36-310 Spring 2004: Estimation II: Some Methods of Estimation

Brian Junker

April 6, 2004

- Review: Parameter, Statistic, Point Estimator, Point Estimate
- Methods of Point Estimation
- Method of Moments (MoM)
- Five Examples
- Properties of MoM Estimators
- Maximum Likelihood (ML)
- Three Examples
- Properties of MLE’s

Review: Parameter, Statistic, Point Estimator, Point Estimate

Let $X_1, X_2, \ldots, X_n$ be a sample. It may be a simple random sample [independent observations from the same distribution $f_X(x)$] or it may not.

A parameter θ of $f_X(x)$ is a free variable that characteristic of $f_X(x)$.

A statistic T is any quantity that can be calculated from a sample. *That is, it is any function of* $X_1, \ldots, X_n$.

A point estimate $\hat{\theta}$ for θ is a single number that is a reasonable value for θ.

A point estimator $\hat{\theta}$ for θ is a *statistic* that gives the formula for computing the point estimate $\hat{\theta}$.

When we report an estimate $\hat{\theta}$, we should also report a standard error (standard deviation), $SE(\hat{\theta}) = \sqrt{\mathrm{Var}\,(\hat{\theta})}$, for the estimate.

Some possible considerations for good estimators include:

- Consistency: $\hat{\theta} \to \theta$ as $n \to \infty$.
- Unbiasedness: $E[\hat{\theta}] = \theta$.
- Efficiency: $SE(\hat{\theta})$ is small.
- Asymptotic Normality: $(\hat{\theta} - \theta)/SE(\hat{\theta}) \approx$ Normal, mean 0, variance 1.

Methods of Point Estimation

All of the estimators we have looked at so far are very “ad hoc”, that is, they look reasonable after you see them, but how would you think of getting them? For example…

- For λ = average number of scratches per CDROM, we chose $\hat{\lambda} = \overline{X}$ by thinking about unbiasedness: $E[\overline{X}] = E[X_1] = \lambda$.
- For the maximum θ of reaction times distributed as uniform on $[0, \theta]$ we considered
  - $\hat{\theta}_1 = 2\overline{X}$, because we were able to verify that it was unbiased: $E[2\overline{X}] = \theta$;
  - $\hat{\theta}_2 = max\{X_1, \ldots, X_n\}$ because it should be “close” to θ; but it was biased: $E[\hat{\theta}_2] = \frac{n}{n+1}\theta$.
  - $\hat{\theta}_3 = \frac{n+1}{n}\hat{\theta}_2$ because that “fixed” the bias in $\hat{\theta}_2$: $E[\hat{\theta}_3] = \theta$.

How can we systematically construct “good” point estimators? There are several methods that have proven useful:

- *Method of Moments (MoM)* A *moment* is the expected value of a power of X. MoM estimators are obtained by combining unbiased estimators for moments of X.
- *Least Squares (LS)* Least squares estimators are obtained by minimizing the mean squared error $\sum_{i=1}^{n}(X_i - E[X_i])^2$. This makes sense when $E[X_i]$ is a function of the parameter of interest, θ. LS turns out to be related to MoM.
- *Maximum Likelihood (ML)* The *likelihood* is the probability of (all) the data. ML estimators (MLE’s) choose the parameter values that makes the data most likely.
- *Bayesian Estimation (Bayes)* In Bayesian estimation we treat the parameters as random variables and use Bayes’ Rule to pick the parameter value that is most likely, or most typical, for the data (it is the “reverse” of ML, though often the answers are very similar!).

In the following slides we will look at MoM and ML. We will see examples of LS and Bayes later in the course.

Method of Moments (MoM)

Let $X_1, \ldots, X_n$ be an s.r.s. from a distribution with pdf or pmf $f_X(x)$, depending on parameter θ.

We will write $f_X(x;\ \theta)$, $E[X;\ \theta]$, etc., to emphasize this.

Definition

- The k^{th} *distribution moment* is the mean of X^k: $E[X^k; \theta]$.
- The k^{th} *sample moment* is the sample average of X^k: $\frac{1}{n}\sum_{i=1}^n X_i^k$.

