TWO (AMONG MANY) WAYS TO FIND ESTIMATORS

The Rao-Blackwell theorem gives good reasons to base estimators on sufficient statistics, but it doesn't tell us what function of a sufficient statistic we should use.

In the following we discuss two methods of constructing estimators. The first, called the method of moments is a bit of a relic, does not generally guarantee good estimators, and is seldom used. The second, called maximum likelihood estimation, always leads to an estimator which is a function of a sufficient statistic, and is one of the most widely used estimation methods.

THE METHOD OF MOMENTS

Let $X_1, \dots, X_n$ be independent r.v.s with the same distribution as X, which depends on the parameters $(\theta_1, \dots, \theta_d) \in \Theta \subset \mathbb{R}^d$. If the first d moments $\mathbb{E}X, \dots, \mathbb{E}X^d$ are finite, then the method of moments (MoMs) estimators $\hat{\theta}_1, \dots, \hat{\theta}_d$ of $\theta_1, \dots, \theta_d$ are the values of $\theta_1, \dots, \theta_d$ which solve the following system of equations:

$$\underbrace{\left\{\begin{array}{l} m_1' := \frac{1}{n}\sum_{i=1}^n X_i \\ \quad\vdots \\ m_d' := \frac{1}{n}\sum_{i=1}^n X_i^d \end{array}\right.}_{\text{Sample moments}} \begin{array}{l} = \mathbb{E}X =: \mu_1'(\theta_1, \dots, \theta_d) \\ \\ = \mathbb{E}X^d =: \mu_d'(\theta_1, \dots, \theta_d) \end{array} \underbrace{\Big\}}_{\text{Moments of } X \text{ as functions of } \theta_1, \dots, \theta_d}$$

So we get the method of moments estimators of $\theta_1, \dots, \theta_d$ by matching the first d moments of the sample to the first d moments of the population distribution.

Example: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. Find the MoMs estimators of μ and σ^2.

The equations

$$m_1' = \mu_1' = \mu$$

$$m_2' = \mu_2' = \sigma^2 + \mu^2 \qquad \left(\mathbb{E}X^2 = \text{Var}\,X + (\mathbb{E}X)^2\right)$$

give $\hat{\mu} = m_1'$ and $\hat{\sigma}^2 = m_2' - (m_1')^2$,

So $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. Find the MoMs estimators of α and β.

We have:

$$m_1' = \mu_1' = \alpha\beta$$

$$m_2' = \mu_2' = \alpha\beta^2 + (\alpha\beta)^2$$

$\Leftrightarrow$

$$\alpha = \frac{1}{\beta} m_1'$$

$$m_2' = \left(\frac{1}{\beta} m_1'\right)\beta^2 + \left[\left(\frac{1}{\beta} m_1'\right)\beta\right]^2 = m_1'\beta + (m_1')^2$$

$\Leftrightarrow$

$$\hat{\beta} = \frac{m_2' - (m_1')^2}{m_1'} = \frac{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2}{\bar{X}_n} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\bar{X}_n}$$

$$\hat{\alpha} = \frac{(m_1')^2}{m_2' - (m_1')^2} = \frac{(\bar{X}_n)^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2} = \frac{(\bar{X}_n)^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Recall that $T(X_1, \ldots, X_n) = \left(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i\right)$ is a sufficient statistic for (α, β).

The MoMs estimator does not involve the quantity $\prod_{i=1}^n X_i$, and therefore, by the Rao-Blackwell theorem, the MoMs estimator cannot be the MVUE.

It can be verified, however, that $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators of α and β.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$. Find the MoMs estimator for θ.

We have $m_1' = \mu_1' = \frac{\theta}{2}$, so $\hat{\theta} = 2m_1' = 2\bar{X}_n$.

We investigated the properties of this estimator before and found that it is much worse, in terms of MSE, than the estimator $X_{(n)}$, although it is unbiased.

The Rao-Blackwell theorem tells us that the best unbiased estimator, the MVUE, of θ is a function of $X_{(n)}$, so we should not use the MoMs estimator in this situation.

Example: Let $Y \sim p_Y(y) = (1-p)^{y-1} p$ be a single realization of a r.v. with the Geometric(p) distribution. Find the MoMs estimator for p.

We have $\underbrace{Y}_{\substack{Y\text{, being the only observation,}\\ \text{is also the sample mean}}} = m_1' = \mu_1' = \frac{1}{p}$,

giving the MoMs estimator $\hat{p} = \frac{1}{Y}$.

