w203: Week 7 HW

Adam Yang

6/21/2018

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?

Answer:

(a)

A normal distribution lives from $-\infty$ to ∞ though the bulk of the data is usually in a much smaller range. The average meat consumption, however, is capped at 0 because you can not have negative meat consumption. On the other end of the distribution, we do not have a cap as someone can theoretically eat 100 lbs of beef per week. Furthermore, there are many people people who do not eat beef, whether for religious reasons or simply because they're vegetarian or vegan, so the density at 0 can be rather high. I think it is highly likely that there is a right skew to the data so we cannot simply approximate it as a normal distribution.

(b)

The central limit theorem tells us that if the sample is large enough, the distribution of the sample mean would be approximately normal. The general rule of thumb is that if n > 30, we can apply the central limit theorem. However, when the data is sufficiently skewed, n would have to be much larger than 30. In this case, we have n = 100 which is large enough under regular circumstances to approximate a normal distribution. However, if the distribution is very skewed, this is not necessarily the case. My answer is, if our distribution is not very skewed, then I would expect the distribution of the sample mean to be approximately normal with a sample size of 100. However, if the distribution is very skewed, then our sample size might not be big enough to approximate a normal distribution.

(c)

Because we only know the sample standard deviation (s), the first thing we need to do is to find the corresponding t-score for 95% confidence interval:

```r
# Find the t-score for a confidence interval of 95% (1-0.25 = 0.975),
# when the sample size is 100 (dof = n-1 = 99)
qt(0.975, 99)
```

```
## [1] 1.984217
```

Now we can plug our values into $(\bar{X} - 1.984217 \cdot \frac{s}{\sqrt{n}}, \bar{X} + 1.984217 \cdot \frac{s}{\sqrt{n}})$ to find the confidence intervals.

```r
xbar <- 2.45
s <- 2
n <- 100
t <- qt(0.975, 99)
c(xbar-t*s/sqrt(n), xbar+t*s/sqrt(n))
```

```
## [1] 2.053157 2.846843
```

Therefore, the confidence interval $= (2.45 - 1.984217 \cdot \frac{2}{\sqrt{100}}, 2.45 + 1.984217 \cdot \frac{2}{\sqrt{100}}) = (2.053157, 2.846843)$.

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:

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

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?

Answer:

In the equation for the confidence interval shown above, the sample standard deviation (s) is used rather than the population standard deviation (σ). This suggests that we should be using the t-score rather than the z-score.

If we used 1.96 as the t-score, with a sample size of 10 (degree of freedom = n-1 = 9):

```r
# Get the cumulative probability density from -infinity to -1.96
x <- pt(-1.96,9)
# Find the cumulative probability density between -1.96 and 1.96
1-x*2
```

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

the real confidence interval would be 91.8%.

If we used a 1.96 as the t-score, with a sample size of 200 (degree of freedom = n-1 = 199):

```r
# Get the cumulative probability density from -infinity to -1.96
x <- pt(-1.96,199)
# Find the cumulative probability density between -1.96 and 1.96
1-x*2
```

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

the real confidence interval would be 94.9%.

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.