For an s.r.s., the sample moment is an unbiased estimator of the distribution moment:

$$E\left[\frac{1}{n}\sum_{i=1}^n X_i^k;\ \theta\right] \;=\; \frac{1}{n}\sum_{i=1}^n E[X_i^k;\ \theta] \;=\; E[X^k;\ \theta]$$

The **method of moments** constructs an estimator $\hat{\theta}$ by setting

$$\frac{1}{n}\sum_{i=1}^n X_i^k \overset{MoM}{\approx} E[X^k;\ \theta]$$

and solving for θ.

Examples

- Let $X_1, \ldots, X_n$ be an s.r.s. of numbers of scratches on CDROMS, following a *Poisson* distribution with parameter λ. Observe that

  $$\lambda = E[X_i] \overset{MoM}{\approx} \overline{X}$$

  and therefore $\hat{\lambda}_{MoM} = \overline{X}$.

- Let $X_1, \ldots, X_n$ be an s.r.s. of lifetimes of computer chips, following an *Exponential* distribution with failure rate λ. We see

  $$1/\lambda = E[X_i] \overset{MoM}{\approx} \overline{X};$$

  solving for λ, we get $\hat{\lambda}_{MoM} = 1/\overline{X}$.

- Let $X_1, \ldots, X_n$ be an s.r.s. of reaction times following a *Uniform* distribution on $[0, \theta]$. Note that

  $$\theta/2 = E[X_i] \overset{MoM}{\approx} \overline{X}$$

  and so, solving for θ, $\hat{\theta}_{MoM} = 2\overline{X}$.

Example: The Sample Variance

Let $X_1, \ldots, X_n$ be an s.r.s. from any distribution with mean μ and variance σ^2. Clearly, $\hat{\mu}_{MoM} = \overline{X}$.

- **What is $\hat{\sigma}^2_{MoM}$?** Note that

  $$\sigma^2 = E[(X-\mu)^2] = E[X^2] - (E[X])^2 \overset{MoM}{\approx} \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2$$

  so, with a little algebra,

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

- **Is $\hat{\sigma}^2_{MoM}$ unbiased?** Using the fact that $\mathrm{Var}\,(\overline{X}) = E[(\overline{X})^2] - (E[\overline{X}])^2$,

  $$E[\hat{\sigma}^2_{MoM}] = E\left\{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2\right\} = \frac{1}{n}\sum_{i=1}^n E[X_i^2] - E[(\overline{X})^2]$$

  $$= E[X^2] - \mathrm{Var}\,(\overline{X}) - (E[\overline{X}])^2 = \sigma^2 - \sigma^2/n = \frac{n-1}{n}\sigma^2$$

  so $\hat{\sigma}^2_{MoM}$ is biased!

Fixing the bias in the sample variance

Since we just showed that

$$E[\hat{\sigma}^2_{MoM}] = \frac{n-1}{n}\sigma^2,$$

we know we can “fix the bias” by using

$$\begin{aligned} S^2 &= \frac{n}{n-1}\hat{\sigma}^2_{MoM} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \end{aligned}$$

This is the usual *unbiased sample variance estimator*. Clearly,

$$E[S^2] = E[\frac{n}{n-1}\hat{\sigma}^2_{MoM}] = \frac{n}{n-1}\left[\frac{n-1}{n}\sigma^2\right] = \sigma^2$$

so S^2 is an *unbiased estimator* for σ^2.

Example: Negative Binomial Distribution

Recall the *negative binomial* distribution, a discrete distribution for the number X of all "failures" in a sequence of Bernoulli trials, before the r^{th} "success". The pmf for X is

$$p(x) = P[X = x] = \binom{x+r-1}{r-1} p^r (1-p)^x,\ x = 0, 1, 2, 3, \ldots$$

and the mean and variance are

$$\mu_X = E[X] = r(1-p)/p\,, \qquad \sigma_X^2 = \mathrm{Var}\,(X) = r(1-p)/p^2$$