E.g., if the first success occurs on the fifth trial, we estimate the success probability to be one-fifth.

MAXIMUM-LIKELIHOOD ESTIMATION

Maximum likelihood estimation is one of the most widely used methods for finding estimators of unknown parameter values.

In essence, having observed some data, we find the values of the unknown parameters under which the observed data were most likely to occur.

To define the maximum likelihood estimator we will refer to the joint pdf/pmf (as the case may be) of the data as the likelihood function.

Defn: Let $X_1, \dots, X_n$ be a random sample from a distribution with pdf $f_X(x; \theta_1, \dots, \theta_d)$ or $p_X(x; \theta_1, \dots, \theta_d)$ depending on some parameters $(\theta_1, \dots, \theta_d) \in \Theta \subset \mathbb{R}^d$. We denote by

$$\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n) = \begin{cases} \prod_{i=1}^n f_X(X_i; \theta_1, \dots, \theta_d), & X_1, \dots, X_n \overset{iid}{\sim} f_X(x; \theta_1, \dots, \theta_d) \\ \prod_{i=1}^n p_X(X_i; \theta_1, \dots, \theta_d), & X_1, \dots, X_n \overset{iid}{\sim} p_X(x; \theta_1, \dots, \theta_d) \end{cases}$$

the joint pdf/pmf (as the case may be) of $X_1, \dots, X_n$, regarded as a function of $\theta_1, \dots, \theta_d$, and we call this function the <u>likelihood function</u> of $X_1, \dots, X_n$.

Moreover, we define the function

$$\ell(\theta_1, \dots, \theta_d; X_1, \dots, X_n) = \log \mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n),$$

and we refer to it as the <u>log-likelihood function</u>.

Remark: When $X_1, \dots, X_n$ are not iid we still define $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ as the joint pdf/pmf of $X_1, \dots, X_n$; then it is no longer the product of marginal pdfs/pmfs. We consider only the case of iid data.

We may now define the maximum likelihood estimator:

Defn: Let $X_1, \dots, X_n$ be r.v.s with likelihood function $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ for some parameters $(\theta_1, \dots, \theta_d) \in \Theta \subset \mathbb{R}^d$. Then the maximum likelihood estimators (MLEs) of $\theta_1, \dots, \theta_d$ are the values $\hat{\theta}_1, \dots, \hat{\theta}_d$ which maximize the likelihood function over all $(\theta_1, \dots, \theta_d) \in \Theta$. That is

$$
\begin{aligned}
(\hat{\theta}_1, \dots, \hat{\theta}_d) &= \underset{(\theta_1, \dots, \theta_d) \in \Theta}{\operatorname{argmax}} \; \mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n) \\
&= \underset{(\theta_1, \dots, \theta_d) \in \Theta}{\operatorname{argmax}} \; \ell(\theta_1, \dots, \theta_d; X_1, \dots, X_n).
\end{aligned}
$$

(The same values maximize the log-likelihood.)

The likelihood function is so named because it is an index of the probability of observing the data at hand. If $X_1, \dots, X_n$ are observed and

$$\mathcal{L}(a_1, \dots, a_d; X_1, \dots, X_n) > \mathcal{L}(b_1, \dots, b_d; X_1, \dots, X_n),$$

then the observed data were more "likely" under the parameter values $a_1, \dots, a_d$ than under $b_1, \dots, b_d$.

The likelihood function is also so named because it does not in general represent a probability; if $X_1, \dots, X_n$ are discrete, $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ is the probability of observing $X_1, \dots, X_n$ under $\theta_1, \dots, \theta_d$, but if $X_1, \dots, X_n$ are continuous, then $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ is equal to the height of a joint density, which is not a probability. Using the word "likelihood" is a way of tip-toeing around the fact that $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ is not in general a probability.

The data $X_1, \dots, X_n$ may also consist of some discrete and some continuous random variables, in which case $\mathcal{L}(\theta_1, \dots, \theta_d; X_1, \dots, X_n)$ is neither a probability nor the height of a joint density. In this case its interpretation is especially problematic, and the name "likelihood," for its ambiguity, is especially appropriate.

