Likelihood Function (or simply Likelihood)

- wikipedia

In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.

It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values.

Likelihood for discrete case:
If X is discrete random variable with pmf p depending on parameter θ, then,

L(θ| x) = p_θ(x) = P_θ (X=x).

The likelihood is considered as a function of θ given the output x.

Continuous case is
Let X be a random variable with pdf f depending on parameter θ, then,

L(θ| x) = f_θ(x)

For example, for Bernoulli distribution, the joint prob function is:

P(X=x|p) = prod i=1 to n p^xi (1-p)^(1-xi)

If we characterized the likelihood as function of p, then,

L(p|x) = prod i=1 to n p^xi (1-p)^(1-xi) (this is same as above)

We also note that, total sum of prob = 1 but may not be for likelihood.

int_p=0_p=1 L(p|x) dp != 1

Example of Likelihood function (coin flip)

- wikipedia

Let's say we have outcome of coin flip as HH. Then, if we assume parameter of fairness p_H = 0.5, then,

likelihood function is,
L(p_H=0.5 | HH) = 0.25

This means, given our outcome of HH, the likelihood that parameter p_H equals 0.5 is 0.25.

Discrete probability distribution [edit]

Let X be a discrete random variable with probability mass function p depending on a parameter θ. Then the function

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

considered as a function of θ, is the *likelihood function*, given the outcome x of the random variable X. Sometimes the probability of "the value x of X for the parameter value θ" is written as $P(X=x \mid \theta)$ or $P(X=x;\theta)$. $\mathcal{L}(\theta \mid x)$ should not be confused with $p(\theta \mid x)$; the likelihood is equal to the probability that a particular outcome x is observed when the true value of the parameter is θ, and hence it is equal to a probability density over the outcome x, not over the parameter θ.

Example [edit]

Consider a simple statistical model of a coin flip: a single parameter p_H that expresses the "fairness" of the coin. The parameter is the probability that a coin lands heads up ("H") when tossed. p_H can take on any value within the range 0.0 to 1.0. For a perfectly fair coin, $p_\text{H} = 0.5$.

Imagine flipping a fair coin twice, and observing the following data: two heads in two tosses ("HH"). Assuming that each successive coin flip is i.i.d., then the probability of observing HH is

$$P(\text{HH} \mid p_\text{H} = 0.5) = 0.5^2 = 0.25.$$

Hence, given the observed data HH, the *likelihood* that the model parameter p_H equals 0.5 is 0.25. Mathematically, this is written as

$$\mathcal{L}(p_\text{H} = 0.5 \mid \text{HH}) = 0.25.$$

This is not the same as saying that the probability that $p_\text{H} = 0.5$, given the observation HH, is 0.25. (For that, we could apply Bayes' theorem, which implies that the posterior probability is proportional to the likelihood times the prior probability.)

Suppose that the coin is not a fair coin, but instead it has $p_\text{H} = 0.3$. Then the probability of getting two heads is

$$P(\text{HH} \mid p_\text{H} = 0.3) = 0.3^2 = 0.09.$$

Hence

$$\mathcal{L}(p_\text{H} = 0.3 \mid \text{HH}) = 0.09.$$

More generally, for each value of p_H, we can calculate the corresponding likelihood. The result of such calculations is displayed in Figure 1.

In Figure 2, the integral of the likelihood over the interval [0, 1] is 1/3. That illustrates an important aspect of likelihoods: likelihoods do not have to integrate (or sum) to 1, unlike probabilities.

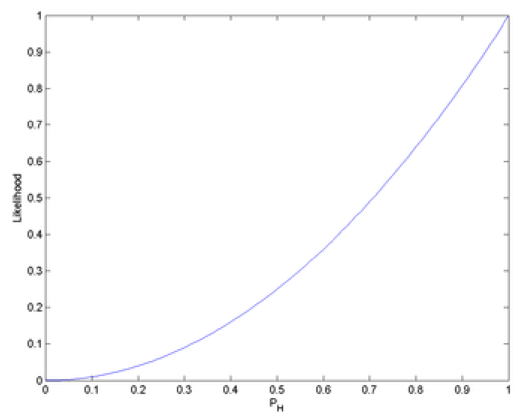
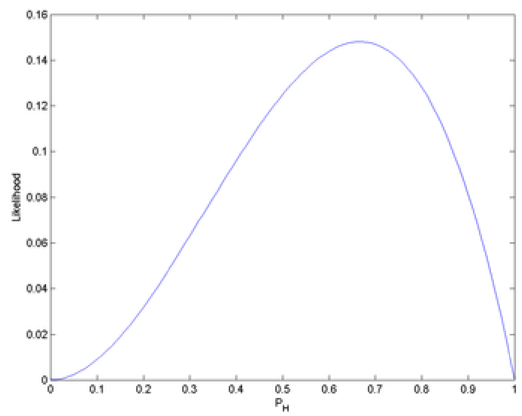


