HW week 7

w203: Statistics for Data Science

The Meat

Suppose that Americans consume an average of 2 pounds of ground beef per month.

(a) Do you expect the distribution of this measure (ground beef consumption per capita per month) to be approximately normal? Why or why not?

(b) Suppose you want to take a sample of 100 people. Do you expect the distribution of the sample mean to be approximately normal? Why or why not?

(c) You take a random sample of 100 Berkeley students to find out if their monthly ground beef consumption is any different than the nation at large. The mean among your sample is 2.45 pounds and the sample standard deviation is 2 pounds. What is the 95% confidence interval for Berkeley students?

GRE Scores

Assume we are analyzing MIDS students' GRE quantitative scores. We want to construct a 95% confidence interval, but we *naively* uses the famous 1.96 threshold as follows:

$$\big(\bar{X} - 1.96 \cdot \frac{s}{\sqrt{n}}, \big(\bar{X} + 1.96 \cdot \frac{s}{\sqrt{n}}\big)$$

What is the real confidence level for the interval we have made, if the sample size is 10? What if the sample size is 200?

Maximim Likelihood Estimation for an Exponential Distribution

A Poisson process is a simple model that statisticians use to describe how events occur over time. Imagine that time stretches out on the x-axis, and each event is a single point on this axis.

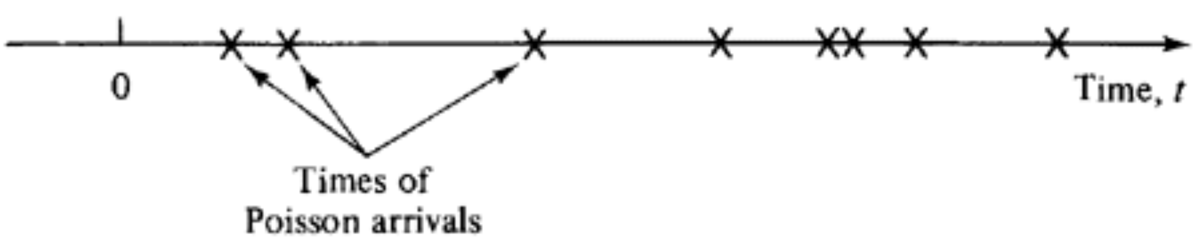


Figure 1: Events over time

The key feature of a Poisson process is that it is *memoryless*. Loosely speaking, the probability that an event occurs in any (differentially small) instant of time is a constant. It doesn't depend on how long ago the previous event was, nor does it depend on when future events occur. Statisticians might use a Poisson process (or more complex variations) to represent:

- The scoring of goals in a world cup match
- The arrival of packets to an internet router
- The arrival of customers to a website

- The failure of servers in a cluster
- The time between large meteors hitting the Earth

In live session, we described a Poisson random variable, a discrete random variable that represents the number of events of a Poisson process that occur in a fixed length of time. However, a Poisson process can be used to generate other random variables.

Another famous random variable is the exponential random variable, which represents the time between events in a Poisson process. For example, if we set up a camera at a particular intersection and record the times between car arrivals, we might model our data using an exponential random variable.

The exponential random variable has a well-known probability density function,

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

Here, λ is a parameter that represents the rate of events.

Suppose we record a set of times between arrivals at our intersection, $x_1, x_2, ...x_n$. We assume that these are independent draws from an exponential distribution and we wish to estimate the rate parameter λ using maximum likelihood.

Do this using the following steps:

a. Write down the likelihood function, $L(\lambda)$. Hint: We want the probility (density) that the data is exactly $x_1, x_2, ..., x_n$. Since the times are independent, this is the probability (density) that $X_1 = x_1$, times the probability (density) that $X_2 = x_2$, and so on.

b. To make your calculations easier, write down the log of the likelihood, and simplify it.

c. Take the derivative of the log of likelihood, set it equal to zero, and solve for λ. How is it related to the mean time between arrivals?

d. Suppose you get the following vector of times between cars:

```
times = c(2.65871285, 8.34273228, 5.09845548, 7.15064545,
          0.39974647, 0.77206050, 5.43415199, 0.36422211,
          3.30789126, 0.07621921, 2.13375997, 0.06577856,
          1.73557740, 0.16524304, 0.27652044)
```

Use R to plot the likelihood function. Then use optimize to approximate the maximum likelihood estimate for λ. How does your answer compare to your solution from part c?

