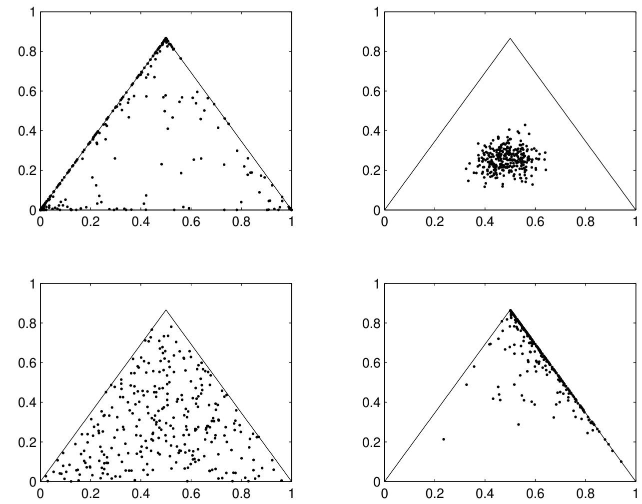
Slides 7: Bayesian inference (1)

- Bayesian vs frequentist approach
- posterior $\propto$ likelihood $\times$ prior
- Maximum Aposteriori Probability
- Conjugate priors
- Binomial-Beta model
- Multinomial-Dirichlet model



Frequentist approach: estimate unknown constant θ by maximising the likelihood function $L(\theta) = f(x|\theta)$.

Bayesian approach treats θ as a random number. New ingredient: a prior distribution $g(\theta)$ which reflects our beliefs on θ before data x is collected.

After the data x is obtained, we update our beliefs on θ using the Bayes formula for the posterior distribution

$$h(\theta|x) = \frac{g(\theta)f(x|\theta)}{\phi(x)},$$

The denominator, the marfinal probability of the data x,

$$\phi(x) = \int f(x|\theta)g(\theta)d\theta \quad \text{or} \quad \phi(x) = \sum_{\theta} f(x|\theta)g(\theta)$$

is treated as a constant and the Bayes formula can be summarised as

$$\boxed{\text{posterior} \propto \text{likelihood} \times \text{prior}}$$

where $\propto$ means proportional.

We define $\hat{\theta}_{\mathrm{map}}$ as the value of θ that maximises $h(\theta|x)$.

With uninformative prior, $g(\theta) = \mathrm{const}$, we get

$$h(\theta|x) \propto f(x|\theta) \text{ so that } \hat{\theta}_{\mathrm{map}} = \hat{\theta}_{\mathrm{mle}}$$

Example. A randomly chosen individual has an unknown IQ value θ. The prior distribution of θ is $\mathrm{N}(100, 15)$ describing the population distribution of IQ with mean of $m = 100$ and standard deviation $v = 15$.

The result x of an IQ measurement is generated by $\mathrm{N}(\theta,\, 10)$. The measurement has a random error of a known size $\sigma = 10$. Since

$$g(\theta) \propto e^{-\frac{(\theta-m)^2}{2v^2}}, \quad f(x|\theta) \propto e^{-\frac{(x-\theta)^2}{2\sigma^2}},$$

and the posterior is proportional to $g(\theta)f(x|\theta)$, we get

$$h(\theta|x) \propto \exp\left\{-\frac{(\theta-m)^2}{2v^2} - \frac{(x-\theta)^2}{2\sigma^2}\right\} \propto \exp\left\{-\frac{(\theta-\gamma m-(1-\gamma)x)^2}{2\gamma v^2}\right\},$$

where $\gamma = \frac{\sigma^2}{\sigma^2+v^2}$ is the so-called shrinkage factor.

We conclude that if the prior is normal and the likelihood is normal, then the posterior distribution is also normal

$$h(\theta|x) \propto e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}, \quad \mu = \gamma m + (1-\gamma)x, \quad \sigma^2 = \gamma v^2$$

In particular, if the observed IQ result is $x = 130$, then the posterior distribution becomes $\mathrm{N}(120.7, 8.3)$. We conclude that

$$\hat{\theta}_{\mathrm{map}} = 120.7$$

lies between the prior expectation $m = 100$ and the observed IQ result $x = 130$.

The posterior variance 69.2 is smaller than that of the prior distribution 225 by the shrinkage factor $\gamma = 0.308$. Our posterior beliefs are less uncertain than the prior beliefs.

Question. What is $\hat{\theta}_{\mathrm{mle}}$ in this example?

Definition. Suppose we have two parametric families of probability distributions $\mathcal{G}$ and $\mathcal{H}$. $\mathcal{G}$ is called a family of conjugate priors to $\mathcal{H}$, if a $\mathcal{G}$-prior and a $\mathcal{H}$-likelihood give a $\mathcal{G}$-posterior.

Below we present five models involving conjugate priors.

Data distribution	Prior	Posterior distribution
$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$	$\mu \sim \mathrm{N}(\mu_0, \sigma_0)$	$\mathrm{N}(\gamma\mu_0 + (1-\gamma)\bar{x}; \sigma_0\sqrt{\gamma})$
$X \sim \mathrm{Bin}(n, p)$	$p \sim \mathrm{Beta}(a, b)$	$\mathrm{Beta}(a + x, b + n - x)$
$(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$	$(p_1, \ldots, p_r) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$	$\mathrm{Dir}(\alpha_1 + x_1, \ldots, \alpha_r + x_r)$
$X_1, \ldots, X_n \sim \mathrm{Geom}(p)$	$p \sim \mathrm{Beta}(a, b)$	$\mathrm{Beta}(a + n, b + n\bar{x} - n)$
$X_1, \ldots, X_n \sim \mathrm{Pois}(\mu)$	$\mu \sim \mathrm{Gam}(\alpha_0, \lambda_0)$	$\mathrm{Gam}(\alpha_0 + n\bar{x}, \lambda_0 + n)$
$X_1, \ldots, X_n \sim \mathrm{Gam}(\alpha, \lambda)$	$\lambda \sim \mathrm{Gam}(\alpha_0, \lambda_0)$	$\mathrm{Gam}(\alpha_0 + \alpha n, \lambda_0 + n\bar{x})$

For the Normal-Normal model, the shrinkage factor

$$\gamma = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

gives the ratio between the posterior variance to the prior variance, and

$$\hat{\mu}_{\mathrm{map}} = \gamma\mu_0 + (1-\gamma)\bar{x}$$

The contribution of the prior distribution becomes smaller for larger n.

