Tutorial on Maximum Likelyhood Estimation: Parametric Density Estimation

Sudhir B Kylasa

03/13/2014

1 Motivation

Suppose one wishes to determine just how biased an unfair coin is. Call the probability of tossing a HEAD is p. The goal then is to determine p.

Also suppose the coin is tossed 80 times: i.e., the sample might be something likex_1 = H, x_2 = T, , x_80 = T, and the count of number of HEADS, "H" is observed.

The probability of tossing TAILS is $1 - p$. Suppose the outcome is 49 HEADS and 31 TAILS, and suppose the coin was taken from a box containing three coins: one which gives HEADS with probability $p = 1/3$, one which gives HEADS with probability $p = 1/2$ and another which gives HEADS with probability $p = 2/3$. The coins have lost their labels, so which one it was is unknown. Clearly the probability mass function for this experiment is binomial distribution with sample size equal to 80, number of successes equal to 49 but different values of p. We have the following probability mass functions for each of the above mentioned cases:

$$Pr(H = 49|p = 1/3) = \binom{80}{49}(1/3)^{49}(1 - 1/3)^3 1 \approx 0.000, \quad (1)$$

$$Pr(H = 49|p = 1/2) = \binom{80}{49}(1/2)^{49}(1 - 1/2)^3 1 \approx 0.012, \quad (2)$$

$$Pr(H = 49|p = 2/3) = \binom{80}{49}(2/3)^{49}(1 - 2/3)^3 1 \approx 0.054 \quad (3)$$

Based on the above equations, we can conclude that the coin with $p = 2/3$ was more likely to be picked up for the observations which we were given to begin with.

2 Definition

The generic situation is that we observe a n-dimensional random vector X with probability density (or mass) function $f(x;\theta)$. It is assumed that θ is a fixed, unknown constant belonging to the set $\Theta \subset \mathbb{R}^n$

For $x \in \mathbb{R}^n$, the likelihood function of θ is defined as

$$L(\theta/x) = f(x/\theta). \tag{4}$$

x is regarded as fixed, and θ is regarded as the variable for L. The log-likelihood function is defined as $l(\theta/x) = logL(\theta/x)$.

The *maximum likelihood estimate* (or MLE) is the value $\hat{\theta} = \theta(\hat{x}) \in \Theta$ maximizing L(θ/x), provided it exists:

$$L(\hat{\theta}/(x)) = \arg\max_{\theta} L(\theta/x) \tag{5}$$

3 What is Likelihood function ?

If the probability of an event X dependent on model parameters p is written as

$$P(X|p)$$

then we talk about the likelihood

$$L(p|X)$$

that is the likelihood of the parameters given the data.

For most sensible models, we will find that certain data are more probable than other data. The aim of maximum likelihood estimation is to find the parameter value(s) that makes the observed data most likely. This is because the likelihood of the parameters given the data is defined to be equal to the probability of the data given the parameters

If we were in the business of making predictions based on a set of solid assumptions, then we would be interested in probabilities - the probability of certain outcomes occurring or not occurring.

However, in the case of data analysis, we have already observed all the data: once they have been observed they are fixed, there is no 'probabilistic' part to them anymore (the word data comes from the Latin word meaning 'given'). We are much more interested in the likelihood of the model parameters that underly the fixed data.

The following is the relation between the likelihood and the probability spaces:

Probability

Knowing paramtres $\rightarrow$ prediction of outcomes

Likelihood

Observation of data $\rightarrow$ estimation of parameters

4 Method

Maximum likelihood (ML) estimates need not exist nor be unique. In this section, we show how to compute ML estimates when they exist and are unique. For computational convenience, the ML estimate is obtained by maximizing the log-likelihood function, log L(θ/x). This is because the two functions log L(θ/x) and L(θ/x) are monotonically related to each other so the same ML estimate is obtained by maximizing either one. Assume that the log-likelihood function is differentiable, if θ_{MLE} exists, it must satisfy the following partial differential equation known as the likelihood equation:

$$\frac{d}{d\theta}\left(logL(\theta/x)\right) = 0 \tag{6}$$

at $\theta = \theta_{MLE}$. This is because maximum or minimum of a continuously differentiable function implies that its first derivatives vanishes at such points.