We very often find the MLE by using the following result, though it does not apply in every setting:

Result: If $\ell(\theta_1,\dots,\theta_d; X_1,\dots,X_n)$ is differentiable and has a single maximum in the interior of Θ, then we may find the MLEs $\hat\theta_1,\dots,\hat\theta_d$ by solving the system of equations

$$\frac{\partial}{\partial\theta_1}\ell(\theta_1,\dots,\theta_d; X_1,\dots,X_n) = 0$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial}{\partial\theta_d}\ell(\theta_1,\dots,\theta_d; X_1,\dots,X_n) = 0.$$

This result will apply in most cases we encounter in this course, with a few notable exceptions.

We often find the log-likelihood easier to differentiate, but this result would still hold with $\ell(\theta_1,\dots,\theta_d; X_1,\dots,X_n)$ replaced by $\mathcal{L}(\theta_1,\dots,\theta_d; X_1,\dots,X_n)$.

Example: Let $X_1,\dots,X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the MLE for λ.

We have

$$\mathcal{L}(\lambda; X_1,\dots,X_n) = \prod_{i=1}^n \frac{1}{\lambda} e^{-\frac{X_i}{\lambda}} \mathbb{1}(X_i>0)$$

so

$$\ell(\lambda; X_1,\dots,X_n) = \sum_{i=1}^n \log\left(\frac{1}{\lambda} e^{-\frac{X_i}{\lambda}} \mathbb{1}(X_i>0)\right)$$

$$= \sum_{i=1}^n \left[\log\frac{1}{\lambda} - \frac{X_i}{\lambda} + \log \mathbb{1}(X_i>0)\right]$$

$$= -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^n X_i + \sum_{i=1}^n \log\mathbb{1}(X_i>0)$$

Now

$$\frac{\partial}{\partial\lambda}\ell(\lambda; X_1,\dots,X_n) = \underbrace{-\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^n X_i = 0}_{\text{solve for } \lambda}$$

$$\iff \quad \lambda = \frac{1}{n}\sum_{i=1}^n X_i = \bar X_n.$$

So the MLE of λ is $\hat\lambda = \bar X_n$.

Example: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Find the MLE for λ.

We have, for $x_1, \dots, x_n \in \{1, 2, \dots\}$

$$L(\lambda; x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{(x_i)!}$$

and

$$\ell(\lambda; x_1, \dots, x_n) = \sum_{i=1}^{n} \left[-\lambda \log(e) + x_i \log \lambda - \log(x_i!) \right]$$

$$= -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!) .$$

Now

$$\frac{\partial}{\partial \lambda} \ell(\lambda; x_1, \dots, x_n) = \underbrace{-n + \sum_{i=1}^{n} x_i \left(\frac{1}{\lambda}\right) = 0}_{\text{solve for } \lambda}$$

$$\iff \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n .$$

So the MLE of λ is $\hat{\lambda} = \bar{X}_n$.

Example: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. Find the MLEs of μ and σ^2.

We have

$$L(\mu, \sigma^2; x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right],$$

and

$$\ell(\mu, \sigma^2; x_1, \dots, x_n) = \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} - \frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

Now solving

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2; x_1, \dots, x_n) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Differentiate w.r.t. σ^2

$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2; x_1, \dots, x_n) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

for μ and σ^2 gives the MLEs

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

* We omit verification of the conditions ensuring that $(\hat{\mu}, \hat{\sigma}^2)$ is the global maximizer of $\ell(\mu, \sigma^2; X_1, \dots, X_n)$.

In the following example we cannot use calculus methods to find the MLE. In this case it is because the support of the r.v.s depends on the parameter so that the likelihood is not differentiable.

Example: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$. Find the MLE of θ.

We have

$$\mathcal{L}(\theta; X_1, \dots, X_n) = \prod_{i=1}^{n} \frac{1}{\theta} \underbrace{\mathbb{1}(0 < X_i < \theta)}_{\text{Not differentiable in } \theta.}$$

$$= \left(\frac{1}{\theta}\right)^n \underbrace{\prod_{i=1}^{n} \mathbb{1}(0 < X_i < \theta)}_{= \begin{cases} 1 & \text{if } 0 < X_{(1)} < X_{(n)} < \theta \\ 0 & \text{otherwise} \end{cases}}$$

$$= \underbrace{\left(\frac{1}{\theta}\right)^n}_{} \mathbb{1}(0 < X_{(1)} < X_{(n)} < \theta)$$

Decreasing in θ, so take smallest value of θ admitted by $0 < X_{(1)} < X_{(n)} < \theta$.

