Statistics for Data Science

W203 Instructional Team

Main Topics Covered in Lecture 7:

- Point estimation
- The Method of Maximum likelihood
- The Method of Momemts (MOM)
- Properties of estimators, such as unbiasedness, and unbiased estimator with minimum variance
- The Invariance Principle (associated with MLE)
- Confidence intervals
- Confidence intervals under a large sample (for μ and σ)

Readings:

Devore, Jay L. *Probability and Statistics for Engineering and the Science* - Chapter 6 - Chapter 7.1, 7.2

Agenda for the Live Session

1. A brief introduction to the lecture (*Estimated Time: Total 5 minute*)
2. Discussion 1: Point estimation, MOM, and the method of ML (*Estimated Time: Total 15 minutes (Breakout 8 minutes, Classwide disscussion: 7 minutes)*)
3. Optimization in R (*Estimated Time: Total 10 minutes (Breakout 3 minutes, Classwide disscussion: 7 minutes)*)
4. Discussion 2: Maximum Likelihood Estimation of Bernoulli Random Variables (*Estimated Time: Total 30 minutes (Breakout 15 minutes, Classwide disscussion: 15 minutes)*)
5. Discussion 3: Maximum Likelihood Estimation of Poisson Random Variables (*Estimated Time: Total 30 minutes (Breakout 15 minutes, Classwide disscussion: 15 minutes)*)

A brief introduction to the lecture

The professor will give an overview of the materials covered this week and the materials focused in the live session today.

Discussion 1: Point Estimation, MOM, and The Method of ML

1. What is point estimation?
2. Given $X_1, X_2, \dots, X_n$ a random sample from a distribution with PMF or PDF $f(x; \theta_1, \dots, \theta_m)$, where $\theta_1, \dots, \theta_m$ are unknown parameters.

   a. In your own words, describe how the method of moments is used to estimate the unknown parameters?
   b. In your own words, describe how the method of ML is used to estimate the unknown parameters?

3. Why do data scientists need to understand likelihood?

Warm-up: Optimization in R

R: Optimize Function

The method of maximum likelihood requires an optimization routine. For a few very simple probability models, a closed-form solution exists and the MLE can be derived by hand. In most cases, however, hand-derivation will be too tedious, if at all possible, and a numerical computation technique is needed. Numerical computaiton techniques often require some sort of optimization. For our purpose in this live session, we will focus on the practice of maximizing likelihood functions using R.

There are many optimizers in R(), including optimize(), optim(), and optimx(). I will use optimize() which is very simple to use, but only works for one dimension.

As always, I encourage you to read the documentation of the functions you are using: optim

An Example:

Suppose that a firm's revenue r from selling a product is related to price p as follows:

$$r = -p^2 + p + 2$$

1. Explain how you would use calculus to find the maximizing price.
2. To solve this numerically in *R*, we can use the *optimize()* function.

```
f <- function(p) {
  -p^2 + p + 2
}
optimize(f, interval = c(0,100), maximum = TRUE)
```

```
## $maximum
## [1] 0.5
##
## $objective
## [1] 2.25
```

Discussion 2: Maximum Likelihood Estimation of Bernoulli Random Variables

```
- Estimated Time: Total 30 minutes
- Breakout 15 minutes, Classwide disscussion: 15 minutes
```

Suppose that you've got a sequence of values

$$1, 0, 0, 1, 0, 1, 1, 1, 1, 1$$

, which, say, indicates whether a printer jams each day, for the last 10 business days. *Business Question: What is the probability (p) that the printer jams in any given day?*

It resembles draws from a Bernoulli disribution. However, even if we want to model this as a Bernoulli distribution, we do not know what the value of the parameter, p, is.

```r
p=.7
s <- rbinom(10, 1, p)
s
```

```
##  [1] 1 1 1 1 0 1 1 1 0 1
```

Let's review the steps to find the maximum likelihood estimate for p.

1. Define your random variables.

Let $X_1, X_2, \ldots, X_{10}$ be a sequence of *i.i.d.* Bernoulli random variables with parameter p, where $X_i = 1$ if the printer was jammed on day i, and $X_i = 0$ otherwise.

