Stat 343: Maximum Likelihood Example

January 23, 2018

Seedlings in a Forest

This example comes from "Introduction to Statistical Thought," by Michael Lavine. Lavine writes:

> Tree populations move by dispersing their seeds. Seeds become seedlings, seedlings become saplings, and saplings become adults which eventually produce more seeds. Over time, whole populations may migrate in response to climate change. One instance occurred at the end of the Ice Age when species that had been sequestered in the south were free to move north. Another instance may be occurring today in response to global warming. One critical feature of the migration is its speed. Some of the factors determining the speed are the typical distances of long range seed dispersal, the proportion of seeds that germinate and emerge from the forest floor to become seedlings, and the proportion of seedlings that survive each year. To learn about emergence and survival, ecologists return annually to forest quadrats (square meter sites) to count seedlings that have emerged since the previous year. One such study was reported in Lavine et al. [2002].

In each year from 1991 to 1997, the ecologists recorded the number of "old" seedlings in each of 60 quadrats, where an old seedling is at least 1 year old (they can identify old seedlings by whether they have a bud mark). These values are recorded in the columns of the data frame named like 91_old. For the years 1992 through 1997, they also recorded the number of "new" seedlings, i.e. the number of seedlings that were less than 1 year old. The number of new seedlings and the number of old seedlings in each quadrat don't add up like you might expect. This might be due to a variety of factors, like seedlings dying or errors in data collection.

In [10]:

