STA286 Lecture 29

Neil Montgomery

Last edited: 2017-03-31 11:06

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.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a smooth 1-1 function.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a smooth 1-1 function.
4. it is asymptotically normal. (Note: convergence can be slow.)

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a smooth 1-1 function.
4. it is asymptotically normal. (Note: convergence can be slow.)
5. if $c\hat{\theta}$ is unbiased for some constant c, then $c\hat{\theta}$ is the unbiased estimator with the smallest variance, or "MVUE" (our "gold standard".)

the normal case

Population $N(\mu, \sigma)$. Observe: $x_1, \ldots, x_n$. The MLEs are (example 9.21):

$$\widehat{\theta} = \left(\hat{\mu}, \widehat{\sigma^2}\right) = \left(\overline{X}, \frac{\sum \left(X_i - \overline{X}\right)^2}{n}\right)$$

the normal case

Population $N(\mu, \sigma)$. Observe: $x_1, \ldots, x_n$. The MLEs are (example 9.21):

$$\widehat{\theta} = \left(\hat{\mu}, \widehat{\sigma^2}\right) = \left(\overline{X}, \frac{\sum \left(X_i - \overline{X}\right)^2}{n}\right)$$

Therefore, $\overline{X}$ and S^2 are the MVUE estimators for μ and σ^2

exponential distributions - I

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta}$$

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta}$$

$$\ell(\beta) = -n \log \beta - \sum x_i/\beta$$

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta}$$

$$\ell(\beta) = -n \log \beta - \sum x_i/\beta$$

$$\frac{d}{d\beta}\ell(\beta) = -\frac{n}{\beta} + \frac{\sum x_i}{\beta^2}$$

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta} \\
\ell(\beta) &= -n \log \beta - \sum x_i/\beta \\
\frac{d}{d\beta}\ell(\beta) &= -\frac{n}{\beta} + \frac{\sum x_i}{\beta^2}
\end{aligned}
$$

(Technicality: so the ML estimat***e*** is $\overline{x}\ldots$)

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta}$$

$$\ell(\beta) = -n \log \beta - \sum x_i/\beta$$

$$\frac{d}{d\beta}\ell(\beta) = -\frac{n}{\beta} + \frac{\sum x_i}{\beta^2}$$

(Technicality: so the ML estimat***e*** is $\overline{x}\ldots$)

…and the ML estimat***or*** is $\hat{\beta} = \overline{X}$. Since $E\left(\overline{X}\right) = \beta$, it is the MVUE for β.

exponential distributions - II

Recall last week when we considered estimating λ directly. We now know immediately that $\hat{\lambda} = n/(\sum X_i)$ (invariance of MLE).

exponential distributions - II

Recall last week when we considered estimating λ directly. We now know immediately that $\hat{\lambda} = n/(\sum X_i)$ (invariance of MLE).

Then I did all that work on the board to show:

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

and that an unbiased estimator for λ was therefore

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

Now we know immediately that this is the MVUE for λ

exponential distributions - III (mind-blowing version)

I said we observed: $x_1, x_2, \ldots, x_n$. 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 distributions - III (mind-blowing version)

I said we observed: $x_1, x_2, \ldots, x_n$. 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 distributions - III (mind-blowing version)

I said we observed: $x_1, x_2, \ldots, x_n$. 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 follows. "Today" I extract the historical data on the equipment I am interested in:

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
$\vdots$	$\vdots$	$\vdots$

exponential distributions - III

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: $\lambda e^{-\lambda x_i}$

exponential distributions - III

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: $P(X > x_i) = e^{-\lambda x_i}$

exponential distributions - III

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: $P(X > x_i) = e^{-\lambda x_i}$

For example:

ID	Age	Status	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	$e^{-1.5\lambda}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

exponential distributions - III

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.

exponential distributions - III

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:

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

exponential distributions - III

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:

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

exponential distributions - III

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:

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)^{c_i} \left(e^{-\lambda x_i}\right)^{1-c_i} \\
&= \lambda^{\sum c_i} e^{-\lambda \sum x_i} \\
\ell(\lambda) &= \log \lambda \sum c_i - \lambda \sum x_i
\end{aligned}
$$

exponential distributions - III

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:

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)^{c_i} \left(e^{-\lambda x_i}\right)^{1-c_i} \\
&= \lambda^{\sum c_i} e^{-\lambda \sum x_i} \\
\ell(\lambda) &= \log \lambda \sum c_i - \lambda \sum x_i \\
\frac{d}{d\lambda}\ell(\lambda) &= \frac{\sum c_i}{\lambda} - \sum x_i
\end{aligned}
$$

exponential distributions - III

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:

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)^{c_i} \left(e^{-\lambda x_i}\right)^{1-c_i} \\
&= \lambda^{\sum c_i} e^{-\lambda \sum x_i} \\
\ell(\lambda) &= \log \lambda \sum c_i - \lambda \sum x_i \\
\frac{d}{d\lambda}\ell(\lambda) &= \frac{\sum c_i}{\lambda} - \sum x_i
\end{aligned}
$$

So $\hat{\lambda} = \frac{\sum c_i}{\sum x_i} = \frac{\text{\# of failures}}{\text{Total Time}}$. This is called an "occurence-exposure rate".

occurrence-exposure example

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

```
##  [1] 9.00 1.29 9.00 3.38 9.00 0.46 9.00 7.83 0.10 4.36 9.00
## [12] 9.00 9.00 2.29 0.63 5.83 9.00 9.00 7.65 2.45 9.00 9.00
## [23] 4.88 9.00 3.73 9.00 6.93 9.00 4.17 5.02 0.77 9.00 9.00
## [34] 9.00 0.77 2.09 5.98 8.05 9.00 9.00 2.74 2.74 6.97 4.03
## [45] 9.00 5.50 9.00 5.00 2.79 9.00
```

The "naive" mean life estimate (the average of the failed units only): 3.872.

The MLE: 10.943.

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 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:

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

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:

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

The likelihood for λ is therefore:

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

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:

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

The likelihood for λ is therefore:

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i} \\
\ell(\lambda) &= -\lambda \sum_{i=1}^{n} t_i(1 - d_i) + \sum_{i=1}^{n} d_i \log\left(1 - e^{-\lambda t_i}\right)
\end{aligned}
$$

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:

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

The likelihood for λ is therefore:

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

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:

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

The likelihood for λ is therefore:

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i} \\
\ell(\lambda) &= -\lambda \sum_{i=1}^{n} t_i(1 - d_i) + \sum_{i=1}^{n} d_i \log\left(1 - e^{-\lambda t_i}\right)
\end{aligned}
$$

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:

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

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.