Serik Sagitov, Chalmers Tekniska Högskola, January 23, 2013

Chapter 8. Estimation of parameters and fitting of probability distributions

Given a parametric model with unknown parameter(s) θ
estimate θ from a random sample $(X_1,\ldots,X_n)$
Two basic methods of finding good estimates
1. method of moments, simple, first approximation for
2. max likelihood method, good for large samples

1 Parametric models

Binomial $\mathrm{Bin}(n,p)$: number of successes in n Bernoulli trials, $f(k)=\binom{n}{k}p^kq^{n-k}$, $0\le k\le n$.
Mean and variance $\mu=np$, $\sigma^2=npq$.
Hypergeometric $\mathrm{Hg}(N,n,p)$: sampling without replacement, $f(k)=\frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}$, $0\le k\le \min(n,Np)$.
Mean and variance $\mu=np$, $\sigma^2=npq(1-\frac{n-1}{N-1})$. Finite population correction FPC$=1-\frac{n-1}{N-1}$.
Geometric $\mathrm{Geom}(p)$: number of trials until the first success, $f(k)=pq^{k-1}$, $k\ge1$, $\mu=\frac{1}{p}$, $\sigma^2=\frac{q}{p^2}$.
Poisson $\mathrm{Pois}(\lambda)$: number of rare events $\approx \mathrm{Bin}(n,\lambda/n)$, $f(k)=\frac{\lambda^k}{k!}e^{-\lambda}$, $k\ge0$, $\mu=\sigma^2=\lambda$.
Exponential $\mathrm{Exp}(\lambda)$: Poisson process waiting times $f(x)=\lambda e^{-\lambda x}$, $x>0$, $\mu=\sigma=\frac{1}{\lambda}$.
Normal $\mathrm{N}(\mu,\sigma^2)$, CLT: many small independent contributions $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, $-\infty<x<\infty$.
Gamma(α,λ): shape α and scale parameter λ, $f(x)=\frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}$, $x\ge0$, $\mu=\frac{\alpha}{\lambda}$, $\sigma^2=\frac{\alpha}{\lambda^2}$.

2 Method of moments

Suppose we are given IID sample $(X_1,\ldots,X_n)$ from $\mathrm{PD}(\theta_1,\theta_2)$ 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 estimates MME $(\tilde\theta_1,\tilde\theta_2)$: solve equations $\bar X=f(\tilde\theta_1,\tilde\theta_2)$ and $\overline{X^2}=g(\tilde\theta_1,\tilde\theta_2)$.

Example. Bird hops. Data X_i = nunber of hops that a bird does between flights, $n=130$:

No. hops	1	2	3	4	5	6	7	8	9	10	11	12	Tot
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: $\bar X\pm z_{0.025}\cdot s_{\bar X}=2.79\pm1.96\cdot0.205=2.79\pm0.40$.

Geometric model $X \sim \text{Geom}(p)$: from $\mu = 1/p$ we find a MME $\tilde{p} = 1/\bar{X} = 0.358$.

Approximate 95% CI for p: $(\frac{1}{2.79+0.40}, \frac{1}{2.79-0.40}) = (0.31, 0.42)$.

Model fit: compare the observed frequencies to expected:

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

$E_j = 130 \cdot (0.642)^{j-1}(0.358)$ and $E_7 = 130 - E_1 - \ldots - E_6$. The chi-square test statistic is small $X^2 = 1.86$ saying that the model is good.

3 Maximum Likelihood method

Before sampling the random vector $X_1, \ldots, X_n$ 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 θ.
To illustrate draw three density curves for three parameter values $\theta_1 < \theta_2 < \theta_3$: the likelihood curve connects the x-values from the three curves.

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

For the $\text{Bin}(n, p)$ model the sample proportion is MME and MLE of p.

Large sample properties of MLE

If sample is iid, then the likelihood function is given by $L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$ due to independence. This implies for large n

$$\boxed{\text{Normal approximation } \hat{\theta} \in \text{N}(\theta, \tfrac{1}{nI(\theta)})}$$

Fisher information in a single observation: $I(\theta) = \text{E}[\frac{\partial}{\partial\theta} \log f(X|\theta)]^2 = -\text{E}[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)]$
MLE $\hat{\theta}$ is asymptotically unbiased, consistent, and asymptotically efficient (has minimal variance).
Cramer-Rao inequality: if θ^* is an unbiased estimate of θ, the $\text{Var}(\theta^*) \geq \frac{1}{nI(\theta)}$.

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

Example. Battery lifetime. 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 $X \sim \text{Exp}(\lambda)$, where λ is the death rate per hour. MME calculation:
$\mu = 1/\lambda$, $\tilde{\lambda} = 1/\bar{X} = \frac{5}{28.5} = 0.175$.
The likelihood function
$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \lambda e^{-\lambda x_3} \lambda e^{-\lambda x_4} \lambda e^{-\lambda x_5} = \lambda^n e^{-\lambda(x_1+\ldots+x_n)} = \lambda^5 e^{-\lambda \cdot 28.5}$
grows from 0 to $2.2 \cdot 10^{-7}$ and then falls down. The likelihood maximum is reached at $\hat{\lambda} = 0.175$.
For the exponential model the MLE $\hat{\lambda} = 1/\bar{X}$ is biased but asymptotically unbiased: $\text{E}(\hat{\lambda}) \approx \lambda$ for large samples, since $\bar{X} \approx \mu$ due to the Law of Large Numbers.
Fisher information can be computed $\frac{\partial^2}{\partial\lambda^2} \log f(X|\lambda) = -1/\lambda^2$, $I(\lambda) = \frac{1}{\lambda^2}$. Thus, $\text{Var}(\hat{\lambda}) \approx \frac{\lambda^2}{n}$ and we get an approximate 95% CI for λ: $0.175 \pm 1.96 \frac{0.175}{\sqrt{5}} = 0.175 \pm 0.153$.

