Mode Choice — Estimation

Philip A. Viton

November 6, 2013
Summary So Far

- Individual $i$ faces $J$ alternatives (modes).
- Mode $j$ has $K$ observable characteristics (for $i$): $x_{ij}$.
- If $i$ selects mode $j$ she gets (conditional) utility $u_{ij} = v_{ij}(x_{ij}) + \eta_{ij}$ where $v_{ij}$ is the systematic part of utility (which depends on the $K$ modal characteristics $x_{ij}$), and $\eta_{ij}$ is the idiosyncratic part, represented by a random variable.
- Since $u_{ij}$ is a random variable, we focus on $i$’s mode choice probabilities: $P_{ij} = \Pr[j \text{ is best for } i]$.
- Once we choose a joint probability model for the $\eta$’s and decide how $v_{ij}(x_{ij})$ depends on data, we determine the form of the $P_{ij}$.
- The usual choice for $v_{ij}(x_{ij})$ is a linear-in-parameters specification: $v_{ij}(x_{ij}) = x_{ij}\beta$. With this choice, only $\beta$ is unknown, and our remaining task is to estimate it.
Clearly, with $K$ components in $\beta = (\beta_1, \beta_2, \ldots, \beta_K)$ it is going to be impossible to estimate $\beta$ from data on a single individual.

We therefore assume that we have a random sample of $I$ individuals, all of whom are assumed to follow the same decision rule (weight the observable characteristics with the same $\beta$).

For the moment, we assume for simplicity that each person in our sample faces the same choice set.

We also assume that our data consists of one observation per individual (a cross-section, not a time-series).
For each individual $i$ we assume that we observe (have data on):

- For each mode $j$ in $i$’s choice set, a $K$-vector of observable characteristics $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijk})$
- An *indicator* of $i$’s actual mode choice decision: $y_{ij} = 1$ if individual $i$ actually chose mode $j$ and $y_{ij} = 0$ otherwise.

So each individual is represented in our data by a block of $J$ rows (one for each mode) and $K + 1$ columns (one column for each observable characteristic, plus one column for the choice indicator).
### Data Setup (II)

<table>
<thead>
<tr>
<th>Indiv ((i))</th>
<th>Mode ((j))</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(\ldots)</th>
<th>(x_K)</th>
<th>(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>(x_{111})</td>
<td>(x_{112})</td>
<td>(x_{113})</td>
<td>(\ldots)</td>
<td>(x_{11K})</td>
<td>(y_{11})</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(x_{121})</td>
<td>(x_{122})</td>
<td>(x_{123})</td>
<td>(\ldots)</td>
<td>(x_{12K})</td>
<td>(y_{12})</td>
</tr>
<tr>
<td></td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\ldots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td></td>
<td>(J)</td>
<td>(x_{1J1})</td>
<td>(x_{1J2})</td>
<td>(x_{1J3})</td>
<td>(\ldots)</td>
<td>(x_{1JK})</td>
<td>(y_{1J})</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>(x_{211})</td>
<td>(x_{212})</td>
<td>(x_{213})</td>
<td>(\ldots)</td>
<td>(x_{21K})</td>
<td>(y_{21})</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(x_{221})</td>
<td>(x_{222})</td>
<td>(x_{223})</td>
<td>(\ldots)</td>
<td>(x_{22K})</td>
<td>(y_{22})</td>
</tr>
<tr>
<td></td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\ldots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td></td>
<td>(J)</td>
<td>(x_{2J1})</td>
<td>(x_{2J2})</td>
<td>(x_{2J3})</td>
<td>(\ldots)</td>
<td>(x_{2JK})</td>
<td>(y_{2J})</td>
</tr>
<tr>
<td>(I)</td>
<td>1</td>
<td>(x_{I11})</td>
<td>(x_{I12})</td>
<td>(x_{I13})</td>
<td>(\ldots)</td>
<td>(x_{I1K})</td>
<td>(y_{I1})</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(x_{I21})</td>
<td>(x_{I22})</td>
<td>(x_{I23})</td>
<td>(\ldots)</td>
<td>(x_{I2K})</td>
<td>(y_{I2})</td>
</tr>
<tr>
<td></td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\ldots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td></td>
<td>(J)</td>
<td>(x_{IJ1})</td>
<td>(x_{IJ2})</td>
<td>(x_{IJ3})</td>
<td>(\ldots)</td>
<td>(x_{IJK})</td>
<td>(y_{IJ})</td>
</tr>
</tbody>
</table>
Once we have decided on a joint distribution for the random variables (the \( \eta \)'s), the only unknown is \( \beta \).

