Serik Sagitov, Chalmers and GU, January 31, 2016

Chapter 8. Estimation of parameters

Main issue: given a parametric model with unknown parameters θ estimate θ from an IID random sample $(X_1, \ldots, X_n)$. Two basic methods of finding good estimates

1. method of moments - simple, can be used as a first approximation for the other method,
2. maximum likelihood method - optimal for large samples.

1 List of parametric models

Bernoulli distribution $\mathrm{Ber}(p)$:
$X = 1$ with probability p, and $X = 0$ with probability $q = 1 - p$, $\quad \mu = p$, $\sigma^2 = pq$.
Binomial distribution $\mathrm{Bin}(n, p)$:
X = number of successes in n Bernoulli trials, p = probability of success, $q = 1 - p$,
$\mathrm{P}(X = k) = \binom{n}{k} p^k q^{n-k}$, $0 \le k \le n$, $\quad \mu = np$, $\sigma^2 = npq$.
Hypergeometric distribution $\mathrm{Hg}(N, n, p)$: sampling n elements out of N without replacement,
$\mathrm{P}(X = k) = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}$, $0 \le k \le \min(n, Np)$, $\quad \mu = np$, $\sigma^2 = npq(1 - \frac{n-1}{N-1})$.
Geometric distribution $\mathrm{Geom}(p)$:
X = number of Bernoulli trials until the first success,
$\mathrm{P}(X = k) = pq^{k-1}$, $k \ge 1$, $\quad \mu = \frac{1}{p}$, $\sigma^2 = \frac{q}{p^2}$.
Poisson distribution $\mathrm{Pois}(\lambda)$, an approximation for $\mathrm{Bin}(n, \lambda/n)$ with large n:
X = number of rare events,
$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k \ge 0$, $\quad \mu = \sigma^2 = \lambda$.
Exponential distribution $\mathrm{Exp}(\lambda)$, a continuous version of geometric distribution:
X = life length without aging,
density function $f(x) = \lambda e^{-\lambda x}$, $x > 0$, $\quad \mu = \frac{1}{\lambda}$, $\sigma^2 = \frac{1}{\lambda^2}$.
Normal distribution $\mathrm{N}(\mu, \sigma^2)$,
Central Limit Theorem predicts for the sums of many small almost independent contributions,
density function $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, $\quad -\infty < x < \infty$.
Gamma distribution $\mathrm{Gamma}(\alpha, \lambda)$: shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$,
density function $f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}$, $x > 0$, $\quad \mu = \frac{\alpha}{\lambda}$, $\sigma^2 = \frac{\alpha}{\lambda^2}$,
$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$, $\quad$ in particular for $k = 1, 2, \ldots$, we have $\Gamma(k) = (k-1)!$

2 Method of moments

Suppose we are given IID sample $(X_1, \ldots, X_n)$ from a parametric population distribution $\mathrm{D}(\theta_1, \theta_2)$ with population moments

$$\mathrm{E}(X) = f(\theta_1, \theta_2) \text{ and } \mathrm{E}(X^2) = g(\theta_1, \theta_2).$$

Method of moments estimators $(\tilde\theta_1, \tilde\theta_2)$ are found after replacing the population moment with sample moments, and then solving the equations $\bar X = f(\tilde\theta_1, \tilde\theta_2)$ and $\overline{X^2} = g(\tilde\theta_1, \tilde\theta_2)$.

Example (geometric model)
Data X_i = number of hops that a bird does between flights, $n = 130$:

Number of hops	1	2	3	4	5	6	7	8	9	10	11	12	Tot
Number of birds (Observed frequency)	48	31	20	9	6	5	4	2	1	1	2	1	130

Summary statistics

$\bar{X} = \frac{\text{total number of hops}}{\text{number of birds}} = \frac{363}{130} = 2.79$,

$\overline{X^2} = 1^2 \cdot \frac{48}{130} + 2^2 \cdot \frac{31}{130} + \ldots + 11^2 \cdot \frac{2}{130} + 12^2 \cdot \frac{1}{130} = 13.20$,

$s^2 = \frac{130}{129}(\overline{X^2} - \bar{X}^2) = 5.47$,

$s_{\bar{X}} = \sqrt{\frac{5.47}{130}} = 0.205$.

An approximate 95% CI for μ, the mean number of hops per bird:

$\bar{X} \pm z_{0.025} \cdot s_{\bar{X}} = 2.79 \pm 1.96 \cdot 0.205 = 2.79 \pm 0.40$.

Geometric model $X \sim \text{Geom}(p)$ assumes that a bird does not "remember" the number of jumps made so far. Method of moment estimate for p:

from $\mu = 1/p$ we build an equation $\bar{X} = 1/\tilde{p}$ and find $\tilde{p} = 1/\bar{X} = 0.358$.

