UNDERSTANDING MAXIMUM LIKELIHOOD ESTIMATION (MLE): THE STATISTICAL MODELLING TECHNIQUES USED IN MANY STATS TECHNIQUES (E.G. LOGISTIC REGRESSION)

Sample Estimates and Population Probabilities

This section is quite simple conceptually (and the technical bits are in boxes). If you grasp what's going on, even if only roughly, then it will really help you to understand: (a) the *process* of executing logistic regression; (b) what to look at in the printed output; and (c) the jargon used in the printout.

Imagine we tested a random sample of 100 right-handed subjects in their twenties in a dart-throwing experiment. We got each subject to take one left-handed throw at the board. We scored the data very simply: did the subject hit the scoring portion of the board? This generated a categorical variable *hitboard* with values: 1=yes; 2=no.

hitboard		
1=yes	2=no	Total
60	40	100

Table 3. The summary data for the *hitboard* variable in the dart-throwing data

So the overall probability of hitting the dartboard was 0.6 (60/100; let's call that probability *q*). The measured value of 0.6, from this particular measurement sample, might be used to estimate a particular *population probability* that we are interested in (e.g., the probability with which right-handed in their twenties would hit a dartboard given a single throw; this population probability will be denoted by the letter *p*). We might ask what is the most likely value of the population probability given the sample value (*q*=0.6) that we obtained. For any hypothetical value of *p*, we can easily calculate the likelihood of getting exactly 60 hits from 100 participants using probability theory. The underlying theory and mechanics of the calculation are described in the box on the following page.

Using Probability Theory To Derive Likelihoods

Take tossing a coin as an easy example, which involves all the processes we are interested in. What is the likelihood of getting exactly 2 heads in 3 tosses of a completely fair coin? This is a binomial problem as there are just two outcomes for each trial (Heads, H; Tails, T). We can count the answer. There are 8 (i.e. 2^3) possible sequences of 3 tosses which are all equally likely:

TTT; TTH; THT; HTT; HHT; HTH; THH; HHH

Only 3 of the sequences have exactly 2 Heads (THH; HTH; HHT), so the likelihood is 3/8 (=0.375). It is important that the outcome on each toss is independent of the outcome on every other toss. Independence therefore means, for example, that tossing an H on one trial does not change the chance of getting an H on the next trial. In this way the 8 possible sequences shown above are equally likely. This binomial problem, for 2 possible outcomes, can be described generally as trying to find the likelihood, *L*, of getting exactly *m* occurrences of outcome 1 in a total of *N* independent trials, when the probability of outcome 1 on a single trial is *p*). For our example above: *N*=3; *m*=2; *p*=0.5. (The value of *p* is 0.5 because it is a fair coin.) The general formula is:

$$L = {}^{N}C_{m} * p^{m} * (1-p)^{(N-m)}$$

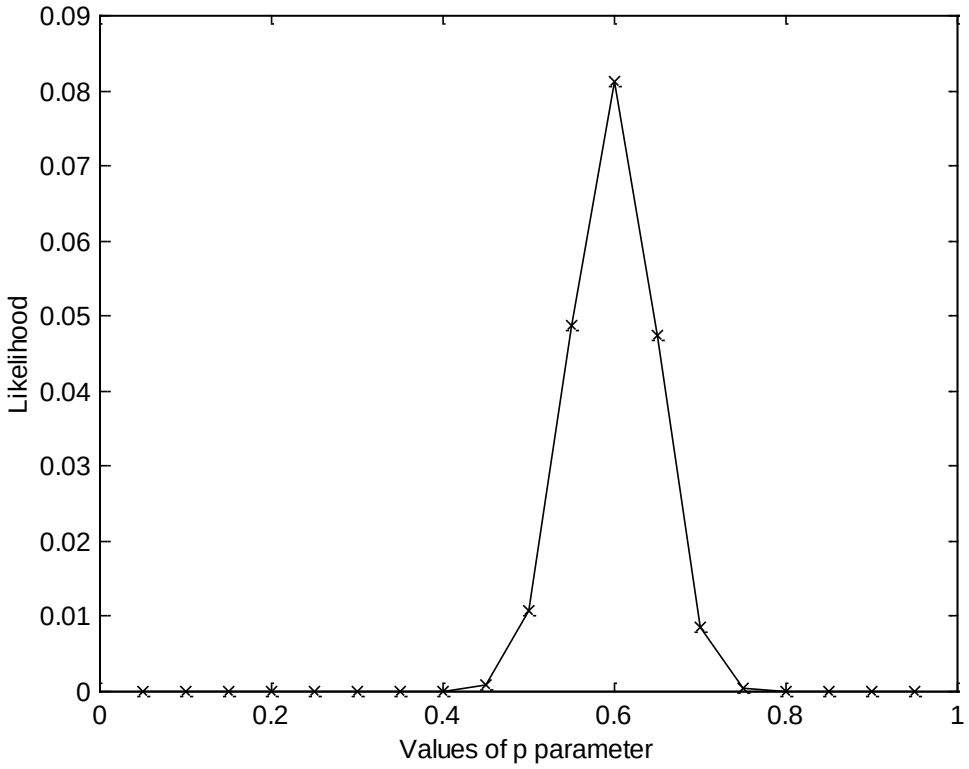
Where ${}^{N}C_{m}$ is the number of ways (combinations) that *m* outcomes of a particular type can be arranged in a series of *N* outcomes (answer=3 in our example). ${}^{N}C_{m}$ is itself given by the following formula:-