This suggests a MoM strategy for estimating p and r from a s.r.s. $X_1, \ldots, X_n$:

$$\begin{aligned} \overline{X} &= \hat{\mu}_{MoM} \overset{MoM}{\approx} \mu_X = r(1-p)/p \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \hat{\sigma}^2_{MoM} \overset{MoM}{\approx} \sigma_X^2 = r(1-p)/p^2 \end{aligned}$$

Sovling for p and r we get

$$\hat{p}_{MoM} = \frac{\hat{\mu}_{MoM}}{\hat{\sigma}^2_{MoM}}\,, \qquad \hat{r}_{MoM} = \frac{(\hat{\mu}_{MoM})^2}{\hat{\sigma}^2_{MoM} - \hat{\mu}_{MoM}}$$

Application

Reep, Pollard and Benjamin (“Skill and Change in Ball Games”, *J. Royal Stat. Soc*, 1971, pp. 623–629) consider the negative binomial distribution as a model for the number of goals per game scored by National Hockey League teams. The data, for 420 games in the 1966–1967 season, is:

Goals	0	1	2	3	4	5	6	7	8	9	10
Frequency	29	71	82	89	65	45	24	7	4	1	3

It is easy to calculate

$$\begin{aligned} \hat{\mu}_{MoM} &= \overline{X} = 2.98 \\ \hat{\sigma}^2_{MoM} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = 3.52 \end{aligned}$$

and therefore

$$\hat{p}_{MoM} = \frac{2.98}{3.52} = 0.85 , \qquad \hat{r}_{MoM} = \frac{(2.98)^2}{3.52 - 2.98} = 16.4$$

(not so easy to see what $SE(\hat{p})$ and $SE(\hat{r})$ should be!)

Properties of MoM Estimators

- They are pretty easy to derive!
- One can show that $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ is a *consistent*, *unbiased*, *asymptotically normal* estimator of $E[X^k]$.

  Since MoM estimators are built out of these ingredients, it can be shown in more advanced courses that

Usually, Method of Moments estimators

- Are unbiased or mildly biased
- Are consistent
- Have reasonably low SE
- Are asymptotically normal

Maximum Likelihood (ML)

Let $X_1, \ldots, X_n$ be an s.r.s. from a distribution with pdf or pmf $f_X(x;\ \theta)$, depending on parameter θ. Let $x_1, \ldots, x_n$ be the observed values in the sample.

Definition The **likelihood** of the sample is the joint pdf (or pmf)

$$\begin{aligned} L(\theta) &= f(x_1, \ldots, x_n;\ \theta) = f(x_1;\ \theta) f(x_2;\ \theta) \cdots f(x_n;\ \theta) \\ &= \prod_{i=1}^{n} f(x_i;\ \theta) \end{aligned}$$

Definition The **maximum likelihood estimate** $\hat{\theta}_{MLE}$ maximizes $L(\theta)$:

$$L(\hat{\theta}_{MLE}) \geq L(\theta) \quad \forall\ \theta$$

If we use X_i's instead of x_i's then $\hat{\theta}$ is the **maximum likelihood estimator**.

Strategy: It is usually (but not always!) easier to work with the **log-likelihood**

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\ \theta)\ .$$

Example: Exponential Distribution

The pdf of an exponential distribution is

$$f_X(x;\ \lambda) = \lambda e^{-\lambda x}$$

so the likelihood for an s.r.s. of size n is

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

and the log-likelihood is $\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$. To find $\hat{\lambda}_{MLE}$, we differentiate the log-likelihood and set it to zero

$$0 \overset{set}{=} \frac{d}{d\lambda}\ell(\lambda) = \frac{d}{d\lambda}\left[n \log \lambda - \lambda \sum_{i=1}^{n} x_i \right] = n/\lambda - \sum_{i=1}^{n} x_i$$

Solving for λ, it is easy to see that

$$\hat{\lambda}_{MLE} = n \Big/ \sum_{i=1}^{n} x_i = 1/\overline{X} = \hat{\lambda}_{MoM}$$

Example: Normal Distribution

The pdf of for a Normal r.v. with mean μ and variance σ^2 is

$$f_X(x;\ \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

so the likelihood for an s.r.s. of size n is

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$$

and the log-likelihood is

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

Once again we would like to find $\hat{\mu}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ by setting the derivative**s** of $\ell(\mu,\sigma^2)$ equal to zero, and solving for μ and σ^2.

