STA261 Lecture 5 — 2017-07-24

Neil Montgomery

Last edited: 2017-07-26 12:19

maximum likelihood summary

The joint pmf/pdf is treated as a function of the parameter(s) θ, given the data.

This function is called a "likelihood" $L(\theta)$.

A likelihood can be thought of as the "probability" of the data.

The parameter value $\hat{\theta}$ that maximizes $L(\theta)$ is the maximum likelihood estimator.

maximum likelihood summary

The joint pmf/pdf is treated as a function of the parameter(s) θ, given the data.

This function is called a "likelihood" $L(\theta)$.

A likelihood can be thought of as the "probability" of the data.

The parameter value $\hat{\theta}$ that maximizes $L(\theta)$ is the maximum likelihood estimator.

The examples we've done so far have all had a closed form solution, but this isn't necessary or even "better" in any sense.

exponential distributions revisited

When we considered estimating the rate parameter λ directly, we got $\hat{\lambda} = n/(\sum X_i)$.
(This is also the M.O.M. estimator.)

exponential distributions revisited

When we considered estimating the rate parameter λ directly, we got $\hat{\lambda} = n/(\sum X_i)$. (This is also the M.O.M. estimator.)

We also found that:

$$E\left(\hat{\lambda}\right) = \frac{n}{n-1}\lambda$$

and that an unbiased estimator for λ was therefore

$$\frac{n-1}{n}\hat{\lambda} = \frac{n-1}{\sum X_i}$$

exponential distributions revisited

When we considered estimating the rate parameter λ directly, we got $\hat{\lambda} = n/(\sum X_i)$. (This is also the M.O.M. estimator.)

We also found that:

$$E\left(\hat{\lambda}\right) = \frac{n}{n-1}\lambda$$

and that an unbiased estimator for λ was therefore

$$\frac{n-1}{n}\hat{\lambda} = \frac{n-1}{\sum X_i}$$

We'll see over the next few classes that this is in one particular sense the best possible estimator for λ.

exponential distribution - different kind of dataset

The observed data: $x_1, x_2, \ldots, x_n$. These often might be times-to-events, such as failure times of equipment, or the death/remission times of people in a medical study.

exponential distribution - different kind of dataset

The observed data: $x_1, x_2, \ldots, x_n$. These often might be times-to-events, such as failure times of equipment, or the death/remission times of people in a medical study.

In real life analyses most stuff doesn't fail, and most people survive. Or at least we don't wait around long enought to see everything actually fail.

exponential distribution - different kind of dataset

The observed data: $x_1, x_2, \ldots, x_n$. These often might be times-to-events, such as failure times of equipment, or the death/remission times of people in a medical study.

In real life analyses most stuff doesn't fail, and most people survive. Or at least we don't wait around long enought to see everything actually fail.

What we would more typically see is data as on the next page. "Today" I extract the historical data on the equipment I am interested in...

“survival” data

ID	Age	Status
A023	6.8	Failed
A324	7.2	Operating
A620	10.1	Taken Out of Service
A092	2.4	Operating
A526	5.5	Operating
A985	8.1	Failed
A723	1.5	Operating
$\vdots$	$\vdots$	$\vdots$

likelihood for "survival data"

The model for failure times is $X \sim \text{Exp}(\lambda)$.

What is the likelihood of the data?

The likelihood for a unit to fail at time x_i is: $\lambda e^{-\lambda x_i}$

likelihood for “survival data”

The model for failure times is $X \sim \text{Exp}(\lambda)$.

What is the likelihood of the data?

