Correlating data

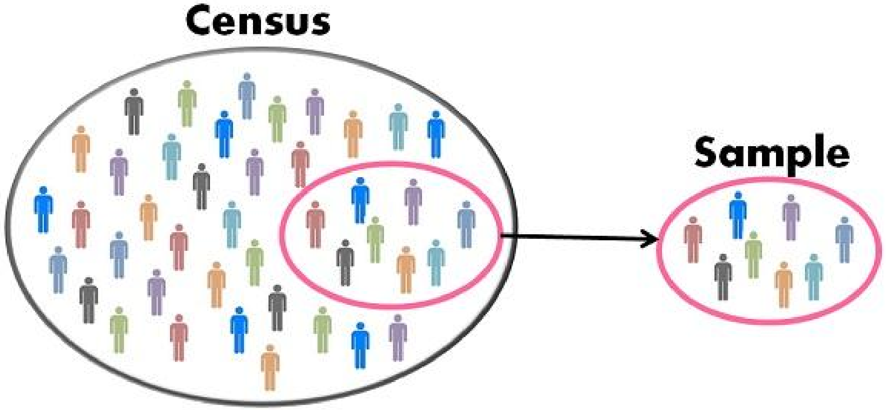
Introduction

Topics

- Surveys
- Mean/Average
- Bell/Normal/Gaussian distribution curve
- Standard Deviation
- Z-Scores
- Scatterplot graph
- Equation of a line
- Line of best fit
- Pearson's r correlation coefficient
- Regression line

Surveys

You have probably heard of a census. This is a measure of information about the population. A full census is expensive, so it is common to survey a subset of the population using a sample.



Surveys

Population distribution symbols are usually written with greek letters, and sample distribution symbols with roman letters, sampling distribution symbols are a combination of greek letters with a subscript of roman letters

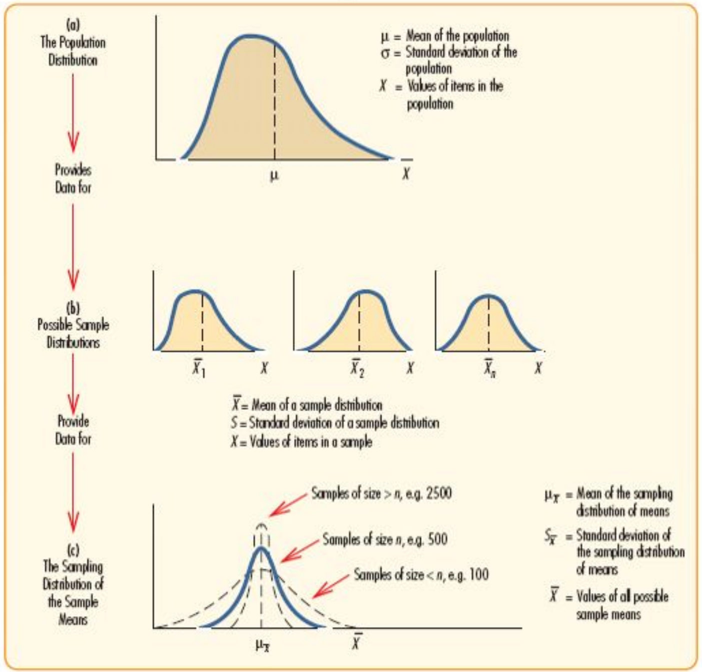
	Population	Sample	Sampling
Mean/Average	μ (mu)	$\bar{x}$ (x-bar)	$\mu_{\bar{x}}$
Standard deviation	σ (sigma)	s (s)	$\sigma_{\bar{x}}$
Number/Size	N (nu)	n (n)	

