Probability & Likelihood

October 24, 2017

1 Rules of Probability

Probabilities are defined for events. An event is some outcome that we could potentially or hypothetically observe or experience. Let event A, we will denote its probability $P(A)$ or $P(X = A)$.

- Probabilities properties :
  1. they are between zero and one, i.e., $0 \leqslant P(A) \leqslant 1$ for any event A.
  2. they add to one, i.e., if we add up the probabilities of all possible events , those probabilities must add to one. $\sum_{i=1}^{n} P(X = i) = 1$
  3. The complement of an event $A(A^C)$ means that the event doesn't happen, therefore can be defined as $P(A^C) = 1 - P(A)$
  4. The probability of 2 events A and B happening (this is an inclusive) is the probability of the union of the events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  .
  5. If a set of events A_{i} are mutually exclusive (only one can happen), then $P(\bigcup_{i=1}^{m} A_i) = \sum_{i=1}^{m} P(A_i)$.

Let A be the event that we roll a "4" on a fair six-sided die and B the event that we roll a "6" with the same die. Find the probability of the event A or event B.

```
In [1]: p_b = 1/6 # probability of event B occuring
        p_a = 1/6 # probability of event A occuring
        p_a_and_b = 0 # probability of event B AND A occuring, it is 0 because A ar
        p_a_or_b = p_a + p_b + p_a_and_b # probability of event B OR A occuring
        print(p_a_or_b)
```

```
[1] 0.3333333
```

2 Odds

Probabilities can be re-expressed in terms of odds. The odds for event A, denoted $O(A)$ is defined as $O(A) = \dfrac{P(A)}{P(A^C)}$

Suppose again that we denote rolling a "4" on a fair six-sided die as the event A.

```
In [2]: o_a = p_a/(1-p_a)
        print(o_a)

[1] 0.2
```

This can also be expressed as 1:5 (or 5:1 "odds against"). Thus, an event with probability $\frac{3}{10}$ has 3:7 odds (7:3 odds against) and an event with probability $\frac{4}{5}$ has 4:1 odds.

Note that we can also calculate probabilities from odds. If an event B has $a : b$ odds (with a > 0 and b > 0), then

$$\frac{P(B)}{1 - P(B)} = \frac{a}{b}$$
$$\Rightarrow P(B)b = a - P(B) \times a$$
$$\Rightarrow P(B) = \frac{a}{a + b}$$

.

Calculate, an event with 2:5 odds has probability.

```
In [4]: a <- 2
        b <- 5
        p_b <- a/(a + b)
        print(p_b)

[1] 0.2857143
```

3 Indicator Functions

The concept of an indicator function is a really useful one. This is a function that takes the value one if its argument is true, and the value zero if its argument is false. Sometimes these functions are called Heaviside functions or unit step functions.

Indicator functions are always first in the order of operations— if the indicator function is zero, you don't try to evaluate the rest of the expression. When taking derivatives they just go along for the ride. When taking integrals, they may affect the range over which the integral is evaluated.

4 Descriptive measures

4.1 Expectation

The expected value of a random variable X is a weighted average of values X can take, with weights given by the probabilities of those values. If X can take on only a finite number of values (say,$x_1, x_2, \ldots x_n$), we can calculate the expected value as

$$E(x) = \sum_{i=1}^{n} x_i P(x_i) = \sum_{i=1}^{n} x_i f(x_i)$$

If X is a continuous random variable with probability density function (PDF) $f(x)$, we replace the summation with an integral:

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx$$

One useful property of expected values is that they are easy to compute for linear functions of random variables. Let $Z = \alpha X + \beta Y + c$, $E(X) = \mu_X$, $E(Y) = \mu_Y$, the expected value for Z will be $E(Z) = \alpha E(X) + \beta E(Y) + c$.

It is also feasible to compute the expected value from a function of X. Let $g(X) = \frac{X+3}{2}$, the expected value of $g(X)$ will be:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{-\infty}^{\infty} \frac{x+3}{2} f(x)dx$$

Be careful, don't make the assumption $E(g(X)) = g(E(X))$.

