

MAXIMUM LIKELIHOOD ESTIMATION

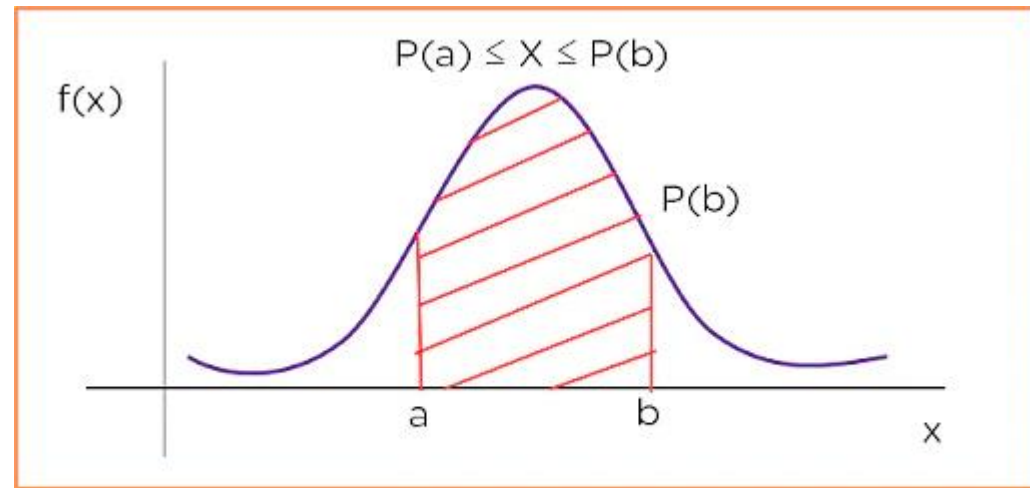
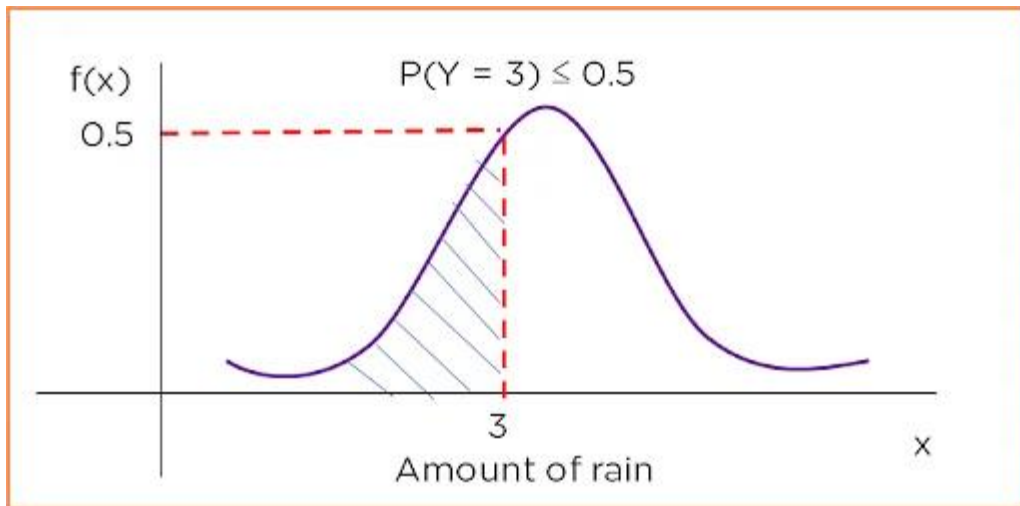
Probability vs Likelihood

- Probability talks about the outcomes(data)/hypothesis, while likelihood talks about the model/evidence.

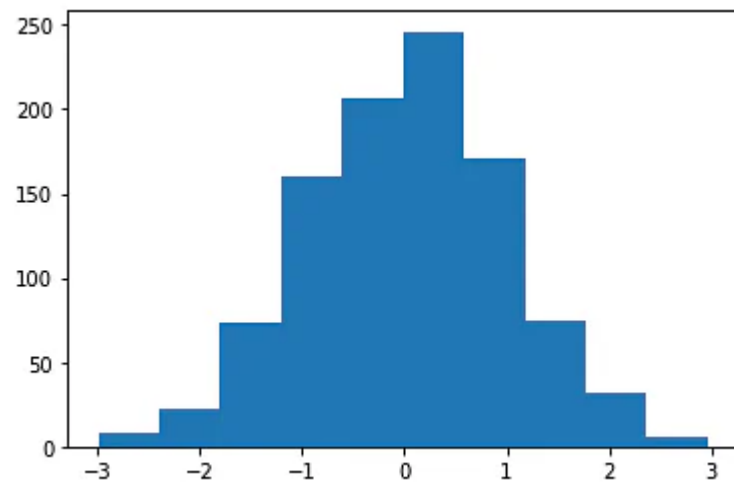
S. No	Likelihood	Probability
1	Refers to the past events with known outcomes	Refers to the occurrence of future events
2	I flipped a coin 10 times and obtained 10 heads. What is the likelihood that the coin is fair? Given the fixed outcomes (data), what is the likelihood of different parameter values?	I flipped a coin 10 times. What is the probability of it landing heads or tails every time? Given the fixed parameter($p=0.5$). What is the probability of different outcomes?
3	Likelihoods doesn't add up to 1	Probabilities add up to 1