2. Write down the likelihood function.

Remember that likelihood is just another way of looking at the probability function:

$L(p) =$ P(The data is (1, 0, 0, 1, 0, 1, 1, 1, 1, 1) given p)

$= P(X_1 = 1 \cap X_2 = 0 \cap ... \cap X_{10} = 1; p)$

Question: How can you simplify the last expression? What property do you use to do this?

We can write down the likelihood for an individual day as follows:

$$P(X_i = x_i; p) = f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

Putting this together, our overall likelihood function is:

$$\begin{aligned} L(p) &= f(x_1, \ldots, x_{10}; p) \\ &= \prod_{i=1}^{10} f(x_i; p) \\ &= \prod_{i=1}^{10} p^{x_i}(1-p)^{10-x_i} \\ &= p^{\sum_{i=1}^{10} x_i}(1-p)^{10-\sum_{i=1}^{10} x_i} \\ &= p^7(1-p)^3 \end{aligned}$$

** 3. (Optional) take the log of the likelihood function

$$\begin{aligned} l(p) = ln(L(p)) &= ln[f(x_1, \ldots, x_{10}; p)] \\ &= \left(\sum_{i=1}^{10} x_i\right) ln(p) + \left(10 - \sum_{i=1}^{10} x_i\right) ln(1-p) \\ &= 7ln(p) + 3ln(1-p) \end{aligned}$$

Question: Why does taking the log help us in this case?

4. Maximize (the log of) likelihood using calculus

Using the maximum likelihood approach, we want to select the parameter with the highest likelihood. For a Bernoulli variable we can search through the space of values for p (i.e $[0, 1]$) that makes the data most probable to have been observed. In the async, Professor Laskowski demonstrated how to find the maximizing value of p using calculus.

Taking the first-order condition, we have

$$\begin{aligned}\frac{d}{dp}l(p) &= \frac{d}{dp}\left(7ln(p) + 3ln(1-p)\right)\\ &= \frac{7}{p} - \frac{3}{1-p}\end{aligned}$$

Set the slope equal to zero and solve. We obtain

$$\hat{p}_{MLE} = \frac{7}{10}$$

4. Alternately, maximize likelihood numerically

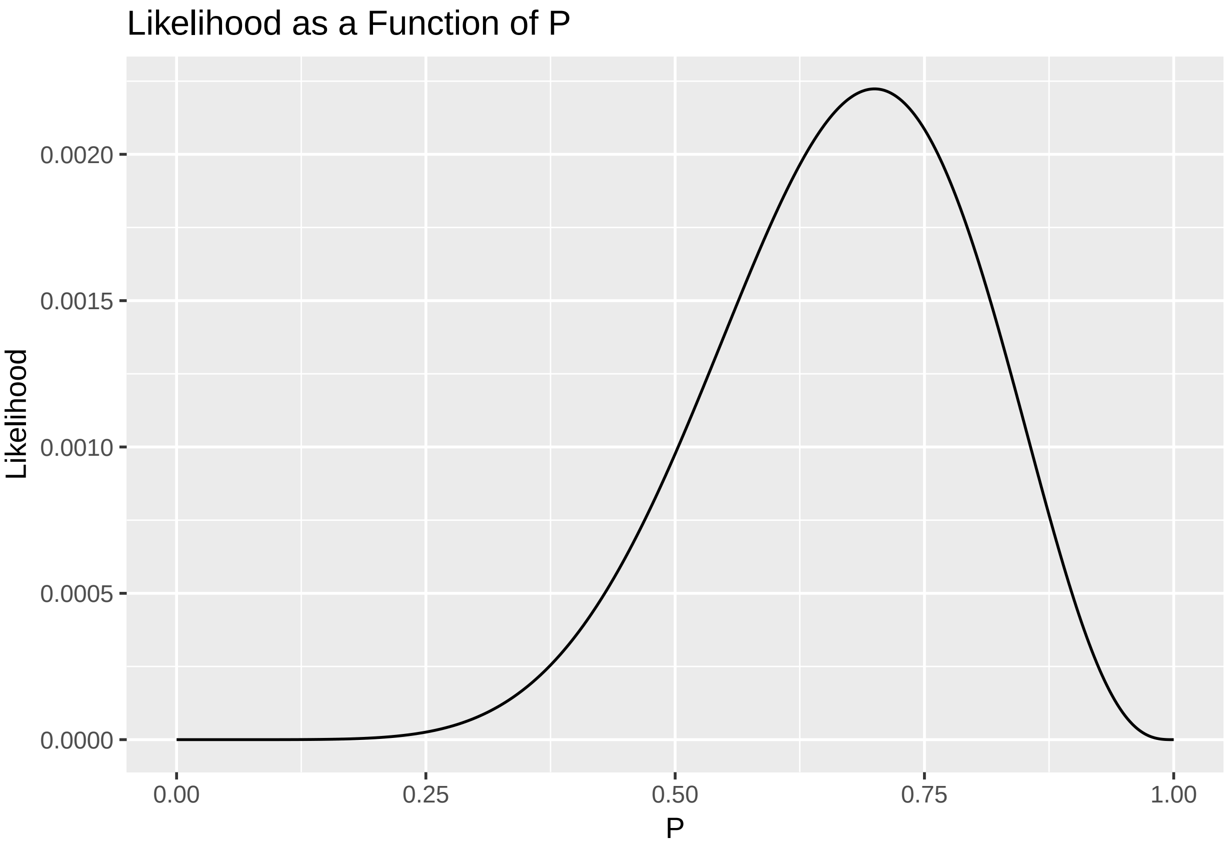
Instead of using calculus, we will find this parameter numerically. We first need to define a function that specifies the probability of our data set. You may do this using code like the following.

