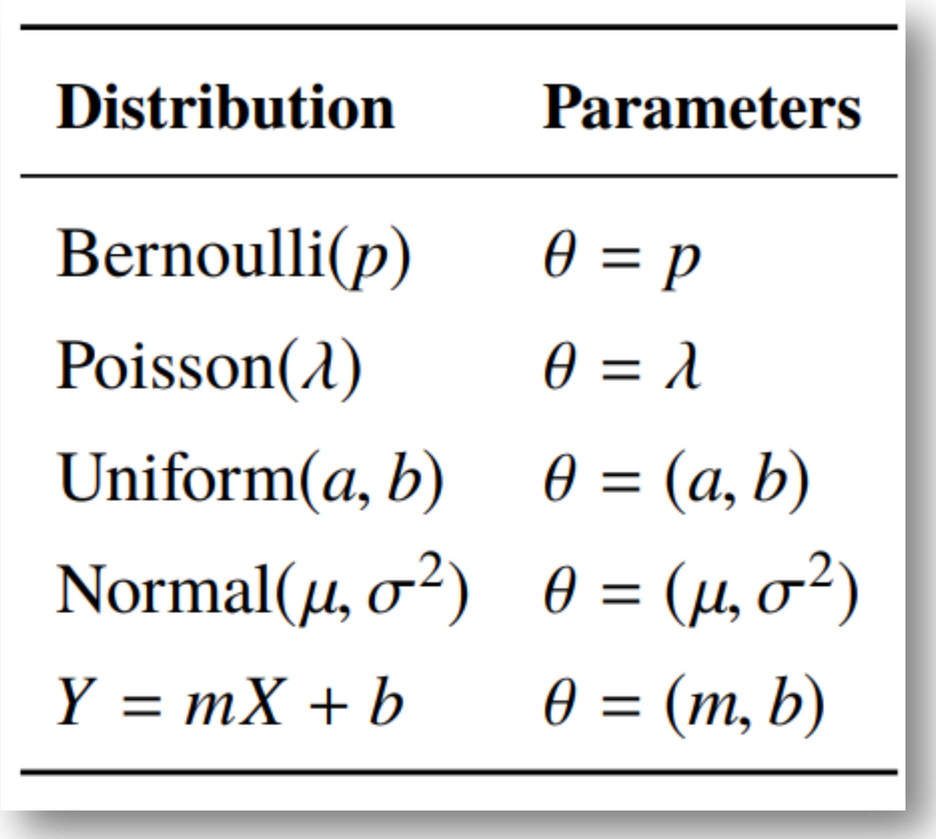
Maximum likelihood estimation

PARAMETERS



Distribution	Parameters
Bernoulli(p)	$\theta = p$
Poisson(λ)	$\theta = \lambda$
Uniform(a, b)	$\theta = (a, b)$
Normal(μ, σ^2)	$\theta = (\mu, \sigma^2)$
$Y = mX + b$	$\theta = (m, b)$

In the real world often you don’t know the “true” parameters, but you get to observe data.

Point estimation-Maximum likelihood estimation

- The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.
- Raised by Gaussian (in 1821), development by Fisher (in 1912).

Probability & Likelihood (function)

Probability

$$P(X;\theta)$$

=

likelihood

$$L(\theta;X)$$

- **Probability** refers to the chance that a particular outcome occurs (X) based on the values of parameters (θ) in a model.

- **Likelihood** refers to how well a sample provides support for particular values of a parameter in a model, is the probability that we see the data we see if we set the parameter equal to θ.

What does likelihood mean and how is "likelihood" different than "probability"? In the case of discrete distributions, likelihood is a synonym for the joint probability of your data. In the case of continuous distribution, likelihood refers to the joint probability density of your data.

EXAMPLE 1:
TURTLE & RABBIT

The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.

A **race** will be held between a **rabbit** and a **turtle**

Guess who will win?

Rabbit

Because the rabbit runs faster than the turtle

EXAMPLE 2:

LU HAN & GUAN XIAOTONG

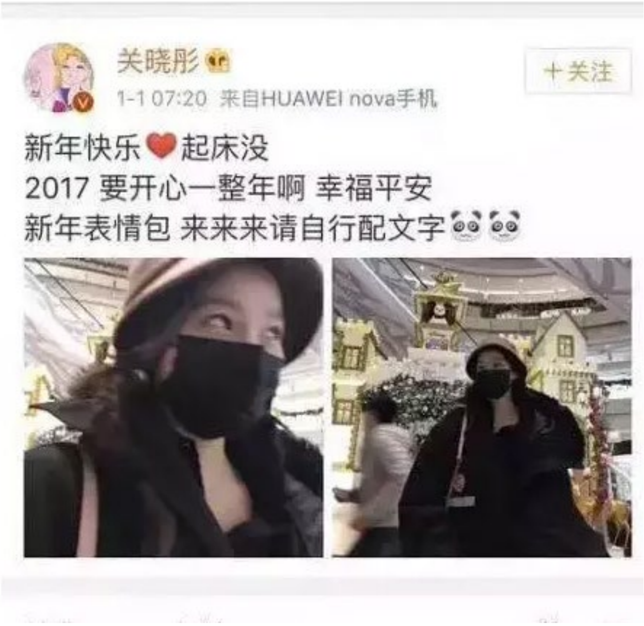
The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.