Fundamental types of distributions

- Population distribution
- Sample distribution(s)
- Sampling distribution

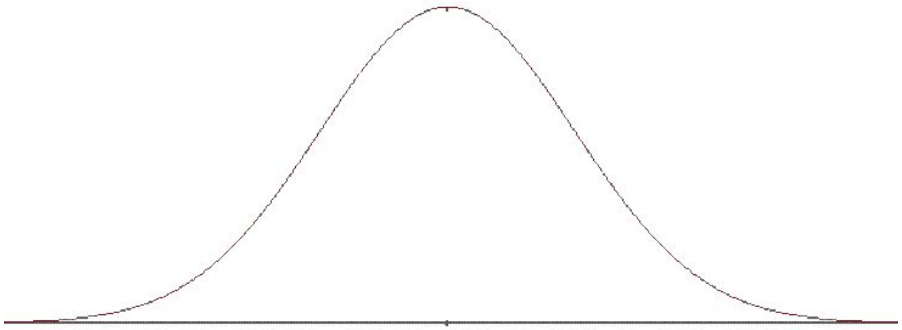
Even if the population or sample distributions are not normally distributed, the sampling distribution will be normally distributed, thanks to the central limit theorem.

EXHIBIT 12.6 Schematic of the three fundamental types of distributions[8]



Normal distribution curve

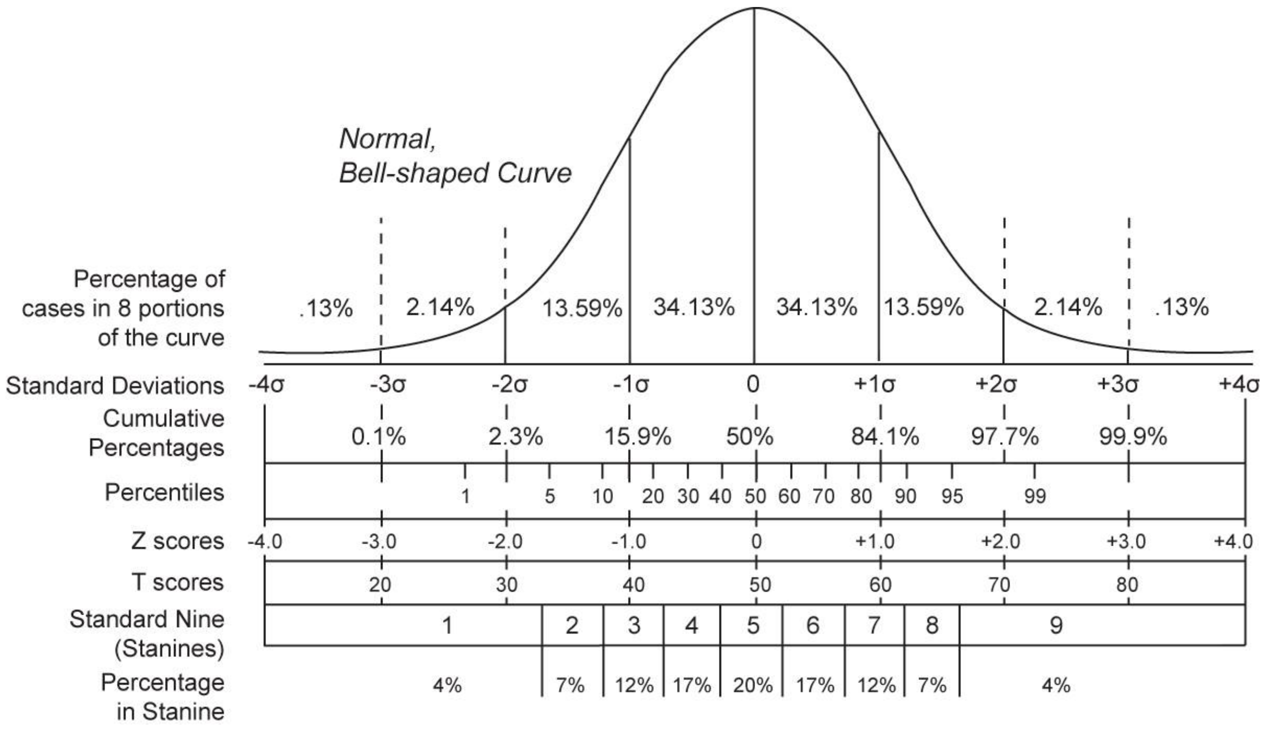
The normal distribution is useful because of the central limit theorem. In its most general form, under some conditions, the central limit theorem states that average of samples of observations independently drawn from separate independent distributions become normally distributed when the number of observations is sufficiently large.