The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that log L(θ/x) is a maximum and not minimum, since the first derivative cannot reveal this. To be a maximum, the shape of the log-likelihood function should be convex in the neighborhood of θ_{MLE}. This can be checked by calculating the second derivatives of the log-likelihoods and showing whether they are all negative at $\theta = \theta_{MLE}$.

$$\frac{d^2}{d\theta^2}\left(logL(\theta/x)\right) < 0 \tag{7}$$

5 Properties

Some general properties (advantages and disadvantages) of the Maximum Likelihood Estimate are as follows:

1. For large data samples (large N) the likelihood function L approaches a Gaussian distribution
2. Maximum Likelihood estimates are usually consistent. For large N the estimates converge to the true value of the parameters which are estimated.
3. Maximum Likelihood Estimates are usually unbiased. For all sample sizes the parameter of interest is calculated correctly.
4. Maximum Likelihood Estimate is efficient: (the estimates have the smallest variance).
5. Maximum Likelihood Estimate is sufficient: (it uses all the information in the observations).
6. The solution from the Maximum Likelihood Estimate is unique.

On the other hand, we must know the correct probability distribution for the problem at hand.

6 Numerical examples using Maximum Likelihood Estimation

In the following section, we discuss the applications of MLE procedure in estimating unknown parameters of various common density distributions.

6.1 Estimating prior probability using MLE

Consider a two-class classification problem, with classes (ω_1, ω_2) and let Prob(ω_1) = p and Prob(ω_2) = $1 - p$ (here p is the unknown parameter). By using MLE, we can estimate p as follows:

Let the sample be $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. Let ω_{ij} be the class of the feature vector x_j (N is the sample size and N_1 is the number of feature vectors belonging to class ω_1). Also assume that samples $x_1, x_2 \ldots, x_N$ are independent events. Then we have the following equations.

$$\begin{aligned} Prob(\mathcal{D}/p) &= \prod_{j=1}^{N} Prob(\omega_{ij}/p) \\ \text{By Independence of the feature vectors} & \\ &= p^{N_1} * (1-p)^{N-N_1} \end{aligned}$$

Please note that $Prob(\mathcal{D}/p)$ is infinitely differentiable function of p, so the local maxima lies where its derivative is zero.

$$\begin{aligned} \frac{d}{dp}\left(p^{N_1} * (1-p)^{N-N_1}\right) &= 0 \\ N_1 * p^{N_1-1} * (1-p)^{N-N_1} - (N-N_1) * (1-p)^{N-N_1-1} * p^N &= 0 \\ p^{N_1} * (1-p)^{N-N_1-1} &= 0 \end{aligned}$$

Solving the above equation for p, we get the following:

$$p^{N_1} * (1-p)^{N-N_1-1} = 0 \quad (8)$$

$$N_1 * (1-p) = (N-N_1) * p \quad (9)$$

So p is either 0 or 1 by eq. 8 and p is (N_1/N) by eq. 9. Hence this proves that taking the frequencies for probabilities of the feature vectors is optimum and using MLE we showed that p is maximized. Hence the class probabilities are optimum (the likelihood function is maximized using MLE).

6.2 Estimating μ of a Gaussian distribution when Σ is known

In this section, we estimate the value of μ (μ_{MLE}), when the covariance matrix (Σ) in known, for gaussian distribution in n-dimensional feature space.

$$\rho(\mathcal{D}/\mu) = \prod_{j=1}^{N} \rho(x_j/\mu)$$

log likelihood function is

$$\begin{aligned} log\rho(\mathcal{D}/\mu) &= \sum_{j=1}^{N} log(\rho(x_j/\mu)) \\ &= \sum_{j=1}^{N} log\left[\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp^{-\frac{(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)}{2}}\right] \\ &= \sum_{j=1}^{N} log\left[\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\right] - \frac{1}{2}\left[(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right] \end{aligned} \quad (10)$$