PDF (Probability Density Function)

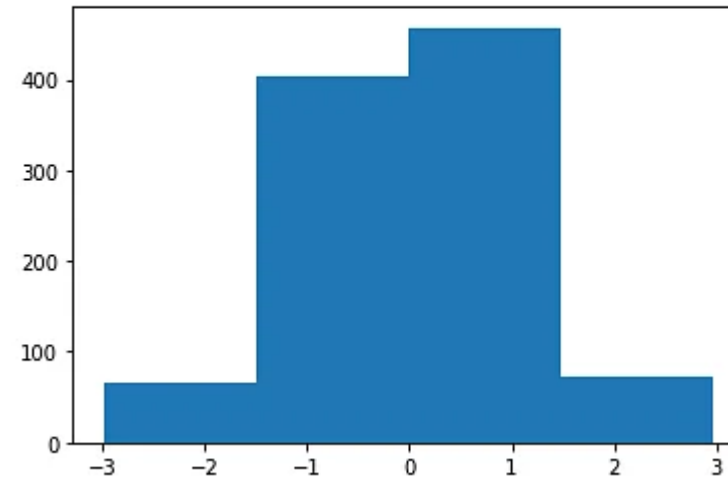
- A function that defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function, is called a Probability Density Function (PDF) in statistics.



Probability Density Function

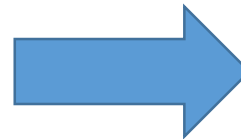


```
pyplot.hist(sample, bins=4)
pyplot.show()
```

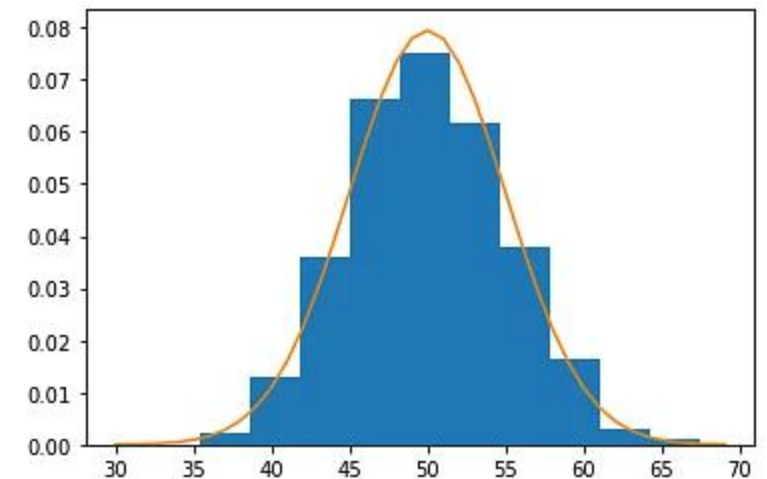


```
# example of parametric probability density estimation
# generate a sample
sample = normal(loc=50, scale=5, size=1000)
sample
```

```
array([48.84805937, 49.64182645, 47.03991025, 46.52308008, 48.7005173 ,
        61.13173104, 54.53233533, 51.40918539, 46.5197703 , 52.10736238,
        57.81844529, 53.77609112, 41.73680375, 53.96261695, 46.71511335,
        49.15080523, 39.83518599, 56.88098253, 50.33099844, 48.14062688,
        51.94422316, 45.54845016, 48.5271123 , 40.14095103, 39.20290513,
        52.11350203, 45.6861659 , 54.3283532 , 51.42856746, 42.80210139,
        48.95240147, 45.39953248, 53.33755656, 54.75719381, 51.07402736,
        64.49656413, 47.12326034, 37.926384 , 58.77177249, 47.12235208,
        57.42906738, 42.83104695, 42.08103957, 45.47214878, 44.01347927,
        53.77143945, 47.84426063, 47.67256084, 53.50008046, 50.60489559,
        47.95142504, 39.90253028, 50.79473712, 54.48727795, 46.68624492,
        59.24840421, 48.59579162, 41.02895684, 46.00291122, 53.09953503,
        55.67726872, 50.42055556, 52.0544431 , 46.7010282 , 44.22272573,
        48.52491116, 44.75760767, 50.06398808, 55.18587567, 49.79294363,
        42.83499723, 49.99654463, 42.91646387, 46.33508915, 60.33454846,
        62.35638223, 51.37871745, 46.10206571, 44.6375268 , 49.05651453,
        55.443026 , 52.78845979, 46.65807886, 46.40127796, 51.46406381,
        52.56133546, 50.41679239, 49.46745663, 51.72878202, 54.73159091,
        56.4692516 , 42.73329658, 42.38763846, 52.40598286, 48.22920091,
```