Figure 1. The likelihood function (p_H^2) for the probability of a coin landing heads-up (without prior knowledge of the coin's fairness), given that we have observed HH.



Continuous probability distribution [edit]

Probability vs Likelihood

Example of Likelihood!

Likelihood calculation involves calculating the best distribution or best characteristics of data given a particular feature value or situation.

Consider the exactly same dataset example as provided above for probability, if their likelihood of height > 170 cm has to be calculated then it will be done using the information shown below:

$$Likelihood(\mu = 170, \sigma = 3.5 | height > 170cm)$$

Likelihood calculation [Image by Author!]

In the calculation of the Likelihood, the equation of the conditional probability flips as compared to the equation in the probability calculation.

Here, the dataset features will be varied, i.e. Mean & Standard Deviation of the dataset will be varied in order to get the maximum likelihood for height > 170 cm.

> The likelihood in very simple terms means to increase the chances of a particular situation to happen/occur by varying the characteristics of the dataset distribution.

Example of Probability!

Consider a dataset containing the heights of the people of a particular country. Let's say the mean of the data is 170 & the standard deviation is 3.5.

When Probability has to be calculated of any situation using this dataset, then the dataset features will be constant i.e. mean & standard deviation of the dataset will be constant, they will not be altered. Let's say the probability of height > 170 cm has to be calculated for a random record in the dataset, then that will be calculated using the information shown below:

$$P(height > 170cm|\mu = 170, \sigma = 3.5)$$

Calculating Probability [Image by Author!]

In the above image, “mu” represents mean & “sigma” represents Standard Deviation.

> While calculating probability, feature value can be varied, but the characteristics(mean & Standard Deviation) of the data distribution cannot be altered.

If in the same dataset, the probability of height > 190 cm has to be calculated, then in the above equation, only the height part would have changed.

8

I think maybe the best way to explain the notion of likelihood is to consider a concrete example. Suppose I have a sample of IID observations drawn from a Bernoulli distribution with unknown probability of success p: $X_i \sim \text{Bernoulli}(p)$, $i = 1, \ldots, n$, so the joint probability mass function of the sample is

$$\Pr[\boldsymbol{X} = \boldsymbol{x} \mid p] = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

This expression also characterizes the likelihood of p, given an observed sample $\boldsymbol{x} = (x_1, \ldots, x_n)$:

$$L(p \mid \boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

But if we think of p as a random variable, this likelihood is not a density:

$$\int_{p=0}^{1} L(p \mid \boldsymbol{x})\, dp \neq 1.$$

It is, however, *proportional* to a probability density, which is why we say it is a likelihood of p being a particular value given the sample--it represents, in some sense, the relative plausibility of p being some value for the observations we made.

For instance, suppose $n = 5$ and the sample was $\boldsymbol{x} = (1, 1, 0, 1, 1)$. Intuitively we would conclude that p is more likely to be closer to 1 than to 0, because we observed more ones. Indeed, we have

$$L(p \mid \boldsymbol{x}) = p^4(1-p).$$

If we plot this function on $p \in [0, 1]$, we can see how the likelihood confirms our intuition. Of course, we do not know the true value of p--it could have been $p = 0.25$ rather than $p = 0.8$, but the likelihood function tells us that the former is much less likely than the latter. But if we want to determine a *probability* that p lies in a certain interval, we have to normalize the likelihood: since $\int_{p=0}^{1} p^4(1-p)\, dp = \frac{1}{30}$, it follows that in order to get a *posterior density* for p, we must multiply by 30:

$$f_p(p \mid \boldsymbol{x}) = 30p^4(1-p).$$

In fact, this posterior is a beta distribution with parameters $a = 5, b = 2$. Now the areas under the density correspond to probabilities.