$${}^{N}C_{m} = N!/(m!*[N-m]!)$$

Where the symbol ! means the *factorial* function. $X! = X*(X-1)*(X-2)\ldots*2*1$. Thus, for example, 3!= 3*2*1 . Check that the above formulae generate 0.375 as the answer to our coin-tossing problem. The multinomial formulae are an extension of the above to deal with cases where there are more than 2 types of outcome.

Applying Probability Theory to the Dart-Throwing Data

For our darts data, the values to plug into the formula thus: there are 100 trials (i.e., *N*=100); and we are interested in the case where we obtained exactly 60 "*hitboard*=yes" outcomes (i.e., *m*=60). We can allow the value of *p* to vary in small steps from 0.05 to 0.95 and calculate a likelihood for each value of *p*. Putting these values into the formula, we get the likelihoods that are shown in the following graph:



It is fairly clear from the graph that the likelihood of getting the result q=0.6 is at a maximum for the value p=0.6. In fact, this might seem intuitively obvious: if the true value of p were, say, 0.5 then to get a sample estimate of 0.6 must mean that the random sample used was slightly better than expected. This seems less likely to occur than getting a sample which performs exactly as expected. A coin-tossing example may help: I think many people would intuitively "know" that the likelihood of getting 10 heads in 20 tosses of a fair coin is greater than the likelihood of getting 8 (or 12) heads (and much greater than the likelihood of getting 2 or 18 heads). This means that our sample value (q=0.6) is the best estimate for p we can make, given the data.

Another, possibly surprising, point that one might notice from the graph is that the likelihoods are all quite low. Even the maximum likelihood (for p=0.6) is only around 0.08. Even if the population probability was really 0.6, we would get sample values which differ from this value 92% of the time. The low values are because here we are talking about the likelihood of p being exactly 0.6 and, in psychology, we are more used to giving ranges of values. For example, we might more usefully give the 95% confidence intervals (CIs) around our sample estimate of q=0.6. These CIs give a range of values which, with 95% confidence, would be expected to contain the true value of p. (How to calculate such CIs is not discussed here.)

Maximum Likelihood Estimation
In general, if one has frequency data of this kind, and an underlying hypothesis (or model) that can be expressed in terms of particular probabilities, then one can create a computer program to estimate the values of those probabilities which are associated with the maximum likelihood of leading to the data values obtained. This is the process called *maximum likelihood estimation*. In the darts example, we would therefore say that 0.6 is the maximum likelihood estimate (MLE) of the underlying population probability parameter (p), given the data obtained. It can also be said that the value p=0.6 provides the *best fit* to the data obtained in the experiment.

For the simple dart-throwing example it was possible to work out the MLE for p by logic/intuition. For a more complex model, with several probabilities, numerical estimation by computer is often the only way to derive MLEs. Statistical packages, such as SPSS, use numerical methods to generate MLEs in several different kinds of analyses, including those involved in logistic regression.

Comparing Likelihoods
Recall that the likelihood was about 0.08 that the true value of p is 0.6, given our sample estimate. Statisticians do *not* usually apply any conventional likelihood values in order to draw conclusions about the real value of p. We do *not* apply the 0.05 convention used in hypothesis testing to evaluate likelihoods in these situations. (CIs, as described earlier, show how the 0.05 convention can be used in this situation.) Instead, statistical modelling works by comparing the likelihoods under two different hypotheses.

Let us suppose that we have calculated the (maximum) likelihood under hypothesis 1 and a (maximum) likelihood under hypothesis 2. We can denote these hypotheses as H_1 and H_2, and the associated likelihoods as L_1 and L_2. It has been demonstrated that -2 times the natural logarithm of the ratio between these two likelihoods (i.e., $-2*\log_e[L_1/L_2]$) has approximately the χ^2 distribution. Thus, we can use the *log likelihood ratio* and the χ^2 distribution to test whether H_1 is significantly less likely than H_2.

In analysing frequency data, this approach is typically used when there is a *hierarchy* of hypotheses of increasing complexity. Hence, loglinear modelling of frequency data is often referred to as *hierarchical loglinear modelling*. Analysis proceeds by finding the simplest hypothesis which is able to account for the observed data with a likelihood that is not significantly lower than the next most complex hypothesis in the hierarchy. We will illustrate this rather abstract and wordy statement with a concrete worked example.

