**Moments**

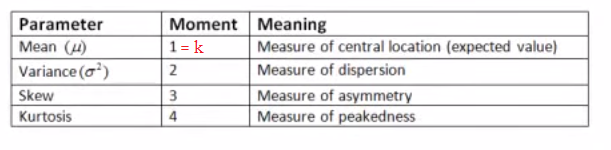
A moment of a random variable is just the expected value of a power of the random variable.

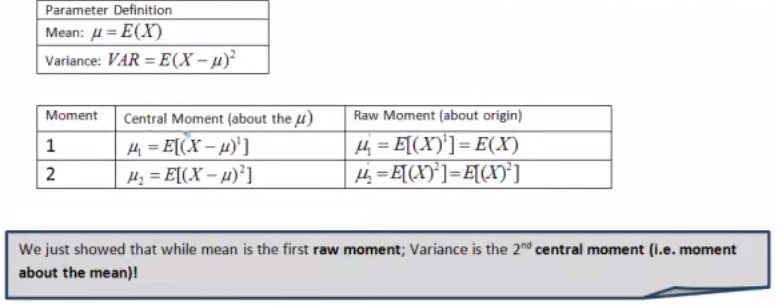
Suppose that we have a set of data with a total of n discrete points. The kth moment of the data set with values x1, x2, x3, … xn is given by the formula:

(x1k + x2k + x3k + . . . + xnk)/n

In mathematics, a moment is a specific quantitative measure, used in both mechanics and statistics, of the shape of a set of points. If the points represent probability density, then the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second moment is the variance, and the third moment is the skewness. The mathematical concept is closely related to the concept of moment in physics.

In physics the moment of a system of point masses is calculated with a formula identical to that above, and this formula is used in finding the center of mass of the points. In statistics the values are no longer masses, but as we will see, moments in statistics still measure something relative to the center of the values.

****

****

**Mean**

Mean is just the average value of the set of numbers

****



**Variance**

Variance measures how far a set of numbers is spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data points tend to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data points are very spread out around the mean and from each other.

V(X)=E(X^2)-E(X)^2

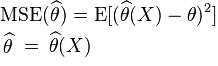


\operatorname{var}(\widehat{\theta}) = \operatorname{E}[(\widehat{\theta} - \operatorname{E}(\widehat{\theta}) )^2]

**Mean squared error**

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated.

The [mean squared error](https://en.wikipedia.org/wiki/Mean_squared_error) of \widehat{\theta} is defined as the expected value (probability-weighted average, over all samples) of the squared errors; that is,



It is used to indicate how far, on average, the collection of estimates are from the single parameter being estimated. Consider the following analogy. Suppose the parameter is the bull's-eye of a target, the estimator is the process of shooting arrows at the target, and the individual arrows are estimates (samples). Then high MSE means the average distance of the arrows from the bull's-eye is high, and low MSE means the average distance from the bull's-eye is low. The arrows may or may not be clustered.

**Estimator**

In statistics, an estimator is a rule for calculating an estimate of a given quantity based on observed data: thus the rule (the estimator), the quantity of interest (the estimand) and its result (the estimate) are distinguished.

Suppose there is a fixed *parameter*  \theta \  that needs to be estimated. Then an "estimator" is a function that maps the [sample space](https://en.wikipedia.org/wiki/Sample_space) to a set of *sample estimates*. An estimator of  \theta \  is usually denoted by the symbol \widehat{\theta}. It is often convenient to express the theory using the [algebra of random variables](https://en.wikipedia.org/wiki/Algebra_of_random_variables): thus if *X* is used to denote a [random variable](https://en.wikipedia.org/wiki/Random_variable) corresponding to the observed data, the estimator (itself treated as a random variable) is symbolised as a function of that random variable, \widehat{\theta}(X). The estimate for a particular observed dataset (i.e. for *X*=*x*) is then \widehat{\theta}(x), which is a fixed value. Often an abbreviated notation is used in which \widehat{\theta} is interpreted directly as a random variable, but this can cause confusion.