Observed (X): The same bracelet/ hoodie

Relationship (θ): intimacy

EXAMPLE 2:
LU HAN & GUAN XIAOTONG

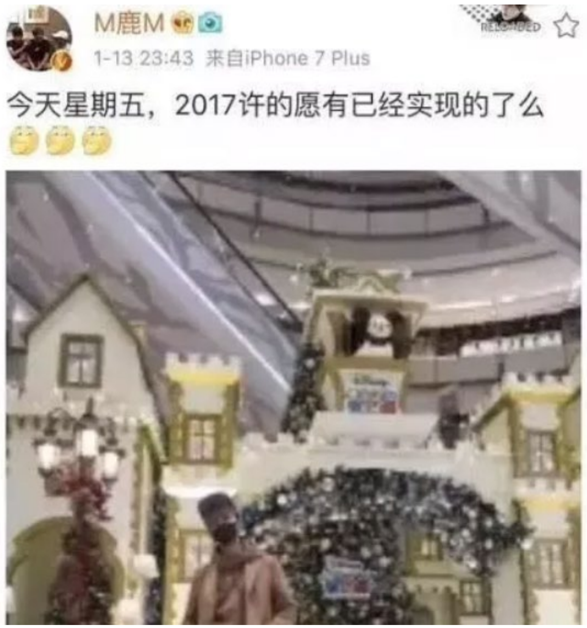
The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.



关晓彤

1-1 07:20 来自HUAWEI nova手机

新年快乐❤起床没
2017 要开心一整年啊 幸福平安
新年表情包 来来来请自行配文字🐼🐼



M鹿M

1-13 23:43 来自iPhone 7 Plus

今天星期五，2017许的愿有已经实现的了么
😳😳😳

Observed (X): vacationed together in new year’s eve

Relationship (θ): dating

EXAMPLE 2:
LU HAN & GUAN XIAOTONG

The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.

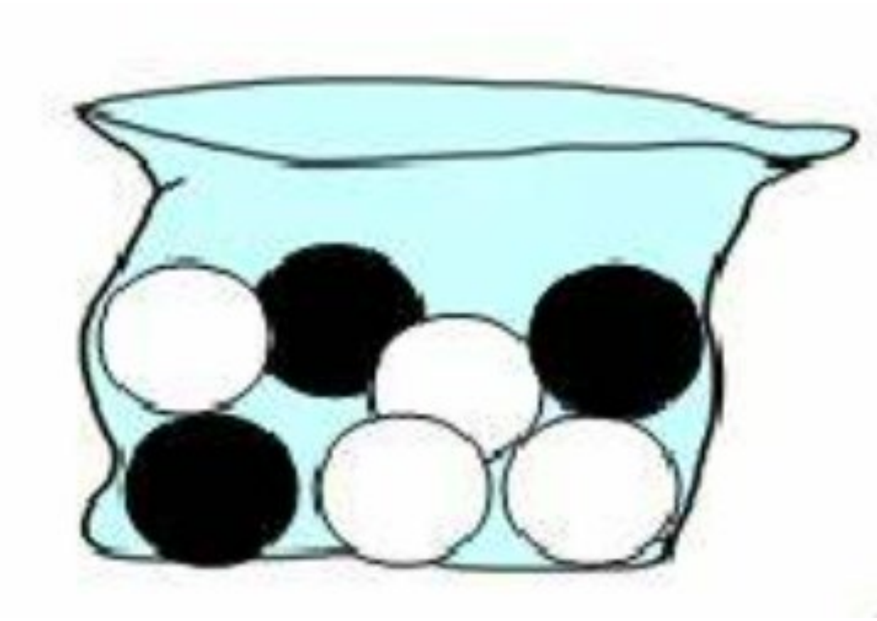
Observed (X): first on-screen kiss

Guan said her first kiss on the big screen would give to her boyfriend.

