Motivating example: Spelling correction

- **Problem**: Someone types '*radom*'.
- **Question**: What did they mean to type? Random?

Ingredients:

- **Data** x: The observed word — *radom*
- **Parameter of interest** θ: The correct word

Comments: To solve this we need

- *background information* on which words are usually typed.
- an idea about how words are typically mistyped.

Example adapted from *Bayesian Data Analysis* by Gelman et al.

Bayesian Idea

- **Data model**: Conditional on θ, data x is distributed according to pf/pdf $\pi(x)$:

$$\pi(x|\theta) \propto L(\theta; x) \quad \leftarrow \text{the likelihood}$$

- **Prior**: Prior knowledge (i.e. *before* collecting data) about θ is summaries by a pf/pdf,

$$\pi(\theta) \quad \leftarrow \text{the prior}$$

- **Posterior** : The updated knowledge about θ *after* collecting data: The conditional distribution of θ given data x:

$$\begin{aligned} \pi(\theta|x) &= \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} \\ &\propto \pi(x|\theta)\pi(\theta) \\ \text{“}&posterior = likelihood \times prior'' \end{aligned}$$

Example: Prior

Without any other prior knowledge Google provies the following *prior probabilities* (for three candidate words):

θ	$\pi(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

Comments

- The relative high probablity for the word *radom* is surprising! Name of Polish airshow and nickname for Polish hand gun...
- In the context of writing a scientific report these prior probailities would look different...?

Example: Likelihood

Google provies the following conditional probabilities *prior probabilities*:

θ	$\pi(x = \text{'radom'}\|\theta)$
random	0.00193
radon	0.000143
radom	0.975

Comments

- This is *not* a probability distribution!
- If one in fact intends to write 'radom' this actually happens in 97.5% of cases.
- If one intends to write either 'random' or 'radon' this is rarely misspelled 'radom'.

Example: Posterior

Combining the prior and likelihhod we obtain the *posterior probabilities*:

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} \propto \pi(x|\theta)\pi(\theta)$$

θ	$\pi(x = \text{'radom'}|\theta)\pi(\theta)$	$\pi(\theta|x = \text{'radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

Conclusion

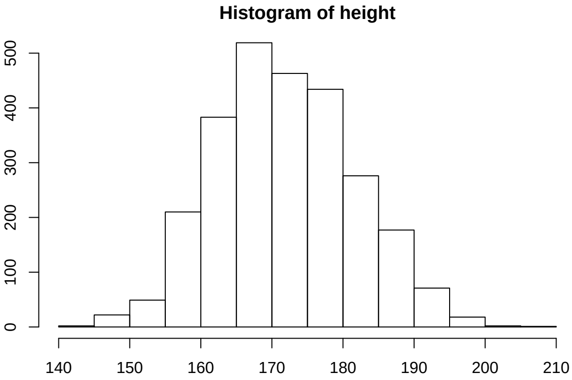
- With the given prior and likelihood the word 'radom' is twice as likely as 'random'.

Criticism

- The posterior probability for 'radom' seems too high.
- Likelihood or prior to blame
- Likelihood is probably ok in this case
- Prior depends on context — and hence might be "wrong".

Another example

Heights of some Copenhageners in 1995

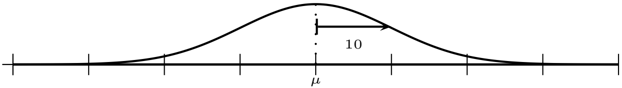


Assume: Heights are normal, $X \sim \mathcal{N}(\mu, \tau)$.

For now: Assume precision τ known.

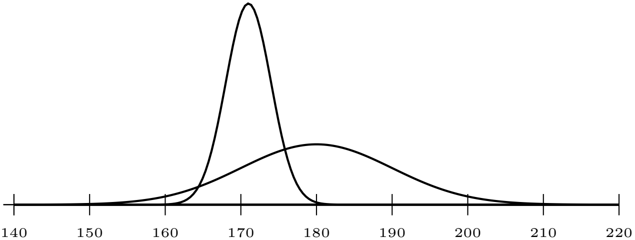
Bayesian Idea: Illustration

Data model: $X \sim \mathcal{N}(\mu, 0.01)$ (i.e. pop. sd = 10)



Prior: We believe that the population mean is most likely between 160 cm and 200 cm. $\pi(\mu) = \mathcal{N}(180, 0.01)$.