The likelihood for a unit to fail at time x_i is: $\lambda e^{-\lambda x_i}$

The likelihood for a unit to not have failed yet at time x_i is: $P(X > x_i) = e^{-\lambda x_i}$

likelihood, line by line

ID	Age	Status	Contribution to Likelihood
A023	6.8	Failed	$\lambda e^{-6.8\lambda}$
A324	7.2	Operating	$e^{-7.2\lambda}$
A620	10.1	Taken Out of Service	$e^{-10.1\lambda}$
A092	2.4	Operating	$e^{-2.4\lambda}$
A526	5.5	Operating	$e^{-5.5\lambda}$
A985	8.1	Failed	$\lambda e^{-8.1\lambda}$
A723	1.5	Operating	$e^{-1.5\lambda}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

censored data, and the likelihood function

When the failure time is unknown, because it hasn't happened yet, we say the failure time is *censored*. Define the *censoring indicator* c_i to be 1 if the unit failed and 0 otherwise.

censored data, and the likelihood function

When the failure time is unknown, because it hasn't happened yet, we say the failure time is *censored*. Define the *censoring indicator* c_i to be 1 if the unit failed and 0 otherwise.

Putting it all together, given times $x_1, \ldots, x_n$ and censoring indicators $c_1, \ldots, c_n$, the likelihood of the data is:

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

censored data, and the likelihood function

When the failure time is unknown, because it hasn't happened yet, we say the failure time is *censored*. Define the *censoring indicator* c_i to be 1 if the unit failed and 0 otherwise.

Putting it all together, given times $x_1, \ldots, x_n$ and censoring indicators $c_1, \ldots, c_n$, the likelihood of the data is:

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

Proposition 5.0: The MLE for λ is $\hat{\lambda} = \sum_{i=1}^{n} c_i \Big/ \sum_{i=1}^{n} X_i$

occurrence-exposure example

Here are 50 simulated “ages” from an Exp(0.1) population, “censored” at 9.0 “years”

```
##  [1] 9.00 9.00 5.66 8.04 4.12 4.22 9.00 2.64 9.00 3.79 9.00
## [12] 1.19 0.15 7.21 1.49 3.00 9.00 9.00 9.00 2.10 6.25 9.00
## [23] 7.57 9.00 9.00 1.27 4.49 9.00 1.39 3.86 0.36 6.73 9.00
## [34] 4.82 4.18 2.73 5.39 2.40 9.00 9.00 8.71 2.91 2.98 7.01
## [45] 0.31 1.56 3.65 9.00 0.74 1.08
```

The “naive” mean life estimate (the average of the failed units only): 3.647.

The MLE: 7.882.

MLE result I published in 2016

The basic “shock and damage model” works like this:

- a unit suffers shock events that occur according to a Poisson process $N(t)$

MLE result I published in 2016

The basic "shock and damage model" works like this:

- a unit suffers shock events that occur according to a Poisson process $N(t)$
- at each shock event, the damage suffered is X_i (in general, random, but not necessarily)

MLE result I published in 2016

The basic “shock and damage model” works like this:

- a unit suffers shock events that occur according to a Poisson process $N(t)$
- at each shock event, the damage suffered is X_i (in general, random, but not necessarily)
- the cumulative damage is a sum of a random number of random damages:

$$Z(t) = \sum_{i=1}^{N(t)} X_i$$

MLE result I published in 2016

The basic "shock and damage model" works like this:

- a unit suffers shock events that occur according to a Poisson process $N(t)$
- at each shock event, the damage suffered is X_i (in general, random, but not necessarily)
- the cumulative damage is a sum of a random number of random damages:

$$Z(t) = \sum_{i=1}^{N(t)} X_i$$

- the unit fails the moment $Z(t)$ reaches some threshold

MLE result I published in 2016

One day I encountered a situation where the company only knew the age of an item, if an item had ever suffered at least one shock event (some items never did), and the total amount of damage.

MLE result I published in 2016

One day I encountered a situation where the company only knew the age of an item, if an item had ever suffered at least one shock event (some items never did), and the total amount of damage.

The company needed an estimate of the Poisson rate λ at which shocks occurred (among other things).

MLE result I published in 2016

One day I encountered a situation where the company only knew the age of an item, if an item had ever suffered at least one shock event (some items never did), and the total amount of damage.

The company needed an estimate of the Poisson rate λ at which shocks occurred (among other things).

So I went looking for the method that everyone used to estimate the rate in these situations. But nobody had ever done this before.

MLE result I published in 2016

One day I encountered a situation where the company only knew the age of an item, if an item had ever suffered at least one shock event (some items never did), and the total amount of damage.