We can compute an approximate 95% CI for p using the above CI for μ:

$(\frac{1}{2.79+0.40}, \frac{1}{2.79-0.40}) = (0.31, 0.42)$.

Model fit question: does the geometric distribution fit the data? To answer, compare the observed frequencies to expected frequencies:

j	1	2	3	4	5	6	7+
O_j	48	31	20	9	6	5	11
E_j	46.5	29.9	19.2	12.3	7.9	5.1	9.1

Expected frequencies are computed using geometric distribution with the estimated parameter value:

$E_j = \text{E}(O_j|\text{model}) = n\tilde{q}^{j-1}\tilde{p} = 130 \cdot (0.642)^{j-1}(0.358), j = 1, \ldots, 6$, and $E_7 = 130 - E_1 - \ldots - E_6$.

The chi-square test statistic is small indicating a good fit of the model:

$X^2 = \sum_{j=1}^{7} \frac{(O_j - E_j)^2}{E_j} = 1.86$.

3 Maximum Likelihood method

Before sampling the vector of future observations $(X_1, \ldots, X_n)$ is random and has a joint distribution $f(x_1, \ldots x_n|\theta)$.

After sampling the observed vector $(x_1, \ldots, x_n)$ has a likelihood $L(\theta) = f(x_1, \ldots x_n|\theta)$, which is a function of the unknown population parameter θ. In general, the likelihood function is not a density function.

To illustrate draw three density curves for three parameter values $\theta_1 < \theta_2 < \theta_3$, then show how for a given x, the likelihood curve connects the x-values from the three curves.

$$\boxed{\text{The maximum likelihood estimate } \hat{\theta} \text{ of } \theta \text{ is the value of } \theta \text{ that maximises } L(\theta).}$$

Example (binomial model)

Consider the binomial distribution model $X \sim \text{Bin}(n, p)$, with a single observation corresponding to n observations in the $\text{Ber}(p)$ model. From $\mu = np$, we see that the method of moment estimator $\tilde{p} = \frac{x}{n}$ is the sample proportion.

Likelihood function $L(p) = \binom{n}{x} p^x q^{n-x}$. To maximise log-likelihood function

$\log L(p) = \log \binom{n}{x} + x \log p + (n - x) \log(1 - p)$,

take its derivative $\frac{d \log L(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p}$, and solve the equation $\frac{d \log L(p)}{dp} = 0$. As a results we again obtain the sample proportion $\hat{p} = \frac{x}{n}$, which is consistent with our earlier notation.

4 Large sample properties of the maximum likelihood estimates

For an IID sample $(X_1, \ldots, X_n)$, the likelihood function is given by the product $L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$ due to independence. This implies that the log-likelihood function can be treated as a sum of independent and identically distributed random variables $Y_i = \log f(X_i|\theta)$. Using the central limit theorem argument one can conclude that for large n, we have a

$$\boxed{\text{Normal approximation } \hat{\theta} \overset{a}{\sim} \mathrm{N}(\theta, \tfrac{1}{nI(\theta)})}$$

Fisher information in a single observation: $I(\theta) = \mathrm{E}[\frac{\partial}{\partial\theta} \log f(X|\theta)]^2 = -\mathrm{E}[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)]$.
Maximum likelihood estimators are

asymptotically unbiased, consistent, and asymptotically efficient (has minimal variance),
Cramer-Rao inequality: if θ^* is an unbiased estimator of θ, then $\mathrm{Var}(\theta^*) \geq \frac{1}{nI(\theta)}$.

$$\boxed{\text{Approximate } 100(1-\alpha)\% \text{ CI for } \theta\text{: } \hat{\theta} \pm \tfrac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}}$$

Example (exponential model)
Lifetimes of five batteries measured in hours

$x_1 = 0.5, x_2 = 14.6, x_3 = 5.0, x_4 = 7.2, x_5 = 1.2$.

Consider an exponential model $X \sim \mathrm{Exp}(\lambda)$, where λ is the death rate per hour.
Method of moment estimate:

from $\mu = 1/\lambda$, we find $\tilde{\lambda} = 1/\bar{X} = \frac{5}{28.5} = 0.175$.

The likelihood function grows from 0 to $2.2 \cdot 10^{-7}$ and then falls down

$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \lambda e^{-\lambda x_3} \lambda e^{-\lambda x_4} \lambda e^{-\lambda x_5} = \lambda^n e^{-\lambda(x_1+\ldots+x_n)} = \lambda^5 e^{-\lambda \cdot 28.5}$
the likelihood maximum is reached at $\hat{\lambda} = 0.175$.