- Suppose we have a guess for \( \beta \), call it \( \beta^0 \).
- Then we can plug \( \beta^0 \) into our expression for the mode choice probabilities and calculate the \( P_{ij}(\beta^0) \), the probability that individual \( i \) will choose mode \( j \).
- We also know which mode she actually selected — this is the single element of her choice set for which \( y_{ij} = 1 \).
- So we can calculate the probability, given \( \beta^0 \), that individual \( i \) will choose the mode she actually chose. Call this probability \( \tilde{P}_i(\beta^0) \) — note the tilde.
Suppose we do this, and suppose it turns that for all individuals \( i \), the probabilities \( \tilde{P}_i(\beta^0) \) of the choices they actually made are very small.

Then it is natural to conclude that our guess \( \beta^0 \) was not a good one: it seems inconsistent with the choices actually made by our sample.

This suggests an estimation principle: choose a value for \( \beta \) that makes the observed (actual) choices of our sample have the highest probability — ie most likely.

This is the principle of maximum-likelihood estimation. See the Appendix for more on this.
Are there reasons to think that this maximum-likelihood strategy is an attractive one, besides its intuitive plausibility?

It turns out that the principle of maximum likelihood estimation has a number of attractive properties.

Before listing them, it is important to note that these hold only for “large” samples (formally, as the sample size tends to infinity). It is generally an open question just how large an actual sample must be in order to qualify, but most people believe that sample sizes of several hundred or more — these are very common in mode-choice settings — qualify as “large”.
Properties of the maximum-likelihood estimates (MLEs) $\hat{\beta}_{\text{ML}}$:

- As the sample size gets large, the sampling distribution of the MLEs shrinks to a spike centered around the true (but unknown) value of the parameter: we say that $\hat{\beta}_{\text{ML}}$ is consistent for $\beta$. This implies that any bias shrinks to zero: the estimators are asymptotically unbiased.
- The sampling distribution of the MLEs is asymptotically normal. This implies that the standard tests from linear regression can be used.
- Optimality: the MLEs have the lowest variances (actually, the smallest variance-covariance matrix) in the class of all asymptotically unbiased estimators of $\beta$. We say that the MLEs are asymptotically efficient.

The first two properties are often summarized by saying that the MLEs are Consistent and Asymptotically Normal, or CAN.
MLE for the Discrete Choice Model (I)

- For individual \( i \), her individual likelihood is just the probability that she selects the mode she was actually observed to choose. In our earlier notation, this is \( \tilde{P}_i(\beta) \), and it depends on \( \beta \).

- It turns out to be convenient to write this as follows (even though it looks more complicated):

\[
\mathcal{L}_i = \tilde{P}_i(\beta) = P_{i1}^{y_{i1}}(\beta) \cdot P_{i2}^{y_{i2}}(\beta) \cdots P_{iJ}^{y_{ij}}(\beta) = \prod_{j=1}^{J} P_{ij}^{y_{ij}}(\beta)
\]

- This works because for any mode not chosen \( y_{ij} = 0 \), and for this mode \( P_{ij}^{y_{ij}} = 1 \), and contributes nothing to the product; while for the chosen mode \( y_{ij} = 1 \) so \( P_{ij}^{y_{ij}} \) will just be \( \tilde{P}_i(\beta) \).
Because we have a random sample, the sample likelihood is the product of the individual likelihoods:

\[
\mathcal{L} = \prod_{i=1}^{I} \mathcal{L}_i = \prod_{i=1}^{I} \prod_{j=1}^{J} P_{ij}^{y_{ij}}(\beta)
\]

For computational reasons, it is often more convenient to maximize the sample log-likelihood, which is:

\[
\ln(\mathcal{L}) = \sum_i \sum_j y_{ij} \ln P_{ij}(\beta)
\]
So the MLE of the discrete choice model involves choosing $\beta$ to maximize

$$L = \prod_{i=1}^{I} \prod_{j=1}^{J} P_{ij}^{y_{ij}}(\beta) \quad \text{or} \quad \ln(L) = \sum_{i} \sum_{j} y_{ij} \ln P_{ij}(\beta)$$

As we previously observed, this is going to be do-able for probit only when $J$ (the number of elements in the choice set) is small (less than about 4), because each $L_i$ is an $J$-fold integral of the multivariate normal distribution.