By studying $\mathcal{L}(\theta; X_1, \dots, X_n)$, we see that the MLE of θ is $\hat{\theta} = X_{(n)}$.

The following is a nice property of MLEs:

Result: The MLE is always a function of a sufficient statistic.

Proof: Suppose $T(X_1,\dots,X_n)$ is a sufficient statistic for $(\theta_1,\dots,\theta_d)$. Then the likelihood function of $X_1,\dots,X_n$, which is the joint pdf/pmf of $X_1,\dots,X_n$, admits a factorization of the form

$$\mathcal{L}(\theta_1,\dots,\theta_d\,;X_1,\dots,X_n) = g(T(X_1,\dots,X_n);\theta_1,\dots,\theta_d)\cdot h(X_1,\dots,X_n)$$

for some functions g and h.

Therefore the maximizer $(\hat\theta_1,\dots,\hat\theta_d)$ of $\mathcal{L}(\theta_1,\dots,\theta_d\,;X_1,\dots,X_d)$ is the same as the maximizer of $g(T(X_1,\dots,X_n);\theta_1,\dots,\theta_d)$ in $\theta_1,\dots,\theta_d$, since $h(X_1,\dots,X_n)$ is constant in $\theta_1,\dots,\theta_d$. The minimizer of $g(T(X_1,\dots,X_n);\theta_1,\dots,\theta_d)$ in $\theta_1,\dots,\theta_d$ is a function of $T(X_1,\dots,X_n)$.

Remark: Note that MoMs estimators do not have this property. In the setting $X_1,\dots,X_n \overset{iid}{\sim} \text{Uniform}(0,\theta)$, we have

$$\text{MoMs:}\quad \hat\theta_{MoM} = 2\bar X_n$$

$$\text{MLE:}\quad \hat\theta_{MLE} = X_{(n)}$$

Recall that $X_{(n)}$ is a sufficient statistic for θ, while $\bar X_n$ is not.

Note that the MLE is biased: $\mathbb{E}\,\hat\theta_{MLE} = \mathbb{E}\,X_{(n)} = \left(\frac{n}{n+1}\right)\theta$.

However, if we modify it, we get an unbiased estimator

$$\hat\theta^{\text{unbiased}} = \left(\frac{n+1}{n}\right)\hat\theta_{MLE} = \left(\frac{n+1}{n}\right)X_{(n)}.$$

Since this estimator is unbiased and is a function of a sufficient statistic, the Rao-Blackwell theorem tells us that it is the MVUE for θ.

The MLE is generally a good first guess when trying to find the MVUE.

Example: Let $X_1,\dots,X_n \overset{ind}{\sim} \text{Gamma}(\alpha,2)$. Find the MLE of α.

The pdf of the $\text{Gamma}(\alpha,2)$ distribution is given by

$$f_X(x;\alpha) = \frac{1}{\Gamma(\alpha)2^\alpha}x^{\alpha-1}e^{-x/2}, \qquad \text{for } x>0,$$

so the likelihood function is

$$L(\alpha; X_1, \dots, X_n) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)2^{\alpha}} X_i^{\alpha-1} e^{-X_i/2} \mathbb{1}(X_i > 0)$$

$$= \left(\frac{1}{\Gamma(\alpha)2^{\alpha}}\right)^{n} \left(\prod_{i=1}^{n} X_i\right)^{\alpha-1} e^{-\sum_{i=1}^{n} X_i/2}$$

and the log-likelihood is

$$\ell(\alpha; X_1, \dots, X_n) = -n \log \Gamma(\alpha) - n\alpha \log 2 + (\alpha-1)\sum_{i=1}^{n} \log X_i - \frac{\sum_{i=1}^{n} X_i}{2}.$$

We have

$$\frac{\partial}{\partial \alpha} \ell(\alpha; X_1, \dots, X_n) = -n \underbrace{\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}}_{= \psi(\alpha), \text{ the digamma function.}} - n \log 2 + \sum_{i=1}^{n} \log X_i \overset{\text{set}}{=} 0.$$

So the MLE $\hat{\alpha}$ for α is the value of α which satisfies

$$-n\psi(\alpha) - n \log 2 + \sum_{i=1}^{n} \log X_i = 0$$

$$\iff \frac{1}{n} \sum_{i=1}^{n} \log\left(X_i/2\right) = \psi(\alpha).$$

Note that $\hat{\alpha}$ is determined by the value of $\prod_{i=1}^{n} X_i$, which is a sufficient statistic for α.

since $\sum_{i=1}^{n} \log \frac{X_i}{2} = \log\left[\prod_{i=1}^{n} \frac{X_i}{2}\right] = 2^{-n} \log\left[\prod_{i=1}^{n} X_i\right]$

The next result tells us how to find the MLE for a function of a parameter. The result is often called the invariance property of MLEs.

