Statistical inference (MVE155/MSG200)

Bayesian inference

Bayesian approach

Frequentistic approach: Data x are generated from some population distribution $f(x|\theta)$, where θ is an unknown (constant) parameter.

Bayesian approach:

- Parameters of interest are treated as random variables and generated from some prior distribution $g(\theta)$.
- Given θ, data has the distribution or likelihood $f(x|\theta)$.
- Parameters are estimated by finding the posterior distribution $h(\theta|x)$.

Bayes theorem

We have two events A and B, where $P(A) \neq 0$ and $P(B) \neq 0$. The Bayes theorem says that

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}.$$

Also, for random variables X and Y with density (or probability mass) functions f_X and f_Y, respectively, and $f_X(x) \neq 0$, $f_Y(y) \neq 0$,

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

Posterior distribution

Given a prior distribution $g(\theta)$ and likelihood $f(x|\theta)$, the posterior distribution $h(\theta|x)$ can be computed by using the Bayes theorem:

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

where

$$\phi(x) = \int f(x|\theta)g(\theta)\, d\theta \text{ or } \phi(x) = \sum \mathrm{P}(X = x|\theta)g(\theta)$$

depending on whether X is continuous or discrete. This gives that

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

Note on the prior

- We choose the prior.
- If we do not have any prior information on the parameter(s), we can choose uninformative, uniform priors.
- If we have some prior information, we can take it into account when choosing the prior.
- The prior should be chosen before the data are collected.

Estimating the mean of normal distribution

A sample $x_1, ..., x_n$ from a normal distribution with known variance σ^2.

We choose $N(\mu_0, \sigma_0)$ as the prior distribution $g(\theta)$ for the mean θ and the likelihood

$$f(x_1, ..., x_n|\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

The posterior distribution $h(\theta|x) \propto f(x_1, ..., x_n|\theta)g(\theta)$ is also normal.

Conjugate priors

Let the data be generated from a parametric model having the likelihood $f(x|\theta)$ and let us have a parametric family of prior distributions $\mathcal{G}$.

Then $\mathcal{G}$ is called a family of conjugated priors for the likelihood function $f(x|\theta)$ if for any prior $g(\theta) \in \mathcal{G}$, the posterior

$$h(\theta|x) \propto f(x|\theta)g(\theta)$$

also belongs to $\mathcal{G}$.

Normal distributions are conjugated priors for normal distributions when estimating the mean.

Conjugate priors

Model for the data	θ	Prior	Posterior
$N(\mu, \sigma)$	μ	$N(\mu_0, \sigma_0)$	$N(\gamma_n \mu_0 + (1-\gamma_N)\bar{x}, \sigma_0\sqrt{\gamma_n})$
$Bin(n, p)$	p	$\text{Beta}(a, b)$	$\text{Beta}(a+x, b+n-x)$
$\text{Pois}(\mu)$	μ	$Gam(\alpha_0, \lambda_0)$	$Gam(\alpha_0 + n\bar{x}, \lambda_0 + n)$
$Gam(\alpha, \lambda)$	λ	$Gam(\alpha_0, \lambda_0)$	$Gam(\alpha_0 + \alpha n, \lambda_0 + n\bar{x})$

Above, $\gamma_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$.

Note that as n increases the posterior becomes less effected by the prior.

Binomial-Beta

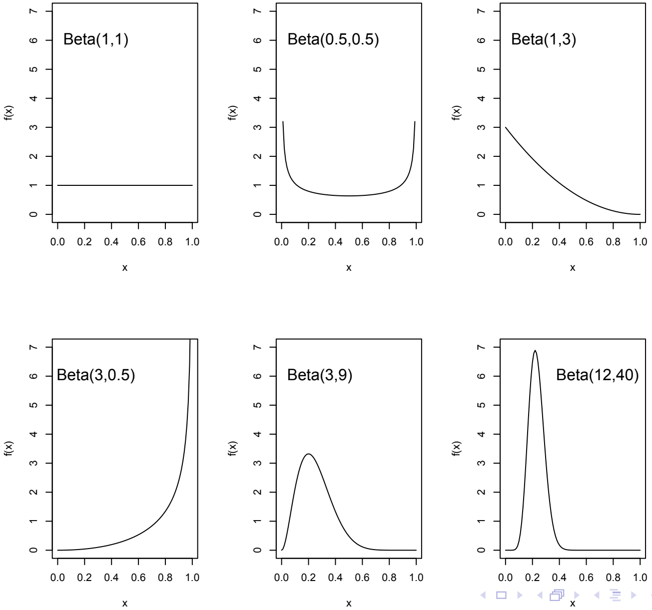
Data: $X \sim Bin(n, p)$, where $X = X_1 + ... + X_n$ and each $X_i \sim Bin(1, p)$.

