Foundations of Medical Statistics

Statistical Inference 3: Likelihood

Aims

The aim or this session is to introduce the concept of likelihood and show how it can be used in estimation through maximum likelihood estimates.

Objectives

At the end of this session you should:

- understand the concept of likelihood
- be able to derive a likelihood in a novel situation
- be able to derive maximum likelihood estimators, and obtain them from the log-likelihood, if appropriate
- know the main properties of maximum likelihood estimators.

3.1 Probability versus Inference

In the context of elementary probability, we are usually given the probability of a simple event, e.g. Prob(coin landing heads)=0.5, and required to calculate the probability of a complex event, e.g. Prob(observing 4 heads from 10 throws).

In the context of statistical inference, the situation is the other way round: we do not know the probability of simple event, e.g. Prob(coin landing heads), but observe a complex event, e.g. 4 heads from 10 throws. From this we then need to obtain the best estimate of the unknown probability of the simple event. We also need to know how well we have estimated the unknown probability: what is the uncertainty associated with our estimate? The concept of **likelihood** provides the best single framework for this task.

3.2 Likelihood and maximum likelihood estimators

We introduce these ideas using an example: suppose we observe 4 events among 10 subjects. We define a **model** which assumes these **data** are drawn from a binomial distribution, with **parameter** π:

Model: we will assume the number of events to be a **random variable** $X \sim \text{Bin}(10, \pi)$. (The model also assumes the events occur independently of one another.)
Data: We have observed the realisation of X, x=4.

Now, the probability of observing these data, based on this model and the unknown parameter π, is

$$\text{Prob}(X=4|\pi) = \pi^4 (1-\pi)^{(10-4)} \binom{10}{4}$$

Since π is unknown, it is natural to consider how the probability of observing these data varies with different values of π:

Value of π	Probability of observing data X=4
0	0
0.2	0.088
0.4	**0.251**
0.5	0.205
0.6	0.111
0.8	0.006
1	0

Clearly, based on the model, the probability of getting the observed data is higher if we choose π to be 0.4 than, say, 0.2, or 0.5. A sensible estimator of π, based on the data, would seem to be that value of π yielding the *highest probability of getting the observed data*. In fact, of the values of π we have chosen in the table, the probability is highest if we choose π = 0.4; intuitively this is also sensible, as it is the sample proportion.

To summarise:

- We have observed the data (4 events out of 10 subjects).
- We have defined a binomial probability model, in terms of parameter π, on the basis of which we assume the data have been sampled; and we have calculated the probability of observing the data for various values of π.
- We have chosen as a sensible estimator of π that yields the highest probability.

Note that the function we are maximising, although taking the same algebraic appearance as the probability, is a function of π, **maximised** with respect to π. This function is called the likelihood for π:

$$L(\pi|X=4) = \pi^4 (1-\pi)^{(10-4)} \binom{10}{4}$$

We can plot the shape of this function for values of π between 0 and 1, as shown in Figure 3.1. Note that the probability distribution which is used to generate the values of this function, for **fixed observed data** X=4, is discrete, and quite distinct from the continuous likelihood function.

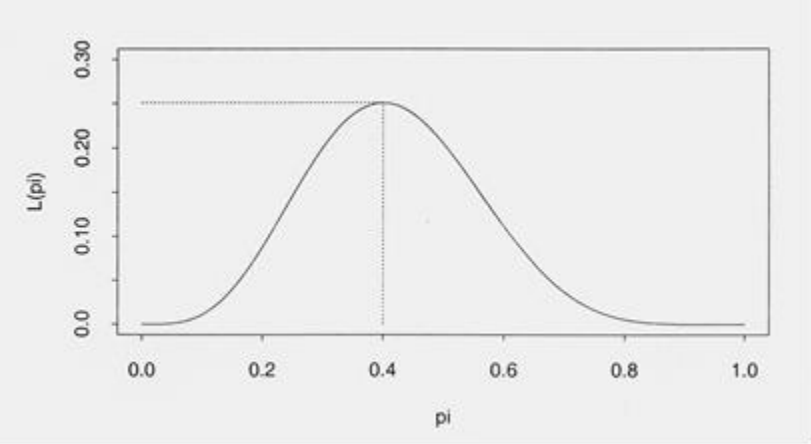


Figure 3.1 Binomial likelihood

The plot confirms graphically that the value of π which maximises the likelihood is indeed 0.4; this value is therefore the maximum likelihood estimate of π.

More generally, if X is a random variable following a binomial distribution, $X \sim \text{Bin}(n, \pi)$, then the statistic X/n is the **maximum likelihood estimator**, $\hat{\Pi}$, of π; and x/n is the maximum likelihood **estimate**, $\hat{\pi}$, of π. You will be asked to prove this result in the practical.