<u>**Result:**</u> If $\hat{\theta}$ is the MLE for θ, then for any function τ, $\tau(\hat{\theta})$ is the MLE for $\tau(\theta)$.

Proof: We first prove the result in the simple case in which τ is a one-to-one function:

Let $\tau : \Theta \to H$ be a one-to-one function with inverse τ^{-1} such that

$$\eta = \tau(\theta) \iff \theta = \tau^{-1}(\eta)$$

and let $\hat{\theta}$ be the MLE for θ.

Define the induced likelihood, induced by the reparameterization $\eta = \tau(\theta)$, as

$$L_\tau(\eta; X_1, \dots, X_n) = L(\tau^{-1}(\eta); X_1, \dots, X_n).$$

Then

$$\begin{aligned}
\sup_{\eta \in H} L_\tau(\eta; X_1, \dots, X_n) &= \sup_{\eta \in H} L(\tau^{-1}(\eta); X_1, \dots, X_n) \\
&= \sup_{\theta \in \Theta} L(\theta; X_1, \dots, X_n) \\
&= L(\hat{\theta}; X_1, \dots, X_n) \\
&= L(\tau^{-1}(\tau(\hat{\theta})); X_1, \dots, X_n) \\
&= L_\tau(\tau(\hat{\theta}); X_1, \dots, X_n),
\end{aligned}$$

so that $\hat{\eta} = \tau(\hat{\theta})$ is the maximizer of $L_\tau(\eta; X_1, \dots, X_n)$.

If $\tau : \Theta \to H$ is not one-to-one, define τ^{-1} as

$$\tau^{-1}(\eta) = \{\theta : \tau(\theta) = \eta\}$$

(Could be many θ values giving the same $\tau(\theta)$ value.)

and the induced likelihood as

$$L_\tau(\eta; X_1, \dots, X_n) = \sup_{\theta \in \tau^{-1}(\eta)} L(\theta; X_1, \dots, X_n).$$

Then

$$\sup_{\eta} L_\tau(\eta; X_1, \dots, X_n) = \sup_{\eta} \left(\sup_{\theta \in \tau^{-1}(\eta)} L(\theta; X_1, \dots, X_n) \right)$$

$$= \sup_{\theta \in \Theta} L(\theta; X_1, \dots, X_n)$$

$$= L(\hat{\theta}; X_1, \dots, X_n)$$

$\hat{\theta}$ belongs to the set $\tau^{-1}(\tau(\hat{\theta}))$ and is the maximizer of $L(\theta; X_1, \dots, X_n)$

$$= \sup_{\theta \in \tau^{-1}(\tau(\hat{\theta}))} L(\theta; X_1, \dots, X_n)$$

$$= L_\tau(\tau(\hat{\theta}); X_1, \dots, X_n),$$

so that $\hat{\eta} = \tau(\hat{\theta})$ is the maximizer of $L_\tau(\eta; X_1, \dots, X_n)$.

Remark: The above result is also true if $\theta = (\theta_1, \dots, \theta_d)$ consists of $d \geq 2$ parameters and τ is a function of $\theta_1, \dots, \theta_d$; If $\hat{\theta}_1, \dots, \hat{\theta}_d$ are the MLEs of $\theta_1, \dots, \theta_d$, then $\tau(\hat{\theta}_1, \dots, \hat{\theta}_d)$ is the MLE of $\tau(\theta_1, \dots, \theta_d)$.

Example: Let $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$. Find the MLE of $\eta = 1/\lambda$.

We previously found that the MLE of λ is $\hat{\lambda} = \bar{X}_n$.

So the MLE of $\eta = \frac{1}{\lambda}$ is $\hat{\eta} = \frac{1}{\hat{\lambda}} = \frac{1}{\bar{X}_n}$.

Note that $\mathbb{E}\hat{\lambda} = \lambda$, but $\mathbb{E}\hat{\eta} \neq \eta$; unbiasedness is not a guaranteed property of MLEs.