Relationship (θ): couple in a romantic relationship

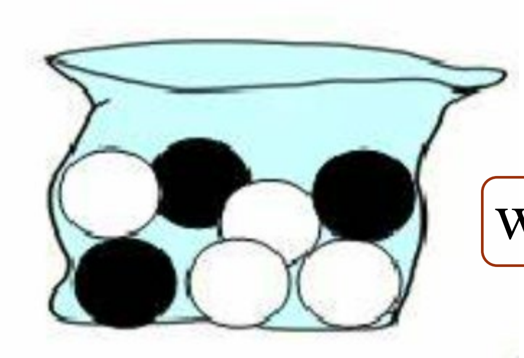
Finally, in 2017, Luhan revealed on his Weibo account that Guan Xiaotong is his girlfriend, saying "Hello everyone, I am introducing my girlfriend," and tagged Guan's account. The female celebrity replied to his post with a heart and said, "Aiya, so awesome!"

Example 3



A box contains several white and black balls, we don’t know their proportions.
Suppose I picked one ball from the box, recorded its color and then put it back to the box.
Eventually, I picked 10 balls and observed 6 white balls.
What is the proportion of white ball (θ) in the box?

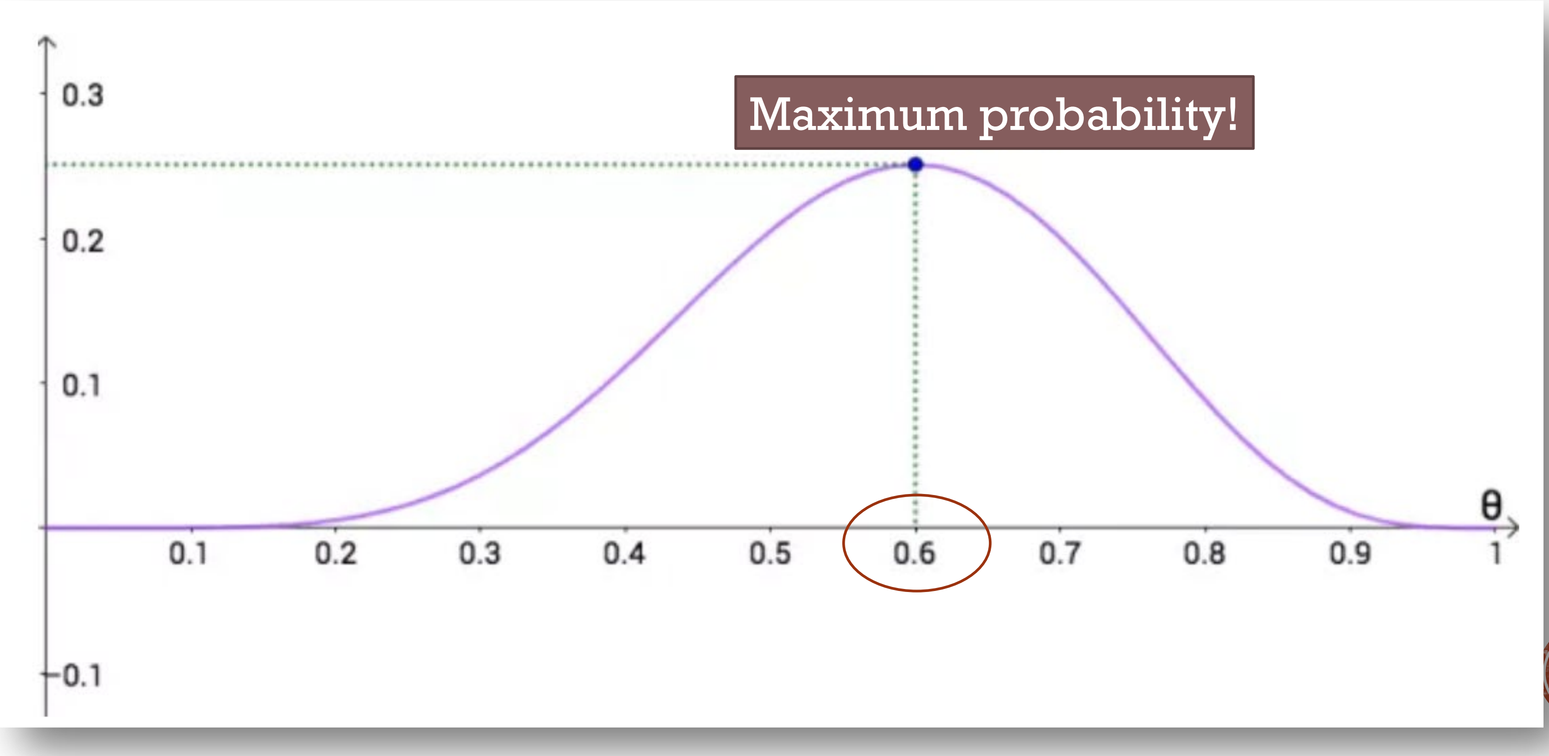
Example 3



white colour of ball X~B(10,θ)

- Suppose θ=0.5, then P(6 of 10) =$C_{10}^{6} \times 0.5^{6} \times (1-0.5)^{4}$ =0.21
- Suppose θ=0.6, then P(6 of 10) =$C_{10}^{6} \times 0.6^{6} \times (1-0.6)^{4}$ =0.25

$$\frac{0.25}{0.21}=1.2>1$$



Maximum probability!