This function is infinitely differentiable function of unknown parameters (μ)'s. To find the maxima, we set the derivative of this eq. 10 to 0.

$$\begin{bmatrix} \frac{d}{d\mu_1} \\ \frac{d}{d\mu_2} \\ \dots \\ \frac{d}{d\mu_N} \end{bmatrix}_{n\times 1} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}_{n\times 1} \tag{11}$$

Differentiating the log likelihood functions yields the following results:

$$\begin{aligned} \nabla_{\vec{\mu}} log(\mathcal{D}/\vec{\mu}) &= \frac{-1}{2}\sum_{j=1}^{N} \nabla_{\vec{\mu}} \left[(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right] \\ &= \frac{-1}{2}\sum_{j=1}^{N} \begin{bmatrix} \frac{d}{d\mu_1}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) \\ \frac{d}{d\mu_2}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) \\ \dots \\ \frac{d}{d\mu_N}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) \end{bmatrix}_{n\times 1} \end{aligned}$$

$$\begin{aligned} \text{But } \frac{d}{d\mu_i}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) &= \left[\frac{d}{d\mu}(x_j-\mu)^T\right]\Sigma^{-1}(x_j-\mu) + (x_j-\mu)^T\Sigma^{-1}\left[\frac{d}{d\mu}(x_j-\mu)\right] \\ &= 2\left[\frac{d}{d\mu}(x_j-\mu)^T\right]\Sigma^{-1}(x_j-\mu) \\ &= -2e_i^T\Sigma^{-1}(x_j-\mu) \end{aligned}$$

where e_i^T = [0,0,0 …, 1, …] and 1 is located at position i in the array

$$\begin{aligned} \nabla_{\vec{\mu}} log(\mathcal{D}/\vec{\mu}) &= \frac{-1}{2}\sum_{j=1}^{N} \begin{bmatrix} -2e_1^T\Sigma^{-1}(x_j-\mu) \\ -2e_2^T\Sigma^{-1}(x_j-\mu) \\ \dots \\ -2e_N^T\Sigma^{-1}(x_j-\mu) \end{bmatrix}_{n\times 1} \\ &= \sum_{j=1}^{N} \begin{bmatrix} e_1^T \\ e_2^T \\ \dots \\ e_N^T \end{bmatrix}_{n\times 1} \Sigma^{-1}(x_j-\mu) \end{aligned}$$

Notice that $\begin{bmatrix} e_1^T \\ e_2^T \\ \dots \\ e_N^T \end{bmatrix}_{n\times 1}$ is the identity matrix.

So the above equation reduces to.

$$\begin{aligned} \nabla_{\vec{\mu}} log(\mathcal{D}/\vec{\mu}) &= \sum_{j=1}^{N} \Sigma^{-1}(x_j - \mu) \\ &= \Sigma^{-1} \sum_{j=1}^{N} (x_j - \mu) \end{aligned} \tag{12}$$

By solving the equation. 12 for μ.

$$\begin{aligned} \Sigma^{-1} \sum_{j=1}^{N} (x_j - \mu) &= 0 \\ \sum \Sigma^{-1} \sum_{j=1}^{N} (x_j - \mu) &= 0 \\ \sum_{j=1}^{N} (x_j - \mu) &= 0 \\ \sum_{j=1}^{N} (x_j) - \sum_{j=1}^{N} (\mu) &= 0 \\ \mu &= \frac{1}{N} (\sum_{j=1}^{N} (x_j)) = \mu_{MLE} \end{aligned} \tag{13}$$

Hence, we proved that using MLE the sample mean is the maximum likelihood estimate of any given sample.

6.3 Estimating μ and σ^2 for 1-D gaussian distribution using MLE

In this subsection, we estimate the μ and σ^2 for one-dimensional gaussian data. Here $\theta = (\theta_1, \theta_2)$ are (μ, σ^2), which are unknown parameters. We estimate these parameter using the procedure discussed in the section 4.