Estimand  \theta \  -> Estimator(Estimate) \widehat{\theta} = \widehat{\theta}(X) -> Estimate \widehat{\theta}(x)

**Bias of an estimator**

The [*bias*](https://en.wikipedia.org/wiki/Bias_of_an_estimator) of \widehat{\theta} is defined as . It is the distance between the average of the collection of estimates, and the single parameter being estimated. It also is the expected value of the error, since . If the parameter is the bull's-eye of a target, and the arrows are estimates, then a relatively high absolute value for the bias means the average position of the arrows is off-target, and a relatively low absolute bias means the average position of the arrows is on target. They may be dispersed, or may be clustered. The relationship between bias and variance is analogous to the relationship between [accuracy and precision](https://en.wikipedia.org/wiki/Accuracy_and_precision).

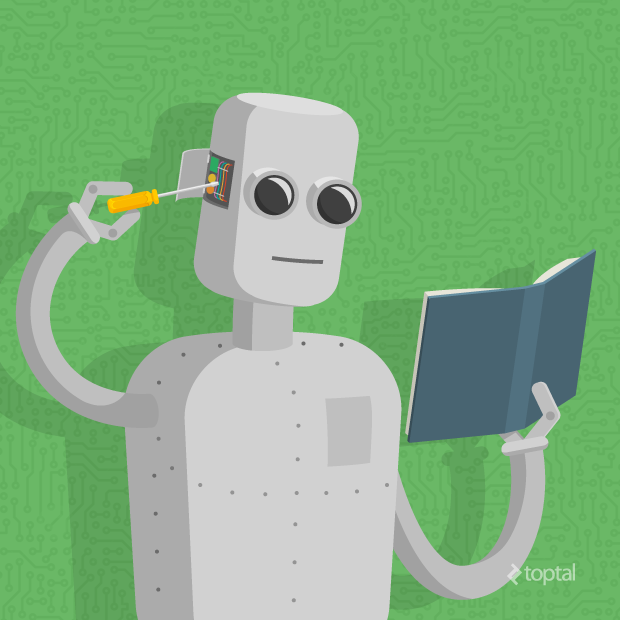
The estimator \widehat{\theta} is an [*unbiased estimator*](https://en.wikipedia.org/wiki/Estimator_bias) of  \theta \  [if and only if](https://en.wikipedia.org/wiki/If_and_only_if) .

**Accuracy Vs. Precision**

In the fields of science, engineering, industry, and statistics, the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value. The precision of a measurement system, related to reproducibility and repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. Although the two words precision and accuracy can be synonymous in colloquial use, they are deliberately contrasted in the context of the scientific method.

A measurement system can be accurate but not precise, precise but not accurate, neither, or both. For example, if an experiment contains a systematic error, then increasing the sample size generally increases precision but does not improve accuracy. The result would be a consistent yet inaccurate string of results from the flawed experiment. Eliminating the systematic error improves accuracy but does not change precision.

**Machine Learning + NLP + Image Processing**



**Artificial intelligence** (**AI**) is the [intelligence](http://en.wikipedia.org/wiki/Intelligence) exhibited by machines or software. It is an academic [field of study](http://en.wikipedia.org/wiki/Field_of_study) which studies the goal of creating intelligence. Major AI researchers and textbooks define this field as "the study and design of intelligent agents",in which an [intelligent agent](http://en.wikipedia.org/wiki/Intelligent_agent) is a system that perceives its environment and takes actions that maximize its chances of success.

**Data mining** is the computational process of discovering patterns in large [data sets](http://en.wikipedia.org/wiki/Data_set) involving methods at the intersection of [artificial intelligence](http://en.wikipedia.org/wiki/Artificial_intelligence), [machine learning](http://en.wikipedia.org/wiki/Machine_learning), [statistics](http://en.wikipedia.org/wiki/Statistics), and [database systems](http://en.wikipedia.org/wiki/Database_system). The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

* Every time a credit card or a store loyalty card is being used, or a warranty card is being filled, data is being collected about the users behavior. Many people find the amount of information stored about us from companies, such as Google, Facebook, and Amazon, disturbing and are concerned about privacy. Although there is the potential for our personal data to be used in harmful, or unwanted, ways it is also being used to make our lives better.