On the other hand, for the logit model the problem is straightforward and, even for large $I$ (number of individuals in the sample) and $J$ (size of the choice set) this can be done extremely quickly on modern workstations.

For this reason, outside the binary choice setting, most analysis of discrete choice assumes that the logit model applies.
Specification

Under a linear-in-parameters specification, we have

\[ v_{ij}(x_{ij}) = x_{ij} \beta = x_{ij1} \beta_1 + x_{ij2} \beta_2 + \cdots + x_{ijK} \beta_K \]

where \( \beta \) is to be estimated. We now look at some issues connected with how the variables enter the specification. We discuss:

- Intercepts
- Alternative-specific dummies
- Inclusion of personal characteristics
Suppose we adopt a linear-in-parameters specification for systematic utility.

And suppose we try to include a constant (intercept) term (as would be natural in a linear regression context).

Then we would write (in the obvious notation):

\[ u_{ij} = \beta_0 + x_{ij} \beta + \eta_{ij} \]

For binary choice (multinomial choice is exactly analogous, though of course there will be more terms involved) we have:

\[ P_{i1} = \Pr[u_{i1} \geq u_{i2}] \]
\[ = \Pr[\beta_0 + x_{i1} \beta + \eta_{i1} \geq \beta_0 + x_{i2} \beta + \eta_{i2}] \]
\[ = \Pr[x_{i1} \beta + \eta_{i1} \geq x_{i2} \beta + \eta_{i2}] \]

(subtracting \( \beta_0 \) from both sides).
Note what has happened: our intercept ($\beta_0$) has dropped out.

In other words, we get the same choice probabilities with and without an intercept. We say that an intercept $\beta_0$ applicable to all modes is not identified in this model.

However, instead of an intercept, we can define alternative-specific or (in the mode-choice case) mode-specific dummies.

These are variables defined to be 1 if we are describing the mode in question and 0 otherwise.

In other words, the dummies provide a separate intercept for each mode’s systematic utility:

$$v_{ij}(x_{ij}) = \beta_j + x_{ij}\beta$$
But we still need to be careful. With an alternative-specific dummy for each mode we would have, in the binary choice case:

\[ P_{i1} = \Pr[\beta_1 + x_{i1}\beta + \eta_{i1} \geq \beta_2 + x_{i2}\beta + \eta_{i2}] \]

But this can be written as

\[ P_{i1} = \Pr[\eta_{i2} - \eta_{i1} \leq \beta_1 - \beta_2 + (x_{i1} - x_{i2})\beta] \]

which is exactly the same as

\[ P_{i1} = \Pr[\eta_{i2} - \eta_{i1} \leq \beta_{12} + (x_{i1} - x_{i2})\beta] \]

with \( \beta_{12} = \beta_1 - \beta_2 \), meaning that only one of the two alternative-specific dummies is identified.
In general, if there are $J$ modes in the choice set, you can estimate at most $J - 1$ alternative-specific dummies.

You must always leave one of them out: *which* one does not matter.

If you try to estimate all $J$ dummies you generate a linear dependency in the data (one column will be a linear combination of the others) and the computer program will be unable to invert some matrices.

Some software will handle this automatically: if you ask for intercepts, the software will generate $J - 1$ of them for you, and decide internally which to leave out (usually the dummy for the first alternative).