Posterior: After observing a number of heights ($n = 10$, $\bar{x} = 169$), our knowledge about μ is updated. Summarised by the posterior.



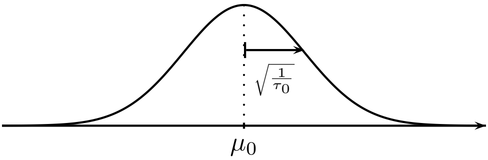
Normal example: One (!) observation

Data model: $X \sim \mathcal{N}(\mu, \tau)$

Assume: Precision τ known.

Interest: The unknown mean μ.

Prior: The prior for μ is specified as $\mu \sim \mathcal{N}(\mu_0, \tau_0)$.



Normal example: Data density

Data: One observation, X, from a normal distribution:

$$\begin{aligned}
\pi(x|\mu) &= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \\
&= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau x^2 - \frac{1}{2}\tau\mu^2 + \tau\mu x\right) \\
&\propto \exp\left(-\frac{1}{2}\tau x^2 + \tau\mu x\right)
\end{aligned}$$

Notice the “pattern” inside the exponential.

Normal example: Posterior density

$$
\begin{aligned}
Posterior &\propto Likelihood \times Prior \\
\pi(\mu|x) &\propto \pi(x|\mu)\pi(\mu) \\
&= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu-\mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\tau\mu^2 + \tau x\mu - \frac{1}{2}\tau_0\mu^2 + \tau_0\mu\mu_0\right) \\
&= \exp\left(-\frac{1}{2}(\tau+\tau_0)\mu^2 + (\tau x + \tau_0\mu_0)\mu\right) \\
&\propto \mathcal{N}\left(\frac{\tau x + \tau_0\mu_0}{\tau+\tau_0}, \tau+\tau_0\right)
\end{aligned}
$$

Notice: Prior for μ was normal, now the posterior for μ is also normal!

Normal example: Posterior mean & variance

The posterior: $\pi(\mu|x) = \mathcal{N}\left(\frac{\tau x + \tau_0 \mu_0}{\tau + \tau_0}, \tau + \tau_0\right)$

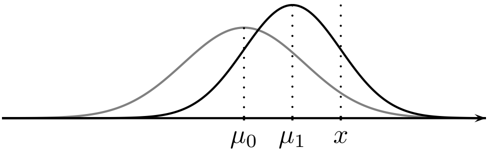
Posterior expectation:

$$\mathbb{E}[\mu|x] = \frac{\tau x + \tau_0 \mu_0}{\tau + \tau_0} = \frac{\tau}{\tau + \tau_0} x + \frac{\tau_0}{\tau + \tau_0} \mu_0 (= \mu_1).$$

Weighted average of prior mean and observation x.

Posterior variance:

$$\mathbb{V}\mathrm{ar}[\mu|x] = \frac{1}{\tau + \tau_0} (= \frac{1}{\tau_1})$$



Posterior as prior — or updating believes

General setup: We are interested in parameter θ.

- Data model: $\pi(x|\theta)$
- Prior: $\pi(\theta)$
- Data: First observation $x_1 \sim \pi(x_1|\theta)$
- Posterior: $\pi(\theta|x_1) \propto \pi(x|\theta)\pi(\theta)$

Assume we have a second independent observation $x_2 \sim \pi(x_2|\theta)$.
Posterior:

$$\begin{aligned}\pi(\theta|x_1, x_2) &\propto \pi(x_1, x_2|\theta)\pi(\theta)\\ &= \pi(x_1|\theta)\pi(x_2|\theta)\pi(\theta)\\ &\propto \underbrace{\pi(x_2|\theta)}_{likelihood}\,\underbrace{\pi(\theta|x_1)}_{prior}\end{aligned}$$

Notice: The posterior after observing x_1 is the prior before observing x_2.

Independent normal case

Posterior mean and precision after one observation x_1:

$$\mu_1 = \frac{x_1\tau + \mu_0\tau_0}{\tau + \tau_0} \quad \text{and} \quad \tau_1 = \tau + \tau_0.$$

Next, μ_1 and τ_1 are prior mean and precision before observing x_2.

