Odds

Odds are a numerical expression, usually expressed as a pair of numbers, used in both gambling and statistics. In statistics, the **odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not.

Odds are expressed in the form X to Y, where X and Y are numbers. Usually, the word "to" is replaced by a symbol for ease of use. This is conventionally either a slash or hyphen, although a colon is sometimes seen. Thus, 6/1, 6-1, and 6:1 are all interchangeable.

Odds against

When the probability that the event will not happen is greater than the probability that it will, then the odds are "against" that event happening. Odds of 6 to 1, for example, are therefore sometimes said to be "6 to 1 *against*".

Odds on

"Odds on" is the opposite of "odds against". It means that the event is more likely to happen than not. This is sometimes expressed with the smaller number first (1 to 2) but more often using the word "on" ("2 to 1 *on*"), meaning that the event is twice as likely to happen as not.

The odds are a ratio of probabilities; an odds ratio is a ratio of odds, that is, a ratio of ratios of probabilities.

Example #1

There are 5 pink marbles, 2 blue marbles, and 8 purple marbles. What are the odds in favor of picking a blue marble?

Answer: The odds in favor of a blue marble are 2:13. One can equivalently say, that the odds are 13:2 *against*. There are 2 out of 15 chances in favor of blue, 13 out of 15 against blue.

In probability theory and statistics, where the variable *p* is the probability in favor of a binary event, and the probability against the event is therefore 1-*p*, "the odds" of the event are the quotient of the two, or p/1-p.

The probability that an event will occur is the fraction of times you expect to see that event in many trials. Probabilities always range between 0 and 1. The odds are defined as the probability that the event will occur divided by the probability that the event will not occur.

If the probability of an event occurring is Y, then the probability of the event not occurring is 1-Y. (Example: If the probability of an event is 0.80 (80%), then the probability that the event will not occur is 1-0.80 = 0.20, or 20%.

The odds of an event represent the ratio of the (probability that the event will occur) / (probability that the event will not occur). This could be expressed as follows:

Odds of event = Y / (1-Y)

So, in this example, if the probability of the event occurring = 0.80, then the odds are 0.80 / (1-0.80) = 0.80/0.20 = 4 (i.e., 4 to 1).

- If a race horse runs 100 races and wins 25 times and loses the other 75 times, the probability of winning is 25/100 = 0.25 or 25%, but the odds of the horse winning are 25/75 = 0.333 or 1 win to 3 loses.
- If the horse runs 100 races and wins 5 and loses the other 95 times, the probability of winning is 0.05 or 5%, and the odds of the horse winning are 5/95 = 0.0526.
- If the horse runs 100 races and wins 50, the probability of winning is 50/100 = 0.50 or 50%, and the odds of winning are 50/50 = 1 (even odds).
- If the horse runs 100 races and wins 80, the probability of winning is 80/100 = 0.80 or 80%, and the odds of winning are 80/20 = 4 to 1.

NOTE that when the probability is low, the odds and the probability are very similar.

	Diseased	**Non-diseased**	
Exposed	a	b	
Non-exposed	c	d	

Odds ratio is computed by taking the ratio of odds, where the odds in each group is computed as follows:

OR = (a/b) / (c/d)

Example:

Consider again the hypothetical pilot study on pesticide exposure and breast cancer:

	Diseased	**Non-diseased**	
Pesticide Exposure	7	10	
Non-exposed	6	57	

We noted above that

OR= (7/10) / (5/57) = 6.6

Interpretation: The odds of breast cancer in women with high DDT exposure are 6.65 times greater than the odds of breast cancer in women without high DDT exposure. **Odds** is the probability that something is so, will occur, or is more likely to occur than something else. Odds is the probability (= how likely it is) that a particular thing will or will not happen.

Ex1: If you drive a car all your life, the odds **are** that you'll have an accident at some point.

Ex2: There are heavy odds **against** people succeeding in such a bad economic climate

Probability is defined as the fraction of desired outcomes in the context of every possible outcome with a value between 0 and 1, where 0 would be an impossible event and 1 would represent an inevitable event. Probabilities are usually given as percentages. [I.e. 50% probability that a coin will land on HEADS.] Odds can have any value from zero to infinity and they represent a ratio of desired outcomes versus the field.

Odds are a ratio, and can be given in two different ways: 'odds in favor' and 'odds against'. 'Odds in favor' are odds describing the if an event will occur, while 'odds against' will describe if an event will not occur. If you are familiar with gambling, 'odds against' are what Vegas gives as odds. More on that later. For the coin flip odds in favor of a HEADS outcome is 1:1, not 50%.

Simple probability of event A occurring is mathematically defined as:

$$P(A) = \frac{Number\ of\ Event\ A}{Total\ Number\ of\ Events}$$

The best way to illustrate this is with the classic marbles-in-a-bag example. The graphic below depicts all the marbles in an opaque bag that one marble will be pulled out of. There are 6 blue, 3 red, 2 yellow, and 1 green for a total of 12 marbles in the bag.

The probability of pulling a red marble would be calculated by taking the total number of red marbles and dividing it by the total number of marbles.

$$P(RED) = \frac{3\ RED\ marbles}{12\ TOTAL\ marbles} = 25\%$$

Notice that the probability calculation includes the red marbles in the denominator of the calculation, because probability considers the context of the entire event space. Odds, on the other hand, are the ratio of favorable outcomes to unfavorable outcomes. The denominator

contains ONLY the marbles that aren't the favorable outcomes. Odds uses the contexts of good outcomes and bad outcomes. Written as fractions, these two values are completely different. Probability is 1/4 while odds in favor are 1/3. You can see how mistakenly interchanging the terms could give the wrong information. The '**odds in favor**' of RED would be mathematically calculated by

OR

$$Odds\_Favor(RED) = \frac{3\ RED\ marbles}{9\ NOT\ RED\ marbles} = 1:3$$

To find '**odds against'** you would simply flip odds in favor upside down and this describes the odds of the event not occurring.

$$Odds\_Against(RED) = \frac{9\ NOT\ RED\ marbles}{3\ RED\ marbles} = 3:1$$

Odds are used to describe the chance of an event occurring. The odds are the ratios that compare the number of ways the event can occur with the number of ways the event cannot occur.

The odds in favor - the ratio of the number of ways that an outcome can occur compared to how many ways it cannot occur.

Odds in favor = Number of successes: Number of failures

The odds against - the ratio of the number of ways that an outcome cannot occur compared to in how many ways it can occur.

Odds against = Number of failures: Number of successes

<u>Example</u>

A jewelry box contains 5 white pearl, 2 gold rings and 6 silver rings. What are the odds of drawing a white pearl from the jewelry box?

Number of successes = 5

Number of failures = 2 + 6 = 8

Numbers of ways to draw a white pearl: number of ways to draw another jewelry.

5:8

The odds are 5:8

Probability compares the number of successes to the total number of attempts made. The **odds in favor** of an event compares the number of successes to the number of failures.

An Example of Probability to Odds

In the past five seasons, crosstown football rivals the Quakers and the Comets have played each other with the Comets winning twice and the Quakers winning three times. On the basis of these outcomes, we can calculate the probability the Quakers win and the odds in favor of their winning. There was a total of three wins out of five, so the probability of winning this year is 3/5 = 0.6 = 60%. Expressed in terms of odds, we have that there were three wins for the Quakers and two losses, so the odds in favor of them winning are **3:2**.

Why Use Odds?

Probability is nice, and gets the job done, so why do we have an alternate way to express it? Odds can be helpful when we want to compare how much larger one probability is relative to another. An event with a probability 75% has odds of 75 to 25. We can simplify this to 3 to 1. This means that the event is three times more likely to occur than not occur.

Probability to express the chance that an event of interest occurs. So a probability of 0.1, or 10% risk, means that there is a 1 in 10 chance of the event occurring. The usual way of thinking about probability is that if we could repeat the experiment or process under consideration a large number of times, the fraction of experiments where the event occurs should be close to the probability (e.g. 0.1).

The odds of an event of interest occurring is defined by odds = p/(1-p) where p is the probability of the event occurring. So if p=0.1, the odds are equal to 0.1/0.9=0.111 (recurring). So here the probability (0.1) and the odds (0.111) are quite similar. Indeed whenever p is small, the probability and odds will be similar. This is because when p is small, 1-p is approximately 1, so that p/(1-p) is approximately equal to p.