*Machine Learning means to build computer algorithms to improve their performance through experience or examples (training sets). System itself can learn from changing environment and adapt the changes. Machine learning is used in wide area including data mining, medical diagnosis, vision, face, pattern recognition, signal processing & control theory in electrical engineering, network optimization in telecommunication, artificial intelligence, ANN, robotics, psychology and neurobiology in bioinformatics.*

**Machine learning** is a [scientific discipline](http://en.wikipedia.org/wiki/Academic_disciplines) that explores the construction and study of [algorithms](http://en.wikipedia.org/wiki/Algorithm) that can [learn](http://en.wikipedia.org/wiki/Learning) from data. Such algorithms operate by building a [model](http://en.wikipedia.org/wiki/Mathematical_model) from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

Machine learning is a type of artificial intelligence ([AI](http://searchcio.techtarget.com/definition/AI)) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

**Other Definitions**

***Arthur Samuel*** *described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." (This is an older, informal definition)*

***Tom Mitchell*** *provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

*Example: playing checkers.*

* *E = the experience of playing many games of checkers*
* *T = the task of playing checkers.*
* *P = the probability that the program will win the next game.*

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension - as is the case in data mining applications - machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and adjust program actions accordingly. For example, Facebook's News Feed changes according to the user's personal interactions with other users. If a user frequently tags a friend in photos, writes on his wall or "likes" his links, the News Feed will show more of that friend's activity in the user's News Feed due to presumed closeness.

(With Machine Learning we are improving our applications actions through extracting data patterns)

**Applications**

* Creating intelligent Robots, Who can learn from human’s actions. In order to copy them.
* Email system intelligently distinguish our mail from spam and filter them each time we mark a particular mail as a spam.
* Web Page Ranking
* Photo Tagging.

**Supervised Learning**

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Supervised learning problems are categorized into **"regression"** and **"classification"** problems. In a **regression** problem, we are trying to predict results within a **continuous** output, meaning that we are trying to map input variables to some **continuous** function. In a **classification** problem, we are instead trying to predict results in a **discrete** output. In other words, we are trying to map input variables into **discrete** categories.

**Example:**

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a *continuous* output, so this is a regression problem.

We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two *discrete* categories.

**Unsupervised Learning**

Unsupervised learning, on the other hand, allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables. (The problem of **unsupervised learning** is that of trying to find hidden structure in unlabeled data.)

We can derive this structure by **clustering** the data based on relationships among the variables in the data.

With unsupervised learning there is no feedback based on the prediction results, i.e., there is no teacher to correct you. It’s not just about clustering. For example, associative memory is unsupervised learning.

**Example:**

*Clustering*: Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number that are somehow similar or related by different variables, such as word frequency, sentence length, page count, and so on.

*Associative*: Suppose a doctor over years of experience forms associations in his mind between patient characteristics and illnesses that they have. If a new patient shows up then based on this patient’s characteristics such as symptoms, family medical history, physical attributes, mental outlook, etc the doctor associates possible illness or illnesses based on what the doctor has seen before with similar patients. This is not the same as rule based reasoning as in expert systems. In this case we would like to estimate a mapping function from patient characteristics into illnesses.

**Linear Regression with One Variable**

**linear regression** is an approach for modeling the relationship between a scalar [dependent variable](http://en.wikipedia.org/wiki/Dependent_variable) *y* and one or more [explanatory variables](http://en.wikipedia.org/wiki/Explanatory_variable) (or independent variable) denoted *X*. The case of one explanatory variable is called [*simple linear regression*](http://en.wikipedia.org/wiki/Simple_linear_regression). For more than one explanatory variable, the process is called *multiple linear regression*.

Univariate linear regression is used when you want to predict a single output value from a single input value. We're doing supervised learning here, so that means we already have an idea what the input/output cause and effect should be.

**The Hypothesis Function**

Our hypothesis function has the general form:

*hθ*(*x*)=*θ*0+*θ*1*x*

We give to hθ values for θ0 and θ1 to get our output 'y'. In other words, we are trying to create a function called hθ that is able to reliably map our input data (the x's) to our output data (the y's).