Find the expected value of a die.

```
In [5]: x <- c(1, 2, 3, 4, 5, 6) # possible values of the die
        e_x = sum(x * (1/6))
        print(e_x)
```

```
[1] 3.5
```

4.2 Variance

The variance of random variable measures how spread out its values are. If X is a random variable with mean $E(X) = \mu$, then the variance is $E[(X-\mu)^2]$. In words, the variance is the expected value of the squared deviation of X from its mean. If X is discrete, this is calculated as:

$$Var(X) = \sum_{\chi} (\chi - \mu)^2 P(X = \chi)$$

and if X is continuous:

$$Var(X) = \int_{-\infty}^{\infty} (\chi - \mu)^2 P(X = \chi)$$

For both discrete and continuous X, a convenient formula for the variance is

$$Var(X) = E[X^2] - (E[X])^2$$

. The square root of variance is called the standard deviation. Variance has a linear property similar to expectation. Again, let X and Y be random variables with $Var(X) = \sigma_X^2$ and $Var(Y) = \sigma_Y^2$. It is also necessary to assume that X and Y are independent. Suppose we are interested in a new random variable $Z = aX + bY + c$ where a, b, and c are any real constants.

The variance of Z is then $Var(Z) = Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 0 = a^2\sigma_X^2 + b^2\sigma_Y^2$. Because c is constant, it has variance 0.

** Exercise**

Let random variable X with PDF $f(X) = X^2 I_{\{0<x<1\}}$, find its variance:

$$E(X) = \int_{-\infty}^{\infty} xx^2 I_{\{0<x<1\}} dx = \int_0^1 xx^2 dx = \int_0^1 x^3 dx = \frac{x^4}{4}\Big|_{x=0}^{x=1} = \frac{(1-0)}{4} = \frac{1}{4}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 x^2 I_{\{0<x<1\}} dx = \int_0^1 x^2 * x^2 dx = \int_0^1 x^4 dx = \frac{x^5}{5}\Big|_{x=0}^{x=1} = \frac{(1-0)}{5} = \frac{1}{5}$$

$$Var(X) = E[X^2] + (E[X])^2 = \frac{1}{5} + \frac{1}{16} = \frac{21}{80}$$

```
In [6]: var_x = sum(1/6* (x - e_x)**2 )
        print(var_x)
```

```
[1] 2.916667
```

5 Likelihood and Maximum Likelihood Estimation

5.1 Likelihood

5.1.1 Definition

Likelihood is a function of the parameters of a statistical model given data.

5.1.2 Example

Consider a hospital where 400 patients are admitted over a month for heart attacks, and a month later 72 of them have died and 328 of them have survived. We can say each patient comes from a Bernoulli distribution with unknown parameter θ. Let's estimate the unknown parameter using the data we have.

$P(Y_i = 1) = \theta$, where sucess is mortality

The probability density function for all patients can be written as

$$P(Y = y|\theta) = P(Y_1 = y_1, Y_2 = y_2 \dots Y_n = y_n|\theta)$$

$$= P(Y_1 = y_1) \dots P(Y_n = y_n|\theta)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

We can now think about this expression as a function of theta. This is a concept of a **likelihood**. So we can write it as:

$$L(\theta|y) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i},\ where\ y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$$

One way to estimate θ is that we choose the θ that gives us the largest value of the likelihood. It makes the data the most likely to occur for the particular data we observed. This is referred to as the **maximum likelihood estimate**, or MLE, maximum likelihood estimate.

5.2 Maximum Likelihood Estimate

5.2.1 Definition

MLE is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.

$$\hat{\theta} = argmax L(\theta|y)$$

In practice it is easier to maximize the natural logarithm of the likelihood, referred as **log likelihood**.

$$l(\theta) = \log(L(\theta|y))$$

We usually drop the conditional notation on y as well.

Since the logarithm is a monotone function, if we maximize the logarithm of the function, we also maximize the original function.

Why did you say that it is easier to maximize the natural logarithm?

$$\begin{aligned} l(\theta) &= \log(L(\theta|y)) \\ &= \log(\prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}) \\ &= \sum \log(\theta^{y_i}(1-\theta)^{1-y_i}) \\ &= \sum [y_i \log\theta + (1-y_i)\log(1-\theta))] \\ &= (\sum y_i)\log\theta + (\sum (1-y_i)\log(1-\theta)) \end{aligned}$$

How do we find the theta that maximizes this function?