Example: Let $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(p)$. Find the MLE of $\text{Var}\, X_1$.

We have $\text{Var}\, X_1 = p(1-p)$. Find MLE $\hat{p}$ of p. Then $\hat{p}(1-\hat{p})$ is the MLE of $p(1-p)$.

The likelihood function is

$$L(p; X_1, \dots, X_n) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i},$$

for $X_1, \dots, X_n \in \{0, 1\}$, giving

$$\ell(p; X_1, \dots, X_n) = \sum_{i=1}^n \left[X_i \log p + (1-X_i) \log(1-p) \right]$$

Next, take

$$\frac{\partial}{\partial p} \ell(p; X_1, \dots, X_n) = \sum_{i=1}^{n} X_i \left(\frac{1}{p}\right) - \left(n - \sum_{i=1}^{n} X_i\right)\left(\frac{1}{1-p}\right) = 0$$

$$\iff \qquad \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{1 - \frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{p}{1-p} .$$

Solving for p gives

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n .$$

So the MLE for $p(1-p)$ is $\hat{p}(1-\hat{p}) = \bar{X}_n(1-\bar{X}_n)$.

ASYMPTOTIC PROPERTIES OF MLEs

In certain "nice" situations, MLEs have the enviable property of being asymptotically Normally distributed.

In particular, let $f_X(x;\theta)$ for $\theta \in \Theta \subset \mathbb{R}$ be a family of pdfs or pmfs. Then suppose

(i) $X_1, \ldots, X_n \overset{ind}{\sim} f_X(x;\theta_0)$, $\theta_0 \in \Theta$ being the <u>true</u> value of θ.

(ii) $\theta = \theta' \iff f_X(x;\theta) = f_X(x;\theta')$ for all $x \in \mathbb{R}$.

(iii) $f_X(x;\theta)$ is differentiable with respect to θ.

(iv) θ_0 is not on the boundary of Θ.

(v), (vi) See Casella & Berger, 2nd Ed, pg. 516. These conditions are a bit ghastly! 

Theorem: Under conditions (i)–(vi), the MLE $\hat{\theta}_n$ satisfies

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} \text{Normal}\left(0, 1/I(\theta_0)\right),$$

where

$$I(\theta_0) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_X(X;\theta)\Big|_{\theta=\theta_0}\right)^2\right],$$

with the expectation based on $X \sim f_X(x;\theta_0)$.

Definition: The quantity $I(\theta_0)$ is called the <u>Fisher information</u>.

Example: Let $X_1, \ldots, X_n \overset{ind}{\sim} \text{Exponential}(\lambda)$. The MLE of λ is $\hat{\lambda} = \bar{X}_n$.

The Fisher information is obtained as follows: First, we find

$$\frac{\partial}{\partial\lambda}\log f_X(x;\lambda) = \frac{\partial}{\partial\lambda}\log\left[\frac{1}{\lambda}e^{-x/\lambda}\right]$$

$$= \frac{\partial}{\partial\lambda}\left[-\log\lambda - \frac{x}{\lambda}\right]$$

$$= -\frac{1}{\lambda} + \frac{x}{\lambda^2}.$$

Then, if λ_0 is the true value of the parameter λ,

$$I(\lambda_0) = \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\log f_X(X;\lambda)\Big|_{\lambda=\lambda_0}\right)^2\right]$$

$$= \mathbb{E}\left(-\frac{1}{\lambda_0} + \frac{X}{\lambda_0^2}\right)^2$$

For $X \sim f_X(x;\lambda_0)$,
$\mathbb{E}X = \lambda_0$, $\mathbb{E}X^2 = 2\lambda_0^2$

$$= \mathbb{E}\left[\left(\frac{1}{\lambda_0}\right)^2 - 2\frac{X}{\lambda_0^3} + \frac{X^2}{\lambda_0^4}\right]$$

$$= \left(\frac{1}{\lambda_0}\right)^2 - 2\frac{\lambda_0}{\lambda_0^3} + \frac{2\lambda_0^2}{\lambda_0^4}$$

$$= \left(\frac{1}{\lambda_0}\right)^2.$$

So we have, by the theorem,

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda_0\right) \xrightarrow{D} \text{Normal}\left(0, \lambda_0^2\right) \quad \text{as} \quad n \to \infty.$$

Note: we knew already by the C.L.T. that

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda_0\right)/\lambda_0 \xrightarrow{D} N(0,1) \quad \text{as} \quad n \to \infty.$$

Example: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(1, \theta)$, so $f_X(x;\theta) = \theta(1-x)^{\theta-1}\mathbb{1}(0 < x < 1)$.