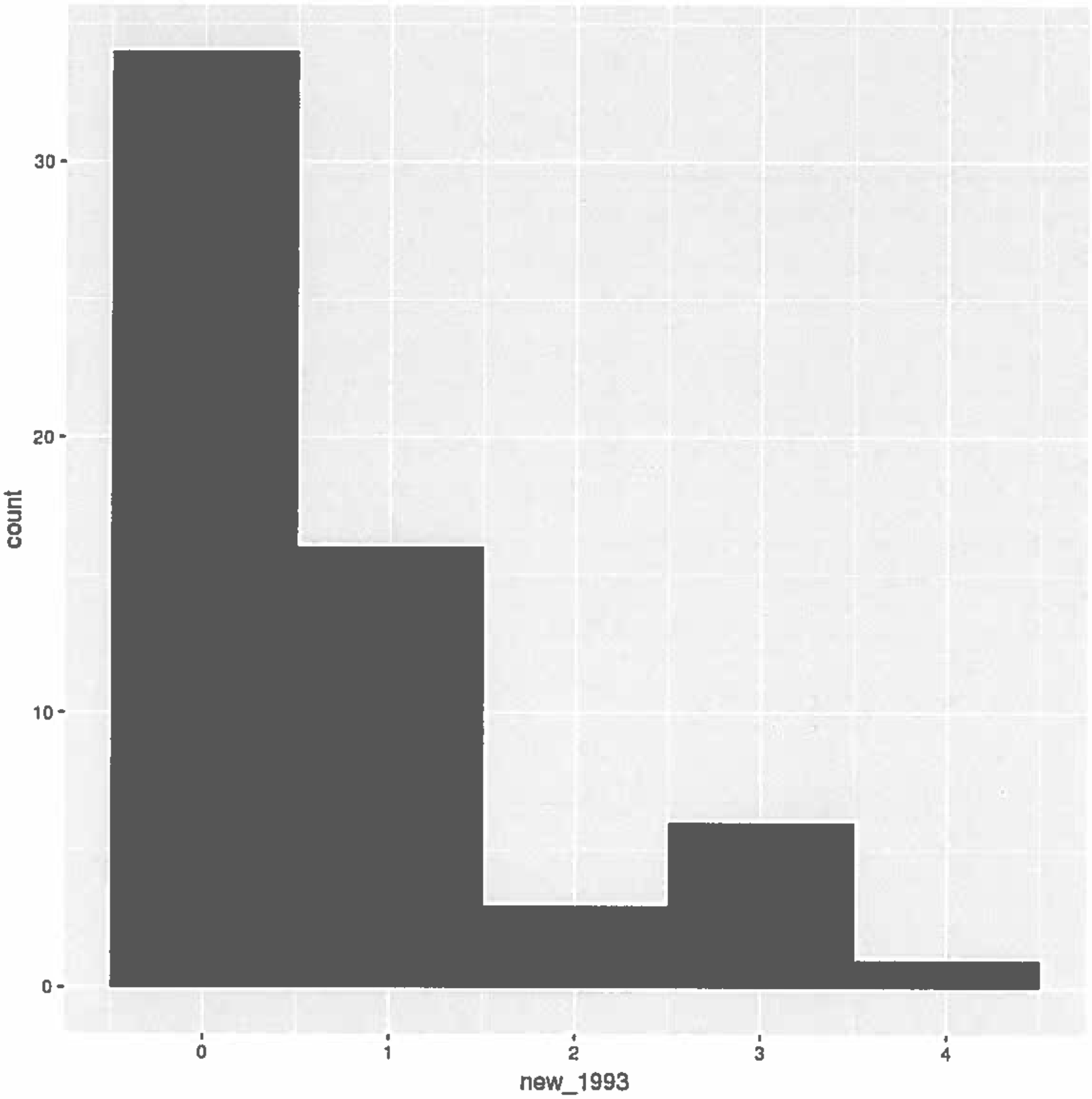
```
library(tidyverse)
seedlings <- read_table("http://www.evanlray.com/data/lavine_intro_stat_thought/seedlings.txt")
seedlings <- seedlings %>%
  select(quadrat = Block,
    old_1991 = `91`,
    old_1992 = `92-t`,
    old_1993 = `93-t`,
    old_1994 = `94-t`,
    old_1995 = `95-t`,
    old_1996 = `96-t`,
    old_1997 = `97-t`,
    new_1992 = `92-1`,
    new_1993 = `93-1`,
    new_1994 = `94-1`,
    new_1995 = `95-1`,
    new_1996 = `96-1`,
    new_1997 = `97-1`
  )
head(seedlings)
```

```
Parsed with column specification:
cols(
  Block = col_integer(),
  `91` = col_integer(),
  `92-t` = col_integer(),
  `93-t` = col_integer(),
  `94-t` = col_integer(),
  `95-t` = col_integer(),
  `96-t` = col_integer(),
  `97-t` = col_integer(),
  Mean = col_double(),
  SD = col_double(),
  `92-1` = col_integer(),
  `93-1` = col_integer(),
  `94-1` = col_integer(),
  `95-1` = col_integer(),
  `96-1` = col_integer(),
  `97-1` = col_integer(),
  `Mean-1` = col_double(),
  `SD-1` = col_double()
)
```

quadrat	old_1991	old_1992	old_1993	old_1994	old_1995	old_1996	old_1997	new_1992	new_1993	new_1994	new_1995	new_1996	ne
1	1	1	1	1	2	1	1	0	0	0	1	0	0
2	1	1	1	0	1	0	0	0	0	0	0	0	0
3	1	1	0	0	1	1	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	2	1	1	1	1	0	0	0	0	0	0
6	1	0	1	0	0	0	0	0	1	0	0	0	0

For today, let's analyze the number of new seedlings observed in each quadrat during 1993. Here's a plot:

```
In [13]: ggplot(data = seedlings, mapping = aes(x = new_1993)) +
           geom_histogram(binwidth = 1)
```



1. What statistical model would be appropriate for the distribution of the number of new seedlings observed in the first quadrat?

For concreteness, let the random variables X_1 denote the number of new seedlings counted in quadrat number 1 in 1993.

Note that the seedlings are very small, so it's reasonable to assume that they do not affect each other. So, an assumption of independence for the seedlings within each quadrat is plausible.

$X_1 \sim \text{Poisson}(\lambda)$

2. Ecologists want to learn about the rate at which new seedlings emerge in a quadrat. How does this relate to the statistical model you wrote down in part 1.?