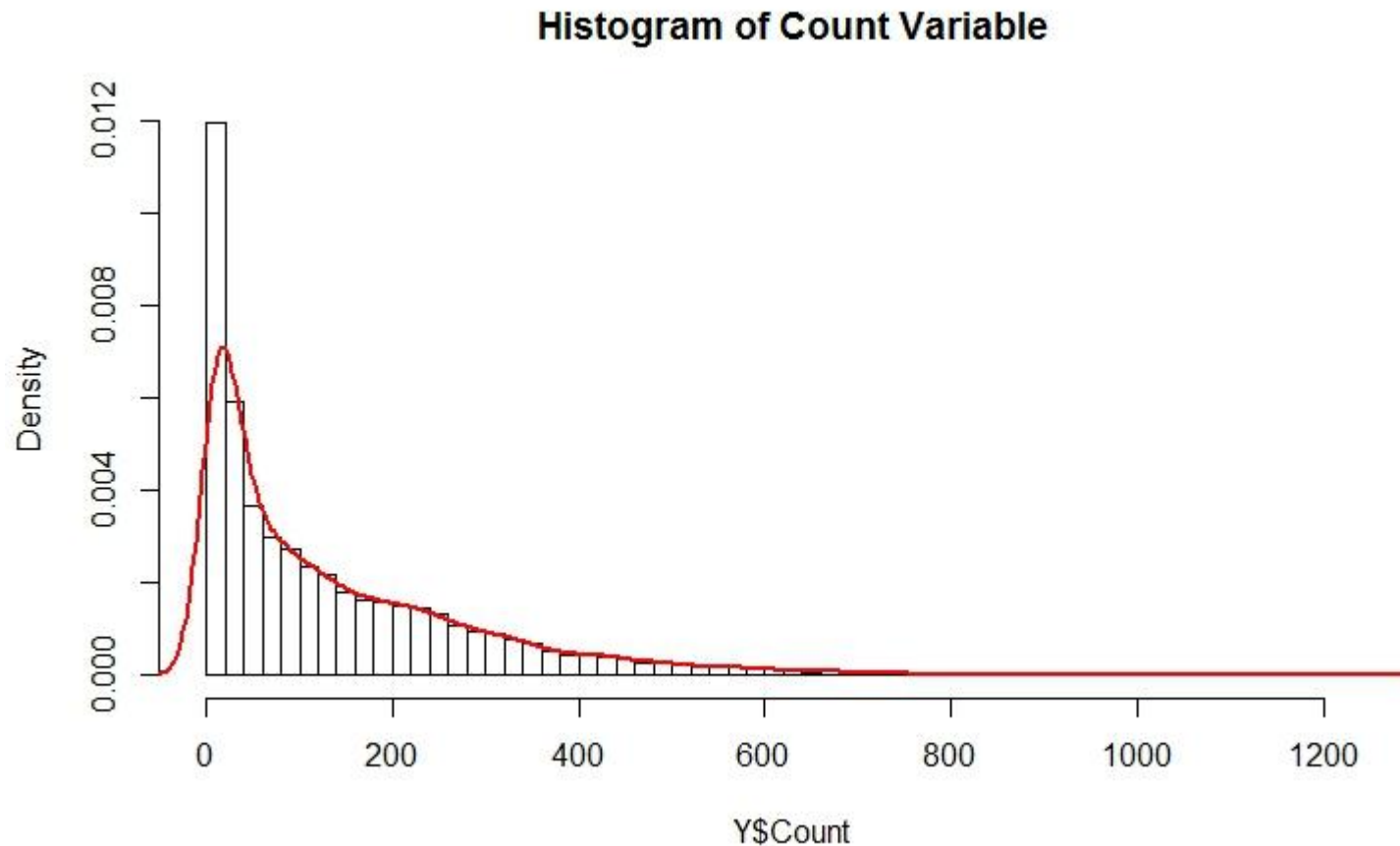
```
# plot the histogram and pdf
pyplot.hist(sample, bins=10, density=True)
pyplot.plot(values, probabilities)
pyplot.show()
```