Task: Estimate p using a Beta(a, b) prior, which has the density function

$$g(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1,$$

where $a > 0$ and $\beta > 0$,

Beta distribution



Point estimate?

A point estimate a for the paramter θ is chosen by minimizing the posterior risk (given the data)

$$R(a|x) = \mathbb{E}(l(\Theta, a)|x)$$

which is computed by using the posterior distribution, i.e.

$$R(a|x) = \int l(\theta, a) h(\theta|x) \, d\theta, \qquad \left(\text{or} \sum_{\theta} l(\theta, a) h(\theta|x) \right)$$

where l is a loss function, for example

- zero-loss function: $l(\theta, a) = \mathbf{1}_{\theta \neq a}$ (maximum a posteriori (map), the value that maximizes the posterior, posterior mode)
- squared loss: $l(\theta, a) = (\theta - a)^2$ (posterior mean)

Credibility intervals

A parameter is a random variable Θ having the (posterior) distribution $h(\theta|x)$ and we can compute $100(1-\alpha)\%$ credibility intervals for Θ. They are of the form

$$J_\theta = (b_1(x), b_2(x))$$

such that

$$\mathrm{P}(b_1(x) < \Theta < b_2(x)|x) = 1 - \alpha.$$

Bayesian hypothesis testing

Consider the case of two simple hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

Likelihood functions connected to these hypotheses are $f(x|\theta_0)$ and $f(x|\theta_1)$ and priors $\mathrm{P}(H_0) = \pi_0$ and $\mathrm{P}(H_1) = \pi_1 = 1 - \pi_0$.

Bayesian hypothesis testing

The rejection region $\mathcal{R}$ and whether to reject the null hypothesis is decided based on the cost function:

	Decision	H_0 true	H_1 true
$x \notin \mathcal{R}$	Do not reject H_0	0	cost_1
$x \in \mathcal{R}$	Reject H_0	cost_0	0

Here,

- cost_0 is the cost for the type I error
- cost_1 the cost for the type II error

Bayesian hypothesis testing

The rejection region is chosen by minimizing the average cost (weighted mean of cost_0 and cost_1)

$$\text{cost}_0 \pi_0 \mathrm{P}(X \in \mathcal{R} | H_0) + \text{cost}_1 \pi_1 \mathrm{P}(X \notin \mathcal{R} | H_1).$$

This leads to rejecting H_0 if

$$\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{\text{cost}_1 \pi_1}{\text{cost}_0 \pi_0},$$

where π_0 / π_1 is called the prior odds and $\text{cost}_1 / \text{cost}_0$ the cost ratio. Equivalently, H_0 is rejected if

$$\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{\text{cost}_1}{\text{cost}_0}.$$

Example (compendium)

The person N, who is charged for rape, is a male of age 37 living in the area not very far from the crime scene. The jury has to decide whether the person is innocent (H_0: N is innocent) or guilty (H_1: N is guilty).

There are three conditionally independent pieces of evidence:

- E1: a DNA match
- E2: defendant N is not recognised by the victim
- E3: an alibi supported by the N's girlfriend.

Example (compendium)

The reliability of E1-E3 was quantified as

- $\mathrm{P}(E1|H_0) = 1/200,000,000$ and $\mathrm{P}(E1|H_1) = 1$
  $\rightarrow$ very strong evidence for H_1,
  $\mathrm{P}(E1|H_0)/\mathrm{P}(E1|H_1) = 1/200,000,000$
- $\mathrm{P}(E2|H_0) = 0.9$ and $\mathrm{P}(E2|H_1) = 0.1$
  $\rightarrow$ strong evidence for H_0, $\mathrm{P}(E2|H_0)/\mathrm{P}(E2|H_1) = 9$
- $\mathrm{P}(E3|H_0) = 0.5$ and $P(E3|H_1) = 0.25$
  $\rightarrow$ some evidence for H_0, $\mathrm{P}(E3|H_0)/\mathrm{P}(E3|H_1) = 2$

The non-informative prior probability

$$\pi_1 = \mathrm{P}(H_1) = 1/200,000$$

was used (thinking about the number of males who theoretically could have committed the crime without any evidence taken into account).

Example (compendium)

Posterior odds become

$$\frac{\mathrm{P}(H_0|E_1,E_2,E_3)}{\mathrm{P}(H_1|E_1,E_2,E_3)} = \frac{\mathrm{P}(E_1|H_0)\mathrm{P}(E_2|H_0)\mathrm{P}(E_3|H_0)\pi_0}{\mathrm{P}(E_1|H_1)\mathrm{P}(E_2|H_1)\mathrm{P}(E_3|H_1)\pi_1} = 0.018.$$

The person N would be found guilty if the cost values assigned by the jury were such that

$$\frac{\text{cost}_1}{\text{cost}_0} = \frac{\text{cost for unpunished crime}}{\text{cost for punishing an innocent}} > 0.018.$$