$

$$P_\phi(\phi) = P_\theta(\theta)\left|\frac{d\theta}{d\phi}\right|$$

$$= \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right| =$$

$$= I(\phi)\frac{d\phi^2}{d\theta} = \sqrt{I(\phi)}$$

$$P_\theta(\theta) \propto \sqrt{I(\theta)}$$

$$\Rightarrow P_\phi(\phi) \propto \sqrt{I(\phi)}$$

$Y \sim Bin(n, \theta)$

$\theta^{\alpha-1}(1-\theta)^{\beta-1}$

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

$Beta(1/2, 1/2)$

(proper prior)

$Y \overset{iid}{\sim} N(\mu, \sigma^2)$ known.

$\sqrt{I(\mu)} \propto \sqrt{\frac{n}{\sigma^2}} \propto$ const

J.P. $p(\mu) \propto$ const, improper!

Leads to a <u>proper posterior</u>!

Weakly informative priors

- A proper prior does not trivially dominate, i
information unless large amounts of data
- Construction:
  - Start with information and add
enough information to make the mo
reasonable
  - Or start with information and r
information
- Example: incidents per year is usually
magnitude less than 10, in g

Prediction

Posterior predictive distribution

- An important feature of Bayesian inference is the use of posterior distributions.
  - Let $\tilde{y}$ be a (unseen) event and $y_1, \ldots, y_n$ the observed data.
  - The posterior predictive distribution $p(\tilde{y} \mid y_1, \ldots, y_n)$ is

No more θ!

$$P(\tilde{y} \mid y_1, \ldots y_n) = \int \underbrace{P(\tilde{y} \mid \theta) P(\theta \mid y_1, \ldots y_n)}_{P(\tilde{y}, \theta \mid y_1, \ldots y_n)} d\theta$$

Posterior predictive distrib

- The predictive distribution depends on parameters
- The predictive distribution over data
- Asks: what is the probability of new observations, given old data

Another Bias Example

- Take a fair coin and toss it 10 times. Let X be the number of heads you think we will get?
- If my true θ is 0.50, $X \sim \text{Bin}(10, 0.50)$
- Is this the correct probability distribution

$\text{Bin}(10, \hat{\theta}_{MLE})$

Posterior Prediction

If you knew θ, the distribution attempts:

$$\tilde{Y} \sim \text{Bin}(10, \theta)$$

$$P(\tilde{y} \mid y=1, n=2) =$$

$$\int_0^1 \underbrace{P(\tilde{y} \mid \theta)}_{\text{Bin}} \underbrace{P(\theta \mid y)}_{\text{Beta}} \, d\theta =$$

$$\int \underbrace{\binom{10}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{10-\tilde{y}}}_{\text{Bin}(10,\,\theta)} \underbrace{\theta^{2-1} (1-\theta)^{4-1} \frac{\Gamma(6)}{\Gamma(2)\Gamma(4)}}_{\text{Beta}(2,\,4)} \, d\theta$$

$$= \binom{10}{\tilde{y}} \frac{\Gamma(6)}{\Gamma(2)\Gamma(4)} \int_0^1 \theta^{\tilde{y}+1} (1-\theta)^{13-\tilde{y}} \, d\theta$$

$$= \binom{10}{\tilde{y}} \frac{\Gamma(6)}{\Gamma(2)\Gamma(4)} \frac{\Gamma(\tilde{y}+2)\Gamma(14-\tilde{y})}{\Gamma(16)}$$

$$= P(\tilde{y} \mid y), \text{ Beta-Binomial.}$$

Posterior Prediction

- We already observed 1 make out of 2 tries
- Assume a (a1, Be3) prior distribution
  - e.g. a supportive mix of high and low makes 25% of shots
- The posterior $p(\theta \mid Y = 1, n = 2)$ is Beta(2, 4) 1/3
- Imagine tossing the coin 10 times $\tilde{Y} \sim \text{Bin}(10, \theta)$ by $p(\theta \mid Y = 1, n = 2)$

Posterior Distribution

If we take 10 more shots how many will I make?

Posterior predictive distribution

$$p(\theta) = \text{Beta}(1, 3), p(\theta \mid y) = \text{Beta}(2, 4)$$

The predictive $p(\tilde{y} \mid y)$ tells us how many makes we might expect in 10 more shots, taking into account the uncertainty in θ.

The posterior predictive distribution

$$p(\tilde{y} \mid y_1, \ldots y_n) = \int p(\tilde{y}, \theta \mid y_1, \ldots y_n) d\theta$$

$$= \int p(\tilde{y} \mid \theta) p(\theta \mid y_1, \ldots y_n) d\theta$$

- The posterior predictive distribution describes uncertainty about future observations
- It incorporates uncertainty about the sampling $p(\tilde{y} \mid \theta)$ and our uncertainty about the generating parameter $p(\theta \mid y_1, \ldots y_n)$

Theta Beta Binomial Distrib

$$\mathrm{Var}(\tilde{y} \mid y) = E[\mathrm{Var}(\tilde{y} \mid \theta, y)] + \mathrm{Var}(E[\tilde{y} \mid \theta, y])$$

$$= E[n\theta(1-\theta) \mid y] + \mathrm{Var}[n\theta \mid y]$$

"expected variance of the samp. distn"

"variance due to uncertainty about θ."

Posterior predictive density

The predictive distribution

$$p(\tilde{y}) = \int p(\tilde{y}, \theta) d\theta$$
$$= \int p(\tilde{y} \mid \theta) p(\theta) d\theta$$

- The predictive distribution reflects our beliefs about two sources of uncertainty: the sampling distribution $p(\tilde{y} \mid \theta)$ and our uncertainty about the parameter $p(\theta)$
- It is important to understand the sampling distribution $p(\tilde{y} \mid \theta)$ and our uncertainty about the parameter $p(\theta)$

Summary

- Conjugate prior
- Credible intervals
- Non-informative prior
- Posterior predictive distribution