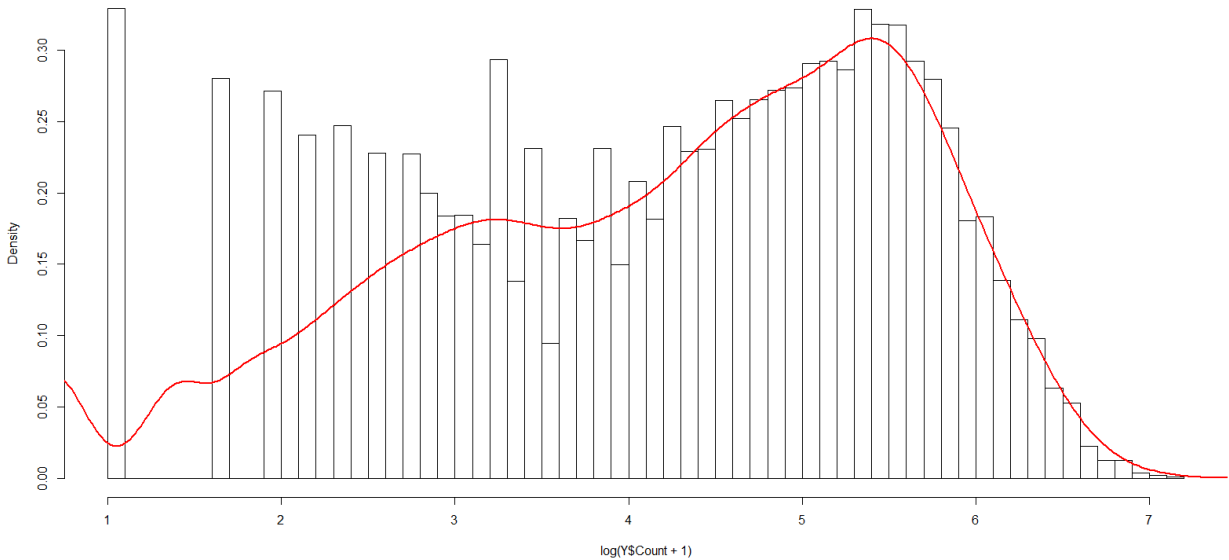
PDF-Issues

- A common modeling problem involves how to estimate a joint probability distribution for a dataset.
- How do you choose the probability distribution function?
- How do you choose the parameters for the probability distribution function?
- **There are many techniques for solving this problem, although two common approaches are:**
- **Maximum a Posteriori (MAP), a Bayesian method.**
- **Maximum Likelihood Estimation (MLE), frequentist method.**

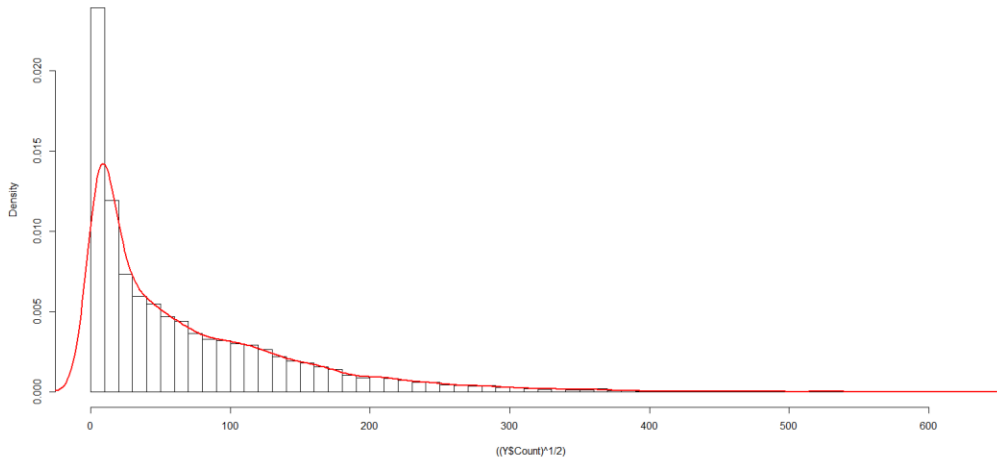
- The goal of maximum likelihood estimation is to make inference about the population, which is most likely to have generated the sample i.e., the joint probability distribution of the random variables.