Let's define the function:

```r
likelihood <- function(s, p)
{
  likelihood <- 1
  for (i in 1:length(s))
  {
    if (s[i] == 1)
    {
      likelihood <- likelihood * p
    }
    else
    {
      likelihood <- likelihood * (1 - p)
    }
  }
  return(likelihood)
}

# Let's graph the likelihood function first
library(ggplot2)
s <- c(1, 0, 0, 1, 0, 1, 1, 1, 1, 1)
p <- seq(0, 1, by = 0.001)

qplot(p,
      sapply(p, function(p) {likelihood(s, p)}),
      geom = 'line',
      main = 'Likelihood as a Function of P',
      xlab = 'P',
      ylab = 'Likelihood')
```



Use R's optimization routine to find the maximum likelihood estimate of p. I use the general optimization routine *optim()* in this exericse.

```
n_event <- sum(s)
n = 10
fn <- function(p) {likelihood(s, p)}
optimize(fn, interval = c(0,1), maximum = T)
```

```
## $maximum
## [1] 0.6999843
##
## $objective
## [1] 0.002223566
```

Discussion 4: MLE for Poisson Random Variables

```
- Estimated Time: Total 30 minutes
- Breakout 15 minutes, Classwide disscussion: 15 minutes
```

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.

[Poisson Time of Arrival.]

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.

Data scientists 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

To understand a Poisson process, imagine an experiment in which you observe the arrival of cars at an intersection. Assume that the probability density that a car arrives in a differentially small interval of time is just a constant. The intersection is no more busy during the day than during the night.

Moreover, the probability density that a car arrives at a particular instant does not depend on when the previous cars arrived, not when future cars are going to arrive. Each moment of time is independent. This is an example of what we call a memory-less process.

Next, suppose we use a camera to record the intersection for a particular length of time, and we write down the number of cars that arrive in that interval. This is what we call a Poisson random variable. It has a well-known probability mass function, given by,