So, we want to solve the equations

$$\begin{aligned} 0 &\overset{set}{=} \frac{\partial}{\partial\mu}\ell(\mu,\sigma^2) = 2\sum_{i=1}^{n}(x_i-\mu) = 2n(\overline{x}-\mu) \\ 0 &\overset{set}{=} \frac{\partial}{\partial(\sigma^2)}\ell(\mu,\sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

for μ and σ^2. When we do, we get

$$\begin{aligned} \hat{\mu}_{MLE} &= \overline{X} \\ \hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \end{aligned}$$

So again we get the MoM estimators back. . .

- MLE and MoM *don't always* produce the same estimators, but in many problems they produce very similar ones.

Example: Uniform Distribution

The pdf for a Uniform distribution on $[0, \theta]$ is

$$f_X(x;\ \theta) = \begin{cases} 1/\theta, & 0 \leq x \leq \theta \\ 0, & else \end{cases}$$

so that the likelihood for an s.r.s. of size n is

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\ \theta) = \begin{cases} 1/\theta^n, & 0 \leq x_i \leq \theta, \underline{for\ all\ i} \\ 0, & else \end{cases}$$

If we proceed as before,

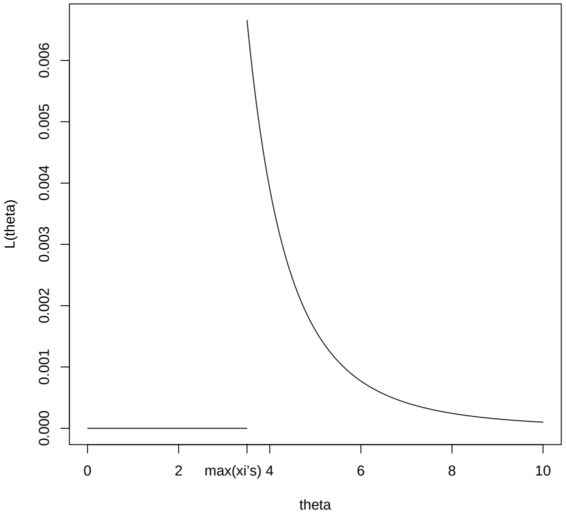
$$\ell(\theta) = \log L(\theta) = \log(1/\theta^n) = -n \log \theta$$

and if we try to solve

$$0 \overset{set}{=} \frac{d}{d\theta} - n \log \theta = -n/\theta$$

we seem to want $\theta = \infty$, which doesn't make any sense. . . .

The graph below shows why using calculus to maximize $L(\theta)$ isn't doing much good...



but it also shows clearly that $\hat{\theta}_{MLE} = \max\{X_1, \ldots, X_n\}$ Note that in this case,

$$\begin{aligned} \hat{\theta}_{MoM} &= 2\overline{X} \\ \hat{\theta}_{MLE} &= \max\{X_1, \ldots, X_n\} \end{aligned}$$

so the two methods generate different estimators.

We've already seen that $\hat{\theta}_{MLE}$ is biased, but $\hat{\theta}_{MoM}$ is not.
We've also seen how to "fix the bias" in $\hat{\theta}_{MLE}$" if we want to!

Properties of MLE's

- They are not as easy to derive as MoM estimators, but still not too bad. . .
- They share many properties with MoM estimators. It can be shown in more advanced courses that

  Usually, Maximum Likeihood estimators

  - Are unbiased or mildly biased
  - Are consistent
  - Have the lowest possible SE, as $n \to \infty$
  - Are asymptotically normal

- Not all MLE's have these properties. For example $\hat{\theta}_{MLE}$ for the Uniform distribution on $[0, \theta]$ is not asymptotically normal. But most of the time, it is safe to assume these properties.
- The SE for an MLE is relatively easy to compute (though perhaps not so easy to see where it's "coming from"). We'll talk about the SE's of MLE's next time.