Prior distributions

- Important: Must have positive prior density (or probability mass) for all plausible values θ!
  (all values of θ for which $p(\theta)=0$, will have $p(\theta|y)=0$)

  ↳ If this is fulfilled, typically with enough data, the information on θ from the data (through the likelihood) will dominate and "wash away" the prior distribution

Binomial data distribution: Finding a convinient prior

Recall the binomial data setting, where y is the number of successes in n identical and indepent trials, with θ as the population proportions. Hence, the sampling distribution

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

Considered as a function of θ, the likelihood becomes

$$p(y|\theta) \propto \theta^y(1-\theta)^{n-y} = \theta^a(1-\theta)^b \quad (*)$$

sep 15-12:14

- We see that if the prior is of the same functional form w.r.t. θ, then the posterior density will also be of the same form

  A prior density of the functional form can be written as

  $$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for some } \alpha,\beta>0$$

  Which is a $\mathrm{Beta}(\theta|\alpha,\beta)$ distribution, where α and β are the parameters

  $$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta\in[0,1], \alpha>0, \beta>0$$

- Comparing this to (*) suggests that this prior density represents $\alpha-1$ prior successes and $\beta-1$ prior failures
- For the special case that $\alpha=1, \beta=1$, this is the $\mathrm{Unif}(0,1)$-distribution
- The parameters of the prior distribution are often called hyperparameters

  ↳ For now, we assume them to be fixed constants

sep 15-12:26

The posterior for θ is now

$$p(\theta|y) \propto \theta^{y}(1-\theta)^{n-y}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

$$\propto \text{Beta}(\theta\,|\,y+\alpha,\ n-y+\beta)$$

$$\boxed{\begin{array}{l}\theta \sim \text{Beta}(\alpha,\beta)\\ E[\theta] = \frac{\alpha}{\alpha+\beta}\end{array}}$$

The posterior mean is hence $E[\theta|y] = \dfrac{\alpha+y}{\alpha+\beta+n}$

Remember that is also the posterior predictive probability of success for a future draw (with $\tilde{n}=1$):

$$\Pr(\tilde{y}=1|y) = \int \Pr(\tilde{y}=1|\theta)\cdot p(\theta|y)\,d\theta = \int \theta\cdot p(\theta|y)\,d\theta = E[\theta|y]$$

$$E[\theta|y] = \frac{\alpha+y}{\alpha+\beta+n} = \underbrace{w}_{\frac{\alpha+\beta}{\alpha+\beta+n}}\cdot\frac{\alpha}{\alpha+\beta} + \underbrace{(1-w)}_{\frac{n}{\alpha+\beta+n}}\cdot\frac{y}{n}$$

← For fixed α,β, as y and $n-y$ grow large: $E[\theta|y] \approx \frac{y}{n}$

The posterior variance is

$$\text{Var}[\theta|y] = \frac{(\alpha+y)\cdot(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{E[\theta|y]\,(1-E[\theta|y])}{\alpha+\beta+n+1}$$

← For fixed α,β, as y and $n-y$ grow large: $\text{Var}[\theta|y] \approx \frac{1}{n}\cdot\frac{y}{n}\cdot\left(1-\frac{y}{n}\right) \xrightarrow[n\to\infty]{} 0$

<u>Conjugacy</u>

If $\mathcal{F}$ is a class of sampling distributions $p(y|\theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is <u>conjugate</u> for $\mathcal{F}$ if

$$p(\theta|y) \in \mathcal{P} \quad \text{for all } p(\cdot|\theta)\in\mathcal{F} \text{ and } p(\cdot)\in\mathcal{P}$$

- If $\mathcal{P}$ is the class of all distributions for θ, then $\mathcal{P}$ is always conjugate, regardless of $\mathcal{F}$
  - ↳ not so interesting
- If $\mathcal{P}$ is the set of all pdfs/pmfs having the same functional form as the likelihood (w.r.t. θ), then $\mathcal{P}$ is a natural conjugate family for $\mathcal{F}$
  - ↳ most interesting

<u>Binomial data with Beta prior:</u>

Since $p(\theta|y)$ is Beta$(\cdot,\cdot)$, and the prior is Beta$(\cdot,\cdot)$, the Beta prior distribution is a (natural) conjugate family for the binomial sampling distribution

Conjugate priors convenient

- Computationally convenient
  - ↳ Single-parameter models (and some others):
    - Posterior distr. is analytically available
  - ↳ Complex models:
    - Convenient building blocks, simplifying computations
      - → We will now look at some of the most commonly used conjugacy building blocks
- Nice interpretation of the prior as "prior" (or "additional") data
- Mainly justified by convenience, which is comparable to using standard sampling distributions
- We can (and will) do computation also with non-conjugate priors!

sep 15-12:57

Normal sampling distribution, unknown mean, known variance

- We consider n iid observations $y_1, \dots, y_n$ with

$$p(y_i|\theta) = N(y_i|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\}, \quad i=1,\dots,n$$

Likelihood as a function of θ

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}$$

$\uparrow$ Prop. to w.r.t. θ

Now $\sum_{i=1}^{n}(y_i-\theta)^2 = \sum_{i=1}^{n}(y_i-\bar{y}-(\theta-\bar{y}))^2 = \sum_{i=1}^{n}\left[(y_i-\bar{y})^2-2(y_i-\bar{y})(\theta-\bar{y})+(\theta-\bar{y})^2\right]$

$$= \sum_{i=1}^{n}\left[(y_i-\bar{y})^2-2(y_i\theta-y_i\bar{y}-\bar{y}\theta+\bar{y}^2)+(\theta-\bar{y})^2\right]$$

$$= \sum_{i=1}^{n}(y_i-\bar{y})^2-2\underbrace{\left[\theta\cdot n\cdot\bar{y}-n\bar{y}^2-n\bar{y}\theta+n\bar{y}^2\right]}_{=0}+n(\theta-\bar{y})^2$$

sep 15-13:16

Hence $p(y|\theta) \propto \exp\left\{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2\right\}$ ← The likelihood