Example:

|  |  |
| --- | --- |
| **x (input)** | **y (output)** |
| 0 | 4 |
| 1 | 7 |
| 2 | 7 |
| 3 | 8 |

Now we can make a random guess about our *hθ* function: *θ*0=2 and *θ*1=2. The hypothesis function becomes *hθ*(*x*)=2+2*x*.

So for input of 1 to our hypothesis, y will be 4. This is off by 3.

***Mean Squared Error****(****MSE****)*

*In*[*statistics*](http://en.wikipedia.org/wiki/Statistics)*, the****mean squared error****(****MSE****) of an*[*estimator*](http://en.wikipedia.org/wiki/Estimator)*measures the*[*average*](http://en.wikipedia.org/wiki/Expected_value)*of the squares of the "errors", that is, the difference between the estimator and what is estimated. MSE is a*[*risk function*](http://en.wikipedia.org/wiki/Risk_function)*, corresponding to the*[*expected value*](http://en.wikipedia.org/wiki/Expected_value)*of the squared error loss or quadratic loss. The difference occurs because of*[*randomness*](http://en.wikipedia.org/wiki/Randomness)*or because the estimator*[*doesn't account for information*](http://en.wikipedia.org/wiki/Omitted-variable_bias)*that could produce a more accurate estimate.*

*If  is a vector of n predictions, and  is the vector of the true values, then the (estimated) MSE of the predictor is:*

**

**Cost Function**

We can measure the accuracy of our hypothesis function by using a cost function. This takes an average (actually a fancier version of an average) of all the results of the hypothesis with inputs from x's compared to the actual output y's.



To break it apart, it is  where  is the mean of the squares of *hθ*(*x*(*i*))−*y*(*i*)), or the difference between the predicted value and the actual value.

This function is otherwise called the "Squared error function", or [*Mean squared error*](https://eventing.coursera.org/api/redirectStrict/sFSQcO8lJd_c7ow3BhRZhBkecyMNCSQ3io20zQ33dzSDT3m0zS1wIo4Gr8doPAPSzqPKm1QHyYuHQ3XBrrVlEg.yIavjrOZMpfy2zta7zYgtA.eueFbBA7wyL7Th-U8HNuly-g9F-2PBF2wuAQwqc4815m-x_5hM0H3mLGLBCzThjl8MUI-xBalNM8kPr9JoU3JnO_exjW5AY959WmQ3x1VuyzTc0DPxiBFfLo-EEBX2QBZo1SL1sOEm-s8iA0ADq03BppB8mZuaIx3_kxwSbv2UmeVbu7N8qoC_eaJJ7PwmaI2z2T1XKrEzkrp9UDcUYUrkkbM2lA_IQro17-3dna8LnD_37HLu2AgfEo5zbTzBukEFqr6gDj2qsVIg5Lqi7KISI1ND4LLCRT38UEhjNMymDhh6vkTKw_BsXm5eZXkGTOqibH9sJRJ6_YCN2ueuADd7Q9QVe89lv7F6s4ruy0KxLHfON3_r2h8pB9yTjLqJ5pNl5b9cgZvkvULf0SwM6f3g). The mean is halved () as a convenience for the computation of the gradient descent, as the derivative term of the square function will cancel out the  term.

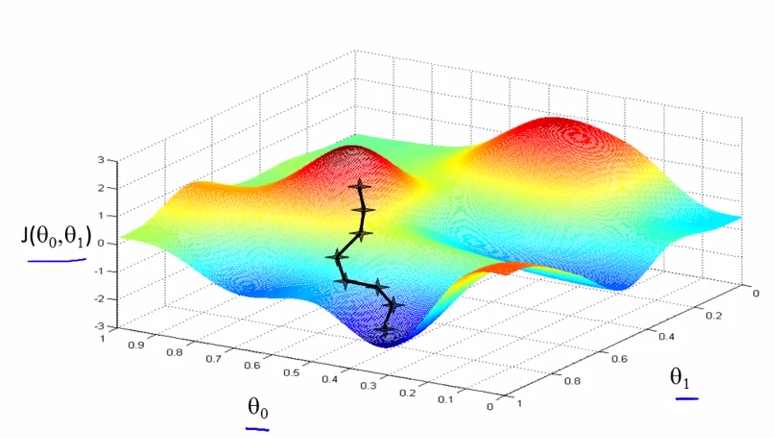
Now we are able to concretely measure the accuracy of our predictor function against the correct results we have so that we can predict new results we don't have.