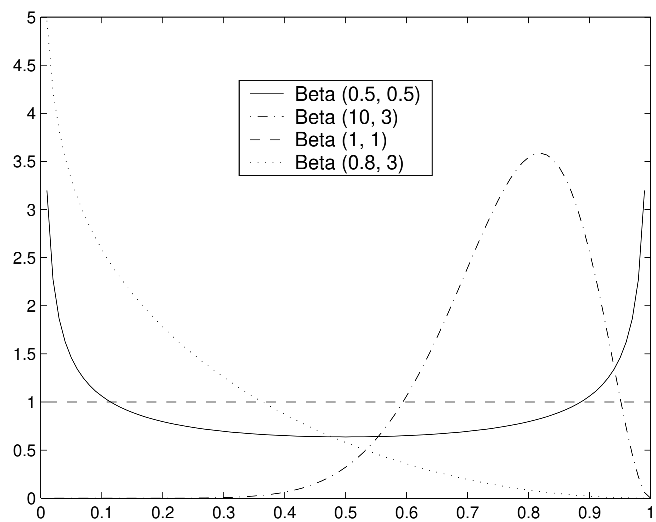
Beta(a, b) distribution is determined by two parameters $a > 0$, $b > 0$ which are called pseudo-counts. It has density,

$$f(p) \propto p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1,$$

with mean and variance having the form

$$\mu = \frac{a}{a+b}, \quad \sigma^2 = \frac{\mu(1-\mu)}{a+b+1}.$$

Beta (a, b) is a rich family of distributions describing a random $p \in (0, 1)$.



For the Binomial-Beta model the update rule has the form

$$\boxed{\text{posterior pseudo-counts} = \text{prior pseudo-counts plus sample counts}}$$

A simple demonstration that beta distribution gives a conjugate prior to the binomial likelihood. If

$$\text{prior } \propto p^{a-1}(1-p)^{b-1},$$

and

$$\text{likelihood } \propto p^{x}(1-p)^{n-x},$$

then obviously posterior is also a beta distribution:

$$\text{postterior } \propto \text{prior } \times \text{likelihood } \propto p^{a+x-1}(1-p)^{b+n-x-1}.$$

You can verify that for $a+x>1$ and $b+n-x>1$, the maximum of the posterior density $\text{Beta}(a+x, b+n-x)$ is attained at

$$\hat{p}_{\text{map}} = \frac{a+x-1}{a+b+n-2}.$$

Question. What is $\text{Beta}(1,1)$-distribution? What is $\hat{p}_{\text{map}}$ if $a=b=1$?

Multinomial distribution is a multivariate extension of the binomial distribution.

Dirichlet distribution is a multivariate extension of the beta distribution.

$\text{Dir}(\alpha_1, \ldots, \alpha_r)$ is a probability distribution over $(p_1, \ldots, p_r)$ with

$$p_1 \geq 0, \ldots, p_r \geq 0, \quad p_1 + \ldots + p_r = 1.$$

Positive $\alpha_1, \ldots, \alpha_r$ are also called pseudo-counts. Dirichlet density

$$f(p_1, \ldots, p_r) \propto p_1^{\alpha_1 - 1} \ldots p_r^{\alpha_r - 1}$$

$\text{Dir}(1, \ldots, 1)$ gives an uninformative prior.

Posterior mean estimates

$$\hat{\theta}_{\text{pme}} = \left(\tfrac{\alpha_1 + x_1}{\alpha_0 + n}, \ldots, \tfrac{\alpha_r + x_r}{\alpha_0 + n}\right)$$

where $\alpha_0 = \alpha_1 + \ldots + \alpha_r$ is the total number of pseudo-counts.

A die is rolled $n = 18$ times, giving 4 ones, 3 twos, 4 threes, 4 fours, 3 fives, and 0 sixes:

$$2, 1, 1, 4, 5, 3, 3, 2, 4, 1, 4, 2, 3, 4, 3, 5, 1, 5.$$

Parameter of interest $\theta = (p_1, \ldots, p_6)$. The MLE

$$\hat{\theta}_{\mathrm{mle}} = (\tfrac{4}{18}, \tfrac{3}{18}, \tfrac{4}{18}, \tfrac{4}{18}, \tfrac{3}{18}, 0)$$

assigns value zero to p_6, effectively excluding future 6 values.

Take uninformative prior $\mathrm{Dir}(1, 1, 1, 1, 1, 1)$ and compare two Bayesian estimates

$$\hat{\theta}_{\mathrm{map}} = (\tfrac{4}{18}, \tfrac{3}{18}, \tfrac{4}{18}, \tfrac{4}{18}, \tfrac{3}{18}, 0), \quad \hat{\theta}_{\mathrm{pme}} = (\tfrac{5}{24}, \tfrac{4}{24}, \tfrac{5}{24}, \tfrac{5}{24}, \tfrac{4}{24}, \tfrac{1}{24}).$$

The latter has an advantage of assigning a positive value to p_6.