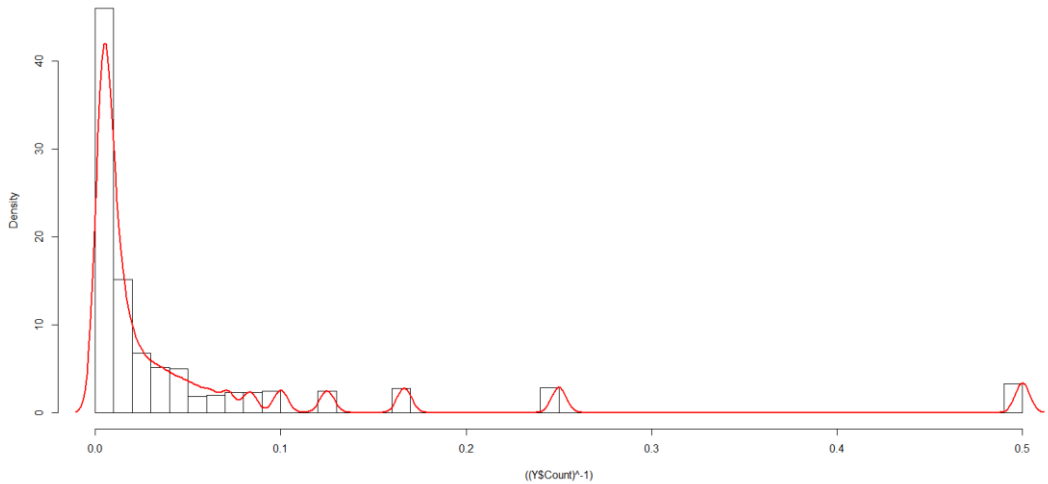
Histogram of Count Variable



Histogram of Count Variable

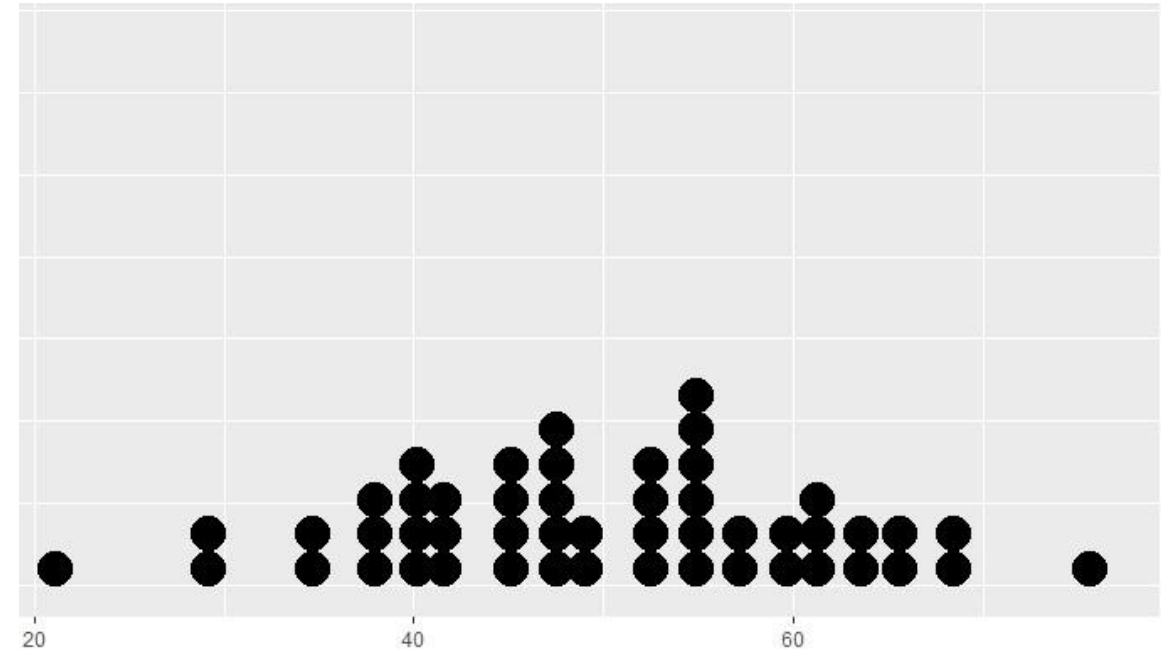


Histogram of Count Variable



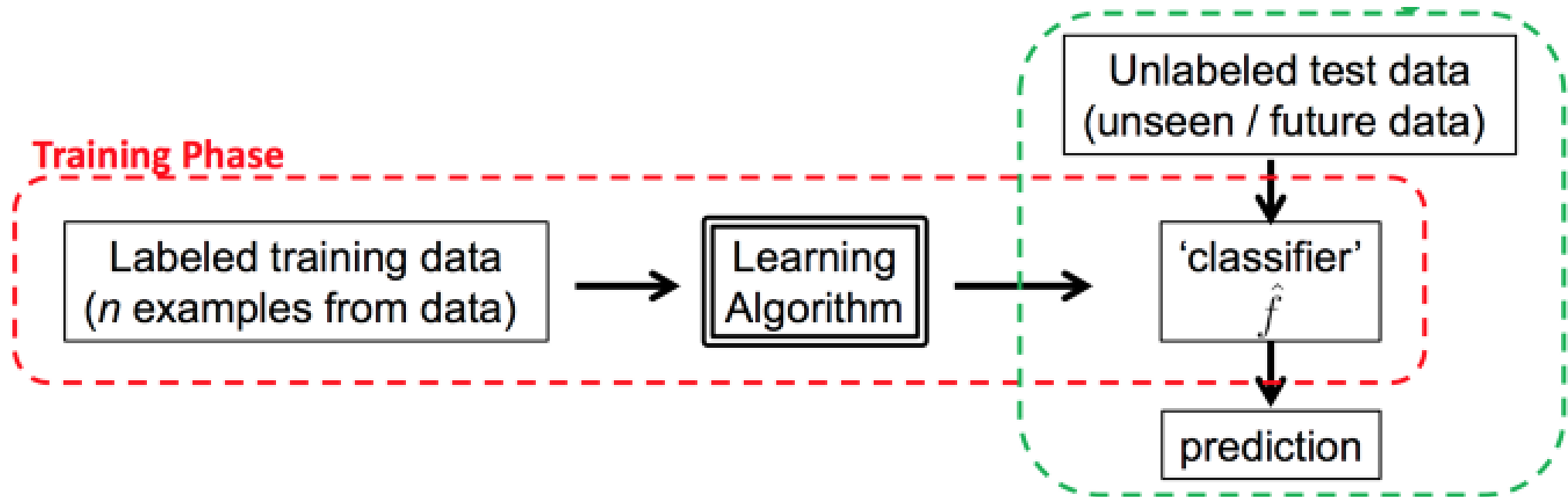
- input is the training data and the output is the parameters that are required for the classifier.
- in order to select parameters for the classifier from the training data, one can use
 - Maximum Likelihood Estimation (MLE),
 - Bayesian Estimation (Maximum a posteriori) or
 - optimization of loss criterion.