3.3 General definition of likelihood

For a probability model with parameter θ, the likelihood of the parameter θ given the observed data $\underline{x}$ is defined as

$$L(\theta \mid \underline{x}) = \text{P}(\underline{x} \mid \theta)$$

Notes:

1. $\text{P}(\underline{x} \mid \theta)$ may be a probability (discrete distribution) or a density (continuous distribution); for this function, θ is fixed, since the probability or density is evaluated over values of $\underline{x}$ conditional on the fixed value of θ.
2. $L(\theta \mid \underline{x})$ is a function of θ; for this function $\underline{x}$ is fixed, since the likelihood is evaluated over values of θ, conditional on the fixed value of $\underline{x}$.
3. Likelihood is **not** a probability density function.

EXERCISE 3.3.1

One observation is made on a discrete random variable X with probability function $f(x|\theta)$. The function is shown in the table below when θ takes the value 1,2 or 3. Find the maximum likelihood estimate of θ when x is observed to be 0, 1, 2, 3 and 4:

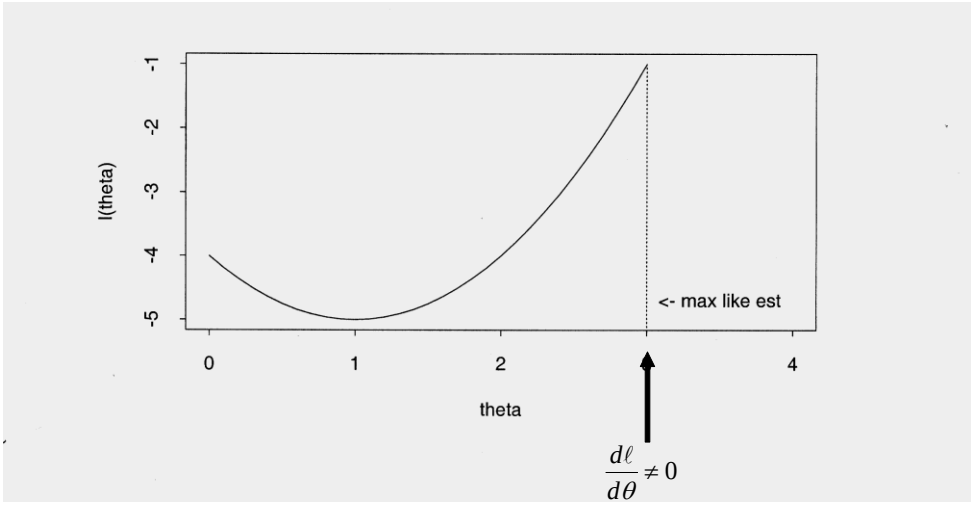
x	$f(x\|1)$	$f(x\|2)$	$f(x\|3)$
0	1/3	1/4	0
1	1/3	1/4	0
2	0	1/4	1/6
3	1/6	1/4	1/2
4	1/6	0	1/3

3.4 Log-likelihood

The maximum likelihood estimate of a parameter θ can be obtained by maximising either the likelihood $L(\theta|\text{data})$, or the log-likelihood $\ell(\theta|\text{data})$ (since ℓ changes in the same direction L; but for a different type of proof see Exercise 3.4.1). Conventionally, the maximum likelihood estimate is denoted by putting a 'hat' on the parameter: $\hat{\theta}$. The log-likelihood is usually the easier function to differentiate, and so the maximum likelihood estimate $\hat{\theta}$ can usually be calculated as the solution of

$$\frac{d\ell}{d\theta} = \ell'(\theta) = 0 \text{ such that } \frac{d^2\ell}{d\theta^2} = \ell''(\theta) < 0$$

Note: differentiation may not yield the maximum likelihood estimate if it is a boundary value of the parameter, e.g. if the likelihood curve is as shown below:



See the Appendix (not examinable) for a practical example.

EXERCISE 3.4.1

Given a likelihood $L(\theta \mid \underline{x})$, with a unique maximum away from the boundary, show that maximum of L is at the same value of θ as the maximum of the log-likelihood $\ell(\theta \mid \underline{x})$.