$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu &= \text{ Mean} \\ \sigma &= \text{ Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$

Normal distribution curve



Central limit theorem

Imagine you roll a single dice, the probability of predicting the outcome is equal for all values. ⅙ chance

However consider rolling 2 dice simultaneously and adding the values. Select a number between 2 and 12.

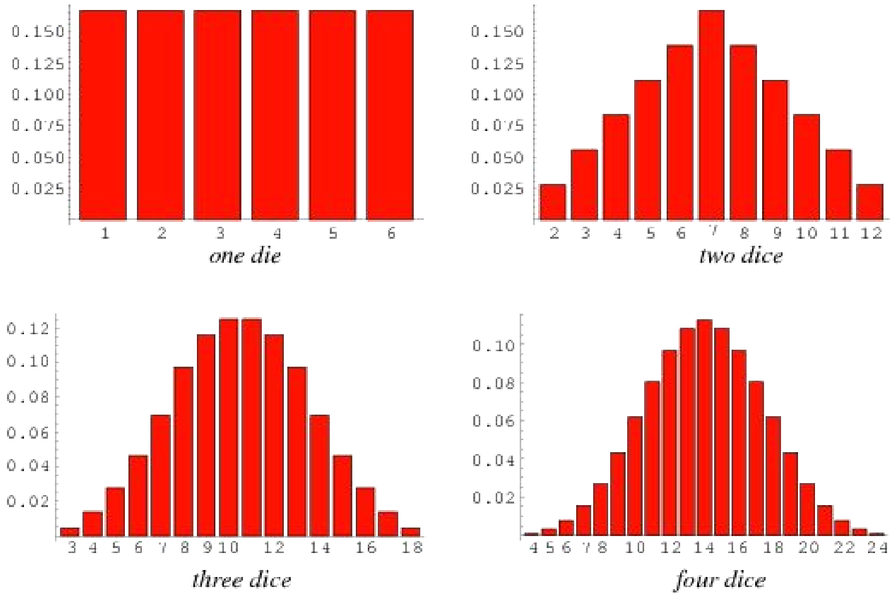
The numbers 2 and 12 can only occur with the combinations 1,1 and 6,6.

The number 10 can occur with 5,5 and 5,5 and 4,6 and 6,4.

Central limit theorem

Consider adding more dice.

As more dice are added
the average sum of the values
approaches a normal distribution.



Average

The average is defined as the sum of the values divided by the total number of values. The full sigma notation is shown below.

Sample Mean	Population Mean
$\bar{x} = \dfrac{\Sigma x}{n}$	$\mu = \dfrac{\Sigma x}{N}$

where Σx is sum of all data values

N is number of data items in population

n is number of data items in sample

$$\text{average} = \frac{\text{sum of values}}{\text{number of values}}.$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Standard deviation

The variance is (average difference)^2

The standard deviation is the sqrt((average difference)^2)

You may notice that the sample standard deviation uses n-1 and the population standard deviation uses N. This is to account for the degrees of freedom.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$n =$ The number of data points

$\bar{x} =$ The mean of the x_i

$x_i =$ Each of the values of the data

Z-score

The standard deviation is an overall measure of the variability of the data and applies to the entire data set.

The Z-score is the number of standard deviations an individual data point lies from the mean.

To calculate Z-score, simply subtract the mean from the data point and divide the result by the standard deviation.

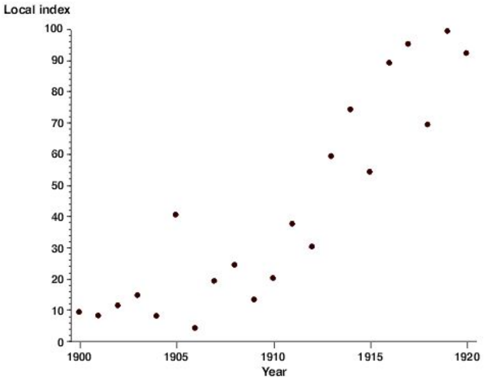
$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

Scatter plot graphs

A scatter plot graph is a type of diagram using (x,y) coordinates to display values for typically two variables of a data set.



A scatter plot can be used either when one variable is under the control of the experimenter (customarily plotted along the horizontal axis) and the other depends on it (customarily plotted along the vertical axis).