W223 Hw 7 Shan He

1. (a) Yes; I'd think that it will be approximately a normal distribution with most Americans consuming around 2 lb of ground beef per month. And as the monthly ground beef consumption deviates from 2 lb, in both directions, there will be fewer and fewer amount of people.

(b) Yes; invoking Central Limit Theorem, the sample size here is larger than 30. Considering that the population distribution won't be heavily skewed, I would expect the sample mean to be approximately normal.

(c) Assuming a t distribution, we can use the following equation:

$$95\%\ CI = \left(\bar{x} - t_{\alpha/2,99} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,99} \cdot \frac{s}{\sqrt{n}}\right) \qquad t_{0.025,99} = 1.984$$

$$= \left(2.45 - 1.984 \cdot \frac{2}{\sqrt{100}},\ 2.45 + 1.984 \cdot \frac{2}{\sqrt{100}}\right)$$

$$= (2.0532,\ 2.8468)$$

2. Referring to $\rightarrow$ table of critical values for t distribution,

① $n = 10$, $t_{\alpha/2,9} = 1.96$, using R: "a ← (1-pt(1.96,9))·2", a = 0.08
So the CI is actually $100(1-0.08)\% = 92\%$ CI

② $n = 200$, $t_{\alpha/2,199} = 1.96$, using R: "a ← (1-pt(1.96,199))·2", a = 0.05
So the CI is actually $100(1-0.05)\% = 95\%$ CI

3. a. $L(\lambda) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot \ldots \cdot \lambda e^{-\lambda x_n}$

b. $\log(L(\lambda)) = \log(\lambda e^{-\lambda x_1}) + \log(\lambda e^{-\lambda x_2}) \ldots + \log(\lambda e^{-\lambda x_n})$

$= \log(\lambda) + \log(e^{-\lambda x_1}) + \log(\lambda) + \log(e^{-\lambda x_2}) \ldots + \log(\lambda) + \log(e^{-\lambda x_n})$

$= n \cdot \log(\lambda) - \lambda x_1 - \lambda x_2 - \ldots - \lambda x_n$ or $\sum_{i=1}^{n} (\log(\lambda) - \lambda x_i)$

c. $d\log(L(\lambda))/d\lambda = n \cdot \frac{1}{\lambda} - x_1 - x_2 - \ldots x_n = 0$

$n \cdot \frac{1}{\lambda} - (x_1 + x_2 + \ldots + x_n) = 0 \Rightarrow n \cdot \frac{1}{\lambda} - n \cdot E(x) = 0$

$\frac{1}{\lambda} = E(x) = \mu. \Rightarrow \lambda = \frac{1}{\mu}$

d. see next page

w203_hw7_q3_SH

Shan He

10/26/2017

1. Observations

```
times = c(2.65871285, 8.34273228, 5.09845548, 7.15064545,
          0.39974647, 0.77206050, 5.43415199, 0.36422211,3.30789126,
          0.07621921, 2.13375997, 0.06577856, 1.73557740, 0.16524304,
          0.27652044)
```