Conjugate prior

A natural conjugate prior must be of the same functional form

$$p(\theta) \propto \exp\left\{-C_1\cdot(\theta-C_2)^2\right\}$$

We choose the following parametisation:

$$p(\theta) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\} = N(\theta|\mu_0,\tau_0^2)$$

μ_0 and τ_0^2 are the hyperparameters, which we for now assume known

We have

$$p(\theta|y) \propto p(\theta)\cdot p(y|\theta)$$

$$\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\cdot\exp\left\{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2\right\}$$

$$= \exp\left\{-\frac{1}{2}\frac{\left[\frac{\sigma^2}{n}\cdot(\theta-\mu_0)^2+\tau_0^2\cdot(\theta-\bar{y})^2\right]}{\tau_0^2\cdot\frac{\sigma^2}{n}}\right\}$$

sep 15-13:24

$$= \exp\left\{-\frac{1}{2}\frac{\left[\frac{\sigma^2}{n}(\theta^2-2\theta\mu_0+\mu_0^2)+\tau_0^2(\theta^2-2\theta\bar{y}+\bar{y}^2)\right]}{\tau_0^2\cdot\frac{\sigma^2}{n}}\right\}$$

$$= \exp\left\{-\frac{1}{2}\frac{\left[\left(\frac{\sigma^2}{n}+\tau_0^2\right)\theta^2-2\theta\left[\frac{\sigma^2}{n}\mu_0+\tau_0^2\bar{y}\right]+\frac{\sigma^2}{n}\mu_0^2+\tau_0^2\bar{y}^2\right]}{\tau_0^2\cdot\frac{\sigma^2}{n}}\right\}$$

$$\cdot\exp\left\{-\frac{1}{2}\frac{\left(\frac{\sigma^2}{n}\mu_0+\tau_0^2\bar{y}\right)^2-\left(\frac{\sigma^2}{n}\mu_0+\tau_0^2\bar{y}\right)^2}{\tau_0^2\cdot\frac{\sigma^2}{n}}\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\frac{(\theta-\mu_n)^2}{\tau_n^2}\right\}$$

where $\mu_n = \dfrac{\frac{1}{\tau_0^2}\cdot\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}$, $\dfrac{1}{\tau_n^2} = \dfrac{1}{\tau_0^2}+\dfrac{n}{\sigma^2}$

The posterior mean is a weighted average of the prior mean and the sample average $\bar{y}$, with weights proportional to the prior precision and the data precision

Posterior precision = Prior precision + data precision

sep 15-13:30

- If $\tau_0^2 = \sigma^2$, then the prior distribution has the same weight as one extra observation with value μ_0

(• If $\tau_0^2 = 0$ and $\frac{\sigma^2}{n} > 0$, or $\bar{y} = \mu_0$, the $\mu_n = \mu_0$)

If $\tau_0 \to \infty$ with n fixed

or if $n \to \infty$ with τ_0 fixed > 0

the $p(\theta | y) \approx N(\theta \mid \bar{y}, \sigma^2/n)$

which is often a good approximation if the prior information is vague for θ-values where the likelihood is large.

sep 15-13:40

Posterior predictive distribution of a future observation $\tilde{y}$

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

$$\propto \underbrace{\int \exp\left\{-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right\} \cdot \exp\left\{-\frac{1}{2\tau_n^2}(\theta - \mu_n)^2\right\}}_{} d\theta$$

Proportional to a bivariate Normal distr. for $(\tilde{y}, \theta) | y$

$\Rightarrow$ Integrating out θ yields a Normal distribution for $\tilde{y} | y$

$$E[\tilde{y} | y] = E[E[\tilde{y} | \theta, y] | y] = E[\theta | y] = \mu_n$$

$$Var[\tilde{y} | y] = E[Var[\tilde{y} | \theta, y] | y] + Var[E[\tilde{y} | \theta, y] | y] =$$

$$E[\sigma^2 | y] + Var[\theta | y] = \sigma^2 + \tau_n^2$$

sep 15-13:46

Poisson sampling distribution

Consider iid count observations y_i, $i=1,\dots,n$
with $p(y_i|\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}$, $i=1,\dots,n$.

The likelihood:

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{y_i} \theta^{y_i} e^{-\theta} \propto \theta^{\sum y_i} \cdot e^{-n\theta}$$

We have already seen (NCH Ex. 1) that the natural conjugate prior is the Gamma(α,β) and that the posterior is Gamma$(\alpha+\sum y_i, \beta+n)$
The Poisson sampling distr. can be extended by included covariate information:

$$y_i \sim \text{Poisson}(x_i \cdot \theta) \quad , x_i > 0 \quad , i=1,\dots,n$$

The likelihood is now

$$p(y|\theta) \propto \theta^{\sum y_i} \cdot \exp\{-\theta \cdot \sum x_i\}$$ and with $\theta \sim \text{Gamma}(\alpha,\beta)$,
we get the posterior distr. Gamma$(\alpha+\sum y_i, \beta+\sum x_i)$

sep 15-13:52