When both variables are independent either can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

Line equation

Given 2 points (x1,y1) and (x2,y2), or (-3,3) and (3,-1)

A line can be defined by the equation of a line y=mx+b and the slope can be calculated using the slope formula.

The constant b is the y intercept is where the line crosses the y-axis, where x=0. This can be done by substituting the values for the slope of the line and the coordinates of a single point (x1, y1) in the line equation, set x=0 and then solve for b.

$$y = \underset{\text{slope}}{m}x + \underset{\text{y-intercept}}{b}$$

$$\text{Slope} = m = \frac{y_2 - y_1}{x_2 - x_1}$$

Regression line equation

A regression line can be defined by its formula $\hat{Y}=bX+a$

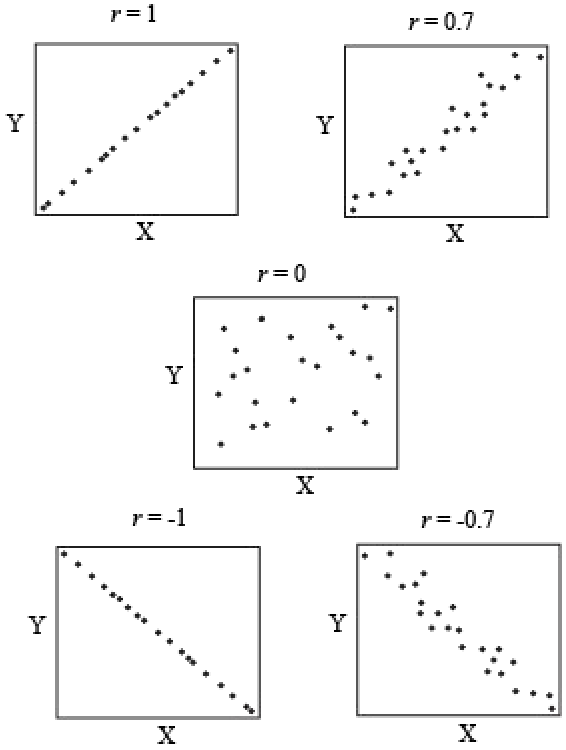
$\hat{Y}$ or y-hat is the predicted value for y based on the estimated variable values for the line of best fit. This is usually used with scatter plot data where the line does not fit perfectly along all the points.

Pearson's correlation coefficient

Pearson's r correlation coefficient is a measure of the linear correlation between two variables X and Y.

The r value can range between 1.0 and -1.0

A value of 1 or -1 denotes a perfect positive or negative linear correlation, and a value of 0 denotes no linear correlation.



Pearson's correlation coefficient

The pearson's correlation formula can be expressed as a relationship of z-scores.

$$\text{Sample: } r = \frac{\sum z_X z_Y}{n-1}$$

$$\text{Population: } \rho = \frac{\sum z_X z_Y}{N}$$

Estimating regression line parameters

If the regression line is defined as Ŷ=bX+a then how can we estimate the variables a and b?

$$b = r\frac{S_Y}{S_X}$$

b can be estimated by multiplying the pearson coefficient value by the dividing the quotient of the standard deviation of the y values divided by the the standard deviation of the x values.

a can be estimated using the estimated b value, the mean of y and the mean of x.

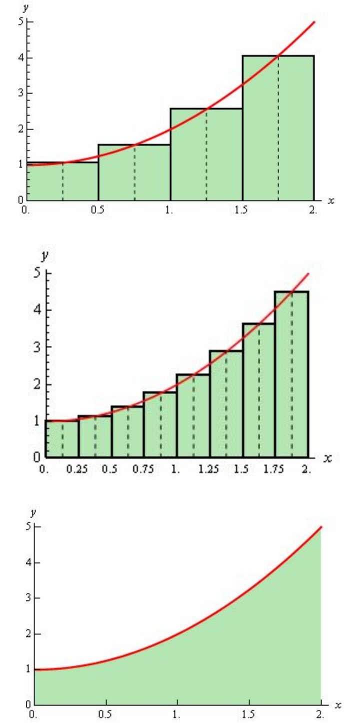
$$a = \bar{Y} - b\bar{X}$$

Integrals for area under the curve

Integration is one of the two main operations of calculus, with its inverse, differentiation.