The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.

MLE

- 1. Likelihood（ function ）
- Suppose we have a countinous random sample $X_1, X_2, X_3, \dots. X_n$ for which the probability density (or mass) function of each Xi is $f(x_i; \theta)$. Since we assumed each data point is independent, the likelihood of all our data is the product of the likelihood of each data point, then the joint probability mass (or density) function of $X_1, X_2, X_3, \dots. X_n$, which we'll call $L(\theta)$ is:

$$L(\theta)=P(X_1=x_1, X_2= x_2, \dots, X_n= x_n)$$

$$= f(x_1; \theta) * f(x_2; \theta) \dots . f(x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i; \theta)$$

To simplify notation, let the vector $\mathbf{X}= (x_1, x_2, \dots. x_n)$ denote the observed sample. Then the joint pdf and likelihood function may be expressed as $f(\mathbf{X}; \theta)$ and $L(\theta;\mathbf{X})$, respectively

MLE

- 1. Likelihood (function)
- If $X_1, X_2, X_3, \dots .X_n$ is discrete random variable with $P(x_i; \theta)$. Since we assumed each data point is independent, then $L(\theta)$ is:

$$L(\theta)=P(X_1=x_1, X_2= x_2, \dots, X_n= x_n)$$

$$= P(x_1; \theta)* P(x_2; \theta) \dots P(x_n; \theta)$$

$$= \prod_{i=1}^{n} P(x_i; \theta)$$

MLE

- 2. Maximization
- In MLE our goal is to chose values of our parameters (θ) that maximizes the likelihood function. We are going to use the notation $\hat{\theta}$ to represent the best choice of values for our parameters. Formally, MLE assumes that:
- $$\hat{\theta} = \underset{\theta}{\arg\max}\, L(\theta)$$

If we find the arg max of the log of likelihood, it will be equal to the arg max of the likelihood. Therefore, for MLE, we first write the log likelihood function (LL)

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

MLE

- 3. taking the derivative of the log-likelihood, and setting it to 0

$$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

Solution for example 3

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

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

X_i =0 or 1, 0<p<1

Therefore, the likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{x1}(1-p)^{1-x1} \times p^{x2}(1-p)^{1-x2} \times \cdots \times p^{xn}(1-p)^{1-xn}$$

Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum_{i=1}^{n} xi}(1-p)^{n-\sum_{i=1}^{n} xi}$$

Solution for example 3

In this case, the natural logarithm of the likelihood function is:

$$LL(p) = (\sum_{i=1}^{n} x_i) \ln p + (n - \sum_{i=1}^{n} x_i) \ln(1-p)$$

Now, taking the derivative of the log-likelihood, and setting it to 0, we get:

$$\frac{\partial LL(p)}{\partial p} = \frac{(\sum_{i=1}^{n} x_i)}{p} - \frac{(n - \sum_{i=1}^{n} x_i)}{1-p} = 0$$

multiplying through by p(1−p), we get:

$$(\sum_{i=1}^{n} x_i)(1-p) - (n - \sum_{i=1}^{n} x_i)p = 0$$

$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - np + p\sum_{i=1}^{n} x_i = 0$$

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

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

EXAMPLE 4. NORMAL MLE ESTIMATION

- Let $X_1, X_2, X_3, \ldots . X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

Solution

In finding the estimators, the first thing we'll do is write the probability density function (to make it simpler, here $\hat{\mu}$ *is wirte as* μ, and $\widehat{\sigma^2}$ is write as σ^2).

$$f(x_i;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

Now, that makes the likelihood function:

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2})$$

and therefore the log of the likelihood function:

$$LL(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$LL(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

EXAMPLE 4. NORMAL MLE ESTIMATION

Solution

Now, upon taking the partial derivative of the log likelihood with respect to μ, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial LL(\mu,\sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

Denominator σ^2 can’t be zero, then

$$\sum_{i=1}^{n}(x_i-\mu)^2 = 0, \sum_{i=1}^{n}x_i - n\mu = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n}x_i}{n} = \overline{x}$$

$$LL(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

EXAMPLE 4. NORMAL MLE ESTIMATION

Solution

Now, for σ^2 . Taking the partial derivative of the log likelihood with respect to σ^2, and setting to 0, we get :

$$\frac{\partial LL(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

Consider σ^2 as a whole

Multiplying through by $2\sigma^4$

$$\frac{\partial LL(\mu,\sigma^2)}{\partial \sigma^2} = \left[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0\right]\times 2\sigma^4$$

We get

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

And, solving for σ^2, and putting on its hat, we have shown that the maximum likelihood estimate of σ^2 is (we already known that $\hat{\mu}=\bar{x}$):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$