Hence, posterior mean and precision after observing (independent) x_1 and x_2 are

$$\begin{aligned}
\mu_2 = \mathbb{E}[\mu_2|x_1, x_2] &= \frac{x_2\tau + \mu_1\tau_1}{\tau + \tau_1} \\
&= \frac{x_2\tau + x_1\tau + \mu_0\tau_0}{\tau + \tau + \tau_0} = \frac{(x_1 + x_2)\tau + \mu_0\tau_0}{2\tau + \tau_0} \\
\tau_2 = 2\tau + \tau_0 &
\end{aligned}$$

This can easily be generalised.

Many independent normal observations

Assume...

- $X_2, X_2, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau)$.
- τ in known.
- Prior $\pi(\mu) = \mathcal{N}(\mu_0, \tau_0)$.

The posterior is then

$$\pi(\mu|x_1, x_2, \ldots, x_n) = \mathcal{N}\left(\frac{\tau \sum_i x_i + \tau_0 \mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right)$$

Posterior mean: Sanity check

The posterior is

$$\pi(\mu|x_1, x_2, \ldots, x_n) = \mathcal{N}\left(\frac{\tau\sum_i x_i + \tau_0\mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right)$$

Does the posterior mean seem "sane"?

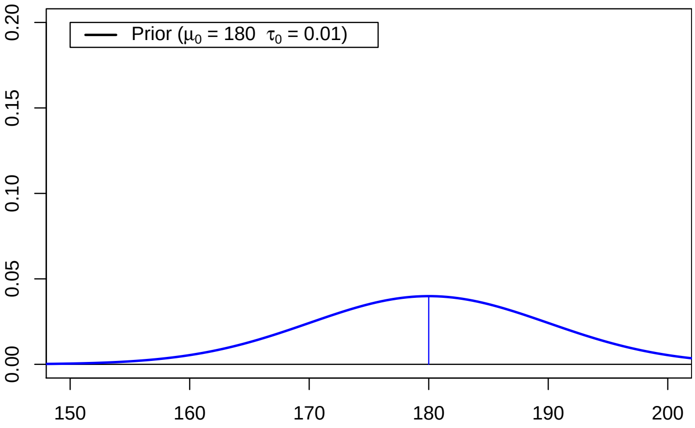
$$\begin{aligned}
\mathbb{E}[\mu|x_1, \ldots, x_n] = \mu_n &= \frac{\tau\sum_i x_i + \tau_0\mu_0}{n\tau + \tau_0} \\
&= \frac{\tau n \frac{1}{n}\sum_i x_i + \tau_0\mu_0}{n\tau + \tau_0} \\
&= \frac{n\tau}{n\tau + \tau_0}\bar{x} + \frac{\tau_0}{n\tau + \tau_0}\mu_0
\end{aligned}$$

Weighted average of sample average $\bar{x}$ and prior mean μ_0.

For n large we have $\mu_n \approx \bar{x}$. Choice of μ_0 of little importance.

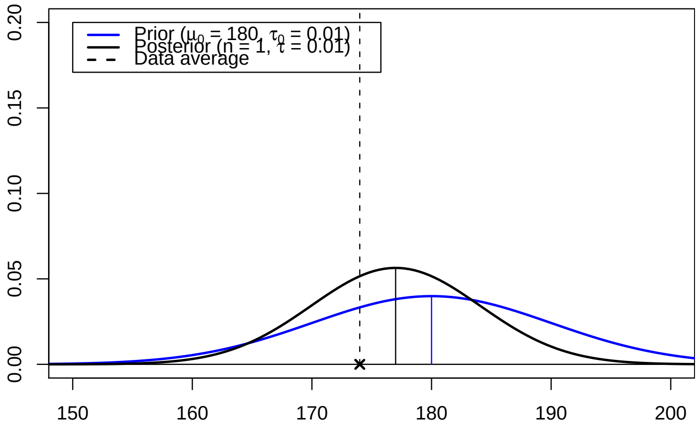
Precision $\tau_n = n\tau + \tau_0$. Knowledge is ever more precise.