## Gradient Descent

So we have our hypothesis function and we have a way of measuring how accurate it is. Now what we need is a way to automatically improve our hypothesis function. That's where gradient descent comes in.

Imagine that we graph our hypothesis function based on its fields *θ*0 and *θ*1 (actually we are graphing the cost function for the combinations of parameters). This can be kind of confusing; we are moving up to a higher level of abstraction. We are not graphing x and y itself, but the guesses of our hypothesis function.

We put *θ*0 on the x axis and *θ*1 on the z axis, with the cost function on the vertical y axis. The points on our graph will be the result of the **cost function** using our hypothesis with those specific theta parameters.



We will know that we have succeeded when our cost function is at the very bottom of the pits in our graph, i.e. when its value is the minimum.

The way we do this is by taking the **derivative** (the line tangent to a function) of our cost function. The slope of the tangent is the derivative at that point and it will give us a direction to move towards. We make steps down that derivative by the parameter *α*, called the learning rate.

The gradient descent equation is:

repeat until convergence:



for j=0 and j=1

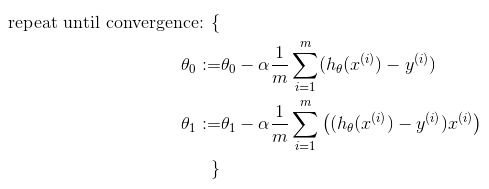
Intuitively, this could be thought of as:

repeat until convergence:



## Gradient Descent for Linear Regression

When specifically applied to the case of linear regression, a new form of the gradient descent equation can be derived. We can substitute our actual cost function and our actual hypothesis function and modify the equation to:



where *m* is the size of the training set, *θ*0 a constant that will be changing simultaneously with *θ*1 and *x*(*i*),*y*(*i*) are values of the given training set (data).

Note that we have separated out the two cases for *θj* and that for *θ*1 we are multiplying *x*(*i*) at the end due to the derivative.

The point of all this is that if we start with a guess for our hypothesis and then repeatedly apply these gradient descent equations, our hypothesis will become more and more accurate.

*It is called “supervised” because of the presence of the outcome variable to guide the learning process. In the unsupervised learning problem, we observe only the features and have no measurements of the outcome.*

*Our task is rather to describe how the data are organized or clustered.*

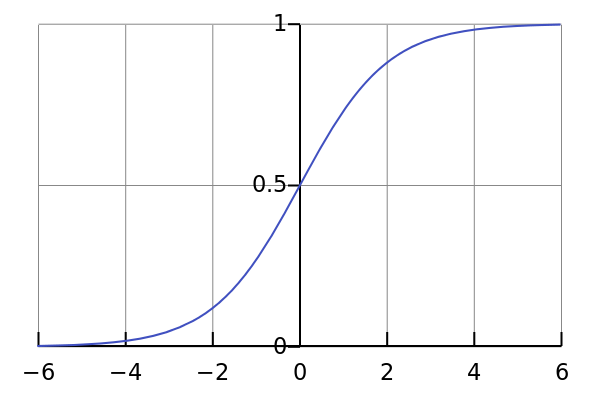
*For each there is a set of variables that might be denoted as inputs, which are measured or preset. These have some influence on one or more outputs. For each example the goal is to use the inputs to predict the values of the outputs. This exercise is called supervised learning.*

*We have used the more modern language of machine learning. In the statistical literature the inputs are often called the predictors, a term we will use interchangeably with inputs, and more classically the independent variables. In the pattern recognition literature the term features is preferred, which we use as well. The outputs are called the responses, or classically the dependent variables.*

**Logistic Regression**

The logistic function is useful because it can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one and hence is interpretable as a probability. The logistic function \sigma (t) is defined as follows:





*The logistic function \sigma (t) \in [0,1] for all t.*