# STA261 Lecture 4 — 2017-07-17

Neil Montgomery

Last edited: 2017-07-17 19:01

# value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.

# value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.
2. They possess the "invariance" property.

# value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.
2. They possess the “invariance” property.
3. They might be the only estimators available (i.e. other techniques we haven’t seen yet don’t work.)

# value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.
2. They possess the "invariance" property.
3. They might be the only estimators available (i.e. other techniques we haven't seen yet don't work.)

# value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.
2. They possess the "invariance" property.
3. They might be the only estimators available (i.e. other techniques we haven't seen yet don't work.)

But they are not usually the best available estimators.

# likelihood methods

# a few facts about maximizing functions

**Proposition 4.0:** Suppose a twice-differentiable function $f(x)$ has a critical value at $x_0$, and $g(x)$ is strictly increasing and twice-differentiable.

Then $g(f(x))$ also has a critical value at $x_0$, and the sign of its second derivative at $x_0$ is the same as the sign of the second derivative of $f$ at $x_0$.

# revisit estimating $p$ from a Bernoulli($p$) distribution

A Bernoulli($p$) has p.m.f. usually expressed as:

$$p(x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

# revisit estimating $p$ from a Bernoulli($p$) distribution

A Bernoulli($p$) has p.m.f. usually expressed as:

$$p(x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

But a more concise way is:

$$p(x) = p^x (1 - p)^{1-x}, \quad x \in \{0, 1\}$$

# revisit estimating $p$ from a Bernoulli($p$) distribution

A Bernoulli($p$) has p.m.f. usually expressed as:

$$p(x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

But a more concise way is:

$$p(x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

Given a sample $X_1, \ldots, X_n$ an intuitive estimator for $p$ is $\hat{p} = \overline{X}$.

# the probability of the data, given $p$

Now suppose $n = 10$ and we observe a particular sequence of 0's and 1's.

Here's a simulated sequence of 0's and 1's from a Bernoulli($p$) distribution. (I know what $p$ is, but you don't.)

```
##  [1] 0 0 0 0 1 1 0 0 0 0
```

# the probability of the data, given $p$

Now suppose $n = 10$ and we observe a particular sequence of 0's and 1's.

Here's a simulated sequence of 0's and 1's from a Bernoulli($p$) distribution. (I know what $p$ is, but you don't.)

```
##  [1] 0 0 0 0 1 1 0 0 0 0
```

The probability of getting this sample exactly is:

$$
\begin{aligned}
L(p) &= (1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot p\cdot p\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p) \\
&= p^2(1-p)^8
\end{aligned}
$$

# what value of $p$ maximizes $L(p)$?

As will often be the case, it is easier to maximize $\log L(p)$, which works due to **Proposition 4.0**.

# what value of $p$ maximizes $L(p)$?

As will often be the case, it is easier to maximize $\log L(p)$, which works due to **Proposition 4.0**.

**Example 4.1:** The maximum is at $\frac{2}{10}$.

## what value of $p$ maximizes $L(p)$?

As will often be the case, it is easier to maximize $\log L(p)$, which works due to **Proposition 4.0**.

**Example 4.1:** The maximum is at $\frac{2}{10}$.

It would have been no harder to work in general, with $k$ 1's out of a sample of size $n$, and maximizing $L(p) = p^k(1-p)^{n-k}$

## what value of $p$ maximizes $L(p)$?

As will often be the case, it is easier to maximize $\log L(p)$, which works due to **Proposition 4.0**.

**Example 4.1:** The maximum is at $\frac{2}{10}$.

It would have been no harder to work in general, with $k$ 1's out of a sample of size $n$, and maximizing $L(p) = p^k(1-p)^{n-k}$

The same calculus gives the maximum at $k/n$, which is what you get when you plug data into the formula $\overline{X}$.