The parameter λ is the average rate at which new seedlings emerge, according to this Poisson model.

3. Write down the probability mass function for X_1, the number of new seedlings in the first quadrat.

$$f(x_1|\lambda) = \frac{e^{-\lambda}\cdot\lambda^{x_1}}{x_1!}$$

4. Write down the joint p.m.f. for $X_1, \dots, X_n$, where X_i is the number of new seedlings in quadrat i.

In this example, $n = 60$ since we have observed data for 60 different quadrats. For now, let's assume that the number of seedlings that emerge in different quadrats are i.i.d., although this may not be realistic (for example, some quadrats may have better than soil than others, so may tend to have more seedlings).

$$f(x_1,\dots,x_n|\lambda) = \prod_{i=1}^{6n} f(x_i|\lambda)$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

5. Prove that the maximum likelihood estimator for the model parameter λ is the sample mean: $\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$

We have observed the values $x_1, \ldots, x_n$, where $n = 60$. We can use those observed values to estimate the model parameters via maximum likelihood.

Here are some definitions:

- **An estimator is**:

- **An estimate is**:

- **The likelihood function is**:

- **The log-likelihood function is**:

Since writing this document, I have switched to using the book's notation: $\mathcal{L}(\lambda|x_1,\ldots,x_n)$ denotes the likelihood and $L(\lambda|x_1,\ldots,x_n)$ the log-likelihood

Our goal is to find a maximum of the likelihood function $L(\lambda|x_1, \ldots, x_n)$, but it's often much easier to work with the *log*-likelihood function $\ell(\lambda|x_1, \ldots, x_n)$. It can be shown that a value λ^* maximizes the likelihood function if and only if it maximizes the log-likelihood function. (I'm not going to prove this in class, but you might be interested in thinking about why this is the case if it's not clear to you. You're also always welcome to ask on Piazza or during office hours!)

Let's work with the log-likelihood. There are two steps to your proof.

(a) Find a critical point.

We know that the maximum of the (log-)likelihood function must occur at a critical point, where $\frac{d}{d\lambda}\ell(\lambda|x_1, \ldots, x_n) = 0$. Calculate the derivative of the log-likelihood function with respect to λ, set the result to 0, and solve for the critical point.

$$\mathcal{L}(\lambda|x_1,\ldots,x_n) = \frac{e^{-n\lambda}\cdot\lambda^{\Sigma x_i}}{\prod_{i=1}^{n} x_i!}$$

$$L(\lambda|x_1,\ldots,x_n) = -n\lambda + \Sigma x_i \cdot \log(\lambda) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{d}{d\lambda} L(\lambda|x_1,\ldots,x_n) = -n + \frac{\Sigma x_i}{\lambda} = 0$$

$$\Rightarrow \hat{\lambda} = \frac{1}{n}\Sigma x_i$$

Technically, we should rewrite this with a capital X_i to emphasize that the estimator $\hat{\lambda}$ is a random variable: $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) Verify that the critical point occurs at a maximum of the likelihood function.

Recall that the critical point you found in part (a) is a maximum if the second derivative of the log-likelihood function is negative. Verify that this condition holds.

$$\frac{d^2}{d\lambda^2} L(\lambda | x_1, \dots, x_n) = \frac{d}{d\lambda}\left[-n + \left(\sum_{i=1}^{n} x_i\right)\lambda^{-1}\right]$$

$$= \left(\sum_{i=1}^{n} x_i\right)\cdot(-1)\,\lambda^{-2}$$

$$= -1 \cdot \frac{\sum_{i=1}^{n} x_i}{\lambda^2} \quad \leq 0$$

The above will generally be negative since $\lambda^2 > 0$ and each $x_i \geq 0$,

6. Find the maximum likelihood estimate of the model parameter λ.

You will find the following output from R to be helpful:

```
In [16]: summarize(seedlings, total_new_seedlings_1993 = sum(new_1993))
```

total_new_seedlings_1993
44

$$\hat{\lambda} = \frac{1}{60} \cdot 44 = \frac{11}{15}$$