But when p is not small, the probability and odds will generally be quite different. For example if p=0.5, we have odds=0.5/0.5=1. As p increases, the odds get larger and larger. For example, with p=0.99, odds=0.99/0.01=99.

For example, odds of 9 to 1 against, said as "nine to one against", and written as 9/1 or 9:1, means the event of interest will occur once for every 9 times that the event does not occur. That is in 10 times/replications, we expect the event of interest to happen once and the event not to happen in the other 9 times. Using odds to express probabilities is useful in a gambling

setting because it readily allows one to calculate how much one would win - with odds of 9/1 you will win 9 for a bet of 1 (assuming your bet comes good!).

Odds ratios

In the statistics world odds ratios are frequently used to express the relative chance of an event happening under two different conditions

For example, in the context of a clinical trial comparing an existing treatment to a new treatment, we may compare the odds of experiencing a bad outcome if a patient takes the new treatment to the odds of a experiencing a bad outcome if a patient takes the existing treatment.

Suppose that the probability of a bad outcome is 0.2 if a patient takes the existing treatment, but that this is reduced to 0.1 if they take the new treatment. The odds of a bad outcome with the existing treatment is 0.2/0.8=0.25, while the odds on the new treatment are 0.1/0.9=0.111 (recurring). The odds ratio comparing the new treatment to the old treatment is then simply the correspond ratio of odds: (0.1/0.9) / (0.2/0.8) = 0.111 / 0.25 = 0.444 (recurring). This means that the odds of a bad outcome if a patient takes the new treatment are 0.444 that of the odds of a bad outcome if they take the existing treatment. The odds (and hence probability) of a bad outcome are reduced by taking the new treatment. We could also express the reduction by saying that the odds are reduced by approximately 56%, since the odds are reduced by a factor of 0.444.

Odds of an event happening is defined as the likelihood that an event will occur, expressed as a proportion of the likelihood that the event will not occur. Therefore, if A is the probability of subjects affected and B is probability of subjects not affected, then odds = A /B.

Therefore, the odds of rolling four on a dice are 1/5 or 20%.

Odds Ratio (OR) is a measure of association between exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

HOW DO I INTERPRET ODDS RATIOS IN LOGISTIC REGRESSION?

When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables.

From probability to odds to log of odds

Everything starts with the concept of probability. Let's say that the probability of success of some event is .8. Then the probability of failure is 1 – .8 = .2. The odds of success are defined as the ratio of the probability of success over the probability of failure. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1.

The transformation from probability to odds is a monotonic transformation, meaning the odds increase as the probability increases or vice versa. Probability ranges from 0 and 1. Odds range from 0 and positive infinity. Below is a table of the transformation from probability to odds and we have also plotted for the range of p less than or equal to .9.

p	odds
.001	.001001
.01	.010101
.15	.1764706
.2	.25
.25	.3333333
.3	.4285714
.35	.5384616
.4	.6666667
.45	.8181818
.5	1
.55	1.222222
.6	1.5
.65	1.857143
.7	2.333333

The transformation from odds to log of odds is the log transformation. Again this is a monotonic transformation. That is to say, the greater the odds, the greater the log of odds and

vice versa. The table below shows the relationship among the probability, odds and log of odds. We have also shown the plot of log odds against odds.

p	odds	logodds
.001	.001001	-6.906755
.01	.010101	-4.59512
.15	.1764706	-1.734601
.2	.25	-1.386294
.25	.3333333	-1.098612
.3	.4285714	-.8472978
.35	.5384616	-.6190392
.4	.6666667	-.4054651
.45	.8181818	-.2006707
.5	1	0
.55	1.222222	.2006707
.6	1.5	.4054651
.65	1.857143	.6190392
.7	2.333333	.8472978

Why do we take all the trouble doing the transformation from probability to log odds? One reason is that it is usually difficult to model a variable which has restricted range, such as probability. This transformation is an attempt to get around the restricted range problem. It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity. Another reason is that among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret. This transformation is called logit transformation.

A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables.

transformed probability as a linear relationship with the predictor variables. More formally, let Y be the binary outcome variable indicating failure/success with $\{0,1\}$ and p be the probability of y to be 1, $p = P(Y = 1)$. Let $x_1, \cdots, x_k$ be a set of predictor variables. Then the logistic regression of Y on $x_1, \cdots, x_k$ estimates parameter values for $\beta_0, \beta_1, \cdots, \beta_k$ via maximum likelihood method of the following equation

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Exponentiate and take the multiplicative inverse of both sides,

$$\frac{1-p}{p} = \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Partial out the fraction on the left-hand side of the equation and add one to both sides,

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Change 1 to a common denominator,

$$\frac{1}{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + 1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Finally, take the multiplicative inverse again to obtain the formula for the probability $P(Y = 1)$,

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

The data set has 200 observations and the outcome variable used will be **hon**, indicating if a student is in an honors class or not. So, our p = prob (**hon**=1).

Logistic regression with no predictor variables

Let's start with the simplest logistic regression, a model without any predictor variables. In an equation, we are modeling

logit(p)= β_0

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -111.35502                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   intercept |   -1.12546   .1644101    -6.85   0.000    -1.447697   -.8032217
------------------------------------------------------------------------------
```

This means log(p/(1-p)) = -1.12546. What is p here? It turns out that p is the overall probability of being in honors class (hon = 1). Let's take a look at the frequency table for **hon**.

hon	Freq.	Percent	Cum.
0	151	75.50	75.50
1	49	24.50	100.00
Total	200	100.00	

So p = 49/200 = .245. The odds are .245/(1-.245) = .3245 and the log of the odds (logit) is log (.3245) = -1.12546. In other words, the intercept from the model with no predictor variables is the estimated log odds of being in honors class for the whole population of interest. We can also transform the log of the odds back to a probability: p = exp (-1.12546)/ (1+exp(-1.12546)) = .245, if we like.

Logistic regression with a single dichotomous predictor variables

Now let's go one step further by adding a binary predictor variable, **female**, to the model. Writing it in an equation, the model describes the following linear relationship.

$$\text{logit}(p) = \beta_0 + \beta_1 * \textbf{female}$$

```
Logistic regression                             Number of obs   =        200
                                                LR chi2(1)      =       3.10
                                                Prob > chi2     =     0.0781
Log likelihood = -109.80312                     Pseudo R2       =     0.0139

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5927822   .3414294     1.74   0.083    -.0764072    1.261972
   intercept |  -1.470852   .2689555    -5.47   0.000    -1.997995   -.9437087
------------------------------------------------------------------------------
```

Before trying to interpret the two parameters estimated above, let's take a look at the crosstab of the variable **hon** with **female**.

hon	female: male	female	Total
0	74	77	151
1	17	32	49
Total	91	109	200

In our dataset, what are the odds of a male being in the honors class and what are the odds of a female being in the honors class? We can manually calculate these odds from the table: for

males, the odds of being in the honors class are (17/91)/(74/91) = 17/74 = .23; and for females, the odds of being in the honors class are (32/109)/(77/109) = 32/77 = .42. The ratio of the odds for female to the odds for male is (32/77)/(17/74) = (32*74)/(77*17) = 1.809. So the odds for males are 17 to 74, the odds for females are 32 to 77, and the odds for female are about 81% higher than the odds for males.

Now we can relate the odds for males and females and the output from the logistic regression. The intercept of -1.471 is the log odds for males since male is the reference group (**female** = 0). Using the odds, we calculated above for males, we can confirm this: log (.23) = -1.47. The coefficient for **female** is the log of odds ratio between the female group and male group: log (1.809) = .593. So, we can get the odds ratio by exponentiating the coefficient for female. Most statistical packages display both the raw regression coefficients and the exponentiated coefficients for logistic regression models.

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       3.10
                                                  Prob > chi2     =     0.0781
Log likelihood = -109.80312                       Pseudo R2       =     0.0139

------------------------------------------------------------------------------
         hon | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.809015   .6176508     1.74   0.083     .9264389    3.532379
------------------------------------------------------------------------------
```

Let's begin with probability. Probabilities range between 0 and 1. Let's say that the probability of success is .8, thus

p = .8

Then the probability of failure is

q = 1 – p = .2

Odds are determined from probabilities and range between 0 and infinity. Odds are defined as the ratio of the probability of success and the probability of failure. The odds of success are

odds(success) = p/(1-p) or p/q = .8/.2 = 4,

that is, the odds of success are 4 to 1. The odds of failure would be

odds(failure) = q/p = .2/.8 = .25.