- How should we model such data so that the basic assumptions of the model are not violated?
- How about modelling this data with a different distribution rather than a normal one? If we do use a different distribution, how will we estimate the coefficients?
- *MLE can be defined as a method for estimating population parameters (such as the mean and variance for Normal, rate (λ) for Poisson, etc.) from sample data such that the probability (likelihood) of obtaining the observed data is maximized.*



What is the Probability of Heads when a single coin is tossed 40 times?





- Data are picked INDEPENDENTLY and IDENTICALLY DISTRIBUTED (i.i.d.)
- Then the data type is checked to decide what probability model can be used.
- For example, if the data is coin tosses, Bernoulli model is used
- if it's dice rolls, multinomial model can be used.
- Gaussian model, which is most common phenomenon, is used.
- In order to make sure the distribution is normal, the [normality test](#) is often done.

Maximum a Posteriori (MAP)

$$H_{MAP} = \arg \max_H P[H | E]$$

$$= \arg \max_H \frac{P[E | H] \cdot P[H]}{P[E]}$$

$$= \arg \max_H P[E | H] \cdot P[H]$$

$$H_{ML} = \arg \max_H P[E | H]$$

Maximum Likelihood (ML)



Major Steps in MLE:

- Perform a certain experiment to collect the data.
- Choose a parametric model of the data, with certain modifiable parameters.
- Formulate the likelihood as an objective function to be maximized.
- Maximize the objective function and derive the parameters of the model.

Using MLE to estimate parameters for the classifier

1. Identify the sample distribution
2. Given a sequence of training data, what is the estimation of σ^2 and μ assuming normal Gaussian population.
3. With statistical approach, we will assume a probability model, meaning we will predict how probable is the data assuming a certain probability distribution model?
4. Then we can find the best fitting probability model via the likelihood estimation.

Suppose we have a random sample X_1, X_2, \dots, X_n whose assumed probability distribution depends on some unknown parameter θ .

For binomial distribution;
unknown parameters

(n, p)

θ_1, θ_2

For Poisson distribution;
unknown parameter


λ
 θ

For Geometric distribution
unknown parameter

p
 θ


For Normal distribution
unknown parameters

μ, σ^2

Suppose we have a random sample X_1, X_2, \dots, X_n whose assumed probability distribution depends on some unknown parameter θ . 

Our primary goal here will be to find a point estimator u ,
such that $u(x_1, x_2, \dots, x_n)$ is
a "good" point estimate of θ ,

where x_1, x_2, \dots, x_n are the observed values of the random sample.

i.e., our goal will be to find a good estimate of θ ,
using the data x_1, x_2, \dots, x_n 
that we obtained from our specific random sample.

Maximum Likelihood Estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data.

For example, if a population is known to follow a normal distribution but the mean and variance are unknown, MLE can be used to estimate them using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the most likely result to have occurred.

- Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ . A random sample of 10 American female college students yielded the following weights (in pounds):
- 115 122 130 127 149 160 152 138 149 180
- Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ , the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

The probability density function of X_i is:

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

for $-\infty < x < \infty$. The parameter space is $\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu, \sigma) = \sigma^{-n} (2\pi)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

$$L(\mu, \sigma) = \sigma^{-n} (2\pi)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to μ , that the maximum likelihood estimator of μ is:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

Based on the given sample, a maximum likelihood estimate of μ is:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} (115 + \cdots + 180) = 142.2$$

It seems reasonable that a good estimate of the unknown parameter θ would be the value of θ that maximizes the probability, that is, the likelihood of getting the data we observed (this is the reason, why we called as likelihood function)

Let x_1, x_2, \dots, x_n be observations from n independent and identically distributed random variables drawn from a Probability Distribution that depends on some parameters θ .

The goal of MLE is to maximize the likelihood function:

$$L = f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

For maximization,
we have

$$\frac{dL}{d\theta} = 0 ; \quad \frac{d^2L}{d\theta^2} < 0$$

Since logarithm is a non-decreasing function, so for maximizing L , it is equivalently correct to maximize $\log L$, i.e.,

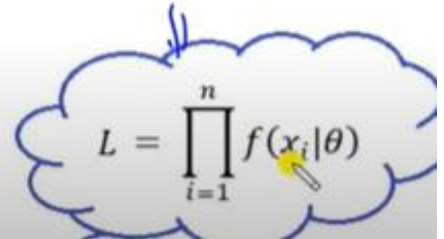
$$\frac{1}{L} \frac{dL}{d\theta} = 0 \Rightarrow \frac{d \log L}{d\theta} = 0$$

In other words, the log-likelihood function is easier to work with:

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

↓
diff.

$$\log(ab) = \log a + \log b$$


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

For Discrete case

The simplest case is when both the distribution and the parameter space (the possible values of the parameters) are discrete, meaning that there are a finite number of possibilities for each.