4 Gamma model example

Male height sample of size $n = 24$

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 model $X \sim \text{Gamma}(\alpha, \lambda)$ is more flexible than the normal model. First we may us the method of moments:

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

Maximum likelihood function

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

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

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

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

Solve numerically two equations

$\log(\hat{\alpha}/\bar{X}) = -\frac{1}{n}\sum \log X_i + \Gamma'(\hat{\alpha})/\Gamma(\hat{\alpha})$,

$\hat{\lambda} = \hat{\alpha}/\bar{X}$, using MME $\tilde{\alpha} = 887.96$, $\tilde{\lambda} = 4.89$ as the initial values.

Mathematica command

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

gives MLE $\hat{\alpha} = 908.76, \hat{\lambda} = 5.01$ which are not far from the MME.

5 Parametric bootstrap

Simulate

1000 samples of size 24 from Gamma(908.76; 5.01)

find 1000 estimates $\hat{\alpha}_j$ and plot a histogram

Use the simulated sampling distribution of $\hat{\alpha}$ and $\hat{\lambda}$

to find $\bar{\alpha} = 1039.0$ and $s_{\hat{\alpha}} = \sqrt{\frac{1}{999}\sum(\hat{\alpha}_j - \bar{\alpha})^2} = 331.29$

large standard error because of small $n = 24$

Bootstrap algorithm to find approximate 95% CI: $(2\hat{\alpha} - c_2, 2\hat{\alpha} - c_1)$

$\hat{\alpha} \to \hat{\alpha}_1, \ldots, \hat{\alpha}_B \to$ sampling distribution of $\hat{\hat{\alpha}} \to$ 95% brackets c_1, c_2.

Explanation of the CI formula:

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

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

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).

6 Exact confidence intervals

Assumption on the PD

an IID sample $(X_1, \ldots, X_n)$ is taken from $\text{N}(\mu, \sigma^2)$ with unspecified parameters μ and σ.

$$\boxed{\text{Exact distributions } \frac{\bar{X}-\mu}{s_{\bar{X}}} \sim t_{n-1} \text{ and } \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}}$$

t_{n-1}-distribution curve looks similar to N(0,1)-curve: symmetric around zero, larger variance $= \frac{n-1}{n-3}$.
If $Z, Z_1, \ldots, Z_k$ are N(0,1) and independent, then $\frac{Z}{\sqrt{(Z_1^2+\ldots+Z_k^2)/n}} \sim t_k$.
If $Z_1, \ldots, Z_k$ are N(0,1) and independent, then $Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2$.
Different shapes of χ_k^2-distribution: $\mu = k$, $\sigma^2 = 2k$. It is a Gamma$(k/2, 1/2)$-distribution.

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ CI for } \mu\text{: } \bar{X} \pm t_{n-1}(\alpha/2) \cdot s_{\bar{X}}}$$

Exact CI for μ is wider than the approximate CI

$\bar{X} \pm 1.96 \cdot s_{\bar{X}}$	approximate CI for large n
$\bar{X} \pm 2.26 \cdot s_{\bar{X}}$	exact CI for $n = 10$
$\bar{X} \pm 2.13 \cdot s_{\bar{X}}$	exact CI for $n = 16$
$\bar{X} \pm 2.06 \cdot s_{\bar{X}}$	exact CI for $n = 25$
$\bar{X} \pm 2.00 \cdot s_{\bar{X}}$	exact CI for $n = 60$

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ CI for } \sigma^2\text{: } \left(\frac{(n-1)s^2}{\chi_{n-1}^2(\alpha/2)}; \frac{(n-1)s^2}{\chi_{n-1}^2(1-\alpha/2)}\right)}$$

A non-symmetric exact confidence interval for σ^2. Examples:

$(0.47s^2, 3.33s^2)$ for $n = 10$	$(0.55s^2, 2.40s^2)$ for $n = 16$
$(0.61s^2, 1.94s^2)$ for $n = 25$	$(0.72s^2, 1.49s^2)$ for $n = 60$
$(0.94s^2, 1.07s^2)$ for $n = 2000$	$(0.98s^2, 1.02s^2)$ for $n = 20000$

7 Sufficiency

Definition: $T = T(X_1, \ldots, X_n)$ is a sufficient statistic for θ, if given $T = t$ conditional distribution of $(X_1, \ldots, X_n)$ does not depend on θ.

$$\boxed{\text{A sufficient statistic } T \text{ contains all the information in the sample about } \theta}$$

Factorization criterium:
if $f(x_1, \ldots, x_n|\theta) = g(t, \theta)h(x_1, \ldots, x_n)$, then $\mathrm{P}(\mathbf{X} = \mathbf{x}|T = t) = \frac{h(\mathbf{x})}{\sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})}$ independent of θ.

$$\boxed{\text{If } T \text{ is sufficient for } \theta, \text{ the MLE is a function of } T}$$

Bernoulli distribution
$\mathrm{P}(X_i = x) = \theta^x(1-\theta)^{1-x}$
$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$.
Sufficient statistic is the number of successes $T = n\bar{X}$. Factorization: $g(t, \theta) = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$.
Normal distribution $\mathrm{N}(\mu, \sigma^2)$ has a two-dimensional sufficient statistic $(t_1, t_2) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$

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

Rao-Blackwell theorem.
Consider two estimates of θ: $\hat{\theta}$ and $\tilde{\theta} = \mathrm{E}(\hat{\theta}|T)$. If $\mathrm{E}(\hat{\theta}^2) < \infty$, then $\mathrm{MSE}(\tilde{\theta}) \leq \mathrm{MSE}(\hat{\theta})$.