This looks a little strange but it is really saying that the odds of failure are 1 to 4. The odds of success and the odds of failure are just reciprocals of one another, i.e., 1/4 = .25 and 1/.25 = 4. Next, we will add another variable to the equation so that we can compute an odds ratio.

Another example

This example is adapted from Pedhazur (1997). Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted. The probabilities for admitting a male are,

p = 7/10 = .7 q = 1 – .7 = .3

If you are male, the probability of being admitted is 0.7 and the probability of not being admitted is 0.3.

Here are the same probabilities for females,

p = 3/10 = .3 q = 1 – .3 = .7

If you are female it is just the opposite, the probability of being admitted is 0.3 and the probability of not being admitted is 0.7.

Now we can use the probabilities to compute the odds of admission for both males and females,

Odds (male) = .7/.3 = 2.33333 odds(female) = .3/.7 = .42857

Next, we compute the odds ratio for admission,

OR = 2.3333/.42857 = 5.44

Thus, for a male, the odds of being admitted are 5.44 times as large as the odds for a female being admitted.

```
Logistic regression                               Number of obs   =         20
                                                  LR chi2(1)      =       3.29
                                                  Prob > chi2     =     0.0696
Log likelihood = -12.217286                       Pseudo R2       =     0.1187

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    1.694596   .9759001     1.74   0.082    -.2181333    3.607325
       _cons |   -.8472979   .6900656    -1.23   0.220    -2.199801    .5052058
------------------------------------------------------------------------------
```

Logistic regression is in reality an ordinary regression using the logit as the response variable. The logit transformation allows for a linear relationship between the response variable and the coefficient.

This means that the coefficients in a simple logistic regression are in terms of the log odds, that is, the coefficient 1.694596 implies that a one unit change in gender results in a 1.694596 unit change in the log of the odds.

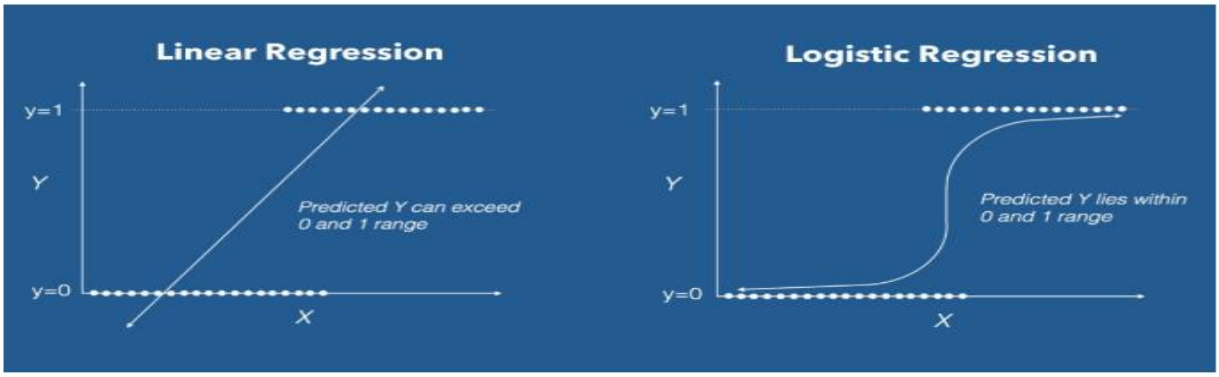
In logistic regression, the dependent variable is a *logit*, which is the natural log of the odds, that is,

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find

logit(P) = a + bX,

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.



Logistic Regression Assumptions:

· Binary logistic regression requires the dependent variable to be binary.
· For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
· Only the meaningful variables should be included.
· The independent variables should be independent of each other. That is, the model should have little or no multi-collinearity.
· The independent variables are linearly related to the log odds.
· Logistic regression requires quite large sample sizes.

Linear versus Logistic Regression

■Linear Regression	■Logistic Regression
■Target is an interval variable.	■Target is a discrete (binary or ordinal) variable.
■Input variables have any measurement level.	■Input variables have any measurement level.
■Predicted values are the mean of the target variable at the given values of the input variables.	■Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables.

Logistic Regression Equation:

The underlying algorithm of Maximum Likelihood Estimation (MLE) determines the regression coefficient for the model that accurately predicts the probability of the binary dependent variable. The algorithm stops when the convergence criterion is met or maximum number of iterations are reached. Since the probability of any event lies between 0 and 1 (or 0% to 100%), when we plot the probability of dependent variable by independent factors, it will demonstrate an 'S' shape curve.

Logit Transformation is defined as follows-

Logit = Log (p/1-p) = log (probability of event happening/ probability of event not happening) = log (Odds)

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (GLM). The fundamental equation of generalized linear model is:

$g(E(y)) = \alpha + \beta x1 + \gamma x2$

Here, g () is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α, β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Key Points:

1. GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
2. The dependent variable need not to be normally distributed.
3. It does not uses OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
4. Errors need to be independent but not normally distributed.

To understand, consider the following example:

We are provided a sample of 1000 customers. We need to predict the probability whether a customer will buy (**y**) a particular magazine or not. As we've a categorical outcome variable, we'll use logistic regression.

To start with logistic regression, first write the simple linear regression equation with dependent variable enclosed in a link function:

$g(y) = \beta o + \beta(Age)$ — — **(a)**

For understanding, consider '*Age*' as independent variable.

In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g () is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure(1-p). p should meet following criteria:

1. It must always be positive (since p >= 0)

2. It must always be less than equals to 1 (since p <= 1)

Now, simply satisfy these 2 conditions and get to the core of logistic regression. To establish link function, we denote g () with 'p' initially and eventually end up deriving this function.

Since probability must always be positive, we'll put the linear equation in exponential form. For any value of slope and dependent variable, exponent of this equation will never be negative.

p = exp(βo + β(Age)) = e^(βo + β(Age)) — — — — **(b)**

To make the probability less than 1, divide p by a number greater than p. This can simply be done by:

p = exp(βo + β(Age)) / exp(βo + β(Age)) + 1 = e^(βo + β(Age)) / e^(βo + β(Age)) + 1 — — — **©**

Using (a), (b) and ©, we can redefine the probability as:

p = e^y/ 1 + e^y — — **(d)**

where p is the probability of success. This (d) is the Logit Function

If p is the probability of success, 1-p will be the probability of failure which can be written as:

q = 1 — p = 1 — (e^y/ 1 + e^y) — — (e)

where q is the probability of failure

On dividing, (d) / (e), we get,

p/(1-p) = e^y

After taking *log* on both side, we get,

log(p/(1-p)) = y

log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.

After substituting value of y, we'll get:

$\log(p/(1-p)) = \beta o + \beta(Age)$

This is the equation used in Logistic Regression. Here (p/1-p) is the odd ratio. Whenever the log of odd ratio is found to be positive, the probability of success is always more than 50%.

A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

In the logistic regression the constant (*b0*) moves the curve left and right and the slope (*b1*) defines the steepness of the curve. Just as ordinary least square regression is the method used to estimate coefficients for the best fit line in linear regression, logistic regression uses **maximum likelihood estimation (MLE)** to obtain the model coefficients that relate predictors to the target. After this initial function is estimated, the process is repeated until **LL** (Log Likelihood) does not change significantly.