We can maximize a function by taking the derivative and setting it equal to 0.

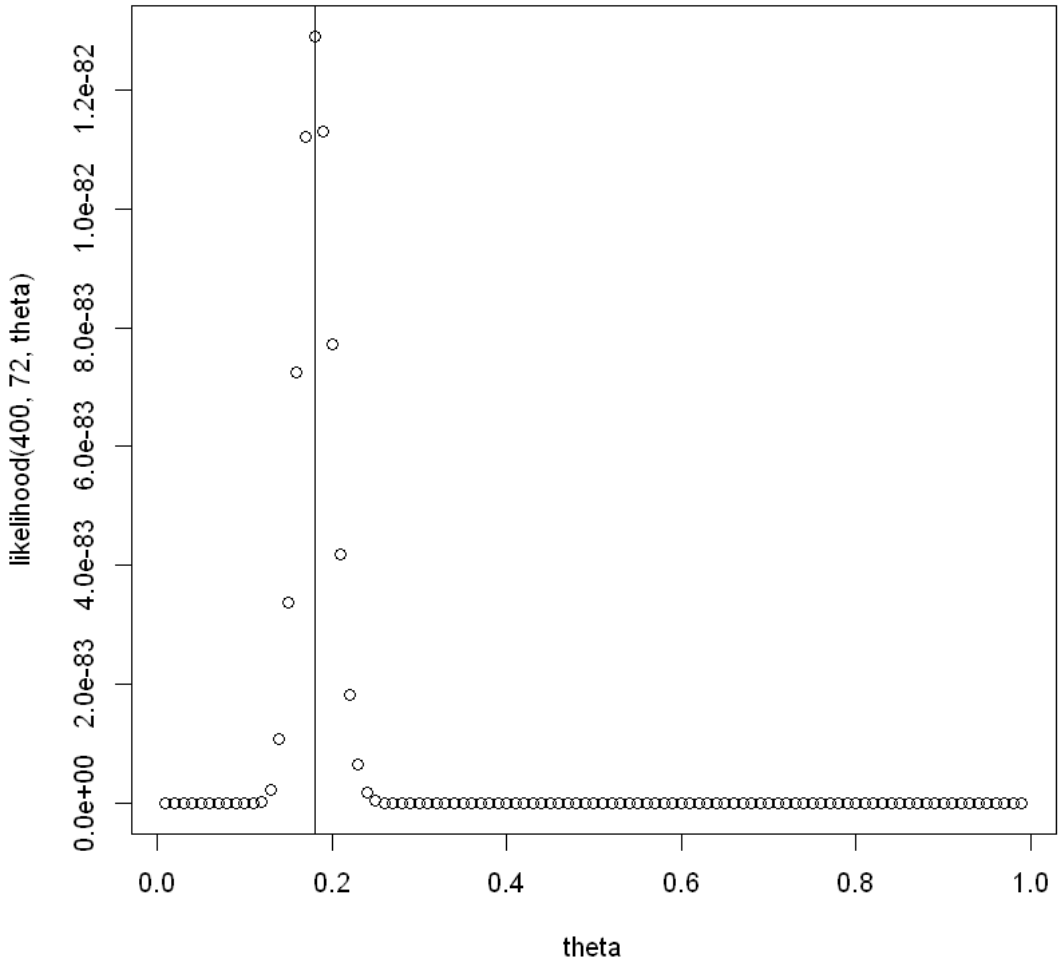
$$\begin{aligned} l'(\theta) = \frac{1}{\theta}\sum y_i - \frac{1}{1-\theta}\sum (1 - y_i = 0 \\ \Rightarrow \frac{\sum y_i}{\hat{\theta}} = \frac{\sum (1-y_i)}{1-\hat{\theta}} \\ \Rightarrow \hat{\theta} = \frac{\sum y_i}{n} \end{aligned}$$

Let's go back to our hospital example. What is the $\hat{\theta}$?