In this case, the MLE can be determined by explicitly trying all possibilities.

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}$, $\frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P(H = 61 | p = \frac{1}{3}) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P(H = 61 | p = \frac{1}{2}) = \binom{100}{61} \left(\frac{1}{2}\right)^{61} \left(\frac{1}{2}\right)^{39} = 0.007$$

$$P(H = 61 | p = \frac{2}{3}) = \binom{100}{61} \left(\frac{2}{3}\right)^{61} \left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x};$$
$$0 \leq p \leq 1$$
$$x = 0, 1, 2, \dots, n;$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follows a Binomial distribution, with parameter p . Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61} (1-p)^{39}$$

For maximization

$$\begin{aligned} \frac{d}{dp} P(H = 61|p) &= 0 \\ \Rightarrow \binom{100}{61} [61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}] &= 0 \\ \Rightarrow p^{60}(1-p)^{38}(61 - 100p) &= 0 \end{aligned}$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0 \checkmark$$

$$P(H = 61|p = \frac{61}{100}) = \binom{100}{61} \left(\frac{61}{100}\right)^{61} \left(\frac{39}{100}\right)^{39}$$

$$P(H = 61|p = 1) = 0$$

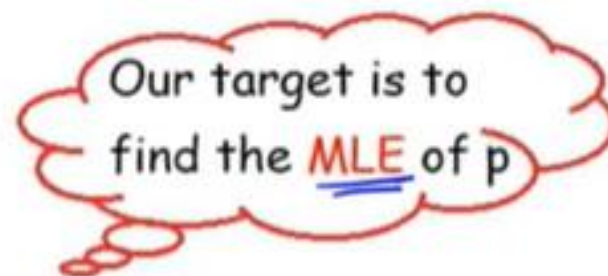
MLE for Bernoulli distribution

Example: For a random sample x_1, x_2, \dots, x_n . Assume that x_i 's are independent Bernoulli random variables of the students picking a course of Statistics with unknown parameter p , find the maximum likelihood estimator of p , the proportion of students who select Statistics subject.

Solution: Define a Bernoulli random variable as

$x_i = 1$; if a randomly selected student selects a Statistics subject

$x_i = 0$; if a randomly selected student does not select a Statistics subject



✓
The p.m.f. of Bernoulli random variable x_i is

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

The **likelihood function** L is defined as

$$\begin{aligned} L &= \prod_{i=1}^n f(x_i|\theta) \\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum x_i}(1-p)^{\sum(1-x_i)} \end{aligned}$$

$$\begin{aligned} \Rightarrow \log L &= \sum x_i \log p + \sum (1-x_i) \log(1-p) \\ &= \log p \sum x_i + (n - \sum x_i) \log(1-p) \end{aligned}$$

- Applying log

To **maximize the** L , we have

$$\begin{aligned} \frac{d}{dp} \log L = 0 &\Rightarrow \frac{1}{p} \sum x_i + (n - \sum x_i) \left(-\frac{1}{1-p} \right) = 0 \\ &\Rightarrow \frac{\sum x_i}{p} = \frac{n - \sum x_i}{1-p} \\ &\Rightarrow \cancel{\sum x_i} - p \cancel{\sum x_i} = n p - \cancel{\sum x_i} p \end{aligned}$$

$$\Rightarrow p = \frac{\sum x_i}{n}$$

Further, $\frac{d^2}{dp^2} \log L < 0$ ✓

Thus,

an estimator of p is $\frac{\sum x_i}{n}$

MLE FOR Binomial Distribution

Example: For a random sample x_1, x_2, \dots, x_n . Assume that x_i 's are independent Binomial random variables with unknown parameter p , find the maximum likelihood estimator of p .

Solution: For binomial distribution, we have

$$p(x_i) = \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \quad ; \quad x_i = 0, 1, 2, \dots, n ; p \in [0, 1]$$

The likelihood function L is defined as

$$L = \prod_{i=1}^n p(x_i | \theta) = \prod_{i=1}^n \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

Our target is to find the MLE of p

The likelihood function L is defined as

$$\begin{aligned} L &= \prod_{i=1}^n p(x_i | \theta) \\ &= \prod_{i=1}^n \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \\ \Rightarrow \log L &= \sum_{i=1}^n \left[\log \binom{n}{x_i} + \log p^{x_i} + \log (1-p)^{n-x_i} \right] \\ \Rightarrow \log L &= \sum_{i=1}^n \left[\log \binom{n}{x_i} + x_i \log p + (n-x_i) \log (1-p) \right] \\ \Rightarrow \log L &= \sum_{i=1}^n \log \binom{n}{x_i} + \log p \sum_{i=1}^n x_i + \log (1-p) \sum_{i=1}^n (n-x_i) \end{aligned}$$