Performance of Logistic Regression Model (Performance Metrics):

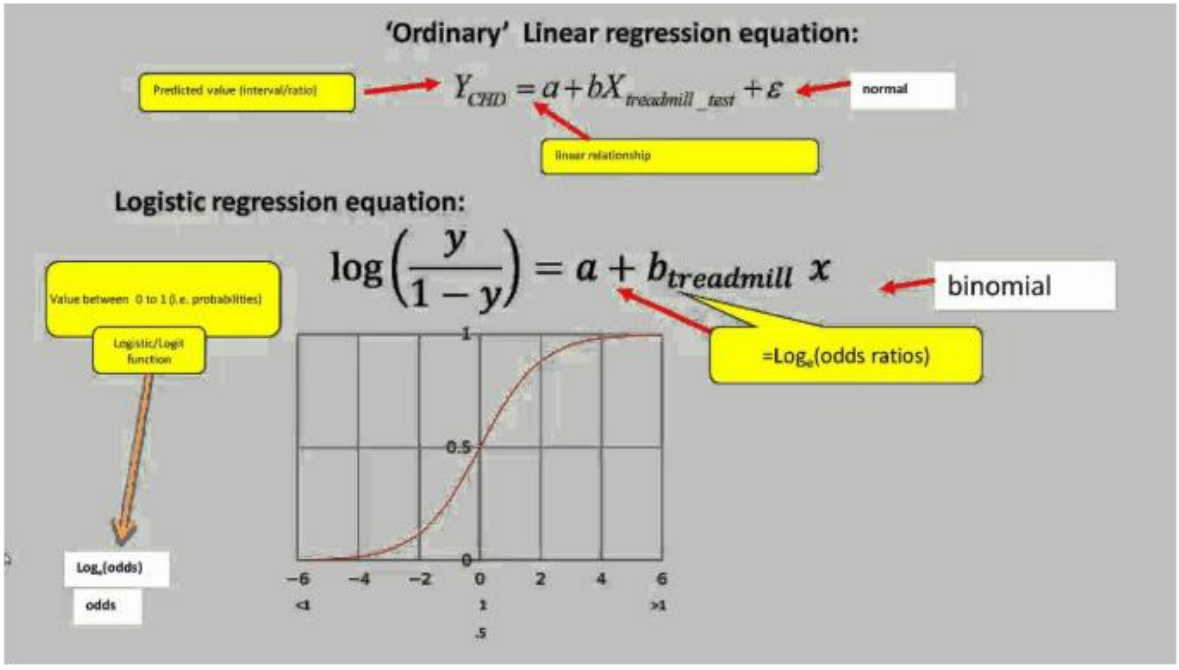
To evaluate the performance of a logistic regression model, we must consider few metrics. Irrespective of tool (SAS, R, Python) we would work on, always look for:

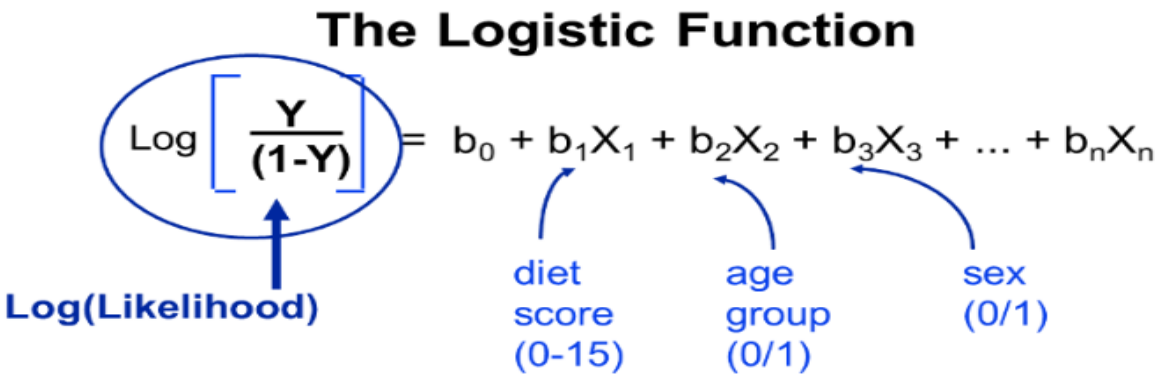
1. **AIC (Akaike Information Criteria)** — The analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

2. **Null Deviance and Residual Deviance** — Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

3. **Confusion Matrix:** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid over-fitting.

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. We can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a **logit** function.





It is widely used for **classification problems**

· Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio

· To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression

· It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square

· The independent variables should not be correlated with each other i.e. **no multi collinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.

· If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**

· If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

Odds and Odds ratio

We will consider a data-set that tells us about depending on the gender, whether a customer will purchase a product or not. We import and check the data-set

We will create a table of frequency of 'yes' and 'no' depending on the gender, using crosstab feature of pandas. The table will be of great use to understand odds and odds ratio later on.

Purchase	No	Yes
Gender		
Female	106	159
Male	125	121

We're now ready to define **Odds,** which describes the ratio of success to ratio of failure. Considering females group, we see that probability that a female will purchase (success) the product is = 159/265 (yes/total number of females). Probability of failure (no purchase) for female is 106/265. In this case the odds is defined as (159/265)/(106/265) = 1.5. **Higher the odds, better is the chance for success**. Range of odds can be any number between *[0 , ∞].* What

happens to the range if we take a natural logarithm of such numbers? log(x) is defined for x≥0 but the range varies from [-∞, ∞].

Odds ratio, which as the name suggests, is the ratio of odds. Considering the example above, **Odds ratio**, represents which group (male/female) has better odds of success, and it's given by calculating the ratio of odds for each group. So odds ratio for females= odds of successful purchase by female / odds of successful purchase by male = (159/106)/(121/125). Odds ratio for males will be the reciprocal of the above number.

We can appreciate clearly that while odds ratio can vary between 0 to positive infinity, log (odds ratio) will vary between [-∞, ∞]. Specifically when odds ratio lies between [0,1], log (odds ratio) is negative.

In linear regression where feature variables can take any values, the output (label) can thus be continuous from negative to positive infinity.

$$Y = a + b_i X_i ; -\infty \leq Y \leq \infty, \ -\infty \leq X_i \leq \infty \qquad (1.1)$$

Range of label and feature in linear regression case

Since logistic regression is about classification, i.e Y is a categorical variable. It's clearly not possible to achieve such output with linear regression model (eq. 1.1), since the range on both sides do not match. Our aim is to transform the LHS in such a way that it matches the range of RHS, which is governed by the range of feature variables, [-∞, ∞].

We will follow some intuitive steps to search how it's possible to achieve such outcome.

$$P = a + b_i X_i ; 0 \leq P \leq 1, \ -\infty \leq X_i \leq \infty \qquad (1.2)$$

- For linear regression, both X and Y ranges from minus infinity to positive infinity. Y in logistic is categorical, or for the problem above it takes either of the two distinct values 0,1. First, we try to predict probability using the regression model. Instead of two distinct values now the LHS can take any values from 0 to 1 but still the ranges differ from the RHS.

$$\frac{P}{1-P} = O = a + b_i X_i ; 0 \leq O \leq \infty, \ -\infty \leq X_i \leq \infty \qquad (1.3)$$

- I discussed above that odds and odds ratio ratio varies from [0, ∞]. This is better than probability (which is limited between 0 and 1) and one step closer to match the range of RHS.
- Many of you have already understood that if we now consider a natural logarithm on LHS of (eq. 1.3) then the ranges on both side matches.

$$ln\,(O) = a + b_i X_i ; \ -\infty \leq ln\,(O) \leq \infty, \ -\infty \leq X_i \leq \infty \qquad (1.4)$$

With this, **we have achieved a regression model, where the output is natural logarithm of the odds , also known as logit**. The base of the logarithm is not important but taking logarithm of odds is.

We can retrieve the probability of success from eq. 1.4 as below.

$$O = e^{a+b_i X_i} ; \frac{P}{1-P} = e^{a+b_i X_i} ; P = \frac{1}{1+e^{-(a+b_i X_i)}} \qquad (1.5)$$

From odds to probability where probability distribution resembles a sigmoid function

If you see the RHS of equation 1.5., which is also known as logistic function, is very similar to the sigmoid function. Just like in linear regression where the constant term denotes the intercept on the Y axis (hence a shift along Y axis), here for logistic function, the constant term shifts the *s* curve along the X axis.

Most importantly we see that the dependent variable in logistic regression follows Bernoulli distribution having an unknown probability P. Therefore, the logit i.e. log of odds, links the independent variables (***Xs***) to the Bernoulli distribution.

In logistic regression the coefficients derived from the model (e.g., b_1) indicate the change in the expected log odds relative to a one unit change in X_1, holding all other predictors constant. Therefore, the antilog of an estimated regression coefficient, $\exp(b_i)$, produces an odds ratio.

Likelihood

In **statistics**, the **likelihood** function (often simply called the **likelihood**) expresses how probable a given set of observations is for different values of the statistical parameters.

Many probability distributions have unknown parameters; We estimate these unknowns using sample data. The Likelihood function gives us an idea of *how well* the data summarizes these parameters.