We have

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} \theta(1-X_i)^{\theta-1} = \theta^n \left[\prod_{i=1}^{n}(1-X_i)\right]^{\theta-1}$$

and

$$\ell(\theta; X_1, \ldots, X_n) = n \log \theta + (\theta-1)\sum_{i=1}^{n} \log(1-X_i).$$

Then

$$\frac{\partial}{\partial\theta}\ell(\theta; X_1, \ldots, X_n) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(1-X_i) \overset{\text{set}}{=} 0$$

$$\iff \qquad \hat{\theta}_{MLE} = -\frac{n}{\sum_{i=1}^{n} \log(1-X_i)}.$$

To find the Fisher information, we need

$$\frac{\partial}{\partial\theta} \log f_X(x;\theta) = \frac{\partial}{\partial\theta} \log\left[\theta(1-x)^{\theta-1}\right]$$

$$= \frac{\partial}{\partial\theta}\left[\log\theta + (\theta-1)\log(1-x)\right]$$

$$= \frac{1}{\theta} + \log(1-x).$$

Then we must find

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\Big|_{\theta=\theta_0}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\theta_0} + \log(1-X)\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{\theta_0^2} + \frac{2}{\theta_0}\log(1-X) + \log^2(1-X)\right],$$

with $X \sim f_X(x;\theta_0)$.

We find via the transformation method that for $Y = \log(1-X)$,

$$Y \sim f_Y(y;\theta) = \theta e^{y\theta} \mathbb{1}(y<0),$$

so that

$$\mathbb{E}\, Y = -\frac{1}{\theta_0} \quad \text{and} \quad \mathbb{E}\, Y^2 = \frac{1}{\theta_0^2} + \left(-\frac{1}{\theta_0}\right)^2 = \frac{2}{\theta_0^2}.$$

The Fisher information $I(\theta_0)$ is thus given by

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_X(X;\theta)\Big|_{\theta=\theta_0}\right)^2\right] = \frac{1}{\theta_0^2} + \frac{2}{\theta_0}\left(-\frac{1}{\theta_0}\right) + \frac{2}{\theta_0^2} = \frac{1}{\theta_0^2}.$$

Finally, by the theorem, we have

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) \xrightarrow{D} \text{Normal}\left(0, \theta_0^2\right)$$

as $n \to \infty$.

$\boxed{\text{APPENDIX}}$

MoMs of Beta parameters: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Beta}(\alpha, \beta)$.

Find the MoMs estimators for α and β.

Set up moment-matching equations:

$$m_1' = \frac{\alpha}{\alpha+\beta}$$

$$m_2' = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \left(\frac{\alpha}{\alpha+\beta}\right)^2$$

$$\Leftrightarrow \quad m_2' = \frac{m_1'(1-m_1')}{(\alpha+\beta+1)} + (m_1')^2 \qquad \left(\frac{\beta}{\alpha+\beta} = 1-m_1'\right)$$

$$\Leftrightarrow \quad \alpha+\beta+1 = \frac{m_1'(1-m_1')}{m_2'-(m_1')^2}$$

$$\Leftrightarrow \quad \alpha+\beta = \frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1$$

$$\Leftrightarrow \quad m_1' = \left[\frac{\alpha}{\frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1}\right]$$

$$\Leftrightarrow$$

$$\alpha = m_1'\left[\frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1\right]$$

$$\beta = \frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1 - m_1'\left[\frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1\right]$$

$$= (1-m_1')\left[\frac{m_1'(1-m_1')}{m_2'-(m_1')^2} - 1\right].$$

So, letting $\hat{\sigma}_n^2 = m_2' - (m_1')^2$, we have

$$\hat{\alpha} = \bar{X}_n \left[\frac{\bar{X}_n(1-\bar{X}_n)}{\hat{\sigma}_n^2} - 1 \right]$$

$$\hat{\beta} = (1-\bar{X}_n)\left[\frac{\bar{X}_n(1-\bar{X}_n)}{\hat{\sigma}_n^2} - 1 \right].$$

The following is a case in which the MoMs estimators do not admit a closed form.