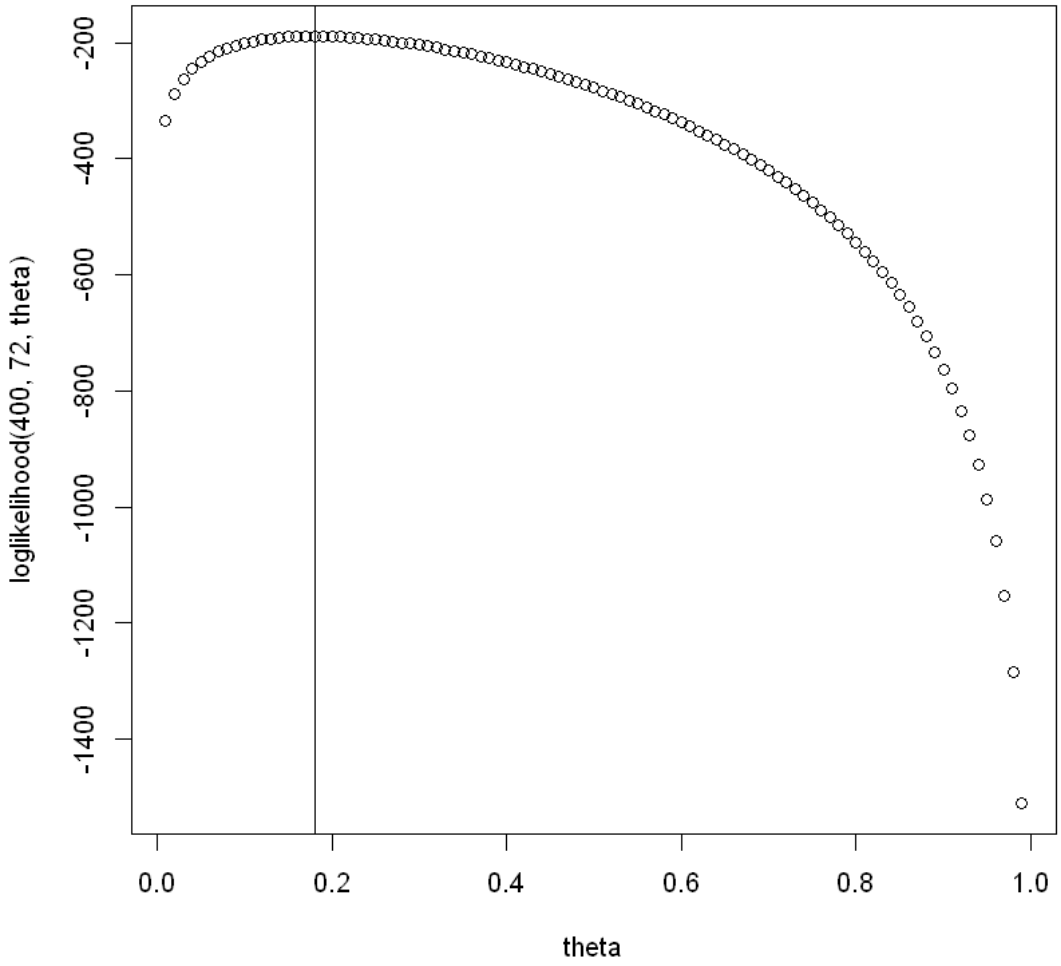
$$\hat{\theta} = \frac{72}{400} = 0.18$$

Notice that MLE have many desirable mathematical properties. They're **unbiased**, they're **consistent**, and they're **invariant**.

```
In [7]: # Let's see how the likelihood seems for the different values of theta
        likelihood <- function(n, y, theta){return(theta^y*(1-theta)^(n-y))}
        theta <- seq(from = 0.01, to = 0.99, by = 0.01)
        plot(theta, likelihood(400, 72, theta))
        abline(v = 72/400) # To make clear where is the MLE
```



In [8]:
```
# Let's take a look for the log likelihood
loglikelihood = function(n, y, theta){return(y*log(theta)+(n-y)*log(1 - the
plot(theta, loglikelihood(400, 72, theta))
abline(v = 72/400)# MLE
```


loglikelihood(400, 72, theta)
-200
-400
-600
-800
-1000
-1200
-1400
0.0
0.2
0.4
0.6
0.8
1.0
theta