Comparing Likelihoods in the Darts Data
In fact the darts data used above were not collected from 100 participants but from a single individual. She was a right-handed woman with no previous experience of darts. The data reflect 100 throws at the board using her right and left hand on alternate throws. (The data from this experiment are in the SPSS file *darts study.sav*.) The full contingency table is illustrated below.

		hitboard		
		1=yes	2=no	Row Totals
Throwing Hand	Right (=1)	40	10	50
	Left (=2)	20	30	50
	Column Totals	60	40	Grand Total=100

Table 4. The overall contingency data for the dart-throwing data

We might generate a hierarchy of two simple hypotheses about the subject's performance:

> H_1: her ability to hit the board is unaffected by the hand she uses to throw
> H_2: her ability to hit the board is affected by the hand she uses to throw[1]

For the full contingency table 2 *independent* probability values were measured in the experiment: the probability of hitting the board measured for her right hand and the probability measured for her left hand. We will denote the measured sample probability values by q_L and q_R for the left and right hand respectively. (The probability values for missing the board are not independent of the probabilities for hitting the board: the probability for a hit plus that for a miss must add up to 1.)

We can represent the hypothesis H_1 and H_2 in terms of underlying population probabilities. H_1 is an *independence* hypothesis ("throwing ability is independent of hand used"). According to H_1 the true probability for a hit with the right hand (denoted p_R) equals that for the left hand (p_L). Because $p_R = p_L$ we can replace these probabilities with a single value (denoted p; $p = p_R = p_L$). Hypotheses have parameters (probabilities in this case) and degrees of freedom. H_1 is thus a "single-parameter" hypothesis (as it specifies only one probability value; i.e., p). The degrees of freedom (df) for a hypothesis are given by the number of independent data points (the independent probabilities measured in this experiment; 2 in this case) minus the number of freely varying parameters of the hypothesis. Thus, for H_1, df=(2-1)=1.

H_2 is a more complex hypothesis than H_1 because it has two parameters. H_2 says that the probabilities for a hit with the left and right hands are not equal; i.e., $p_R \neq p_L$. Thus, 2 separate probabilities are needed to specify the hypothesis.

It also follows that H_2 has df=0. A hypothesis produces a specific model when values are provided for the parameters of the hypothesis. A hypothesis with df=0 can always generate a model that is described as *saturated*. Saturated models are not very interesting because they describe (or *fit*) the data perfectly (in the sense that there is no discrepancy between the observed frequencies and those expected according to the model). The saturated model under hypothesis H_2 would have the following values: p_R=0.8 and p_L=0.4. This model is the best-fitting version of hypothesis H_2, given the data obtained: it is the version of H_2 that has the maximum likelihood of generating the observed data.

[1] This is a nondirectional hypothesis. Given that the subject is right-handed, we might have had a directional hypothesis specifying that her performance is better with her right, than her left, hand.

Question 1
Can you explain why the best-fitting parameter values for H_2 are p_R=0.8 and p_L=0.4? By looking at the formulae given in textbooks work out what the values of the χ^2 and G^2 goodness of fit statistics would be under the expected frequencies generated by these probability values (*Hint*: you do not need a calculator to do this, as long as you remember what log(1) is.)

From these probability values given above one can work out the likelihood of the data for the sample of right hand throws (=0.140) and the likelihood of the data for the left hand throws (=0.115). Assuming that the observed probabilities in the two samples of throws are independent of one another (i.e., the success or failure of the throwing trials for the left hand does not influence the success or failure of the trials for the right hand, nor vice versa), then the overall likelihood across both samples can be worked out by multiplying the likelihoods for the two separate samples[2]. The overall likelihood is therefore 0.016 (=0.140*0.115). This independence assumption must be met in order to apply any kind of analysis of categorical data (from simple χ^2 tests to logistic regression). In addition, as noted earlier, the probability of hitting the board with any throw (with left or right hand) must be independent of the probability of hitting the board with any other throw; if this independence assumption is violated then the maximum likelihood estimation process (described above) will not estimate the true likelihoods, and the test statistics will not follow a chi-squared distribution.

Question 2
(The questions in this box illustrate the fact that the use of a particular statistical analysis technique may inform the choice between similar, but subtly different designs, for the same experiment.)
Is the assumption of independence between the left hand and right hand darts data samples justified? Would it have been more or less justified if the subject had taken all her right hand throws first, followed by all her left hand throws? Is the probability of success with each throw of the dart likely to be independent of the probability of success of any other throw? From the independence point of view, would a better design have been to test 100 separate right-handed participants for one throw each, with half of them (selected at random) being asked to used their left hand?

