Introduction to Data Science With Probability and Statistics

Lecture 25: Maximum Likelihood

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

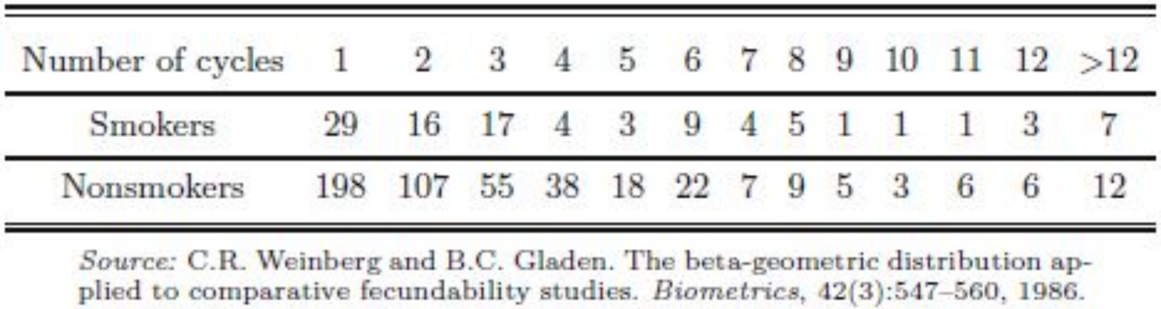
Maximum likelihood

- In previous lectures, we have seen that we could construct estimators for various parameters of interest.

- These parameters had a natural sample analogue: expectation versus sample mean, probabilities versus relative frequencies, etc.

- In some situations such analogues doesn’t exist, then, we need to use a general principle to construct estimators, which is called the maximum likelihood principle.

Why a general principle?



Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

parameter 'p' : probability of becoming pregnant after cycle.

Estimator 1 : $S = \dfrac{\# X_i = 1}{n}$

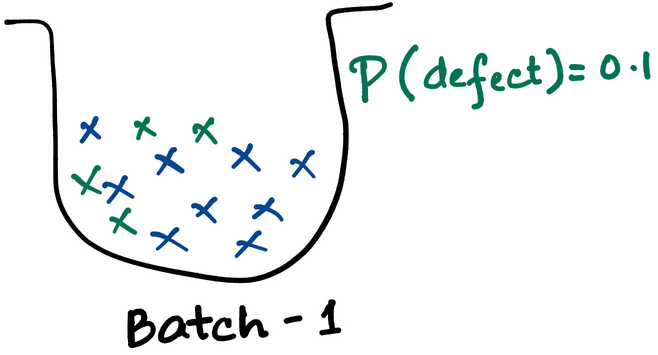
Estimator 2 : $T = \dfrac{1}{\bar{x}_n}$

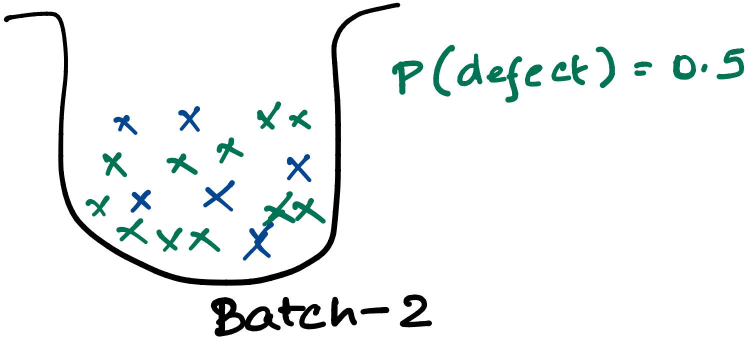
$$\left\{ \begin{array}{l} X_i \sim Geo(p) \\ \therefore \underbrace{E[X_i]}_{\substack{\bar{x}_n \\ \text{estimator}}} = 1/p \end{array} \right.$$

The maximum likelihood principle

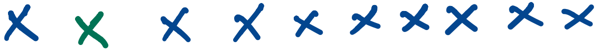
- Thought Experiment :- A dealer of computer chips is offered 2 batches of 10000 chips each. According to the seller, in one batch about 50% of the chips are defective, while this percentage is about 10% in the other batch. Our dealer is only interested in this last batch. Unfortunately the seller cannot tell the two batches apart. To help him to make up his mind, the seller offers our dealer one batch, from which he is allowed to select and test 10 chips. After selecting 10 chips arbitrarily, it turns out that only the second one is defective. Our dealer at once decides to buy this batch. Is this a **wise** decision?

