


In the [CMU Machine Learning Lecture](#), likelihood function is denoted by  $P(D|\theta)$

In the [Cornell lecture note](#), likelihood function is denoted by  $P(D; \theta)$

are semicolon and vertical bar the same here?

  $f(x; \theta)$  is the same as  $f(x|\theta)$ , (simply meaning that  $\theta$  is a fixed parameter and the function  $f$  is a function of  $x$ .)  
 $f(x, \Theta)$ , OTOH, is an element of a family (set) of functions, where the elements are indexed by  $\Theta$ . A subtle distinction, perhaps, but an important one, esp. (when it comes time to estimate an unknown parameter  $\theta$  (on the basis of known data  $x$ )) (at that time,  $\theta$  varies and  $x$  is fixed, (resulting in the "likelihood function") Usage of  $"|"$  is more common among statisticians,  $";"$  among mathematicians. – [jbowman](#) Jun 20, 2012 at 19:20

What is the difference in meaning between the notation  $P(z; d, w)$  and  $P(z|d, w)$  which are commonly used in many books and papers?

I believe the origin of this is the likelihood paradigm (though I have not checked the actual historical correctness of the below, it is a reasonable way of understanding how it came to be).

Let's say in a regression setting, you would have a distribution:

$$p(Y|x, \beta)$$

Which means: the distribution of  $Y$  if you know (conditional on) the  $x$  and  $\beta$  values.

If you want to estimate the betas, you want to maximize the likelihood:

$L(\beta; y, x) = p(Y|x, \beta)$  :  $L$ 함수를  $p$ 함수로 나타낼 수 있음.

Essentially, you are now looking at the expression  $p(Y|x, \beta)$  as a function of the beta's, but apart from that, there is no difference (for mathematical correct expressions that you can properly derive, this is a necessity --- although in practice no one bothers).

Then, in bayesian settings, the difference between parameters and other variables soon fades, so one started to you use both notations intermixedly.

So, in essence: there is no actual difference: they both indicate the conditional distribution of the thing on the left, conditional on the thing(s) on the right.

Although some writers are a bit sloppy and inconsistent, technically, no, they are not the same. They are similar, but not quite identical.

The Sheffer stroke (**vertical bar**) is read "given that" or "contingent upon." Thus  $p(x|y)$  means the probability density of  $x$  *given that* some event or state of affairs  $y$  has occurred. For instance if  $x$  represents a height in centimeters, the term  $p(x|\text{woman})$  means the probability density we find the height value  $x$  *given that* the person we are measuring is a woman. This is classical terminology of *contingent probability density*. This use is restricted to probability and statistics. <sup>조건부 확률 필드</sup>

The **semicolon** is slightly different, and can apply (in cases unrelated to probability and statistics.) Suppose you have a mathematical function (that depends upon two (or more) variables, say  $f(x, y)$ .) Suppose you're primarily interested in the behavior of  $f$  based (on the value of  $x$ .) You want to take derivatives with respect to  $x$ , (limits for large and small  $x$ , count the zeros, and such.) When you write  $f(x; y)$  you're (in essence) creating a new function *OF ONE VARIABLE* ( $x$ ), which we might call  $g(x) = f(x; y)$ . This notation recognizes that there is an *implied* or chosen value of  $y$  (when you examine  $g(x)$ ), but it is not "part of the function." For instance, it is meaningless to take the derivative of  $f(x; y)$  with respect to  $y$  any more than it would to take the derivative of  $g(x)$  with respect to  $y$ . The value of  $y$  is a "setting" or a "parameter" (that is fixed.)

This notation helps avoid confusion when one asks about the intrinsic dimension of a function.

A particularly useful application of this semicolon notation in probability is in the Expectation-Maximization (or EM) Algorithm. In the algorithm some steps involve the optimization of a function (the expectation) *with respect to one of the many variables*. It is natural, then, to use  $f(x; y)$  during this step to optimize over  $x$ , where the value of  $y$  is fixed and "cannot be touched."

Another use in statistics and pattern recognition is the following. Suppose you have training data,  $x_1, x_2, \dots, x_d$ . You create the dot product with some real-valued weight vector  $\mathbf{w} = \{w_1, w_2, \dots, w_d\}$ , as in  $\mathbf{w}^t \mathbf{x} = w_1 x_1 + w_2 x_2 + \dots + w_d x_d$ . You want to optimize the value of  $\mathbf{w}$  with respect to some classification. You create a function  $h(\mathbf{w}; \mathbf{x})$ , meaning you can take derivatives with respect to the weights but it makes *NO SENSE* to take a derivative with respect to the data  $\mathbf{x}$ , nor is  $\mathbf{w}$  "contingent upon"  $\mathbf{x}$ . Nevertheless, the particular functional form of  $h$  has buried behind it the particular values of the data at hand.