If the program unexpectedly reports that it cannot complete the estimation, or generates obviously nonsense results (coefficients very large or very small), this is something to check.
Alternative-Specific Dummys (V)

In terms of our previous data setup, suppose we want to define the first variable as a mode-1 alternative-specific dummy and variable 2 as a mode-2 dummy. Then we would have:

\[
\text{Indiv} (i) \quad \text{Mode} (j) \quad x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_1 \quad y
\]

<table>
<thead>
<tr>
<th>Indiv (i)</th>
<th>Mode (j)</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$\cdots$</th>
<th>$x_1$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$x_{113}$</td>
<td>$\cdots$</td>
<td>$x_{11K}$</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$x_{123}$</td>
<td>$\cdots$</td>
<td>$x_{12K}$</td>
<td>$y_{13}$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$x_{133}$</td>
<td>$\cdots$</td>
<td>$x_{13K}$</td>
<td>$y_{13}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>$\cdots$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$J$</td>
<td>$J$</td>
<td>0</td>
<td>0</td>
<td>$x_{1J3}$</td>
<td>$\cdots$</td>
<td>$x_{1JK}$</td>
<td>$y_{1J}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$x_{213}$</td>
<td>$\cdots$</td>
<td>$x_{21K}$</td>
<td>$y_{21}$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>$x_{223}$</td>
<td>$\cdots$</td>
<td>$x_{22K}$</td>
<td>$y_{22}$</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>$x_{233}$</td>
<td>$\cdots$</td>
<td>$x_{23K}$</td>
<td>$y_{23}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>$\cdots$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$J$</td>
<td>$J$</td>
<td>0</td>
<td>0</td>
<td>$x_{2J3}$</td>
<td>$\cdots$</td>
<td>$x_{2JK}$</td>
<td>$y_{2J}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>$\cdots$</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Suppose we believe that an individual’s income is one determinant of which mode she selects.

Then it is natural to try to include income among the $x$ variables.

But we now run into a similar problem. If $m_i$ is individual $i$’s income, we would have (again, for binary choice):

\[
P_{i1} = \Pr[u_{i1} \geq u_{i2}]
\]

\[
= \Pr[\beta_1 m_i + x_{i1} \beta + \eta_{i1} \geq \beta_1 m_i + x_{i2} \beta + \eta_{i2}]
\]

\[
= \Pr[x_{i1} \beta + \eta_{i1} \geq x_{i2} \beta + \eta_{i2}]
\]

Since $m_i$ could be any individual-specific descriptor — any variable that describes the individual, and therefore will take the same value for each mode in the individual’s choice set — we see that its coefficient is not identified.
Individual-Specific Data (II)

- The usual solution is to “attach” the individual-specific variables to the data for one arbitrary mode, say the first one.
- If we attach the individual’s income to the first mode as variable 1 we would have:

<table>
<thead>
<tr>
<th>Indiv (i)</th>
<th>Mode (j)</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$\cdots$</th>
<th>$x_K$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$m_1$</td>
<td>$x_{112}$</td>
<td>$x_{113}$</td>
<td>$\cdots$</td>
<td>$x_{11K}$</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>$x_{122}$</td>
<td>$x_{123}$</td>
<td>$\cdots$</td>
<td>$x_{12K}$</td>
<td>$y_{12}$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$J$</td>
<td>0</td>
<td>$x_{1J2}$</td>
<td>$x_{1J3}$</td>
<td>$\cdots$</td>
<td>$x_{1JK}$</td>
<td>$y_{1J}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$m_2$</td>
<td>$x_{212}$</td>
<td>$x_{213}$</td>
<td>$\cdots$</td>
<td>$x_{21K}$</td>
<td>$y_{21}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>$x_{222}$</td>
<td>$x_{223}$</td>
<td>$\cdots$</td>
<td>$x_{22K}$</td>
<td>$y_{22}$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$J$</td>
<td>0</td>
<td>$x_{2J2}$</td>
<td>$x_{2J3}$</td>
<td>$\cdots$</td>
<td>$x_{2JK}$</td>
<td>$y_{2J}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>
Individual-Specific Data (III)

- Note that this idea provides a way of partially relaxing the restriction that everyone in the sample is assumed to use the same decision-rule (same $\beta$).

