Stat 343 MLE Practice: Saplings

Introduction

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.

```
library(tidyverse)

seedlings <- read_table("http://www.evanlray.com/data/lavine_intro_stat_thought/seedlings.txt") %>%
  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)
```

```
## # A tibble: 6 x 14
##   quadrat old_1991 old_1992 old_1993 old_1994 old_1995 old_1996 old_1997
##     <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1       1        1        1        1        1        2        1        1
## 2       2        1        1        1        0        1        0        0
## 3       3        1        1        0        0        1        1        0
## 4       4        0        0        0        0        0        0        0
## 5       5        1        1        2        1        1        1        1
```

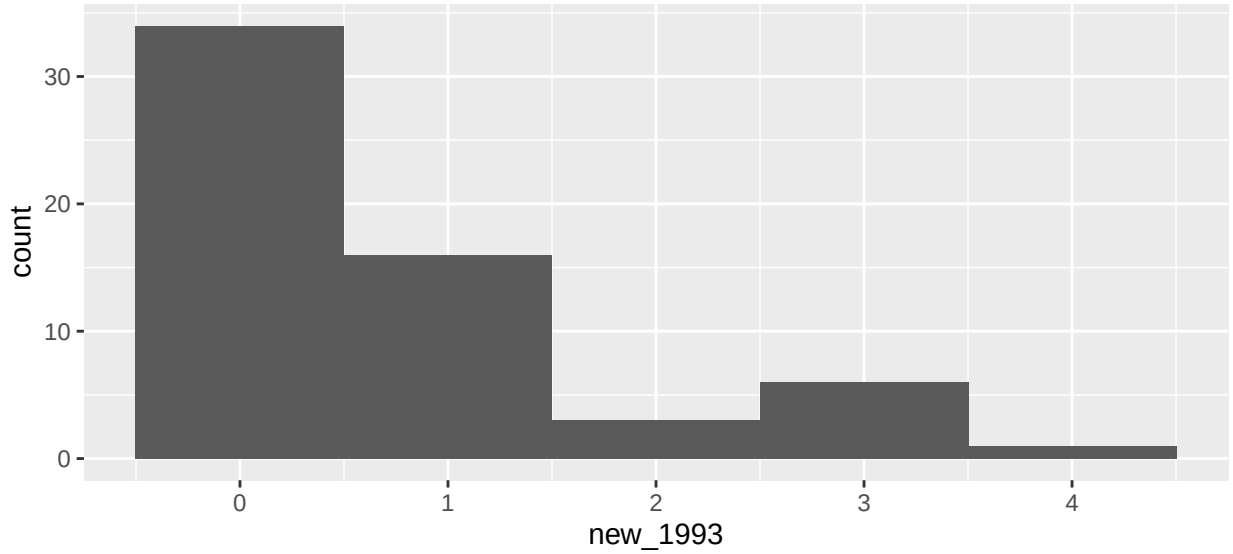
```
## 6       6       1       0       1       0       0       0       0
## # ... with 6 more variables: new_1992 <dbl>, new_1993 <dbl>, new_1994 <dbl>,
## #   new_1995 <dbl>, new_1996 <dbl>, new_1997 <dbl>
```

```
dim(seedlings)
```

```
## [1] 60 14
```

For today, let's analyze the number of new seedlings observed in each of the 60 quadrats during 1993. Here's a plot:

```
ggplot(data = seedlings, mapping = aes(x = new_1993)) +
  geom_histogram(binwidth = 1)
```



Define the random variables X_i, $i = 1, \ldots, 60$ to represent the number of new seedlings observed in each quadrat i in 1993.

1. What statistical model would be appropriate for the distribution of each X_i?

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 roughly plausible.

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

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?

This rate is the parameter λ for the Poisson distribution.

3. Write down the probability mass function for X_i, the number of new seedlings in the first quadrat number i, based on the model you specified in part 1.

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

4. Write down the joint p.m.f. for $X_1, \ldots, X_{60}$.

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). It will be helpful later if you simplify this joint pmf as much as you reasonably can before moving on.

$$\begin{aligned} f(x_1, \ldots, x_{60}|\lambda) &= \prod_{i=1}^{60} f(x_i|\lambda) \\ &= \prod_{i=1}^{60} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \\ &= e^{-60\lambda}\lambda^{\left(\sum_{i=1}^{60} x_i\right)} \prod_{i=1}^{60}\frac{1}{x_i!} \end{aligned}$$

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

We will break this down into a few steps.

i. Find the log-likelihood function.

$$\begin{aligned} \ell(\lambda|x_1, \ldots, x_{60}) &= \log\left\{\mathcal{L}(\lambda|x_1, \ldots, x_{60})\right\} \\ &= \log\left\{e^{-60\lambda}\lambda^{\left(\sum_{i=1}^{60} x_i\right)} \prod_{i=1}^{60}\frac{1}{x_i!}\right\} \\ &= -60\lambda + \left(\sum_{i=1}^{60} x_i\right)\log(\lambda) + \log\left\{\prod_{i=1}^{60}\frac{1}{x_i!}\right\} \end{aligned}$$

ii. Find a critical point of the log-likelihood function.

$$\begin{aligned}\frac{d}{d\lambda}\ell(\lambda|x_1,\ldots,x_{60}) &= \frac{d}{d\lambda}\left[-60\lambda + \left(\sum_{i=1}^{60} x_i\right)\log(\lambda) + \log\left\{\prod_{i=1}^{60}\frac{1}{x_i!}\right\}\right] \\ &= -60 + \frac{1}{\lambda}\left(\sum_{i=1}^{60} x_i\right)\end{aligned}$$

Setting equal to 0, we obtain

$$0 = -60 + \frac{1}{\lambda}\left(\sum_{i=1}^{60} x_i\right)$$

Solving for λ, a critical point is therefore

$$\lambda = \frac{1}{60}\left(\sum_{i=1}^{60} x_i\right)$$

iii. Verify that the critical point occurs at a global maximum of the log-likelihood function.

The second derivative of the log-likelihood function with respect to λ is:

$$\begin{aligned}\frac{d^2}{d\lambda^2}\ell(\lambda|x_1,\ldots,x_{60}) &= \frac{d}{d\lambda}\left[-60 + \frac{1}{\lambda}\left(\sum_{i=1}^{60} x_i\right)\right] \\ &= \frac{-1}{\lambda^2}\left(\sum_{i=1}^{60} x_i\right) \\ &< 0\end{aligned}$$

Since the second derivative of the log-likelihood is negative for all values of λ, the log-likelihood function is concave and the critical point identified in part ii above is a global maximum of the log-likelihood.

The maximum likelihood estimator is $\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This is a random variable.

For a particular sample, the maximum likelihood estimate will be $\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

6. In RStudio, find the maximum likelihood estimate and overlay a visualization of the pmf on a density histogram of the data.