For the exponential model the maximum likelihood estimator $\hat{\lambda} = 1/\bar{X}$

is biased but asymptotically unbiased:
$\mathrm{E}(\hat{\lambda}) \approx \lambda$ for large samples, since $\bar{X} \approx \mu$ due to the Law of Large Numbers.

Fisher information for the exponential model is easy to compute:

$\frac{\partial^2}{\partial\lambda^2} \log f(X|\lambda) = -1/\lambda^2, \qquad I(\lambda) = -\mathrm{E}[\frac{\partial^2}{\partial\lambda^2} \log f(X|\lambda)] = \frac{1}{\lambda^2}$.

Thus, $\mathrm{Var}(\hat{\lambda}) \approx \frac{\lambda^2}{n}$ and we get an approximate 95% CI for λ: $0.175 \pm 1.96 \frac{0.175}{\sqrt{5}} = 0.175 \pm 0.153$.

5 Gamma model example

Male height sample of size $n = 24$ in an ascending order:

170,175,176,176,177,178,178,179,179,180,180,180,180,180,181,181,182,183,184,186,187,192,192,199.

Summary statistics: $\bar{x} = 181.46$, $\overline{x^2} = 32964.2$, $\overline{x^2} - \bar{x}^2 = 37.08$.
Gamma distribution model $X \sim \mathrm{Gamma}(\alpha, \lambda)$ is more flexible than the normal distribution model.
First, we may apply the method of moments:

$\mathrm{E}(X) = \frac{\alpha}{\lambda}$ and $\mathrm{E}(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}$ imply $\tilde{\alpha} = \bar{x}^2/(\overline{x^2} - \bar{x}^2) = 887.96$, $\tilde{\lambda} = \tilde{\alpha}/\bar{x} = 4.89$.

Likelihood function

$$L(\alpha, \lambda) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x_i^{\alpha-1} e^{-\lambda x_i} = \frac{\lambda^{n\alpha}}{\Gamma^n(\alpha)} (x_1 \cdots x_n)^{\alpha-1} e^{-\lambda(x_1+\ldots+x_n)},$$

notice that $t_1 = x_1 + \ldots + x_n$ and $t_2 = x_1 \cdots x_n$ are a pair of sufficient statistics containing all information from the data needed to compute the likelihood function.

Maximisation of the log-likelihood function: set two derivatives equal to zero

$\frac{\partial}{\partial\alpha}\log L(\alpha,\lambda) = n\log(\lambda) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log t_2,$

$\frac{\partial}{\partial\lambda}\log L(\alpha,\lambda) = \frac{n\alpha}{\lambda} - t_1.$

Solve numerically two equations

$\log(\hat\alpha/\bar x) = -\frac{1}{n}\log t_2 + \Gamma'(\hat\alpha)/\Gamma(\hat\alpha)$ and $\hat\lambda = \hat\alpha/\bar x,$

using the method of moment estimates $\tilde\alpha = 887.96$, $\tilde\lambda = 4.89$ as the initial values.

Mathematica command

$$\text{FindRoot[Log[a] == 0.00055+Gamma}'[a]\text{/Gamma[a], \{a, 887.96\}]}$$

gives the maximum likelihood estimates $\hat\alpha = 908.76$, $\hat\lambda = 5.01$.

6 Parametric bootstrap

What is the standard error $s_{\hat\alpha}$ of the maximum likelihood estimate $\hat\alpha = 908.76$? No analytical formula is available. If we could simulate from the true population distribution Gamma(α,λ), then B samples of size $n = 24$ would generate B independent estimates $\hat\alpha_j$. The standard deviation of the sampling distribution is the desired standard error:

$\bar\alpha = \frac{1}{B}\sum_{j=1}^B \hat\alpha_j, \qquad s_{\hat\alpha}^2 = \frac{1}{B-1}\sum_{j=1}^B(\hat\alpha_j - \bar\alpha)^2.$

Parametric bootstrap approach: use Gamma$(\hat\alpha,\hat\lambda)$ as a substitute of Gamma(α,λ).

Bootstrap algorithm for finding an approximate 95% CI for α:

$\hat\alpha$ as a substitute for $\alpha \to \hat\alpha_1,\ldots,\hat\alpha_B \to$ sampling distribution of $\hat{\hat\alpha} \to$ 95% brackets c_1, c_2.