The maximum likelihood principle (Intuition)





Sample drawn of 10 chips :



which batch is more likely?

The maximum likelihood principle

Let the chips be $R_1, R_2, \ldots, R_{10}$. where $R_i \sim Ber(p)$

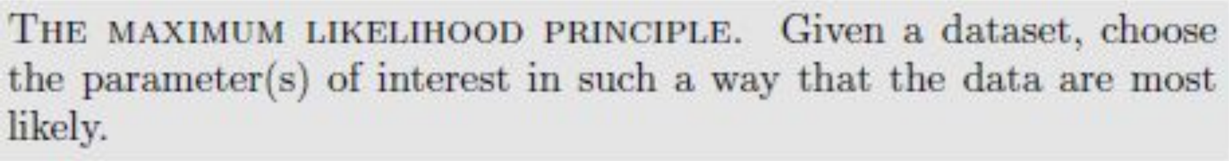
Prob. of observed data = $\mathbb{P}(R_1=0, R_2=1, R_3=0, \ldots R_{10}=0) = \underline{p(1-p)^9}$

for Batch 1 : $P(data) = (0.1)(0.9)^9 = 0.03$

for Batch 2 : $P(data) = (0.5)(0.5)^9 = 0.0009$

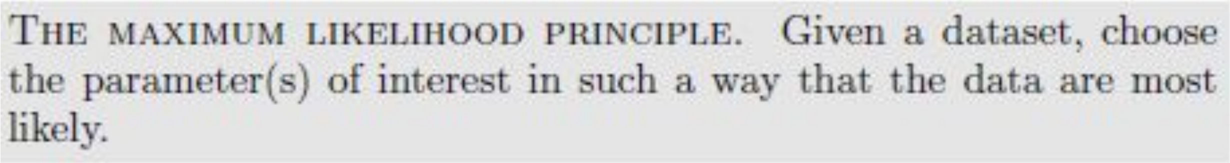
Batch 1 is 40 times more likely, than Batch-2

The maximum likelihood principle



THE MAXIMUM LIKELIHOOD PRINCIPLE. Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

The maximum likelihood principle



THE MAXIMUM LIKELIHOOD PRINCIPLE. Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

Quiz: Which batch should the dealer choose if only the first three chips are defective?

The maximum likelihood principle

$$\mathbb{P}(\text{data}) = p^3(1-p)^7.$$

Batch-1: $(0.1)^3(0.9)^7 = 0.00048$

Batch-2: $(0.5)^3(0.5)^7 = 0.00098$

Batch-2 twice as more likely than Batch-1.

The maximum likelihood principle

Returning to the example of the number of cycles up to pregnancy,

$$X_i \sim Geo(p) \quad \Rightarrow \quad P(X_i = k) = (1-p)^{k-1} p.$$

$$\& \quad P(X_i > 12) = (1-p)^{12}$$

Total no. of ways to acheive this configuration.

Then

$$L(p) = C \cdot P(X_i = 1)^{29} \cdot P(X_i = 2)^{16} \cdots P(X_i > 12)^{7}$$

$$= C \cdot p^{29} \cdot ((1-p)p)^{16} \cdots ((1-p)^{12})^{7}$$

$$= C \cdot p^{93} \cdot (1-p)^{322}$$

likelihood function

The maximum likelihood principle

Returning to the example of the number of cycles up to pregnancy,

$$L'(p) = c\left[93p^{92}(1-p)^{322} - 322p^{93}(1-p)^{321}\right]$$

$$= c\cdot p^{92}(1-p)^{321}\cdot(93-415p)$$

So, for $L'(p)=0$

we get $p=0$, $p=1$, $p = \underbrace{93/415 = 0.224}_{\text{max. likelihood estimate of } p.}$

Likelihood and log-likelihood

Suppose we have a dataset $x_1, x_2, \ldots, x_n$, modeled as a realization of a random sample from a distribution characterized by a parameter θ.

We write probability mass function as $p_\theta(x)$ and probability density function as $f_\theta(x)$.

Likelihood and log-likelihood

For a dataset $x_1, x_2, \ldots, x_n$ modeled as the realization of a random sample $X_1, \ldots, X_n$ from a ***discrete*** distribution, the *likelihood function* is:

$$L(\theta) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = p_\theta(x_1) \cdots p_\theta(x_n)$$

The maximum likelihood estimate of θ is the value for which the likelihood function is maximal.