To maximize L , we have

$$\begin{aligned} \frac{d}{dp} \log L &= 0 \Rightarrow \frac{1}{p} \sum x_i - \frac{1}{1-p} \left(\sum (n-x_i) \right) = 0 \\ &\Rightarrow \frac{1}{p} \sum x_i - \frac{1}{1-p} \sum (n-x_i) = 0 \end{aligned}$$

$$\begin{aligned} \sum (n-x_i) &= \sum n - \sum x_i \\ &= n \sum 1 - \sum x_i \\ &= n^2 \end{aligned}$$

To maximize L , we have

(i)

$$\frac{d}{dp} \log L = 0 \Rightarrow \frac{1}{p} \sum x_i - \frac{1}{1-p} \sum (n - x_i) = 0$$

$$\Rightarrow \frac{1}{p} \sum x_i - \frac{n^2}{1-p} + \frac{1}{1-p} \sum x_i = 0$$

$$\Rightarrow \frac{1}{p(1-p)} \sum x_i = \frac{n^2}{1-p}$$

$$\Rightarrow \frac{1}{p} \sum x_i = n^2$$

$$\Rightarrow p = \frac{\sum x_i}{n^2}$$

$$\frac{d^2}{dp^2} \log L = -\frac{1}{p^2} \sum x_i + \frac{1}{(1-p)^2} \sum (n - x_i) < 0$$

Bernoulli
Binomial
 $n \cdot p$
 $\frac{\sum x_i}{n}$

Hence, the MLE of p is $\frac{\sum x_i}{n^2}$

Poission distribution

- The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time

Minimum Description Length Principle

- The minimum description length (MDL) criteria in machine learning says that the best description of the data is given by the model which compresses it the best.
- Put another way, learning a model for the data or predicting it is about capturing the regularities in the data and any regularity in the data can be used to compress it. Thus, the more we can compress a data, the more we have learnt about it and the better we can predict it

Occam's Razor

- Given a choice of theories that are equally good the simplest theory should be chosen
- Physical sciences: any theory should be consistent with all empirical observations
- Data mining:
 - theory = predictive model
 - good theory = good prediction
 - What is good? Do we minimize the error rate?

Information Function

- Maximizing $P[T/E]$ equivalent to minimizing

$$-\log P[T | E] =$$

$$-\log P[E | T] - \log P[T] + \log P[E]$$

Number of bits it takes
to submit the exceptions



Number of bits it takes
to submit the theory

- That is, the MDL principle!

Minimum Description Length

- MDL principle:
 - Minimize
 - size of theory + info needed to specify exceptions
- Suppose trainings set E is mined resulting in a theory T
- Want to minimize

$$L[T] + L[E | T]$$

Most Likely Theory

- Suppose we want to maximize $P[T/E]$
- Bayes' rule

$$P[T | E] = \frac{P[E | T]P[T]}{P[E]}$$

- Take logarithms

$$\log P[T | E] = \log P[E | T] + \log P[T] - \log P[E]$$

Applications to Learning

- Classification, association, numeric prediction
 - Several predictive models with 'similar' error rate (usually as small as possible)
 - Select between them using Occam's razor
 - Simplicity subjective
 - Use MDL principle
- Clustering
 - Important learning that is difficult to evaluate
 - Can use MDL principle

Outlook	Temperature	Humidity	Wind	Play
x_1	x_2	x_3	x_4	y
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

$$\begin{aligned}
 E = \{ & \{x_1 = \textit{sunny}, x_2 = \textit{hot}, x_3 = \textit{high}, x_4 = \textit{weak}\} \\
 & \{x_1 = \textit{sunny}, x_2 = \textit{hot}, x_3 = \textit{high}, x_4 = \textit{strong}\} \\
 & \{x_1 = \textit{overcast}, x_2 = \textit{hot}, x_3 = \textit{high}, x_4 = \textit{weak}\} \\
 & \dots \\
 & \}
 \end{aligned}$$

$$L(E) = ?$$

Excluding the class attribute, the total number of different attribute value pairs is $ts = 10$.

In each tuple we have 4 pairs involving independent attributes, which may be chosen in $\binom{10}{4} = \frac{10!}{4! \times 6!} = 210$ different ways, and one pair with the class attribute, which may be chosen in 2 different ways.

Thus the probability of choosing a tuple is $\frac{1}{210} \times \frac{1}{2}$.

The code length in bits for this choice is $-\log_2 \frac{1}{210} - \log_2 \frac{1}{2} = \log_2 210 + \log_2 2 = 7.715 + 1 = 8.715$, which is also the number of bits to encode a tuple with its class value.

Then the code length of the whole data set is

$$L(E) = 8.715 \times 14 = 122.01$$