The log-likelihood function for this case is given by the following equation:

$$\begin{aligned} log\rho(x_k/\mu,\sigma^2) &= log\left[\frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{x_k-\mu}{2\sigma^2})\right] \\ &= -\frac{1}{2}log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(x_k-\mu)^2 \\ log\rho(\mathcal{D}/\mu,\sigma^2) &= \prod_{k=1}^{N} log\rho(x_k/\mu,\sigma^2) \end{aligned} \tag{14}$$

Since $(x_1, x_2, \ldots, x_N)$ are I.I.D's, the density function can be written in product form as follows:

$$\begin{aligned} \rho(\mathcal{D}/\mu,\sigma^2) &= \rho((x_1,x_2,\ldots,x_N)/\mu,\sigma^2) \\ &= \rho(x_1/\mu,\sigma^2)\rho(x_2/\mu,\sigma^2)\ldots\rho(x_N/\mu,\sigma^2) \\ log\rho(\mathcal{D}/\mu,\sigma^2) &= log\prod_{k=1}^{N}\rho(x_k/\mu,\sigma^2) \\ &= \sum_{k=1}^{N}\left[-\frac{1}{2}log(1\pi\sigma^2)-\frac{1}{2\sigma^2}(x_k-\mu)^2\right] \end{aligned} \tag{15}$$

Differentiating and setting the eq. 15 to 0, yields the following equations:

$$\begin{aligned} \nabla_{\mu,\sigma^2}\rho(\mathcal{D}/\mu,\sigma^2) &= \begin{bmatrix} \frac{d}{d\mu}log\rho(\mathcal{D}/\mu,\sigma^2) \\ \frac{d}{d\sigma^2}log\rho(\mathcal{D}/\mu,\sigma^2) \end{bmatrix}_{2\times 1} \\ &= \begin{bmatrix} \frac{d}{d\mu}(-\frac{N}{2}log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)) \\ \frac{d}{d\sigma^2}(-\frac{N}{2}log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)) \end{bmatrix}_{2\times 1} \\ &= \begin{bmatrix} \frac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\mu) \\ -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{k=1}^{N}(x_k-\mu)^2 \end{bmatrix}_{2\times 1} \end{aligned}$$

Solving the eq. 16 for μ and σ^2 yields the following:

$$\begin{aligned}
\frac{1}{\sigma^2}\left(\sum_{k=1}^{N}(x_k-\mu)\right) &= 0 \\
\sum_{k=1}^{N}(x_k-\mu) &= 0 \\
\hat{\mu} = \mu &= \frac{1}{N}\sum_{k=1}^{N}(x_k) \qquad (16) \\
-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{k=1}^{N}(x_k-\mu)^2 &= 0 \\
\sum_{k=1}^{N}(x_k-\mu)^2 &= N\sigma^2 \\
\hat{\sigma^2} = \sigma^2 &= \frac{\sum_{k=1}^{N}(x_k-\hat{\mu})^2}{N}
\end{aligned}$$

which are the mean and standard deviation of the empirical data.

In general for N-dimensional Gaussian data, when X $\sim$ N(μ, Σ) where X $\in \mathcal{R}^n$, with $\vec{\mu}$ and Σ are unknown parameters, then MLE for μ and Σ are as follows:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{N}\sum_{k=1}^{N}(\vec{x_k}) \\
\hat{\sigma^2} &= \frac{\sum_{k=1}^{N}(\vec{x_k}-\vec{\mu})(\vec{x_k}-\vec{\mu})^T}{N}
\end{aligned}$$

6.4 How far does the $\hat{\mu}$ deviate from the true μ when MLE is used.

As discussed in section 6.3, we know that

$$\begin{aligned}
E[\hat{\mu}] &= \frac{1}{N}\sum_{k=1}^{N}(\vec{x_k}) \\
&= \mu
\end{aligned}$$

Now we compute the expected value of $(|\hat{\mu}-\mu|)^2$ as follows:

$$
\begin{aligned}
E[(|\hat{\mu}-\mu|)^2] &= E[(\hat{\mu}-\mu)(\hat{\mu}-\mu)] \\
&= E[\hat{\mu}\mu - \mu\hat{\mu} - \hat{\mu}\mu + \mu\mu] \\
&= E[\hat{\mu}\hat{\mu}] - 2E[\mu\hat{\mu}] + E[\mu\mu]
\end{aligned}
$$