Likelihood and log-likelihood

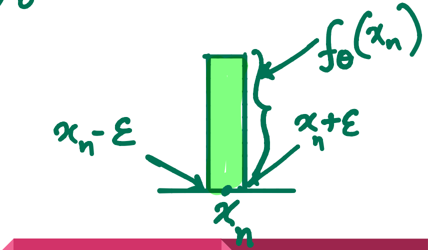
For continuous distribution, we choose θ in such a way that the below probability is maximal (where, $\epsilon > 0$)

$$P(x_1 - \varepsilon \leq X_1 \leq x_1 + \varepsilon, \ldots, x_n - \varepsilon \leq X_n \leq x_n + \varepsilon)$$

Since the X_i are independent, we find that:

$$\begin{aligned} &P(x_1 - \varepsilon \leq X_1 \leq x_1 + \varepsilon, \ldots, x_n - \varepsilon \leq X_n \leq x_n + \varepsilon) \\ &\quad = P(x_1 - \varepsilon \leq X_1 \leq x_1 + \varepsilon) \cdots P(x_n - \varepsilon \leq X_n \leq x_n + \varepsilon) \\ &\quad \approx f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) (2\varepsilon)^n, \end{aligned}$$

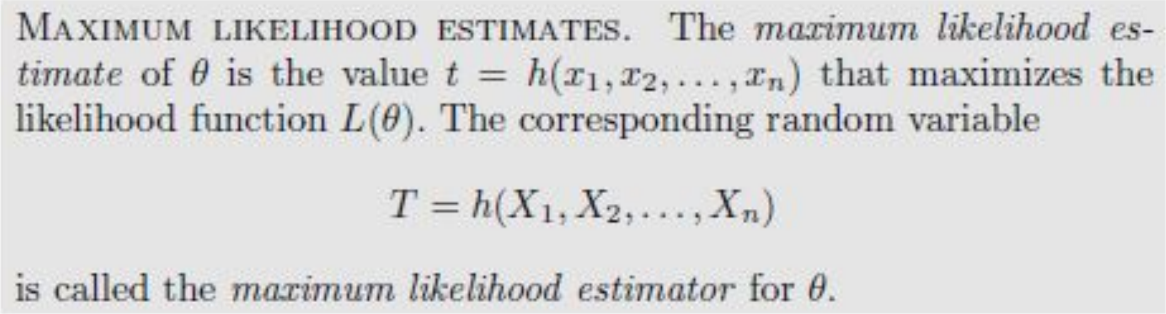
$$\int_{x_n - \varepsilon}^{x_n + \varepsilon} f_\theta(x_n)\, dx \approx 2\varepsilon f_\theta(x_n)$$



Hence, for *continuous* distribution, the *likelihood* function can be written as:

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$$

Likelihood and loglikelihood



Maximum likelihood estimates. The *maximum likelihood estimate* of θ is the value $t = h(x_1, x_2, \ldots, x_n)$ that maximizes the likelihood function $L(\theta)$. The corresponding random variable

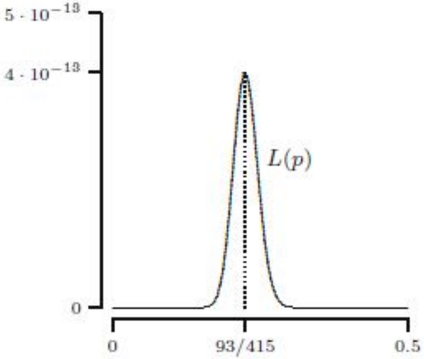
$$T = h(X_1, X_2, \ldots, X_n)$$

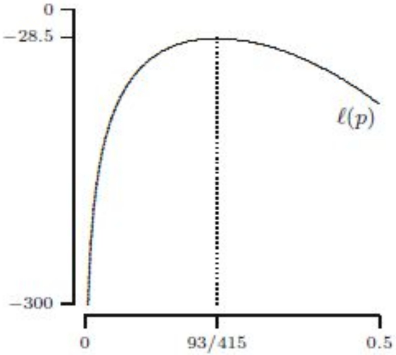
is called the *maximum likelihood estimator* for θ.

Log-likelihood

- ➔ We saw that it was easy to find the value of the parameter for which the likelihood is maximal.
- ➔ Usually one can find the maximum by differentiating the likelihood function *L(θ)*.
- ➔ To differentiate *L(θ)* we have to apply the product rule from calculus.
- ➔ The logarithm of L(θ) changes the product of the terms involving θ into a sum of logarithms of these terms, which makes the process of differentiating easier.
- ➔ Hence, this is the importance of *loglikelihood function* which is given as:

$$\ell(\theta) = \ln(L(\theta))$$





Properties of maximum likelihood estimators

Invariance principle: In general the principle says that if *T* is the maximum likelihood estimator of a parameter *θ* and *g(θ)* is an invertible function of *θ*, then *g(T)* is the maximum likelihood estimator for *g(θ)*.