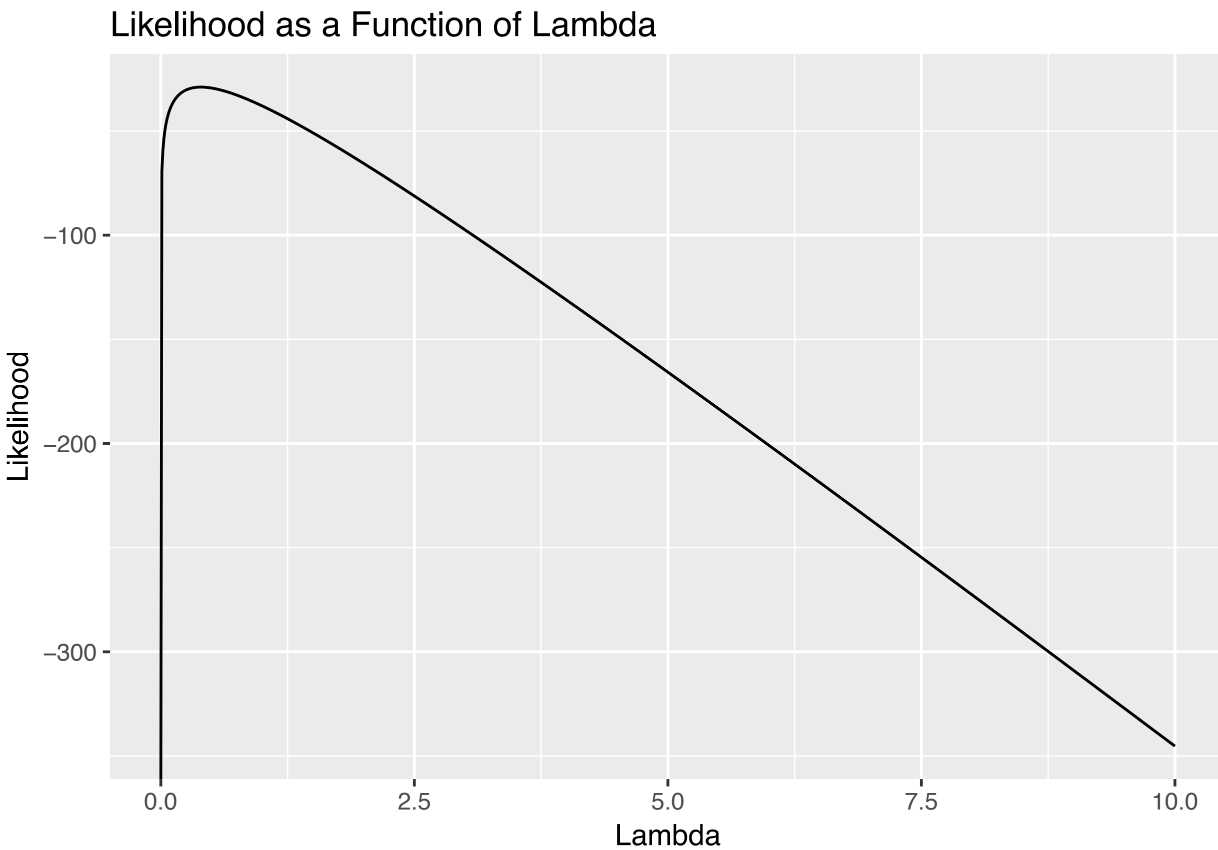
2. Likelihood Function

```
log.lklh.poisson <- function(x, lambda){
   sum(log(lambda) - x * lambda)
}
```

3. Plot Likelihood Function

```
library(ggplot2)

lambda <- seq(0 ,10 ,by = 0.01)
qplot(lambda,
      sapply(lambda, function(lambda){log.lklh.poisson(times, lambda)}),
      geom = 'line',
      main = 'Likelihood as a Function of Lambda',
      xlab = 'Lambda',
      ylab = 'Likelihood'
      )
```



4. Optimize Likelihood Function

```
fn <- function(lambda){log.lklh.poisson(times, lambda)}
optimize(fn, interval=c(0,10), maximum = T)
```

```
## $maximum
## [1] 0.3949309
##
## $objective
## [1] -28.93582
```

5. Compare MLE for Lambda to Mean

```
lambda_optim = optimize(fn, interval=c(0,10), maximum = T)$maximum
lambda_optim
```

```
## [1] 0.3949309
```

```
1/mean(times)
```

```
## [1] 0.3949269
```

The MLE for Lambda is the approximately the same as $1/E(x)$ just as expected.