Statistical inference (MVE155/MSG200)

Parameter estimation

Parameter estimation

Given a parametric model (distribution) which depends on some unknown parameters $\theta = (\theta_1, ..., \theta_k)$, we would like to estimate the parameters from the sample $(x_1, ..., x_n)$.

The two main methods to estimate the parameters are

- method of moments (compares the distribution and sample moments)
- maximum likelihood method (maximises the so-called likelihood function with respect to the parameters).

Method of moments

We have a model (distribution) with, say, two parameters θ_1 and θ_2 and we assume that

$$\mathbb{E}(X) = f(\theta_1, \theta_2), \quad \mathbb{E}(X^2) = g(\theta_1, \theta_2).$$

For example, for the normal distribution $N(\mu, \sigma)$,

- $\mathbb{E}(X) = \mu$
- $\mathbb{E}(X^2) = \sigma^2 + \mu^2$.

Method of moments

The sample moments

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

are consistent estimators for $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$.

The method of moment estimates, $\tilde{\theta}_1$ and $\tilde{\theta}_2$, for the parameters θ_1 and θ_2, respectively, can be found be setting

$$\bar{x} = f(\tilde{\theta}_1, \tilde{\theta}_2), \quad \overline{x^2} = g(\tilde{\theta}_1, \tilde{\theta}_2).$$

Method of moments: normal distribution

For normal distribution $N(\mu, \sigma)$,

$$\mathbb{E}(X) = \mu, \quad \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Method of moment estimates $\tilde{\mu}$ and $\tilde{\sigma^2}$ are

$$\tilde{\mu} = \bar{x}, \quad \tilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Method of moments: geometric distribution

For $X \sim \text{Geom}(p)$,

$$P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \ldots$$

and $\mathbb{E}(X) = \frac{1}{p}$. We will find the method of moments estimate for p by setting

$$\bar{x} = \frac{1}{\tilde{p}}$$

which gives $\tilde{p} = \frac{1}{\bar{x}}$.

Maximum likelihood method

We have a sample $(x_1, ..., x_n)$ (realization of $(X_1, ..., X_n)$) from a population with the population density (or frequency) function $f(x|\theta)$. The joint distribution of the random sample

$$L(\theta) = f(x_1, ..., x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$$

is called a likelihood function. Note that it is treated as a function of the parameter vector θ.

For discrete distributions, the joint frequency or likelihood function gives the probability of observing the given data as a function of θ.

The maximum likelihood (ML) estimate for θ is the one that maximises the likelihood function L. It is denoted by $\hat{\theta}$.

Maximum likelihood method: normal distribution

Let us have a sample $x_1, ...x_n$ from the population distribution $N(\mu, \sigma)$. The likelihood function becomes

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} exp(-\frac{1}{2} \cdot \frac{(x_i - \mu)^2}{\sigma^2}).$$

Often, it is easier to differentiate the log likelihood function $l(\theta) = \ln L(\theta)$ than $L(\theta)$. In our case,

$$l(\mu, \sigma^2) = -\frac{n}{2} ln(n\pi\sigma^2) - \frac{1}{2} \cdot \frac{1}{\sigma^2} \sum_{i+1}^{n} (x_i - \mu)^2.$$

Maximising with respect to μ and σ^2 gives

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Sufficient statistics

Let us have a statistic (a function of the sample $(x_1, ..., x_n)$) $t = g(x_1, ..., x_n)$ such that

$$L(\theta) = f(x_1, ..., x_n|\theta) = h(t, \theta) \cdot c(x_1, ..., x_n) \propto h(t, \theta),$$

where $c(x_1, ..., x_n)$ does not depend on θ. Then, the ML estimate $\hat{\theta}$ depends on the data only through t.
$\rightarrow$ t is called a sufficient statistic for θ.

Sufficient statistics: normal distribution

Let $(x_1, ..., x_n)$ be a sample from $N(\mu, \sigma)$. Then,

$$
\begin{aligned}
L(\mu, \sigma) &= (2\pi\sigma^2)^{-n/2} exp(\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2) \\
&= (2\pi\sigma^2)^{-n/2} exp(\frac{1}{2\sigma^2} (\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2)) \\
&= (2\pi\sigma^2)^{-n/2} exp(\frac{1}{2\sigma^2} (t_2 - 2\mu t_1 + n\mu^2)),
\end{aligned}
$$

where

$$
t_1 = \sum_{i=1}^{n} x_i, \quad t_2 = \sum_{i=1}^{n} x_i^2.
$$

Statistics t_1 and t_2 are sufficient statistics for μ and σ^2. Therefore, if we have two samples with the same t_1 and t_2, they result in the same ML estimates for μ and σ^2.

Sufficient statistics: geometric distribution

For geometric distribution Geom(p), the likelihood function becomes

$$
\begin{aligned}
L(p) &= \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \\
&= p^n \prod_{i=1}^{n} (1-p)^{x_i - 1} \\
&= p^n (1-p)^{\sum_{i=1}^{n} x_i - n} = p^n (1-p)^{t-n},
\end{aligned}
$$

where

$$
t = \sum_{i=1}^{n} x_i
$$

is a sufficient statistic for p.

Large sample properties of ML estimators

Let us have a sample $x_1, ..., x_n$ from the population distribution f with a single parameter θ and the log likelihood function

$$l(\theta) = ln(f(x_1|\theta)) + ... + ln(f(x_n|\theta)).$$

It can be shown that the ML estimator is approximatively normally distributed when the sample size n is large, i.e.

$$\hat{\theta} \approx N(\theta, \frac{\sigma_\theta}{\sqrt{n}}),$$

where σ_θ^2 is the inverse of the so-called Fisher information (variance of the first derivative of the log likelihood function $l(\theta)$, i.e. the expectation of the derivative squared).

Large sample properties of ML estimators: Cramér-Rao inequality

ML estimators are asymptotically efficient estimators in the sense of the Cramér-Rao inequality: If θ^* is an unbiased estimator of θ, then

$$\mathrm{Var}(\theta^*) \geq \frac{\sigma_\theta^2}{n},$$

i.e. the variance is at least the "large sample" variance of the ML estimator.
$\rightarrow$ ML estimator has the smallest variance among all the unbiased estimators.

Also, estimators based on sufficient statistics are more efficient (have smaller variance) than other estimators.

Method of moments and ML estimators for Gam(α, λ)

Method of moments:

$$\tilde{\alpha} = \frac{\bar{x}^2}{\overline{x^2} - \bar{x}^2} \quad \text{and} \quad \lambda = \frac{\bar{x}}{\overline{x^2} - \bar{x}^2}$$

Maximum likelihood:

$$\hat{\alpha} = \hat{\lambda}\,\bar{x} \quad \text{and} \quad n \ln\left(\frac{\hat{\alpha}}{\bar{x}}\right) = n \cdot \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \ln\left(\prod_{i=1}^{n}(x_i)\right)$$

(needs to be computed numerically).