Eg. of invertible func. :

$$y = f(x) = 2x + 3$$

$$x = g(y) = \frac{y-3}{2}$$

Properties of maximum likelihood estimators

Asymptotic unbiasedness: Under mild conditions on the distribution of the random variables X_i under consideration, maximum likelihood estimators are unbiased. By this we mean that if $T_n = h(X_1, X_2, \ldots, X_n)$ is the maximum likelihood estimator for a parameter θ, then:

$$\lim_{n \to \infty} \mathrm{E}[T_n] = \theta.$$

Eg: sample variance of $N(\mu, \sigma^2)$ is estimated as $D_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ by MLE. $E[D_n^2] = E\left[\frac{n-1}{n} S_n^2\right] = \frac{n-1}{n}\sigma^2$

$$\lim_{n \to \infty} \frac{n-1}{n}\sigma^2 = \lim_{n \to \infty}\Big(1 - \underbrace{\frac{1}{n}}_{0}\Big)\sigma^2 = \sigma^2$$

Properties of maximum likelihood estimators

Asymptotic minimum variance: The variance of an unbiased estimator for a parameter θ is always larger than or equal to a certain positive number, known as the *Cramer-Rao lower bound*.

- Again under mild conditions one can show that maximum likelihood estimators have asymptotically the smallest variance among unbiased estimators.

- That is, asymptotically the variance of the maximum likelihood estimator for a parameter θ attains the *Cramer-Rao lower bound*.

$$\mathrm{Var}(T) \geq \frac{1}{n \cdot E\left[\left(\frac{\partial}{\partial \theta} \ln f_\theta(x)\right)^2\right]}$$

Maximum Likelihood Estimators

Consider the following situation. Suppose we have two fair dice, D_1 with 5 red sides and 1 white side and D_2 with 1 red side and 5 white sides. We pick one of the dice randomly, and throw it repeatedly until *red* comes up for the first time. With the same die this experiment is repeated two more times. Suppose the following happens:

First experiment: first red appears in 3rd throw

Second experiment: first red appears in 5th throw

Third experiment: first red appears in 4th throw.

Show that for die D_1 this happens with probability $5.7424 \cdot 10^{-8}$, and for die D_2 the probability with which this happens is $8.9725 \cdot 10^{-4}$. Given these probabilities, which die do you think we picked?

OR. find the relation between the likehoods of D_1 & D_2

Since, indiv. experiments $X_i \sim Geo(p)$.

$$L(p) = P(X_1=3, X_2=5, X_3=4) = (1-p)^2 p \cdot (1-p)^4 p \cdot (1-p)^3 p \cdot$$

$$= p^3 (1-p)^9.$$

for D_1 : $p = 5/6$ $\therefore$ $L(5/6) = \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right)^9$

for D_2 : $p = 1/6$ $\therefore$ $L(1/6) = \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^9 = 5^6 \cdot L(5/6)$

$$\underbrace{L(1/6)}_{D_2} = 5^6 \cdot \underbrace{L(5/6)}_{D_1}$$

we picked D_2

Maximum Likelihood Estimators

We throw an unfair coin repeatedly until heads comes up for the first time. We repeat this experiment three times (with the same coin) and obtain the following data:

First experiment: heads first comes up in 3rd throw

Second experiment: heads first comes up in 5th throw

Third experiment: heads first comes up in 4th throw.

Let p be the probability that heads comes up in a throw with this coin. Determine the maximum likelihood estimate $\hat{p}$ of p.

$L(p) = p^3(1-p)^9$ from last problem.

$\ell(p) = 3\ln p + 9\ln(1-p)$ {optional}

$\ell'(p) = \frac{3}{p} - \frac{9}{1-p} = 0$

$\Rightarrow 3 - 3p - 9p = 0$

$\Rightarrow p = 1/4 = \hat{p}$

Maximum Likelihood Estimators

Let $x_1, x_2, \ldots, x_n$ be a dataset that is a realization of a random sample from a distribution with probability density $f_\delta(x)$ given by:

$$f_\delta(x) = \begin{cases} \mathrm{e}^{-(x-\delta)} & \text{for } x \geq \delta \\ 0 & \text{for } x < \delta. \end{cases}$$

a. Draw the likelihood $L(\delta)$.

b. Determine the maximum likelihood estimate for δ.