Although a likelihood function might look just like a probability density function, it's fundamentally different. A probability density function is a function of x, your data point, and it will tell you how likely it is that certain data points appear. A likelihood function, on the other hand, takes the data set as a given, and represents the likeliness of different parameters for your distribution.

Likelihood is an informal way of discussing the likeliness that something will happen, without specific reference to numerical probability.

Likelihood function (often simply called the **likelihood**) expresses the plausibilities of different parameter values for a given sample of data. While not to be interpreted as a probability, it is equal to the joint probability distribution of a random sample. However, whereas the latter is a density function defined on the sample space for a particular choice of parameter values, the likelihood function is defined on the parameter space while the random variable is fixed at the given observations.

- Probability is the percentage that a success occurs. For example, we do the binomial experiment by tossing a coin. We suppose that the event that we get the face of coin in success, so the probability of success now is 0.5 because the probability of face and back of a coin is equal. **0.5 is the probability of a success.**
- Likelihood is the conditional probability. The same example, we toss the coin 10 times, and we suppose that we get 7 success (show the face) and 3 failed (show the back). The likelihood is calculated (for binomial distribution, it can be varying depend on the distributions)
  - L(0.5|7)= 10C7 * 0.5^7 * (1-0.5)^3 = 0.1171

Meaning: 0.1171 is the probability that the **above event will happen** (got 7 successes out of 10 trials) by knowing that the probability of one success is **0.5** (toss one time).

Therefore,

- Likelihood is the probability (conditional probability) of an event (a set of success) occur by knowing the probability of a success occur.

- Probability is the percentage that a success occurs.

Likelihood is the probability that an event that has already occurred would yield a specific outcome. Probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes.

Probability is used when describing a function of the outcome given a fixed parameter value. For example, if a coin is flipped 10 times and it is a fair coin, what is the probability of it landing heads-up every time?

Likelihood is used when describing a function of a parameter given an outcome. For example, if a coin is flipped 10 times and it has landed heads-up 10 times, what is the likelihood that the coin is fair?

So, the likelihood of a set of parameter values, θθ given outcomes xx is given by
$L(\theta|x)=P(x|\theta)$

Suppose you have a probability model with parameters θ.
p(x | θ) has two names.
It can be called the **probability of x** (given θ),
or the **likelihood of θ** (given that x was observed).

The likelihood is a function of θ. Here are a couple of simple uses:

If you observe x and want to estimate the θ that gave rise to it, the maximum-likelihood principle says to choose the maximum-likelihood θ -- in other words, the θ that maximizes p(x | θ).

Likelihood is a specific probability. Given some data and a set of parameters for a model, likelihood means the probability of getting the data given the model parameters.

Example: You roll a die twice and get 1 each time. If your model assumes equal probability of the die landing on each side, there is 1/6 for each, so the likelihood is 1/6 * 1/6 = 1/36. However, if you assume that the die is rigged so there's a 10% chance of rolling 2 through 6 and 50% for rolling 1, then the likelihood is 1/4.

In a probability distribution, the **model parameters are known and the data set is to be found**. For example, in a binomial distribution, you would know the number of trials and the probability of success in each trial. Based on this, you can find the probability of occurrence of a particular data set, say a fixed number of successes.

In a likelihood function, **the data set is given and the model parameters are to be found**. For example, in a binomial likelihood function, you would know the number of successes and would be asked to figure out either the number of trials and or the probability of success in each trial.

Assume we have a probability distribution with density f(x). (It could also be a probability mass function, but for this discussion let's assume it is a density). f(x)dx is the probability that a draw from this distribution lands in a small neighborhood around x.

f(x) may be characterized by some other parameters. For example, if **f** is a Gaussian density, it is characterized by the mean μ and standard deviation σ. So instead of f(x), we could write f(x,μ,σ). In general, we can pack all such parameters into a single vector θ and write the density as
f(x,θ). When we view **f** as a density, we have some constant values of θ in mind and think of the function as varying in x.

When we think of f(x,θ) as a **likelihood**, we instead hold xx constant and let θ vary.

A common application of the likelihood function is in estimation. In this case, θ is unknown and we want to estimate it from some given data x. A standard approach is **maximum likelihood** estimation: estimate θ by the value which maximizes f(x,θ) for the given observations x.

What is log likelihood in logistic regression

It is the sum of the likelihood residuals. At record level, the natural log of the error (residual) is calculated for each record, multiplied by minus one, and those values are totaled. That total is then used as the basis for deviance (2 x ll) and likelihood (exp(ll)).

The log-likelihood is the expression that Minitab maximizes to determine optimal values of the estimated coefficients (β).

Log-likelihood values cannot be used alone as an index of fit because they are a function of sample size but can be used to compare the fit of different coefficients. Because you want to maximize the log-likelihood, the higher value is better. For example, a log-likelihood value of -3 is better than -7.

```
Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =      71.05
                                                  Prob > chi2     =     0.0000
Log likelihood =  -80.11818                       Pseudo R2       =     0.3072

------------------------------------------------------------------------------
     honcomp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.482498   .4473993     3.31   0.001     .6056111    2.359384
        read |   .1035361   .0257662     4.02   0.000     .0530354    .1540369
     science |   .0947902   .0304537     3.11   0.002      .035102    .1544784
       _cons |   -12.7772    1.97586    -6.47   0.000    -16.64982   -8.904589
------------------------------------------------------------------------------
```

Iteration Log

```
Iteration 0:   log likelihood = -115.64441
Iteration 1:   log likelihood = -84.558481
Iteration 2:   log likelihood = -80.491449
Iteration 3:   log likelihood = -80.123052
Iteration 4:   log likelihood = -80.118181
Iteration 5:[a] log likelihood = -80.11818
```

a. This is a listing of the log likelihoods at each iteration. (Remember that logistic regression uses maximum likelihood, which is an iterative procedure.) The fi iteration (called iteration 0) is the log likelihood of the "null" or "empty" model; that is, a model with no predictors. At the next iteration, the predictor(s) are included in the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have "converged", the iterating is stopped and the results are displayed. For more information on this

Parameter Estimates

honcomp[g]	Coef.[h]	Std. Err.[i]	z[j]	P>|z|[j]	[95% Conf. Interval][k]	
female	1.482498	.4473993	3.31	0.001	.6056111	2.359384
read	.1035361	.0257662	4.02	0.000	.0530354	.1540369
science	.0947902	.0304537	3.11	0.002	.035102	.1544784
_cons	-12.7772	1.97586	-6.47	0.000	-16.64982	-8.904589

g. **honcomp** – This is the dependent variable in our logistic regression. The variables listed below it are the independent variables.

h. **Coef.** – These are the values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is

log(p/1-p) = b0 + b1*female + b2*read + b3*science

where p is the probability of being in honors composition. Expressed in terms of the variables used in this example, the logistic regression equation is

log(p/1-p) = -12.7772 + 1.482498*female + .1035361*read + 0947902*science

These estimates tell you about the relationship between the independent variables and the dependent variable, where the dependent variable is on the logit scale. These estimates tell the amount of increase in the predicted log odds of honcomp = 1 that would be predicted by a 1 unit increase in the predictor, holding all other predictors constant. Note: For the independent variables which are not significant, the coefficients are not significantly different from 0, which should be taken into account when interpreting the coefficients. (See the columns with the z-values and p-values regarding testing whether the coefficients are statistically significant). Because these coefficients are in log-odds units, they are often difficult to interpret, so they are often converted into odds ratios. You can do this by hand by exponentiating the coefficient, or by using the **or** option with **logit** command, or by using the **logistic** command.

female – The coefficient (or parameter estimate) for the variable **female** is 1.482498. This means that for a one-unit increase in **female** (in other words, going from male to female), we expect a 1.482498 increase in the log-odds of the dependent variable **honcomp**, holding all other independent variables constant.

read – For every one-unit increase in reading score (so, for every additional point on the reading test), we expect a .1035361 increase in the log-odds of **honcomp**, holding all other independent variables constant.