<u>MoMs of Weibull parameters:</u> Let $X_1, \dots, X_n \overset{iid}{\sim}$ Weibull(a, b).

The Weibull(a,b) pdf is

$$f_X(x; a, b) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left[-\left(\frac{x}{b}\right)^a\right] \mathbb{1}(x > 0).$$

To find the MoMs of a and b, we need expressions for the first two moments μ_1' and μ_2' of the Weibull(a,b) distribution. Let's go ahead and derive an expression for the k^{th} moment μ_k' of the Weibull(a,b) distribution for $k \geq 1$.

We have

$$\mu_k' = \int_0^\infty x^k \cdot \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left[-\left(\frac{x}{b}\right)^a\right] dx$$

$$= \int_0^\infty \left(\frac{a}{b}\right) b^k \left(\frac{x}{b}\right)^{(a+k)-1} \exp\left[-\left(\frac{x}{b}\right)^a\right] dx$$

let $u = \left(\frac{x}{b}\right)^a$, $x = b u^{\frac{1}{a}}$, $dx = b \frac{1}{a} u^{\frac{1}{a}-1} du$, $0 < u < \infty$

$$= b^k \int_0^\infty \left(\frac{a}{b}\right) \left(u^{\frac{1}{a}}\right)^{(a+k)-1} \exp\left[-u\right] \frac{b}{a} u^{\frac{1}{a}-1} du$$

$$= b^k \int_0^\infty x^{1+\frac{k}{a}-\frac{1}{a}+\frac{1}{a}-1} e^{-x}\,dx$$

$$= b^k \int_0^\infty x^{\left(1+\frac{k}{a}\right)-1} e^{-x}\,dx$$

Recall the definition of the gamma function

$$= b^k\, \Gamma\!\left(1+\frac{k}{a}\right).$$

Now, the MoMs estimators of a and b are the solutions to the set of equations

$$m_1' = b\,\Gamma\!\left(1+\frac{1}{a}\right)$$
$$m_2' = b^2\,\Gamma\!\left(1+\frac{2}{a}\right).$$

The above can be manipulated to give

$$b = m_1' \Big/ \Gamma\!\left(1+\frac{1}{a}\right)$$

$$0 = \frac{m_2'}{(m_1')^2} - \frac{\Gamma\!\left(1+\frac{2}{a}\right)}{\left[\Gamma\!\left(1+\frac{1}{a}\right)\right]^2},$$

so that the MoMs estimator of a is the root of a complicated function. We can find the root $\bar{a}$ of the function

$$\frac{m_2'}{(m_1')^2} - \frac{\Gamma\!\left(1+\frac{2}{a}\right)}{\left[\Gamma\!\left(1+\frac{1}{a}\right)\right]^2}$$

using a numerical procedure such as is implemented by the `uniroot()` function in R. Having found $\bar{a}$, we have

$$\bar{b} = m_1' \Big/ \Gamma\!\left(1+\frac{1}{\bar{a}}\right).$$

NOTE: The MLEs for the Weibull(a,b) parameters must also be found numerically.

Expected value of $\hat{p} = 1/Y$ when $Y \sim \text{Geometric}(p)$.

$$\mathbb{E}\,\frac{1}{Y} = \sum_{y=1}^{\infty} \frac{1}{y} (1-p)^{y-1} p$$

$$= \frac{p}{1-p} \underbrace{\sum_{y=1}^{\infty} \frac{(1-p)^y}{y}}_{= -\log p}$$

$$= -\frac{p}{1-p} \log p .$$

Find Taylor expansion of $\log x$ around $x = 1$ and evaluate at p:

$$\log p = \log 1 + \left.\frac{1}{p}\right|_{p=1} (p-1)^1 + \left.\left(-\frac{1}{p^2}\right)\right|_{p=1} \frac{(p-1)^2}{2}$$

$$+ \left.\left(\frac{2}{p^3}\right)\right|_{p=1} \frac{(p-1)^3}{3!} + \left.\left(-\frac{3\cdot 2}{p^4}\right)\right|_{p=1} \frac{(p-1)^4}{4!} + \cdots$$

$$= 0 - \frac{(1-p)^1}{1} - \frac{(1-p)^2}{2} - \frac{(1-p)^3}{3} - \frac{(1-p)^4}{4} - \cdots$$

$$= -\sum_{y=1}^{\infty} \frac{(1-p)^y}{y}$$