Substituting E[$\hat{\mu}$] = μ we have

$$
\begin{aligned}
&= E[\hat{\mu}\hat{\mu}] - \mu\mu \\
&= E(\frac{1}{N}\sum_{k=1}^{N} X_k * \sum_{j=1}^{N} X_j) - \mu\mu \\
&= \frac{1}{N^2}\sum_{j,k=1}^{N} E[X_j X_k] - \mu\mu
\end{aligned}
$$

Treating X_j as random variables in the above equations, we have

$$
\begin{aligned}
E[(|\hat{\mu}-\mu|)^2] &= \frac{1}{N^2}\left[\sum_{j,k=1}^{N} E(X_j)E(X_k) + \sum_{j,k=1}^{N} E(X_j * X_k)\right] - \mu\mu \\
&= \frac{1}{N^2}\left[N(N-1)\mu\mu + N * E[X^2]\right]\mu\mu \\
E[(|\hat{\mu}-\mu|)^2] &= \frac{1}{N}E[X^2] - \frac{1}{N}\mu\mu
\end{aligned}
$$

But, we know that

$$
\begin{aligned}
E[|X-\mu|^2] &= E[(X-\mu)(X-\mu)] = \sigma^2 \\
E[X*X] - \mu^2 &= \sigma^2 \\
E[X*X] = \sigma^2 * \mu^2
\end{aligned}
$$

Therefore, we have the following result:

$$
\begin{aligned}
E[(|\hat{\mu}-\mu|)^2] &= \frac{1}{N}(\sigma^2 * \mu^2) - \frac{1}{N}(\mu\mu) \\
&= \frac{1}{N}\sigma^2
\end{aligned}
$$

So the expected value of $(|\hat{\mu}-\mu|)^2$ is proportional to the true standard deviation.

6.5 Estimate for Σ is biased, when MLE is used.

In the following section, we will show that the estimate for covariance is biased (when μ is unknown) when MLE is used to estimate its value and equals to true covariance when μ is known.

$$\begin{aligned}
E[\hat{\Sigma}] &= E[\frac{1}{N}\sum_{k=1}^{N}(X_k-\mu)(X_k-\mu)^T] \\
&= \sum_{k=1}^{N}E[\frac{1}{N}(X_k-\mu)(X_k-\mu)^T] \\
&= \frac{1}{N}\sum_{k=1}^{N}E[X_kX_k^T - X_K\mu^T - \mu X^T + \mu\mu^T] \\
&= \frac{1}{N}NE[X_KX_K^T] - \mu\mu^T \\
&= E[XX^T - \mu\mu^T] \\
&= E[XX^T - 2\mu\mu^T - \mu\mu^T] \\
&= E[(X-\mu)(X-\mu)^T] = \Sigma
\end{aligned}$$

If the μ is known, then it turns out that Estimated value of $\hat{\Sigma}$ is equal to true Σ. We now show the derivation that in case where μ is not known, then estimated value of $\hat{\Sigma}$ is not equal to true Σ

$$\begin{aligned}
E[\hat{\Sigma}] &= \frac{1}{N}E[\sum_{k=1}^{N}(X_k-\hat{\mu})(X_k-\hat{\mu})^T] \\
&= \frac{1}{N}\sum_{k=1}^{N}E[X_kX_k^T - X_K\hat{\mu}^T - \hat{\mu}X^T + \hat{\mu}\hat{\mu}^T] \\
&= \frac{1}{N}\sum_{k=1}^{N}\left(E[X_kX_k^T] - E[X_K\hat{\mu}^T] - E[\hat{\mu}X^T] - E[\hat{\mu}\hat{\mu}^]\right) \\
&= \frac{1}{N}\sum_{k=1}^{N}\left(E[X_kX_k^T] - E[(\frac{1}{N}\sum_{l=1}^{N}X_l)X_k]\right) \\
&\quad -E[\frac{1}{N}\sum_{l=1}^{N}X_kX_l^T] + E[\frac{1}{N}\sum_{l=1}^{N}X_l\frac{1}{N}\sum_{m=1}^{N}X_m^T]
\end{aligned}$$