Heights in Copenhagen: Prior



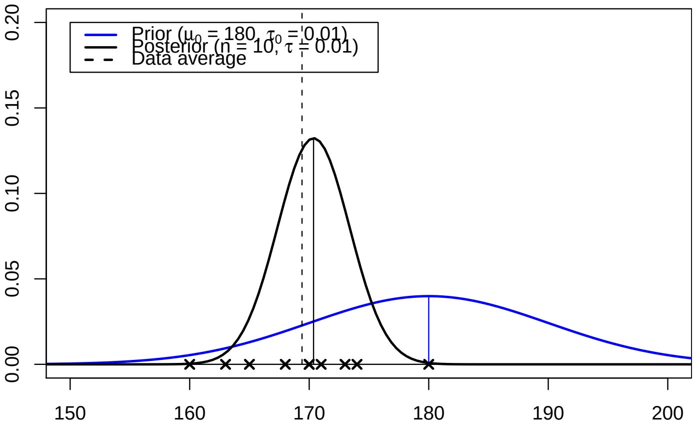
Heights in Copenhagen: Posterior

One observation



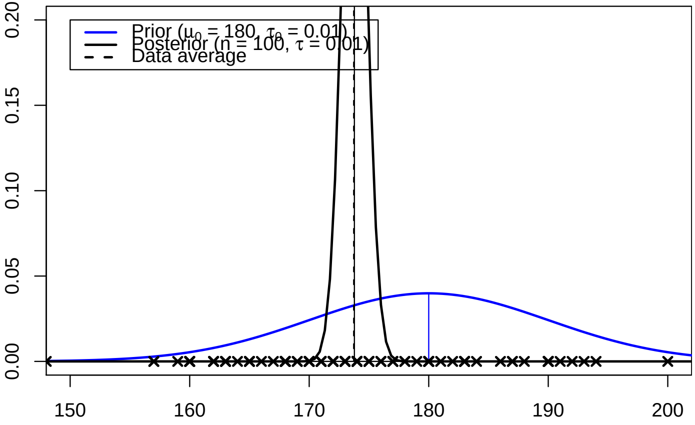
Heights in Copenhagen: Posterior

Ten observations



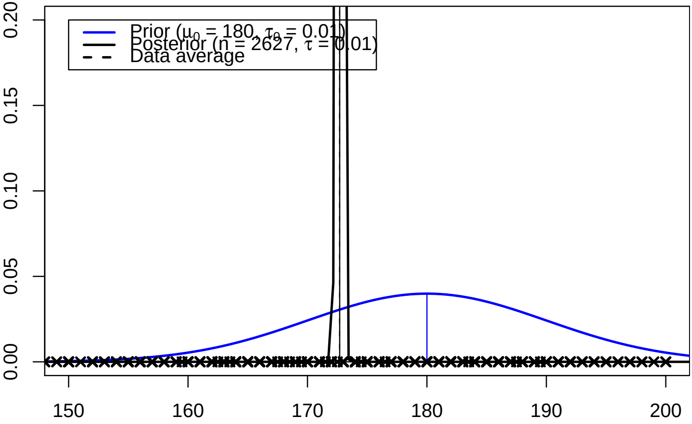
Heights in Copenhagen: Posterior

100 observations



Heights in Copenhagen: Posterior

2627 observations



How to summaries the posterior $\pi(\theta|x)$?

The posterior is usually summaries using one or more of the following:

- Plot of posterior density $\pi(\theta|x)$. See previous slides.
- Posterior mean and variance/precision.
- Central Posterior Interval (CPI). See next slide.
- Maximum A Posteriori (MAP) estimate

$$MAP(\theta) = \underset{\theta}{\text{argmax}}\ \pi(\theta|x).$$

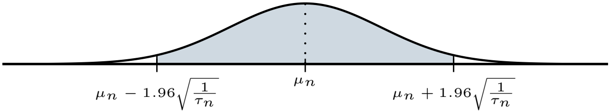
Central Posterior Interval

- The CPI is an interval estimate.
- Also known as credibility interval.
- A 95% CPI for a parameter θ is the shortest (connected) interval which contains θ with 95% posterior probability.
- In case of the normal example, we have

$$P\left(\mu_n - 1.96\sqrt{\frac{1}{\tau_n}} \leq \mu \leq \mu_n + 1.96\sqrt{\frac{1}{\tau_n}}\right) = 0.95$$

Hence, a 95% CPI for μ is

$$95\% \text{ CPI: } \mu_n \pm 1.96\sqrt{\frac{1}{\tau_n}}.$$



CPI compared to confidence interval

The classical 95% confidence interval for μ is

$$95\% \text{ CI: } \bar{x} \pm 1.96\sqrt{\frac{1}{n\tau}}.$$

For CPI: Assume the prior precision is $\tau_0 = 0$, i.e. infinite variance. Then $\mu_n = \bar{x}$ and $\tau_n = n\tau$, i.e.

$$95\% \text{ CPI: } \bar{x} \pm 1.96\sqrt{\frac{1}{n\tau}}.$$

Same interval. Different interpretations.