3.5 Properties of maximum likelihood estimators

Maximum likelihood estimators based on a sample of data size *n* can be shown to have some very useful properties, listed below:

1. Asymptotically unbiased	$n \to \infty \Rightarrow \ \mathrm{E}(\hat{\Theta}) \to \theta$
2. Asymptotically efficient	$n \to \infty \Rightarrow \ \mathrm{Var}(\hat{\Theta})$ is smallest variance amongst all unbiased estimators
3. Asymptotically Normal	$n \to \infty \Rightarrow \ \hat{\Theta} \sim \mathrm{N}(\theta, \mathrm{Var}(\hat{\Theta}))$
4. Transformation invariant	$\hat{\Theta}$ is the MLE of $\theta \Rightarrow g(\hat{\Theta})$ is the MLE of $g(\theta)$ for any function *g*.
5. Sufficient	$\hat{\Theta}$ contains all the imformation from the data relevant to estimating θ.
6. Consistent	$n \to \infty \Rightarrow \hat{\Theta} \to \theta$ in probability. [More formally, given any $\varepsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(\mid \hat{\Theta} - \theta \mid > \varepsilon) = 0$]

3.6 Example: likelihood for a rate

If subjects in a study are followed up for different lengths of time, then we need to express the incidence of events as a rate (e.g. per person year of observation) rather than a risk. If the rate parameter is denoted λ, and the total observation time for all subjects is p person-years, then the expected number of events is $\mu = \lambda p$. As a model for the observed number of events D, assume that D is a random variable drawn from a Poisson distribution with mean μ, and assume events occur independently of each other:

$$D \sim \mathrm{Po}(\mu).$$

Suppose we observe $D = d$ events; the probability of getting the observed data based on the model is given by

$$\mathrm{Prob}(D = d) = e^{-\mu}\mu^d/d! = e^{-\lambda p}\lambda^d p^d/d!$$

so the likelihood for λ is

$$L(\lambda|\text{observed data}) = e^{-\lambda p}\lambda^d p^d/d!$$

and the log-likelihood for λ is

$$\ell(\lambda|\text{observed data}) = d\log(\lambda) - \lambda p + d\log(p) - \log(d!)$$

Solving $\ell'(\theta) = 0$ gives $\hat{\lambda} = d/p$, so the maximum likelihood estimator for the rate parameter is D/p, the total number of events divided by the total observation time.

Note:
Expressions in the log-likelihood which do not involve the parameter do not contribute to the shape of the likelihood, vanish on differentiation, and so are not relevant to obtaining a maximum likelihood estimate. Usually these terms are omitted from the log-likelihood. Thus in the example above the last two terms do not involve λ: $d\log(p) - \log(d!)$ can be omitted from the log-likelihood.

3.7 Likelihood and log-likelihood with *n* independent observations

The overall likelihood from **independent** observations is the **product** of the likelihood from each observation, since the observations are jointly observed. If $X_1,..., X_n \overset{iid}{\sim} f(.|\theta)$

$$L(\theta|x_1,..., x_n) = f(x_1,..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$\Rightarrow \quad \ell(\theta|x_1,..., x_n) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

EXERCISE 3.7.1

Let $x_1,..., x_n$ be a random sample from a population, where each random variable X has probability function $f(x|\alpha) = \alpha x^{\alpha-1}$, for $0 \le x \le 1$ and $\alpha > 0$.

1. Assuming $n = 1$, sketch the probability function for x for (i) $\alpha = 1$ and (ii) $\alpha = 0.5$.
2. Assuming $n = 1$, sketch the likelihood function for α for (i) $x = 1$ and (ii) $x = 0.5$.
3. Find the maximum likelihood estimator for α, still assuming n=1.
4. Find the maximum likelihood estimator for α, no longer assuming n=1.
5. If $n = 3$ and the sample values are 0.2, 0.3 and 0.7, what value does the maximum likelihood estimate take?

Note: You may notice that $f(x|\alpha)$ and therefore $L(x|\alpha)$ can take values greater than 1. Although for a discrete random variable the values of the probability function are probabilities, for continuous random variables the values as you know are densities; and while for any 'small' range of values δx, the product (ie area) $f(x|\alpha).\delta x$ is a probability, and therefore less than 1, the density $f(x|\alpha)$ itself can be greater than 1: in such cases the density is changing rapidly, and even though some of the density values can be very large, the area, given by the integral over the relevant range, is always less than 1. For the relevant values of x the likelihood function (for the particular parameter value) also takes values greater than 1. It is therefore sometimes possible, with continuous random variables, to see positive log likelihoods.

3.8 Method of moments for finding estimators

The method of maximum likelihood estimation was developed by R.A. Fisher in the period 1912 to 1922. Before that a more pragmatic but intuitive method was in use, the method of moments, in which essentially a sample quantity is found to correspond with the population parameter, and then the arithmetic mean of the sample quantity is used as estimator. Methods of moments estimators (MMEs) and maximum likelihood estimators (MLEs) may coincide (e.g. sample mean) but in general MMEs do not have all the optimal properties of MLEs. However, MMEs have computational advantages, and are still in use in areas where this is an issue, or to provide starting values for iterative computational procedures.

The method of moments is described in more detail in the Appendix (*non-examinable*).

In general, the Rao-Blackwell theorem (details given in Rice, p310, Hogg and Craig, p. 223 and Cox and Hinkley, p. 258) may be used to determine a best estimator for a parameter, but likelihood-based methods make this unnecessary in practice.