The key feature of a Poisson process is that it is memory-less. Loosely speaking, the probability that an event occurs in any (deferentially 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, x1,x2,…xn. 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(λ). Hint: We want the probability (density) that the data is exactly x1, x2, …, xn. Since the times are independent, this is the probability (density) that X1 = x1, times the probability (density) that X2 = x2, 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:

```r
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?

Answer:

(a)

The likelihood function is:

$$L(\lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})...(\lambda e^{-\lambda x_n}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

(b)

$$\begin{gathered}
\ln[L(\lambda)] = \ln[(\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})...(\lambda e^{-\lambda x_n})] \\
= \ln(\lambda e^{-\lambda x_1}) + \ln(\lambda e^{-\lambda x_2}) + ... + \ln(\lambda e^{-\lambda x_n}) \\
= \ln(\lambda) + \ln(e^{-\lambda x_1}) + \ln(\lambda) + \ln(e^{-\lambda x_2}) + ... + \ln(\lambda) + \ln(e^{-\lambda x_n}) \\
= n\ln(\lambda) + \ln(e^{-\lambda x_1}) + \ln(e^{-\lambda x_2}) + ... + \ln(e^{-\lambda x_n}) \\
= n\ln(\lambda) + -\lambda x_1 \ln(e) + -\lambda x_2 \ln(e) + ... + -\lambda x_n \ln(e) \\
= n\ln(\lambda) - \lambda[x_1 + x_2 + ... + x_n] = n\ln(\lambda) - \lambda \sum_{i=1}^{n} x_i
\end{gathered}$$

(c)

$$\frac{\partial}{\partial \lambda} L(\lambda) = \frac{\partial}{\partial \lambda}[n\ln(\lambda) - \lambda * (x_1 + x_2 + ... + x_n)]$$

$$= \frac{n}{\lambda} - [x_1 + x_2 + ... + x_n]$$

$$\frac{n}{\lambda} - [x_1 + x_2 + ... + x_n] = 0$$

$$\frac{n}{\lambda} = [x_1 + x_2 + ... + x_n]$$

$$\lambda = \frac{n}{[x_1 + x_2 + ... + x_n]} = \frac{n}{\sum_{i=1}^{n} x_i}$$

The maximum likelihood estimate for λ is the inverse of the mean time between arrivals.

(d)

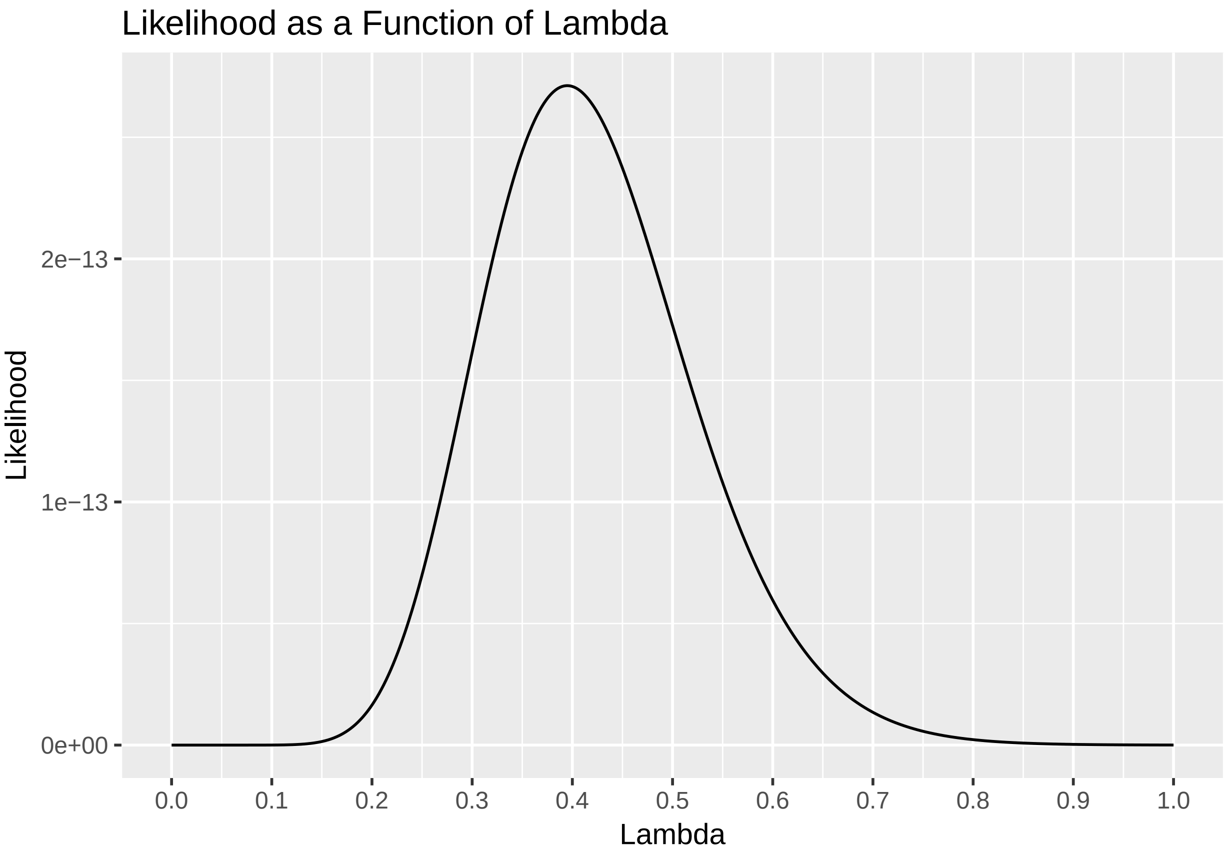
```
  times = c(2.65871285, 8.34273228, 5.09845548, 7.15064545,
            0.39974647, 0.77206050, 5.43415199, 0.36422211,
            3.30789126, 0.07621921, 2.13375997, 0.06577856,
            1.73557740, 0.16524304, 0.27652044)
```

First, lets plot the likelihood function:

```
# Likelihood function
likelihood <- function(x, lambda) {
  l <- 1
  for (i in 1:length(x)) {
    l = l*lambda*exp(-1*lambda*x[i])
  }
  return(l)
}

# Plotting the likelihood function
library(ggplot2)
lambda <- seq(0, 1, by = 0.001)
options(repr.plot.height = 10, repr.plot.width = 15, repr.plot.pointsize = 32)

p <- qplot(lambda,
      sapply(lambda, function(lambda) {likelihood(times, lambda)}),
      geom = 'line',
      main = 'Likelihood as a Function of Lambda',
      xlab = 'Lambda',
      ylab = 'Likelihood')
p + scale_x_continuous(breaks=seq(0,1, by=0.1), limits=c(0,1))
```



From our graph above, it looks like the maximum likelihood estimate for λ is around 3.9.

The maximum likelihood estimate formula I calculated for λ in part c is $\lambda = \frac{n}{\sum\limits_{i=1}^{n} x_i}$. Lets plug in our sample data into this equation:

```
# solve for n/[x1+x2+...+xn]
length(times)/sum(times)
```

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

Our resulting maximum likelihood estimate for λ is 0.3949269.

Now, lets use the optimize function to approximate the maximum likelihood estimate for λ:

```
optimize(function(lambda) {likelihood(times, lambda)}, interval = c(0,1), maximum = T)
```

```
## $maximum
## [1] 0.3949072
##
## $objective
## [1] 2.712269e-13
```

The resulting maximum likelihood estimate for λ in this case is 0.3949072 which is equivalent to our solution in part c to the 0.00001 decimal.