Conjugate priors

In the normal example: Both prior and posterior were normal! Very convenient!

We say that the normal distribution is conjugate.

Definition: Conjugate priors

Let $\pi(x|\theta)$ be the data model.
A class Π of prior distributions for θ is said to be conjugate for $\pi(x|\theta)$ if

$$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta) \in \Pi$$

whenever $\pi(\theta) \in \Pi$. I.e. prior and posterior are in the same class of distributions.

Notice: Π should be a class of “natural” distributions for this to be useful.

Improper priors

If we have no prior knowledge we may be tempted to use a "flat" prior, i.e.

$$\pi(\theta) \propto k$$

If $\theta \in \mathbf{R}$ this is an example of an improper prior, as

$$\int_{-\infty}^{\infty} \pi(\theta) \mathrm{d}\theta = \int_{-\infty}^{\infty} k = \infty.$$

Problematic, but ok, if posterior is proper, i.e. if

$$\int \pi(\theta|x) \mathrm{d}\theta = \int \pi(x|\theta)\pi(\theta) \mathrm{d}\theta < \infty.$$

Notice: If $\pi(\theta) \propto 1$ then MAP estimator = Maximum likelihood estimator.

Normal example: Unknown precision, known mean

- **Data model**: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$:

$$\pi(x|\tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\tau \sum_{i=1}^{n}(x_i - \mu)^2\right)$$

- **Prior**: Gamma distribution: $\pi(\tau) = Gamma(\alpha, \beta)$

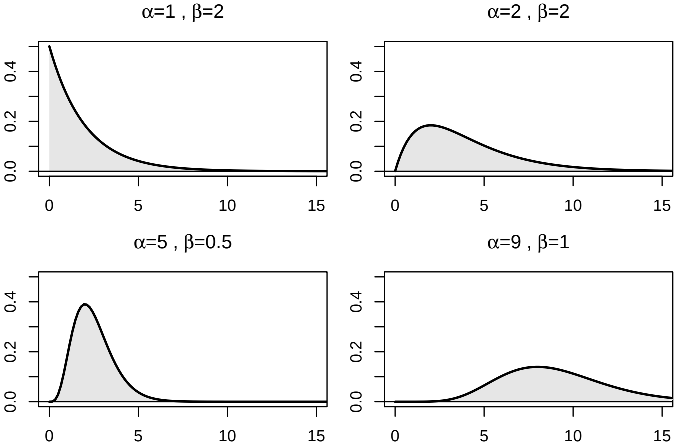
$$\pi(\tau) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \tau^{\alpha-1} \exp\left(-\frac{\tau}{\beta}\right)$$

  Shape parameter α and scale parameter β.

  Properties of the gamma distribution:

$$\mathbb{E}[\tau] = \alpha\beta \qquad \mathbb{V}\text{ar}[\tau] = \alpha\beta^2.$$

Gamma distribution



Normal example: Posterior precision

- **Data model**: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$:
- **Prior**: $\pi(\tau) = Gamma(\alpha, \beta)$
- **Posterior**:

$$\pi(\tau|x) = Gamma\left(\frac{n}{2} + \alpha, \left\{\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta}\right\}^{-1}\right)$$