Appendix: Method of moments (*not examinable*)

First, some definitions:
Suppose $X_1,..., X_n \overset{iid}{\sim} (\mu, \sigma^2)$. Then

- the first *population* moment is E[X], which is μ;
- the second *population* moment is E[X^2], which is $\sigma^2 + \mu^2$;
- ...
- in general the pth population moment is E[X^p];

- the first *sample* moment is $\sum_{i=1}^{n} X_i / n$ usually written $\overline{X}$;
- the second *sample* moment is $\sum_{i=1}^{n} X_i^2 / n$;
- ...
- in general the pth sample moment is $\sum_{i=1}^{n} X_i^p / n$.

Key principle

1. We express the quantity we wish to estimate as a function of *population* moments. For example,
   (i) to estimate parameter θ = mean, then θ = E[X], the first population moment;

   (ii) to estimate θ = variance, then θ = E[X^2] – { E[X]}^{2} , i.e. the second population moment minus the square of the first population moment.

   In general, we express the parameter of interest

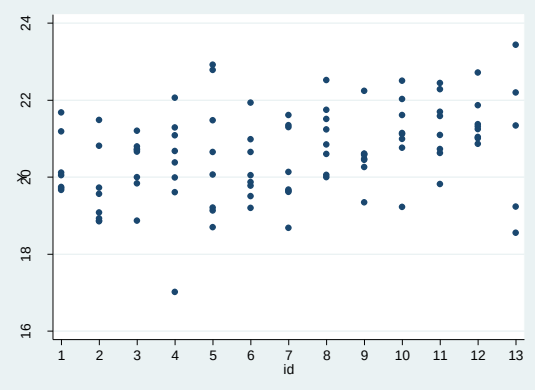
   $$\theta = T(\text{E}[X], \text{E}[X^2], \text{E}[X^3], ..., \text{E}[X^p])$$

2. Obtain the *estimator* by replacing *population* moments with *sample* moments. For example, to estimate θ = E[X], write $\hat{\Theta} = \sum_{i=1}^{n} X_i / n$, the first sample moment. In general, form

   $$\hat{\Theta} = T\left[\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \frac{1}{n}\sum_{i=1}^{n} X_i^3, ..., \frac{1}{n}\sum_{i=1}^{n} X_i^p\right]$$

Appendix: Parameter boundary values (*not examinable*)

In the data below, there are 13 subjects, each with repeated measurements of variable Y:



For this type of data, which you will meet in the Hierarchical course (Term 2), two sources of variation may be of interest:
i) the variability between subjects (this could be captured by the variance of the subject means); and
ii) the variability within subjects (the variance of a subject's points around their mean).

The first model below estimates both sources of variation: var(_cons) is i) and var(Residual) is ii), recognising that the 100 data points belong to 13 'groups': the 13 subjects (id). The second model ignores the 'ownership' of the data points by the subjects, and treats all 100 data points as independent. A natural question now is: does the model recognising ownership, Model 1 below ('Hierarchical') fit the data better than Model 2 ('Simple'), which ignores the data hierarchy? The log likelihood of Model 1 is slightly higher than that of Model 2, and a test is performed to compare the models, with a null hypothesis of no difference between the models. This test gives P=0.2021, insufficient evidence at 5% of Model 1's superiority. However, the null hypothesis for this test says, in effect, that the between-subject variance is zero: then Model 1 would be identical to simple Model 2, which estimates all the variance as within-subject variance. However, the zero value of the between-variance under the null hypothesis is on the *boundary* of the variance parameter space. Stata knows this is a possibility in this context, and warns us that under these conditions we have to exercise caution if our null hypothesis forces one of the parameters onto its boundary.

Model 1 'Hierarchical'

```
. mixed Y ||id:, var

Mixed-effects ML regression                     Number of obs      =       100
Group variable: id                              Number of groups   =        13

Log likelihood = -153.88076
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                  var(_cons) |   .0988462   .1006901      .0134239    .7278473
-----------------------------+------------------------------------------------
             var(Residual) |   1.192107   .1803009      .8862874    1.603451
------------------------------------------------------------------------------
. estimates store hierarchical
```

Model 2 'Simple'

```
. mixed Y ||

Mixed-effects ML regression                     Number of obs      =       100

Log likelihood = -154.69442
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
             var(Residual) |   1.291767   .1826835      .9790531    1.704364
------------------------------------------------------------------------------
. estimates store simple

. lrtest hierarchical simple

Likelihood-ratio test                                 LR chi2(1)  =      1.63
(Assumption: simple nested in hierarchical)           Prob > chi2 =    0.2021

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary
of the parameter space.  If this is not true, then the reported test is conservative.
```