HW week 7

w203: Statistics for Data Science

Tako Hisada

10/29/2017

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?

Yes. CLT states that when n is sufficiently large, a suitable normal curve will approximate the actual distribution of $\bar{X}$. Since n in this case is the entire population of the US which is obviously quite large, we can expect the distribution of this measure to be approximately normal.

(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?

Yes. The rule of thumb is that CLT can be generally applied when $n > 30$.

(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?

```
compCI95 <- function(sd, n) {
  ci = 1.96*sd/sqrt(n)
}
ci <- compCI95(2, 100)

(ci_l <- 2.45 - ci)
```

```
## [1] 2.058
```

```
(ci_r <- 2.45 + ci)
```

```
## [1] 2.842
```

$CI = \bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{n}} = 2.45 \pm (0.392) = (2.058, 2.842)$

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:

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

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?

The sample size of 10 is not large enough for us to subsititute σ with S. Therefore what we are working with here is a t-curve rather than a z-curve. For a sample size of 10, v (df) would be 9.

```
t = 1.96
v = 10 - 1
compCL <- function(t, v) {
  1-2*(1-pt(t, v))
}
(conf_10 <- compCL(t, v))
```

```
## [1] 0.9183556
```

For a sample with 9 DF, this would present a 91.84% confidence interval.

Similarly, for when the sample size is 200, we could compute the confidence level as below:

```
v = 200 - 1
(conf_200 <- compCL(t, v))
```

```
## [1] 0.9486082
```

However the sample size 200 is sufficiently larger than the recommended sample size of $n > 40$. Therefore we can say the interval gives us the confidence level of 95% in this case.

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.

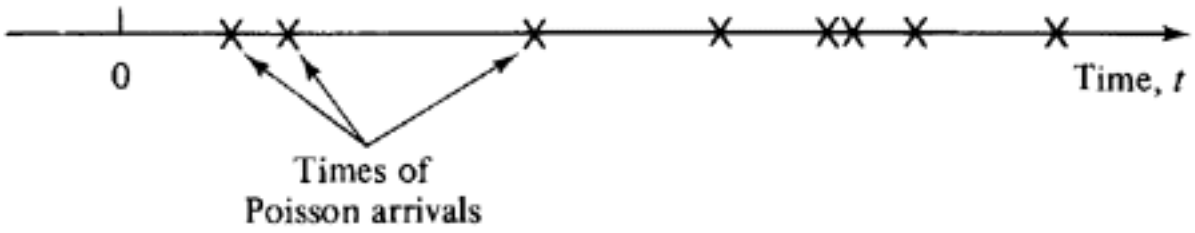


Figure 1: Events over time

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.

$$L(\lambda) = f(x_1, ..., x_n; \lambda) = (\lambda e^{-\lambda x_1})...(\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$

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

$$\ln L(\lambda) = ln(\lambda^n e^{-\lambda \sum x_i}) = ln(\lambda^n) + ln(e^{-\lambda \sum x_i}) = n \ln(\lambda) - \lambda \sum x_i$$

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?

$$\frac{d}{d\lambda} \ln L(\lambda) = \frac{d}{d\lambda}\left(n \ln(\lambda) - \lambda \sum x_i\right) = \frac{n}{\lambda} - \sum x_i = 0$$

$$\frac{n}{\lambda} = \sum x_i$$

$$\frac{1}{\lambda} = \frac{\sum x_i}{n}$$

$$\lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

From this, we can say λ is the inverse of the mean time between arrivals.

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

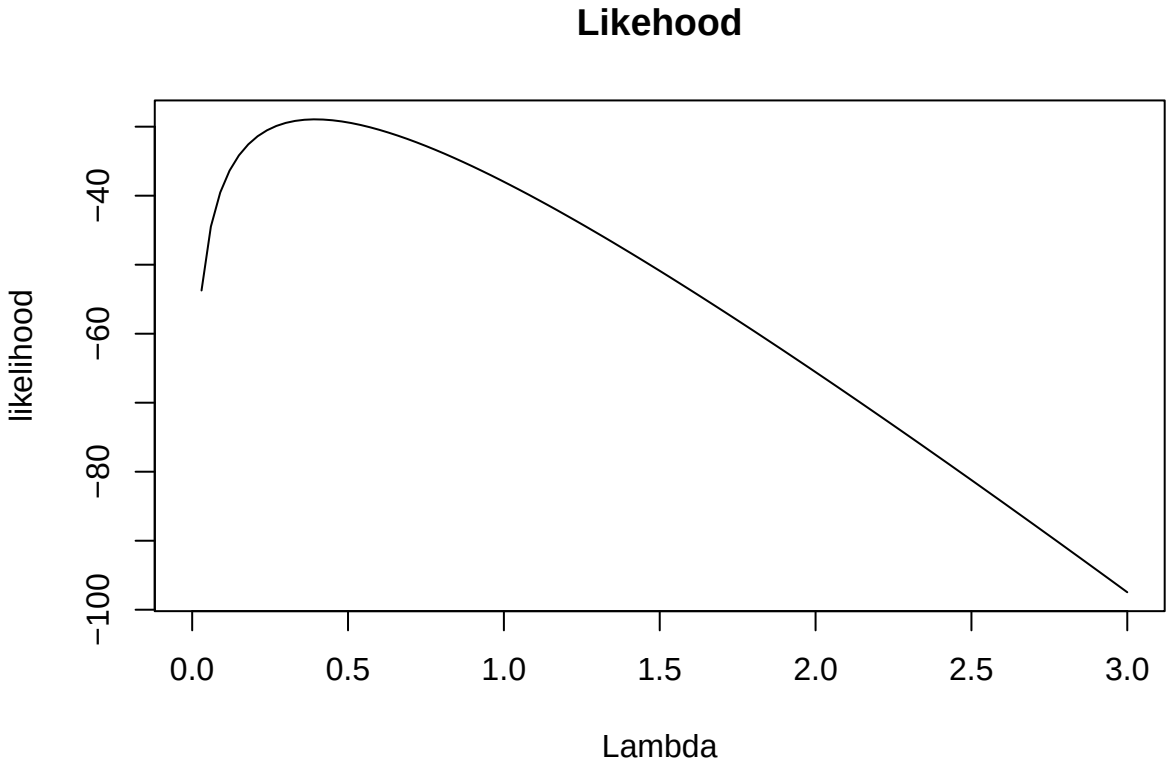
```
times = c(2.65871285, 8.34273228, 5.09845548, 7.15064545,
          0.39974647, 0.77206050, 5.43415199, 0.36422211,
          3.30789126, 0.07621921, 2.13375997, 0.06577856,
          1.73557740, 0.16524304, 0.27652044)
```

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?

```
likelihood <- function(l) {
  length(times)*log(l)-l*sum(times)
}

lambda=seq(0,3,.01)

plot(likelihood, lambda, xlim=c(0,3),xlab='Lambda',main="Likehood")
```




```
(opt_lh <- optimize(likelihood,interval=c(0,3), maximum = TRUE))
```

```
## $maximum
## [1] 0.3949192
##
## $objective
## [1] -28.93582
```

```
mean <- mean(times)
(opt_mean <- 1/mean)
```

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

The maximum likelihood estimate we get from the optimize function is very close to the value we get from the solution for c $\frac{1}{\bar{X}}$.