It is defined informally as the area of the region in the xy-plane that is bounded by the graph of f(), the x-axis and the vertical lines x = a and x = b.

The principles of integration were formulated independently by Isaac Newton and Gottfried Leibniz in the late 17th century, who thought of the integral as an infinite sum of rectangles of infinitesimal width.



How to find normally distributed probabilities

The standard normal distribution table provides the probability that a normally distributed random variable Z, with mean equal to 0 and variance equal to 1, is less than or equal to z.

Each cell in the table represents the area P under the standard normal curve, below the respective z-statistic.

The label for rows contains the integer part and the first decimal place of Z.
The label for columns contains the second decimal place of Z

The standard normal table (z-table)

The z-table allows us to calculate probabilities of encountering values within range on a normally distributed variable.

Probabilities are calculated as a specified area under the standard normal curve.

As computing integrals is a lengthy process, tables were introduced to accelerate and simplify the procedure.

How to find normally distributed probabilities

Since probability tables cannot be printed for every normal distribution, as there are an infinite variety of normal distributions, it is common practice to convert a normal to a standard normal and then use the standard normal table to find probabilities.

To convert any value to it's standardised value, simply calculate it's z-score. The collection of values is it's distribution.

T-distributions

The t-distribution can be used to say how confident you are that any given range of values would contain the true population mean, in situations where the sample size is small (<30) and population standard deviation is unknown.

The t-distribution is similar to the standard normal distribution but it has thicker trails to account for more variance. The t-distribution distribution takes a parameter called df or degrees of freedom. This is simply the number of samples n-1. As the value for df increases the shape of the distribution approaches the shape of the standard normal distribution and converges at around df=30.

T-tables

As computing integrals is a lengthy process, t-tables were introduced to accelerate and simplify the procedure.

Each row represents the degrees of freedom (n-1)

Each column represents the offset from a tail end (upper or lower)

Each cell represents the area under the curve between the tail end and the specified offset. The area represents the probability of capturing the mean within that range.

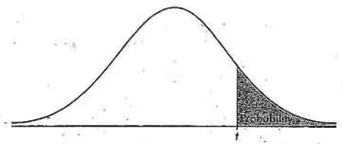


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

T-values

To solve the problem of calculating a t-values with an unknown population standard deviation, the value is estimated using the sample standard deviation.

This leads to the following formula: X-bar plus and minus the z-score for the 95% confidence level times the estimated standard deviation of the sampling distribution of the sample mean, which equals the sample standard deviation divided by the square root of the sample size. We call this estimated standard deviation of the sampling distribution the standard error. But because we now estimate the standard deviation we add extra error in our computation. For that reason we employ another distribution than the standard normal distribution (also called the z-distribution) we employed previously. Because of the extra error we now use the T-distribution. That leads to this formula: X-bar plus and minus the t-score for the 95% confidence level times the estimated standard deviation of the

Logistic Regression

Sum the values of the log likelihood for each value of x and y

$$\begin{matrix}\beta_0 = 0.0 \\ \beta_1 = 0.0\end{matrix},\ \text{-}2\sum_{i=1}^{n} y_i log_e\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) + (1-y_i)log_e\left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

To begin we give the parameters B_0 and B_1 some arbitrary values with the understanding that these starting value will be replaced by optimized values later on. We start with small value and the estimation process will increase them gradually. After these initial parameters are estimated, the search process repeats using newton's method until the Log Likelihood is maximised and does not change significantly.

Maximum likelihood estimation

The regression coefficients are usually estimated using maximum likelihood estimation. Unlike linear regression with normally distributed values, it is not possible to find a closed-form expression for the coefficient values that maximize the likelihood function, so iterative processes must be used instead.

Typically, the log likelihood is maximized using a something like a gradient descent algorithm such as Generalized Reduced Gradient (GRG) or Newton's method.

These processes begins with a tentative solution, revises it slightly to see if it can be improved, and repeats this revision until no more improvement is made, at which point the process is said to have converged.

Logistic Regression

To graph y, simply plug in the constant, coefficient and x value

$$P(y|x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$