In general, a more complex hypothesis (such as H_2) will be able to fit a set of data better than a simpler hypothesis with fewer parameters (such as H_1). Using the *log-likelihood ratio* technique, outlined earlier, we can see if the best-fitting version of the simpler hypothesis (H_1) can fit the observed darts data with a likelihood that is not significantly lower than the likelihood calculated for H_2. If the likelihood for the best-fitting H_1 model is not significantly lower than that for the best-fitting H_2 model, then we adopt H_1 as the best-fitting hypothesis and conclude that dart-throwing accuracy was independent of the hand used. However, if the fit of the H_1 model is significantly poorer than that of H_2 (i.e., the likelihood of H_1 is significantly lower than that of H_2) then we can reject H_1 and conclude that throwing accuracy was not independent of the hand used. (The details of the likelihood ratio calculation is given below and then checked out using SPSS.)

[2] A basic axiom of probability theory states that if event A (occurring with probability p_A) and event B (occurring with probability p_B) are independent, then the occurrence of both A and B is given by (p_A * p_B**).**

To emphasise the analogy with ANOVA, one can think of the likelihood ratio statistic as testing the *interaction* between the variables *Hand* and *hitboard*. A significant interaction would simply mean that the probability of hitting the board was affected by the hand used (i.e. supporting H_2); the lack of significant interaction therefore supports H_1. This way of thinking of the data is particularly helpful when we later analyse tables with more than two variables. Note also that we are usually not interested in the *main effects* under such analyses. The main effects in the darts data (i.e., for *Hand* and *hitboard*) would correspond to questions about the distributions of categories in the row and column totals. Specifically, the main effect for *hitboard* would tell us whether the ratio of *hitboard*=yes: *hitboard*=no responses, across the whole experiment, deviated from 50:50. This is not something of particular interest. Because we sampled the data such that there were equal numbers of left hand and right hand throws, the *Hand* main effect is completely meaningless. When contingency table data are sampled with a clear separation between DVs and IVs (and thus are suitable for logistic analysis) it will generally be the case that the main effects of the IVs will be meaningless.

Calculating Log-Likelihood Ratios for the Darts Data
We already calculated that the likelihoood for the best fitting model under hypothesis H_2 was 0.016. We denote this value by L_2. The corresponding log-likelihood is –4.135. This model has 2 independent parameters (i.e., 2 probabilities). Hypothesis H_1 has only a single parameter, the probability of hitting the board (independent of hand used). It turns out that the best estimate we have for this probability is the overall probability of hitting the board in Table 4 (i.e., 60/100 = 0.6). We can use the likelihood formulae given earlier to calculate the likelihood of getting 40 hits out of 50 with the right hand if the true probability were 0.6. This likelihood is 0.0014. Similarly, the likelihood of getting 20 hits out of 50 for the left hand (if the true probability were 0.6) is 0.002. The overall likelihood (L_1) for the table is therefore (0.0014*0.002), i.e. 2.9×10^{-6} (this is 2.9 in a million). The corresponding log-likelihood is –12.764.

The ratio of the log-likelihood for the simpler model divided by the log-likelihood of the more complex model is thus L_1/L_2. We already noted that, if the simpler model were true, then the statistic $-2*\log_e(L_1/L_2)$ would be distributed approximately as χ^2, with *df* equal to the difference in number of parameters for the two models (here H_2 has two parameters and H_1 has 1; $df = 1$).
But, $-2*\log_e(L_1/L_2) = (-2*\log_e[L_1]) - (-2* \log_e[L_2])$. Therefore, the test statistic for the darts data is (–2*-12.764) – (-2*-4.135) = 17.258. This is very much greater than the critical value for χ^2 with 1 *df* and so we can reject H_1 in favour of H_2. There is a highly significant effect of *Hand* on ability to hit the dartboard.

Checking the result with SPSS
We can run a logistic regression on the darts data using the **Analyze > REGRESSION >>** *MULTINOMIAL LOGISTIC* procedure SPSS. The key part of the printed output is shown below. The final model corresponds to best-fitting probabilities under hypothesis H_2. This model is found to have a –2*log-likelihood (*-2LL*) of 8.268. When the simpler model (H_1) is fitted to the data, this reduced model corresponds to omitting the effect of the *Hand* variable from the full model. The reduced model (with *Hand* omitted) is found to have a *–2LL* of 25.529. The likelihood ratio test involves subtracting the *–2LL* value for the full model from the *–2LL* value for the reduced model. The resulting value (in this case 17.261) is tested against the χ^2 distribution, with *df* equal to the difference in number of parameters between the two models (1 in this case). The result is highly significant. The values are the same as we got by hand earlier (within rounding errors).

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	25.529			
Final	8.268	17.261	1	.000

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	8.268	.000	0	.
HAND	25.529	17.261	1	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.