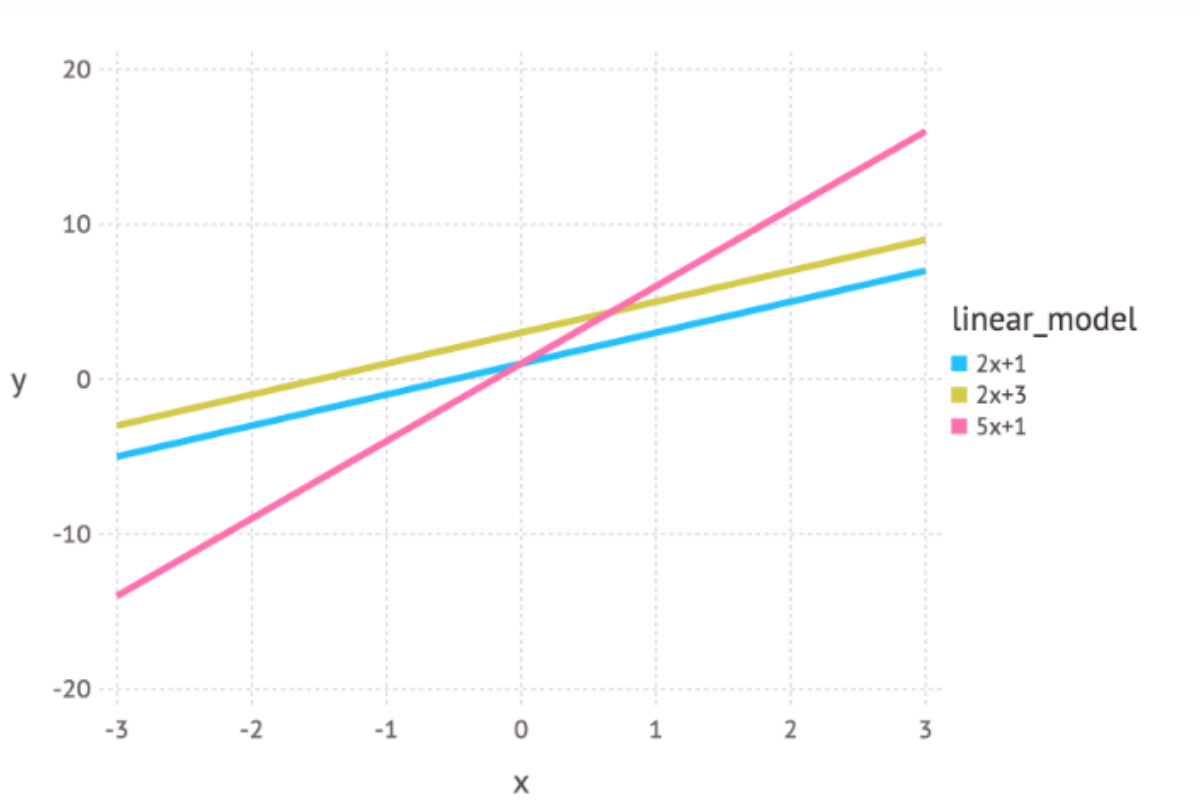
science – For every one-unit increase in science score, we expect a .0947902 increase in the log-odds of **honcomp**, holding all other independent variables constant.

constant – This is the expected value of the log-odds of **honcomp** when all of the predictor variables equal zero. In most cases, this is not interesting. Also, oftentimes zero is not a realistic value for a variable to take.

What are parameters?

Often in machine learning we use a model to describe the process that results in the data that are observed. Each model contains its own set of parameters that ultimately defines what the model looks like.

For a linear model we can write this as $y = mx + c$. In this example x could represent the advertising spend and y might be the revenue generated. m and c are parameters for this model. Different values for these parameters will give different lines (see figure below).



Three linear models with different parameter values.

So parameters define a blueprint for the model. It is only when specific values are chosen for the parameters that we get an instantiation for the model that describes a given phenomenon.

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.

Ex: Let's suppose we have observed 10 data points from some process. For example, each data point could represent the length of time in seconds that it takes a student to answer a specific exam question.

We first have to decide which model we think best describes the process of generating the data. For these data we'll assume that the data generation process can be adequately described by a Gaussian (normal) distribution. Gaussian distribution is plausible because most of the 10 points are clustered in the middle with few points scattered to the left and the right.

Gaussian distribution has 2 parameters. The mean, μ, and the standard deviation, σ. Different values of these parameters result in different curves (just like with the straight lines above). **We**

want to know *which curve was most likely responsible for creating the data points that we observed?*

Maximum likelihood estimation is a method that will find the values of μ and σ that result in the curve that best fits the data. The values that we find using Maximum likelihood estimation are called the maximum likelihood estimates (MLE).

Again, we'll demonstrate this with an example. Suppose we have three data points this time and we assume that they have been generated from a process that is adequately described by a Gaussian distribution. These points are 9, 9.5 and 11. ***How do we calculate the maximum likelihood estimates of the parameter values of the Gaussian distribution μ and σ****?*

What we want to calculate is the total probability of observing all of the data, i.e. the joint probability distribution of all observed data points. To do this we would need to calculate some conditional probabilities, which can get very difficult. So it is here that we'll make our first assumption. *The assumption is that each data point is generated independently of the others*. This assumption makes the math's much easier. If the events (i.e. the process that generates the data) are independent, then the total probability of observing all of data is the product of observing each data point individually (i.e. the product of the marginal probabilities).

The probability density of observing a single data point *x that* is generated from a Gaussian distribution is given by:

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The semi colon used in the notation *P(x; μ, σ)* is there to emphasize that the symbols that appear after it are parameters of the probability distribution. So, it shouldn't be confused with a conditional probability (which is typically represented with a vertical line e.g. *P (A| B)).*

In our example the total (joint) probability density of observing the three data points is given by:

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

We just have to figure out the values of μ and σ that results in giving the maximum value of the above expression.

If you've covered calculus in your math's classes then you'll probably be aware that there is a technique that can help us find maxima (and minima) of functions. It's called *differentiation.* All we have to do is find the derivative of the function, set the derivative function to zero and then rearrange the equation to make the parameter of interest the subject of the equation. And voilà, we'll have our MLE values for our parameters

The log likelihood

The above expression for the total probability is actually quite a pain to differentiate, so it is almost always simplified by taking the natural logarithm of the expression. This is absolutely fine because the natural logarithm is a monotonically increasing function. This means that if the value on the x-axis increases, the value on the y-axis also increases. This is important because it ensures that the maximum value of the log of the probability occurs at the same point as the original probability function. Therefore, we can work with the simpler log-likelihood instead of the original likelihood.

Taking logs of the original expression gives us:

$$\ln(P(x; \mu, \sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

This expression can be simplified again using the laws of logarithms to obtain:

$$\ln(P(x; \mu, \sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

This expression can be differentiated to find the maximum. In this example we'll find the MLE of the mean, μ. To do this we take the partial derivative of the function with respect to μ, giving

Finally, setting the left hand side of the equation to zero and then rearranging for μ gives:

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

And there we have our maximum likelihood estimate for μ. We can do the same thing with σ too but I'll leave that as an exercise for the keen reader.

So why maximum likelihood and not maximum probability?

Well this is just statisticians being pedantic (but for good reason). Most people tend to use probability and likelihood interchangeably but statisticians and probability theorists distinguish between the two. The reason for the confusion is best highlighted by looking at the equation.

$$L(\mu, \sigma; data) = P(data; \mu, \sigma)$$

These expressions are equal! So what does this mean? Let's first define *P (data; μ, σ)*? It means *"**the probability density of observing the data with model parameters μ and σ**"***.* It's worth noting that we can generalize this to any number of parameters and any distribution.

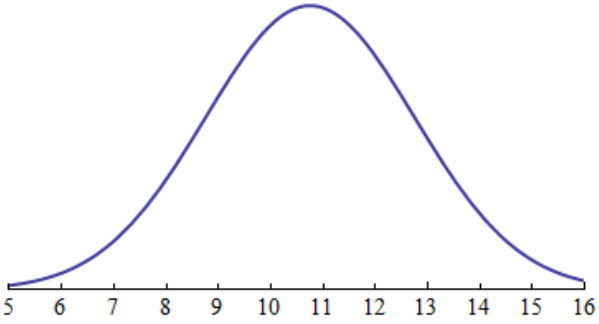
On the other hand *L(μ, σ; data)* means *"***the likelihood of the parameters μ and σ taking certain values given that we've observed a bunch of data***."*

The equation above says that the probability density of the data given the parameters is equal to the likelihood of the parameters given the data. But despite these two things being equal, the likelihood and the probability density are fundamentally asking different questions — one is asking about the data and the other is asking about the parameter values. This is why the method is called maximum likelihood and not maximum probability.