The company needed an estimate of the Poisson rate λ at which shocks occurred (among other things).

So I went looking for the method that everyone used to estimate the rate in these situations. But nobody had ever done this before.

(Many OR / stats professors like to propose models, but often do not dirty themselves with actual data.)

MLE result I published in 2016

MLE result I published in 2016

I introduced a “shock indicator” d_i which is 1 when one or more shocks occurred, and 0 otherwise.

MLE result I published in 2016

I introduced a "shock indicator" d_i which is 1 when one or more shocks occurred, and 0 otherwise.

The probabilities of having endured 0, or 1+ shocks by age t_i are:

$$
\begin{aligned}
P(N(t_i) = 0) &= e^{-\lambda t_i} \\
P(N(t_i) > 0) &= 1 - e^{-\lambda t_i}
\end{aligned}
$$

likelihood

The likelihood for λ is therefore:

$$L(\lambda) = \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i}$$

likelihood

The likelihood for λ is therefore:

$$L(\lambda) = \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i}$$

$$\ell(\lambda) = -\lambda \sum_{i=1}^{n} t_i(1 - d_i) + \sum_{i=1}^{n} d_i \log\left(1 - e^{-\lambda t_i}\right)$$

likelihood

The likelihood for λ is therefore:

$$L(\lambda) = \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i}$$

$$\ell(\lambda) = -\lambda \sum_{i=1}^{n} t_i(1 - d_i) + \sum_{i=1}^{n} d_i \log\left(1 - e^{-\lambda t_i}\right)$$

This can only be maximized numerically. *As usual.*

properties of MLEs

why maximum likelihood is so popular

They are easy to develop, and under a few conditions (most often satisfied), the method of maximum likelihood produces estimators that are:

1. consistent

why maximum likelihood is so popular

They are easy to develop, and under a few conditions (most often satisfied), the method of maximum likelihood produces estimators that are:

1. consistent
2. asymptotically normal

why maximum likelihood is so popular

They are easy to develop, and under a few conditions (most often satisfied), the method of maximum likelihood produces estimators that are:

1. consistent
2. asymptotically normal
3. invariant

the score, and information functions

Likelihood theory deals so much with the following functions that they are given names:

Score:

$$S(\theta) = S(\theta; \mathbf{X}) = \frac{\partial}{\partial \theta} \log L(\theta; \mathbf{X})$$

Information:

$$\mathcal{I}(\theta) = \mathcal{I}(\theta; \mathbf{X}) = E\left(\left(\frac{\partial}{\partial \theta} \log L(\theta; \mathbf{X})\right)^2\right)$$

the score, and information functions

Likelihood theory deals so much with the following functions that they are given names:

Score:

$$S(\theta) = S(\theta; \mathbf{X}) = \frac{\partial}{\partial\theta} \log L(\theta; \mathbf{X})$$

Information:

$$\mathcal{I}(\theta) = \mathcal{I}(\theta; \mathbf{X}) = E\left(\left(\frac{\partial}{\partial\theta} \log L(\theta; \mathbf{X})\right)^2\right)$$

Technical note: when θ is a vector, the score is the gradient vector, and the information is a matrix of all the partial second derivatives.

MLEs are consistent

properties of score and information

Likelihood theory is littered with "under certain regularity conditions" statements, handed down from one generation to the next. From my failing hands I pass the torch...

Proposition 5.1: $E(S(\theta)) = 0$

Theorem 5.2: $\mathcal{I}(\theta) = \text{Var}(S(\theta))$ (CORRECTED) and $\mathcal{I}(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \log L(\theta; \mathbf{X})\right)$

MLEs are consistent and asymptotically normal

Proposition 5.3: An MLE for θ obtained from an i.i.d. sample is consistent for θ.

This next propostion has been altered from when shown in class the first time. The issue was a difference between definitions of $\mathcal{I}$ that I gave, and that the book gives. We will stick with what I gave. I'll explain more in the next class.

Proposition 5.4: $\sqrt{\mathcal{I}(\theta)}(\hat{\theta} - \theta)$ converges (in distribution) to a standard normal distrubution.