# the likelihood function

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable $X$ with pmf or pdf $f(x;\theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter $\theta$ is defined as (for any positive $g$):

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta)$$

# the likelihood function

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable $X$ with pmf or pdf $f(x;\theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter $\theta$ is defined as (for any positive $g$):

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta)$$

We will tend to work with the logarithm $\ell(\theta) = \log L(\theta)$. Note that $\theta$ could be a vector.

**Notes 4.2:** I noticed a few notation issues in the textbook...

# interpretation of $L(\theta)$

When $X$ is discrete, $L(\theta)$ is literally the probability of the data that were observed.

# interpretation of $L(\theta)$

When $X$ is discrete, $L(\theta)$ is literally the probability of the data that were observed.

When $X$ is continuous, $L(\theta)$ is not a probability, but it still provides what I call an "index" to compare different $\theta$ values.

# interpretation of $L(\theta)$

When $X$ is discrete, $L(\theta)$ is literally the probability of the data that were observed.

When $X$ is continuous, $L(\theta)$ is not a probability, but it still provides what I call an "index" to compare different $\theta$ values.

**Note 4.3:** Recall (or, welcome to…) the interpretation of density $f(x)$ as a "local" probability very near to $x$.

# interpretation of $L(\theta)$

When $X$ is discrete, $L(\theta)$ is literally the probability of the data that were observed.

When $X$ is continuous, $L(\theta)$ is not a probability, but it still provides what I call an "index" to compare different $\theta$ values.

**Note 4.3:** Recall (or, welcome to...) the interpretation of density $f(x)$ as a "local" probability very near to $x$.

So instead of probability, we say likelihood. It's a good shortcut to think of likelihood as ("like a") probability.

# the method of maximum likelihood

The likelihood $L(\theta)$ is a function of $\theta$ and some observed data $\mathbf{x} = x_1, \ldots, x_n$, which can be thought of as a "realization" of the model for the idea of "sample", which is $\mathbf{X} = X_1, \ldots, X_n$.

# the method of maximum likelihood

The likelihood $L(\theta)$ is a function of $\theta$ and some observed data $\mathbf{x} = x_1, \ldots, x_n$, which can be thought of as a "realization" of the model for the idea of "sample", which is $\mathbf{X} = X_1, \ldots, X_n$.

Suppose $\hat{\theta}(\mathbf{x})$ is a value of $\theta$ at which $L(\theta)$ is maximized.

# the method of maximum likelihood

The likelihood $L(\theta)$ is a function of $\theta$ and some observed data $\mathbf{x} = x_1, \ldots, x_n$, which can be thought of as a "realization" of the model for the idea of "sample", which is $\mathbf{X} = X_1, \ldots, X_n$.

Suppose $\hat{\theta}(\mathbf{x})$ is a value of $\theta$ at which $L(\theta)$ is maximized.

Then $\hat{\theta}(\mathbf{X})$ is a *maximum likelihood estimator* for $\theta$, or MLE for short.

# the method of maximum likelihood

The likelihood $L(\theta)$ is a function of $\theta$ and some observed data $\mathbf{x} = x_1, \ldots, x_n$, which can be thought of as a "realization" of the model for the idea of "sample", which is $\mathbf{X} = X_1, \ldots, X_n$.

Suppose $\hat{\theta}(\mathbf{x})$ is a value of $\theta$ at which $L(\theta)$ is maximized.

Then $\hat{\theta}(\mathbf{X})$ is a *maximum likelihood estimator* for $\theta$, or MLE for short.

As usual, $\theta$ can be a vector.

# MLE examples

**Example 4.4:** Exponential with rate $\lambda$

**Example 4.5:** Poisson with rate $\lambda$

**Example 4.6:** Uniform$(0, \theta)$

**Proposition 4.7:** For any numbers $x_1, \ldots, x_n$, and $a$, the expression $\sum_{i=1}^{n}(x_i - a)^2$ is minimized at $a = \overline{x}$.

**Example 4.8:** $N(\mu, \sigma^2)$