Probability

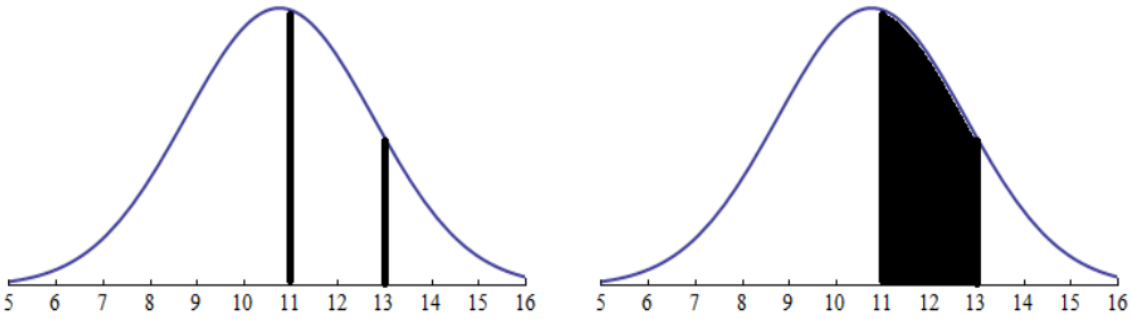
Probability is the measure of the likelihood that an event will occur. The basic idea is out of all given occurrences, what is the certainty that a specific event will occur?

Let us say we have a normal distribution graph of the average marks of students in a surprise test. (This concept will apply to all continuous distributions).



here, 5/16 is the lowest marks scored and 16/16 is the highest.

Now, the probability that a randomly selected student will have marks between 11–13 marks is the area under the curve between those 2 points.



In this case, the area under the curve is around **0.31 or a 31%** chance of the randomly selected student having marks between 11 and 13 marks.

mathematically,

P(marks between 11 and 13 marks | mean=11 and std = 3) = 0.31

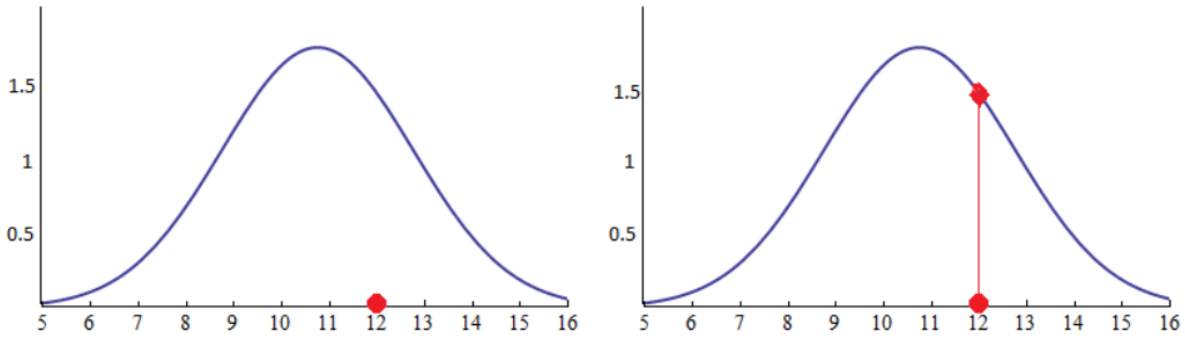
To find out the probability of marks less than 8 marks,

P(marks < 8 | mean=11 and std = 3)

Likelihood

A L**ikelihood function** (often simply a **likelihood**) is a function of parameters within the parameter space that describes the probability of obtaining the observed data.

So for this, we choose a student whose marks we already know. Say, we choose a student who scored 12 marks.



So in this scenario,

L(mean=11 and std = 3 | student scored 12 marks) = 1.48

so, 1.48 is the probability of obtaining the student scoring 12 marks within this parameter space.

In summary

Probabilities are the areas under fixed distribution

P(data|distribution)

Likelihoods are the y-axis values for fixed data points with distributions that can be moved.

L(distribution| data)

Finally, **Probability quantifies anticipation (of outcome), likelihood quantifies trust (in the model).**

Maximum Likelihood Estimation (Analytics Vidhya)

The variable is not normally distributed and is asymmetric and hence it violates the assumptions of linear regression. A popular way is to transform the variable with log, sqrt, reciprocal, etc. so that the transformed variable is normally distributed and can be modelled with linear regression.
If None of these (log, sqrt etc.) are close to a normal distribution. 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?

This is where **Maximum Likelihood Estimation (MLE)** has such a major advantage.

While studying stats and probability, you must have come across problems like – What is the probability of x > 100, given that x follows a normal distribution with mean 50 and standard deviation (sd) 10. In such problems, we already know the distribution (normal in this case) and its parameters (mean and sd) but in real life problems these quantities are unknown and must be estimated from the data. MLE is the technique which helps us in determining the parameters of the distribution that best describe the given data.

Let's understand this with an example: Suppose we have data points representing the weight (in kgs) of students in a class. The data points are shown in the figure below (the R code that was used to generate the image is provided as well):

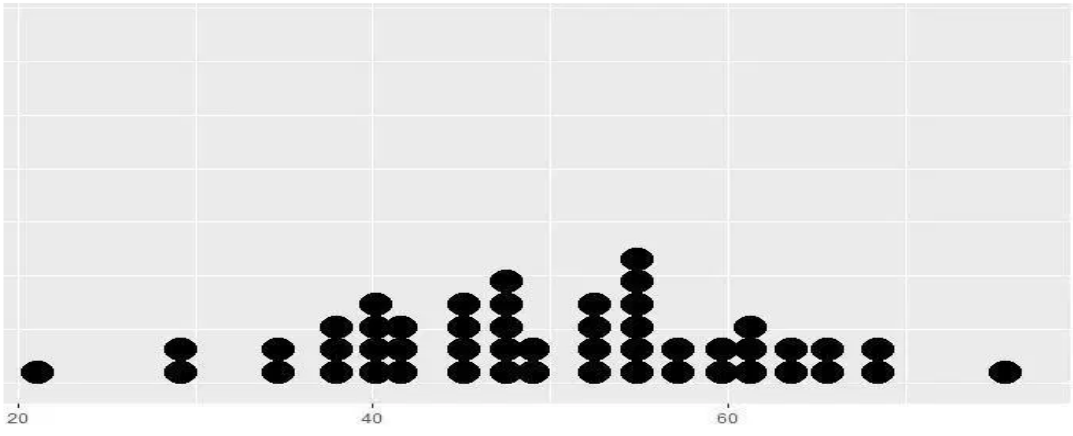


Figure 1

This appears to follow a normal distribution. But how do we get the mean and standard deviation (sd) for this distribution? One way is to directly compute the mean and sd of the given data, which comes out to be 49.8 Kg and 11.37 respectively. These values are a good representation of the given data but may not best describe the population.

We can use MLE in order to get more robust parameter estimates. *Thus,* ***MLE can be defined as a method for estimating population parameters (such as the mean and variance for Normal, rate (lambda) for Poisson, etc.) from sample data such that the probability (likelihood) of obtaining the observed data is maximized.***

Distribution Parameters

Let us first understand distribution parameters. Wikipedia's definition of this term is as follows: "It is a quantity that indexes a family of probability distributions". It can be regarded as a numerical characteristic of a population or a statistical model.