- For example, if we believe that males and females differ in the ways that they summarize the modal characteristics, we could include in our specification:
  
  - A gender dummy ($M_i = 1$ if $i$ is male, 0 otherwise), for example.
  - A gender dummy interacted with another variable, say trip cost: create a new variable as the product of trip cost and the gender dummy, and include both it and the original cost variable. The model would then have

$$v_{ij} = \beta_C C_{ij} + \beta_{CM} (C_{ij} M_i) + \ldots$$

where $C_{ij}$ is cost, and $M_i$ is the gender dummy. Then males would weight costs by $\beta_C + \beta_{CM}$ while females would use $\beta_C$. Of course, we could also do this for other variables.
Finally, we look at two other issues:

1. What happens if different individuals have different choice sets (for example, a low-income individual may not own a car).
2. Why we might want to relax the assumption of random sampling.
Varying Choice Sets

- A setting in which different individuals have different choice sets poses no essential difficulties.
- We define a *master choice set*, consisting of all modes (alternatives) available to anyone in our sample, and say that *i*'s choice set is a subset of this.
- Conceptually, if $S_i$ is the set of modes available to $i$, then
  \[
  \mathcal{L}_i = \prod_{j \in S_i} P_{ij}^{y_{ij}}(\beta)
  \]
- Then the only problem is to keep track of which elements in the master choice set apply to each individual (ie the $S_i$). This is often done via a second indicator variable, taking the value 1 if mode $j$ is available to individual $i$. But this is a programming (book-keeping) issue, not one of model formulation: it can be left to the software.
Suppose it turns out that very few people in the population use one of the available modes (for example, a Segway).

If we sampled randomly in the population, it is likely that this mode will not be represented often enough to be able to make any inferences about how people view it.

One solution is to deliberately over-sample from users of this mode to ensure that it is represented in our data: this is known as *choice-based* sampling.

With choice-based sampling we sample based on which modes were actually utilized, rather than randomly. For example, we might draw our sample by surveying individuals on buses or at parking lots. This would not be random sampling.
Non-Random Sampling

- Of course, if we do this, then we can no longer count on the properties of random sampling for our inference. Fortunately, there is a simple solution.
- Define

  \[ \pi_j^* = \text{proportion using mode } j \text{ in the population} \]
  \[ \pi_j = \text{proportion using mode } j \text{ in the sample} \]
  \[ w_j = \frac{\pi_j^*}{\pi_j} \]

- Now maximize the weighted likelihood function (or its log):

  \[ \mathcal{L} = \prod_{i=1}^{I} \prod_{j=1}^{J} w_j p_{ij}^{y_{ij}} \]

  The resulting estimates are the MLEs for this sampling scheme.
Software Considerations

- Virtually every statistical software package contains facilities to estimate the multinomial logit model or the binary probit model.
- However, if you want to estimate the most recent models in this area, your choices are more limited, and you may want to become familiar with a package that will allow you to do so without having to learn a different package, even if initially you are interested only in the standard model formulations.
- See separate notes on the course website for some guides to the LIMDEP/NLOGIT and R packages. Both of these are good choices for more advanced work. LIMDEP/NLOGIT is available only for Windows (and is commercial software), while R is free, open source, and available for most operating systems, including Windows, the Mac, and Unix/Linux.
This appendix attempts to motivate the idea of maximum likelihood estimation. Consider the following setting:

- You draw a random sample of size 2 from a normal distribution. 
- You know that the variance of the distribution is 1, but you do not know the mean. 
- From your perspective, you have a random sample from the $N(\mu, 1)$ distribution, with $\mu$ unknown. 
- Your task is to use the data — the sample values — to estimate the mean.
The Normal Distribution

- Recall that the probability density function (pdf) of the normal distribution with mean $\mu$ and variance $\sigma^2$ is:

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

- In our case, since we know that the variance is 1, this reduces to:

  $$f(x; \mu, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$
The Likelihood of a Random Sample

- For any $x$, the function $f(x; \mu, 1)$ can be interpreted as the probability of drawing the value $x$ from this distribution.
  You may remember from your course in probability theory that this last statement isn't quite rigorous: if $f$ is the pdf of a continuous random variable, then, strictly, the probability of obtaining any real number $x$ is zero. But for our purposes we can still think of the pdf as giving the probability of obtaining $x$.