Splitting the summation above into two parts, l = k and l $\neq$ k, we have the following:

$$\begin{aligned}
E[\hat{\Sigma}] &= \frac{1}{N}\sum_{k=1}^{N}\{E[XX^T] - \frac{1}{N}\Sigma_{l=1,l\neq k}^{N}E(X_l)E(X_k^T) - \frac{1}{N}\Sigma_{l=1,l=k}^{N}E(X_kX_k^T) \\
&\quad -\frac{1}{N}\Sigma_{l=1,l\neq k}^{N}E(X_k)E(X_l^T) - \frac{1}{N}\Sigma_{l=1,l=k}^{N}E(X_kX_k^T) \\
&\quad +\frac{1}{N^2}\Sigma_{l,m=1,l\neq m}^{K}E(X_l)E(X_m^T) + \frac{1}{N^2}\Sigma_{l,m=1,l=m}^{K}E(X_l)E(X_l^T)\} \\
&= \frac{1}{N}\sum_{k=1}^{N}\{E[XX^T] - \frac{1}{N}\Sigma_{l=1,l\neq k}^{N}\mu\mu^T - \frac{1}{N}\Sigma_{l=1,l=k}^{N}E(X_kX_k^T) \\
&\quad -\frac{1}{N}\Sigma_{l=1,l\neq k}^{N}\mu\mu^T - \frac{1}{N}\Sigma_{l=1,l=k}^{N}E(X_kX_k^T) \\
&\quad +\frac{1}{N^2}\Sigma_{l,m=1,l\neq m}^{K}\mu\mu^T + \frac{1}{N^2}\Sigma_{l,m=1,l=m}^{K}E(X_l)E(X_l^T)\}
\end{aligned}$$

$$\begin{aligned}
E[\hat{\Sigma}] &= \frac{1}{N}\left[N(1-\frac{1}{N}E(XX^T) + N(\frac{1}{N}-1)\mu\mu^T\right] \\
&= (1-\frac{1}{N})\left[E(XX^T) - \mu\mu^T\right] \\
&= (\frac{N-1}{N})E(XX^T - \mu\mu^T)) \\
&= (\frac{N-1}{N})E[(X-\mu)(X-\mu)^T] \neq \Sigma
\end{aligned}$$

6.6 Binomial Distribution

This section discusses the estimation of p in a typical binomial distribution. Assuming that parameter p, is unknown in a binomial distribution give by the equation below:

$$
\begin{aligned}
Prob(r/p) &= \binom{n}{r} p^r (i-p)^{n-r} \\
logProb(r/p) &= log\binom{n}{r} = rlog(p) + (n-r)log(1-p) \\
logProb(\mathcal{D}/p) &= log\prod_{i=1}^{N} Prob(r_i/p) \\
&= \sum_{i=1}^{N} log[Prob(r_i/p)] \\
&= \sum_{i=1}^{N} [log\binom{n}{r_i} + r_i log? + (n-r_i)log(1-p)]
\end{aligned}
$$

To find the maximum, we set the derivative of log-likelihood function to be 0:

$$
\begin{aligned}
\frac{d}{dp} log(\mathcal{D}/p) &= 0 \\
\sum_{i=1}^{N} [\frac{r_i}{p} + \frac{n-r_i}{1-p}(-1)] &= 0 \\
\frac{1}{p}\sum_{i=1}^{N} r_i &= \frac{1}{1-p}\sum_{i=1}^{N}(n-r_i) \\
(1-p)\sum_{i=1}^{N} r_i &= p\sum_{i=1}^{N}(n-r_i) \\
p &= \frac{1}{N}\sum_{i=1}^{N} r_i/n
\end{aligned}
$$

Hence p, which is the unknown parameter is equal to the average of number of successes in the sample space.