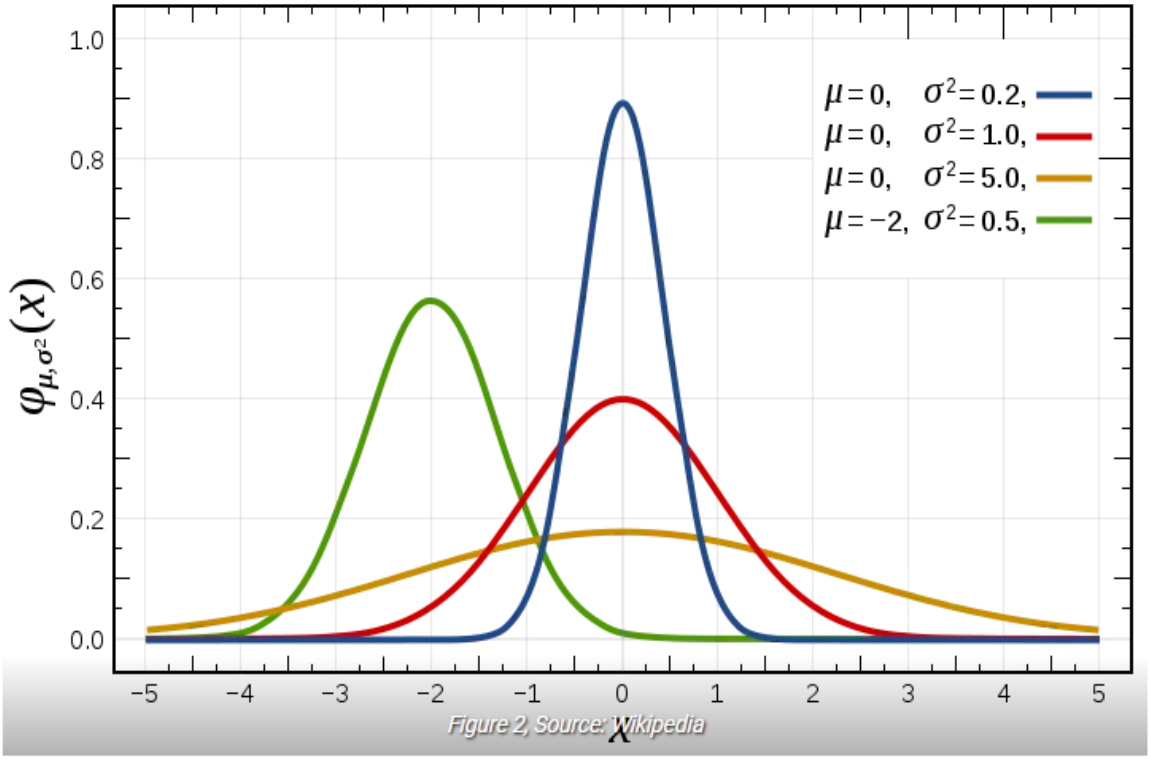


Figure 2, Source: Wikipedia

The width and height of the bell curve is governed by two parameters – mean and variance. These are known as distribution parameters for normal distribution. Similarly, Poisson distribution is governed by one parameter – lambda, which is the number of times an event occurs in an interval of time or space.

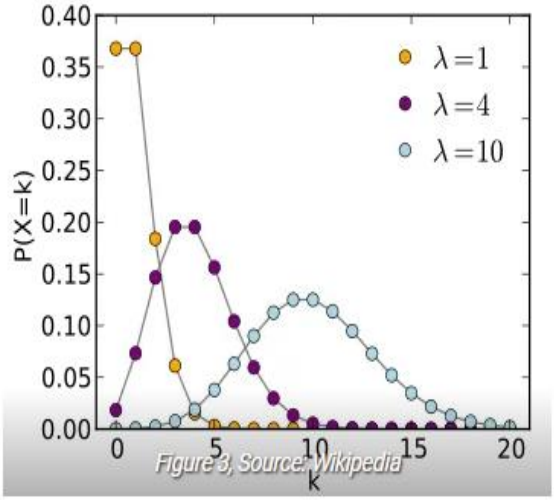


Figure 3, Source: Wikipedia

Most of the distributions have one or two parameters, but some distributions can have up to 4 parameters, like a 4 parameter beta distribution.

Likelihood

From Fig. 2 and 3 we can see that given a set of distribution parameters, some data values are more probable than other data. From Fig. 1, we have seen that the given data is more likely to occur when the mean is 50, rather than 100. In reality however, we have already observed the data. Accordingly, we are faced with an inverse problem: *Given the observed data and a model of interest, we need to find the one Probability Density Function/Probability Mass Function (f(x|ϑ)), among all the probability densities that are most likely to have produced the data*.

To solve this inverse problem, we define the likelihood function by reversing the roles of the data vector x and the (distribution) parameter vector θ in f(x| θ), i.e.,

$L(\theta;x) = f(x|\ \theta)$

In MLE, we can assume that we have a likelihood function L(θ;x), where θ is the distribution parameter vector and x is the set of observations. We are interested in finding the value of θ that maximizes the likelihood with given observations (values of x).

Log Likelihood

The mathematical problem at hand becomes simpler if we assume that the observations (xi) are independent and identically distributed random variables drawn from a Probability

Distribution, *f0 (where f0 = Normal Distribution for example in Fig.1)*. This reduces the Likelihood function to:

$$L(\theta; x) = f_0(x_1, x_2, x_3, \ldots x_n \mid \theta) = f_0(x_1 \mid \theta) \cdot f_0(x_2 \mid \theta) \cdot f0(x3 \mid \theta) \ldots . f_0(x_n \mid \theta)$$

To find the maxima/minima of this function, we can take the derivative of this function w.r.t θ and equate it to 0 (as zero slope indicates maxima or minima). Since we have terms in product here, we need to apply the chain rule which is quite cumbersome with products. **A clever trick would be to take log of the likelihood function and maximize the same.** This will convert the product to sum and since log is a strictly increasing function, it would not impact the resulting value of θ. So we have:

$$LL(\theta; x) = \log[f_0(x_1 \mid \theta) \cdot f_0(x_2 \mid \theta) \cdot f0(x3 \mid \theta) \ldots . f_0(x_n \mid \theta)]$$

$$= \log(f_0(x_1 \mid \theta) + \log(f_0(x_2 \mid \theta) + \ldots . + \log(f_0(x_n \mid \theta)$$

Maximizing the Likelihood

To find the maxima of the log likelihood function LL(θ; x), we can:

- Take first derivative of LL(θ; x) function w.r.t θ and equate it to 0
- Take second derivative of LL(θ; x) function w.r.t θ and confirm that it is negative

MLE:

Density estimation is the problem of estimating the probability distribution for a sample of observations from a problem domain.

There are many techniques for solving density estimation, although a common framework used throughout the field of machine learning is maximum likelihood estimation. Maximum likelihood estimation involves defining a likelihood function for calculating the conditional probability of observing the data sample given a probability distribution and distribution parameters. This approach can be used to search a space of possible distributions and parameters.

Problem of Probability Density Estimation

A common modeling problem involves how to estimate a joint probability distribution for a dataset.

For example, given a sample of observation (*X*) from a domain (*x1, x2, x3, …, xn*), where each observation is drawn independently from the domain with the same probability distribution (so-called independent and identically distributed, i.i.d., or close to it).

Density estimation involves selecting a probability distribution function and the parameters of that distribution that best explain the joint probability distribution of the observed data (*X*).

- How do you choose the probability distribution function?
- How do you choose the parameters for the probability distribution function?

This problem is made more challenging as sample (*X*) drawn from the population is small and has noise, meaning that any evaluation of an estimated probability density function and its parameters will have some error.

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 main difference is that MLE assumes that all solutions are equally likely beforehand, whereas MAP allows prior information about the form of the solution to be harnessed.

Maximum Likelihood Estimation

One solution to probability density estimation is referred to as Maximum Likelihood Estimation, or MLE for short.

Maximum Likelihood Estimation involves treating the problem as an optimization or search problem, where we seek a set of parameters that results in the best fit for the joint probability of the data sample (*X*).

First, it involves defining a parameter called *theta* that defines both the choice of the probability density function and the parameters of that distribution. It may be a vector of numerical values whose values change smoothly and map to different probability distributions and their parameters.

In Maximum Likelihood Estimation, we wish to maximize the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters, stated formally as:

- P(X | theta)

This conditional probability is often stated using the semicolon (;) notation instead of the bar notation (|) because *theta* is not a random variable, but instead an unknown parameter. For example:

- P(X ; theta)

or

- P(x1, x2, x3, …, xn ; theta)

This resulting conditional probability is referred to as the likelihood of observing the data given the model parameters and written using the notation *L()* to denote the likelihood function. For example:

- L(X ; theta)

The objective of Maximum Likelihood Estimation is to find the set of parameters (*theta*) that maximize the likelihood function, e.g. result in the largest likelihood value.

- maximize L(X ; theta)

We can unpack the conditional probability calculated by the likelihood function.

Given that the sample is comprised of n examples, we can frame this as the joint probability of the observed data samples *x1, x2, x3, …, xn* in *X* given the probability distribution parameters (*theta*).

- L(x1, x2, x3, …, xn ; theta)

The joint probability distribution can be restated as the multiplication of the conditional probability for observing each example given the distribution parameters.