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

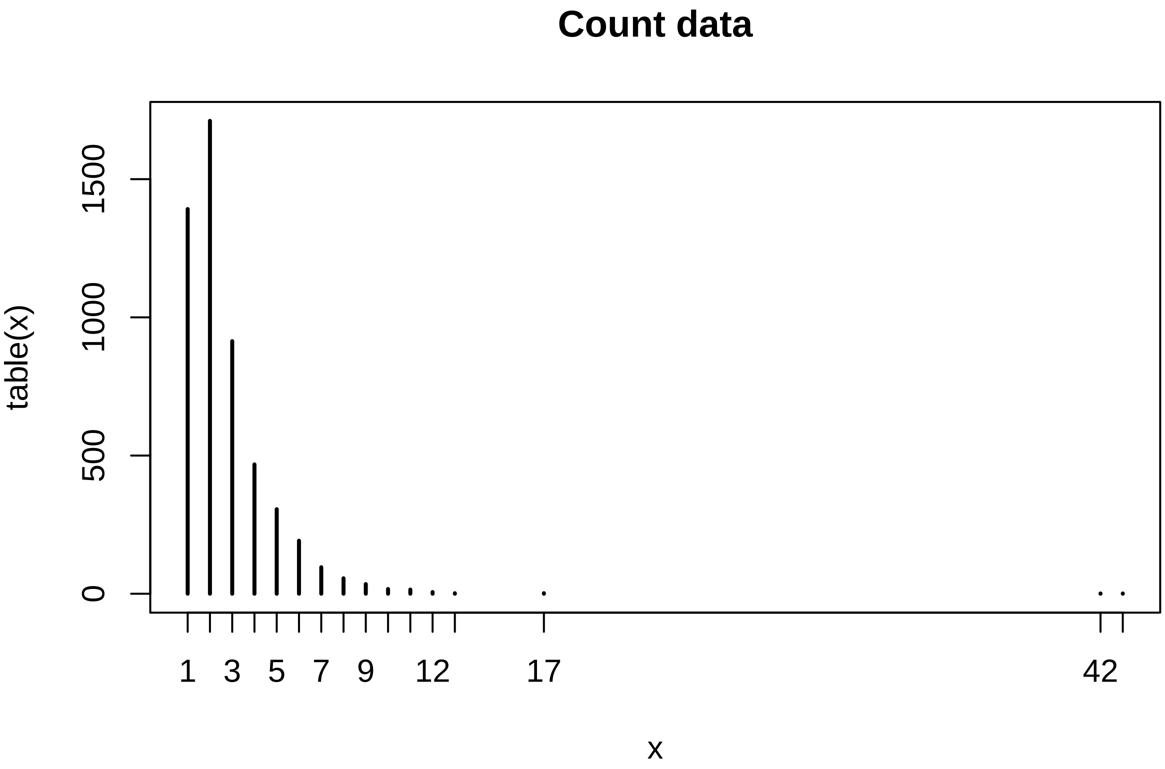
Here, λ is a parameter, which represents the mean number of cars in an interval. (You may take the expectation to check this). The following graph, "freely" borrowed from Wikipedia shows the probability mass function for different values of λ.

[Poisson Distribution]

Suppose we take a random sample, and the data appears as below.

```r
obs = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 42, 43)
freq = c(1392, 1711, 914, 468, 306, 192, 96, 56, 35, 17, 15, 6, 2, 2, 1, 1)

x <- rep(obs, freq)
plot(table(x), main="Count data")
```



In a breakout room discussion, perform a maximum likelihood estimation for the unknown parameter λ. Be sure to carefully follow all 5 steps:

1. Define your random variables.

Let each X_i for i between 1 and n be an independent Poisson variable with rate parameter λ.

2. Write down the likelihood function.

Since each X_i is independent,

$$L(\lambda) = P(X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_n = x_x|\lambda) = P(X_1 = x_1|\lambda)P(X_2 = x_2|\lambda) \cdots P(X_n = x_n|\lambda)$$

Substituting in the equation for the Poisson probability mass function,

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

3. (Optional) take the log of the likelihood function

Take the Log of the Likelihood

$$\sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!))$$

4. Maximize (the log of) likelihood using calculus

Take the 1st derivative with respect to λ and set equal to zero

$$\frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

Solve for λ

$$\hat{\lambda}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

It turns out the MLE estimate is just the average observed value. This explains why λ is called the rate parameter.

5. Maximize likelihood numerically

```r
lik.poisson <- function(x, lambda) lambda^x/factorial(x) * exp(-lambda)

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

#optim(par = 2, log.lklh.poisson, x = x)
optim(par = 2, log.lklh.poisson, x = x, method = "Brent", lower = 2, upper = 3)$par
```

```
## [1] 2.703682
```