- If $(x_1, x_2)$ is a random sample of size 2 from this distribution, then the probability of obtaining both these particular values is:

$$L(x_1, x_2; \mu, 1) = f(x_1; \mu, 1) \cdot f(x_2; \mu, 1)$$

- This is called the Likelihood of the random sample. Think of it as the probability of obtaining the sample values $x_1$ and $x_2$ when sampling from the distribution $f$. 
The fundamental observation is that if we make a guess (estimate) as to the value of $\mu$, then we can calculate the probability of obtaining the data we actually got. If our guess is $\mu = \mu_0$ then the probability of getting our observed data is the sample likelihood $\mathcal{L}(x_1, x_2; \mu_0, 1)$.

Now, suppose this probability turns out to be very small.

This suggests that our guess as to the value of $\mu$ (ie $\mu_0$) was not a good one. So suppose we try another guess, say $\mu_1$.

And suppose that under this guess it is more likely that we would obtain the sample that we actually got: that is, suppose

$$\mathcal{L}(x_1, x_2; \mu_1, 1) > \mathcal{L}(x_1, x_2; \mu_0, 1)$$

This suggests that $\mu_1$ is a better guess (estimate) than $\mu_0$. 
The principle of Maximum Likelihood Estimation is the obvious extension of this idea. It says that we should take as our best guess (estimate) as to the value of an unknown parameter that value that makes it most likely that, under that guess, we obtain the data that we actually did obtain. In other words, the maximum-likelihood estimator of a parameter $\mu$ is the value of $\mu$ that maximizes the sample likelihood.
Return to our motivating example (estimating the unknown mean of a normal distribution), and suppose that our 2-sample gave us the values $x_1 = 4$ and $x_2 = 6$.

Our task is to use these two values to come up with an estimate of the unknown mean $\mu$.

We shall do so by maximizing the sample log-likelihood. Of course, since the logarithm is a monotonically increasing function, maximizing the log of the likelihood and maximizing the likelihood itself will give the same answers.
Suppose we guess $\mu = 0$, ie that the unknown distribution is standard normal.

The picture opposite shows the situation.

Under the hypothesis that $\mu = 0$, getting our observed data ($x_1 = 4$ and $x_2 = 6$) is possible but very unlikely.

In fact, we can calculate that the sample likelihood is a tiny $8.1 \times 10^{-13}$, or that the log-likelihood is $-27.838$. 
What can we do to make the probability of our sample more likely?

Clearly, we want to shift the distribution rightwards.

So let's try $\mu = 2$. The picture shows the new situation.

Under this new hypothesis, we can calculate that

$$\ln L(x_1, x_2; 2, 1) = -11.838.$$  

This is an improvement, but the sample likelihood is still quite low: $7.224735 \times 10^{-6}$.
But the solution is clear: shift even further to the right.

Let’s try \( \mu = 4.5 \).

In this case

\[
\ln \mathcal{L}(x_1, x_2; 4.5, 1) = -3.088, \text{ or sample likelihood } = 4.559305 \times 10^{-2}.
\]
We can see where this is going: we can assemble a table of log-likelihood values corresponding to different guesses for $\mu$:

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\ln \mathcal{L}(x_1, x_2; \mu, 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>-3.087877</td>
</tr>
<tr>
<td>4.7</td>
<td>-2.927877</td>
</tr>
<tr>
<td>4.9</td>
<td>-2.847877</td>
</tr>
<tr>
<td>5.0</td>
<td>-2.837877</td>
</tr>
<tr>
<td>5.1</td>
<td>-2.847877</td>
</tr>
</tbody>
</table>
Example (VI)

- It looks as though the value of $\mu$ that maximizes the log-likelihood is about $\mu = 5$.
- It turns out that in this particular case we can solve the problem analytically.
- For a random sample of size $N$ from a normal distribution with unknown mean and known variance, we can show that the maximum-likelihood estimator of the population mean $\mu$ is the sample mean $\bar{x}$.
- In our example, the sample mean is $\frac{1}{2}(4 + 6) = 5$, just as our table lead us to conclude.
- In more difficult cases, like the logit or probit models, we cannot obtain analytical results. In these cases we have to rely on a computer program to find the parameter value(s) that maximize the sample log-likelihood.
- For the logit model, this is a well-behaved problem, that can be solved very quickly.