Compute a confidence interval as $(2\hat\alpha - c_2, 2\hat\alpha - c_1)$. Explanation of the CI formula:

$$0.95 \approx \mathrm{P}(c_1 < \hat{\hat\alpha} < c_2) = \mathrm{P}(c_1 - \hat\alpha < \hat{\hat\alpha} - \hat\alpha < c_2 - \hat\alpha)$$
$$\approx \mathrm{P}(c_1 - \hat\alpha < \hat\alpha - \alpha < c_2 - \hat\alpha) = \mathrm{P}(2\hat\alpha - c_2 < \alpha < 2\hat\alpha - c_1).$$

Example (male heights)

I simulated $B = 1000$ samples of size $n = 24$ from Gamma(908.76; 5.01) and found $\bar\alpha = 1039.0$, $s_{\hat\alpha} = \sqrt{\frac{1}{999}\sum(\hat\alpha_j - \bar\alpha)^2} = 331.29$. The standard error is large because of small sample size $n = 24$.

Matlab commands for the male heights example:

gamrnd(908.76*ones(1000,24), 5.01*ones(1000,24)),

prctile(x,2.5), prctile(x,97.5).

7 Exact confidence intervals

A restrictive assumption on the population distribution: an IID sample $(X_1,\ldots,X_n)$ is taken from the normal distribution N(μ,σ^2) with unspecified parameters μ and σ.

Exact distribution $\frac{\bar X-\mu}{s_{\bar X}} \sim t_{n-1}$ gives an exact $100(1-\alpha)$% CI for μ: $\bar X \pm t_{n-1}(\alpha/2)\cdot s_{\bar X}$

A t_k-distribution curve looks similar to N(0,1)-curve. Its density function is symmetric around zero:

$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad k \ge 1.$

It has larger spread. If the number of degrees of freedom $k \ge 3$, then the variance is $\frac{k}{k-2}$.

Connection to the standard normal distribution:

if $Z, Z_1, \ldots, Z_k$ are N(0,1) and independent, then $\frac{Z}{\sqrt{(Z_1^2+\ldots+Z_k^2)/n}} \sim t_k$.

Let $\alpha = 0.05$. The exact CI for μ is wider than the approximate confidence interval $\bar{X} \pm 1.96 \cdot s_{\bar{X}}$ valid for the very large n. For example

$\bar{X} \pm 2.26 \cdot s_{\bar{X}}$ for $n = 10$ $\quad$ $\bar{X} \pm 2.13 \cdot s_{\bar{X}}$ for $n = 16$

$\bar{X} \pm 2.06 \cdot s_{\bar{X}}$ for $n = 25$ $\quad$ $\bar{X} \pm 2.00 \cdot s_{\bar{X}}$ for $n = 60$

Exact distribution $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ gives an exact $100(1-\alpha)\%$ CI for σ^2: $\left(\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}; \frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$

The chi-square distribution with k degrees of freedom is the gamma distribution with $\alpha = \frac{k}{2}, \lambda = \frac{1}{2}$.
Connection to the standard normal distribution:
if $Z_1, \ldots, Z_k$ are N(0,1) and independent, then $Z_1^2 + \ldots + Z_k^2 \sim \chi^2_k$.
The exact confidence interval for σ^2 is non-symmetric . Examples of 95% confidence intervals for σ^2:

$(0.47s^2, 3.33s^2)$ for $n = 10$ $\quad$ $(0.55s^2, 2.40s^2)$ for $n = 16$ $\quad$ $(0.61s^2, 1.94s^2)$ for $n = 25$

$(0.72s^2, 1.49s^2)$ for $n = 60$ $\quad$ $(0.94s^2, 1.07s^2)$ for $n = 2000$ $\quad$ $(0.98s^2, 1.02s^2)$ for $n = 20000$

Under the normality assumption $\mathrm{Var}(s^2) = \frac{2\sigma^4}{n-1}$, estimated standard error for s^2 is $\sqrt{\frac{2}{n-1}}s^2$.

8 Sufficiency

Definition: $T = T(X_1, \ldots, X_n)$ is a sufficient statistic for θ, if no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter θ.

If T is sufficient for θ, then the maximum likelihood estimator is a function of T.

Factorisation criterium: T is a sufficient statistic for θ, if and only if
$f(x_1, \ldots, x_n|\theta) = g(t, \theta)h(x_1, \ldots, x_n)$, where $t = T(x_1, \ldots, x_n)$.

Examples

Bernoulli distribution. Since for a single observation, $\mathrm{P}(X = x) = \theta^x(1-\theta)^{1-x}$, it follows that
$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$,
thus the number of successes $T = n\bar{X}$ is a sufficient statistic.

Bernoulli model Ber(p) with n observations = binomial model Bin(n, p) with a single observation.

Normal distribution $\mathrm{N}(\mu, \sigma^2)$ has a two-dimensional sufficient statistic $(t_1, t_2) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$

$$\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{t_2 - 2\mu t_1 + n\mu^2}{2\sigma^2}}.$$

Also recall the gamma distribution model discussed earlier.