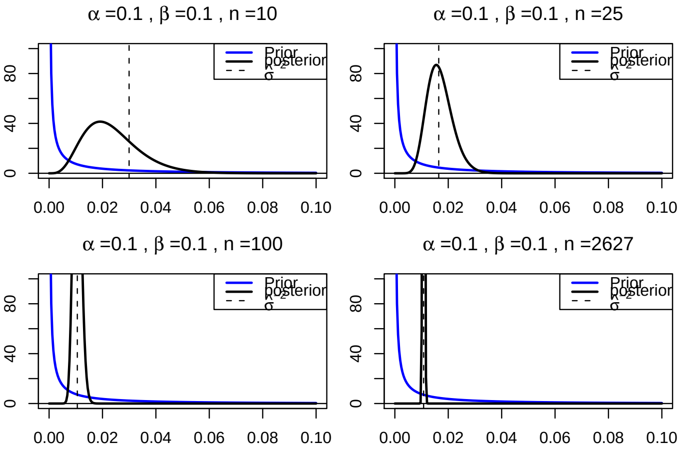
Posterior mean and variance

$$\mathbb{E}[\tau|x] = \frac{\frac{n}{2} + \alpha}{\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta}} \qquad \mathbb{V}\mathrm{ar}[\tau|x] = \frac{\frac{n}{2} + \alpha}{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta}\right)^2}$$

For large n we have

$\mathbb{E}[\tau|x] \approx \dfrac{1}{\hat{\sigma}^2}$ where $\hat{\sigma}^2$ is the usual ML variance estimate.

Known mean: Priors and posteriors



Binomial example

- **Data model**: Binomial, $X \sim \mathsf{B}(n,p)$, n known.

$$\pi(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad 0 \le p \le 1.$$

- **Prior**: Beta distribution,

$$\pi(p) = \mathsf{Be}(\alpha, \beta) \quad , \alpha, \beta > 0.$$

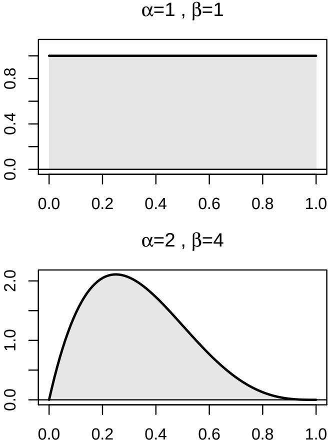
  Where we have to specify α and β.

The Beta distribution has density

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1} & \text{for } 0 \le p \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $\alpha = \beta = 1$ then $\pi(p) = 1$ for $0 \le p \le 1$, i.e. uniform.

Beta distribution: Examples



Mean: $\mathbb{E}[p] = \dfrac{\alpha}{\alpha + \beta}$ Variance: $\mathbb{V}\mathrm{ar}[p] = \dfrac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

Binomial example — cont.

- **Data model**:$X \sim B(n, p)$
- **Prior**: $\pi(p) = Be(\alpha, \beta)$, that is

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1} & \text{for } 0 \leq p \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

- **Posterior**:

$$\begin{aligned} \pi(p|x) &\propto \pi(x|p)\pi(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \quad \cdot \quad \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \\ &= Be(x+\alpha, n-n+\beta). \end{aligned}$$

Posterior mean & Variance

Posterior

$$\pi(p|x) = Be(x+\alpha, n-n+\beta).$$

Posterior mean

$$\mathbb{E}[p|x] = \frac{x+\alpha}{(x+\alpha)+(n-x+\beta)} = \frac{x+\alpha}{\alpha+\beta+n}$$

If $x, n \gg \alpha, \beta$ then $\mathbb{E}[p|x] \approx \frac{x}{n}$.

Posterior variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}[p|x] &= \frac{(x+\alpha)(n-x+\beta)}{(x+\alpha+n-x+\beta)^2(x+\alpha+n-x+\beta+1)} \\ &= \frac{(x+\alpha)(n-x+\beta)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = O\left(\frac{1}{n}\right)\end{aligned}$$

Example: Placentia Previa (PP)

- **Question**: Is the sex ratio different for PP births compared to normal births?
- **Prior knowledge**: 51.5% of new-borns are boys.
- **Data**: Of $n = 980$ cases of PP x=543 were boys (543/980=55.4%).
- **Data model**: $X \sim B(n, p)$.
- **Prior**: $\pi(p) = Be(\alpha, \beta)$.
- **Posterior**:

$$\begin{aligned}\pi(p|x) &= Be(x + \alpha, n - x + \beta) \\ &= Be(543 + \alpha, 437 + \beta)\end{aligned}$$

How to choose α and β, and what difference does it make?

Placenta Previa: Beta priors and posteriors