University of Glasgow
School of Mathematics & Statistics

Statistical Inference: Level M

Chapter 3

Parametric Inference

Part 1 - Point Estimation and Likelihood

February 2021

Dr Benn Macdonald
Room 225, Maths & Stats Building
Benn.Macdonald@glasgow.ac.uk

3.1 Introduction

Up until now, we have used approaches to statistical inference that are called nonparametric. This means we have only made assumptions about (a) the design of the study, e.g. that the observations were collected independently of one another, and (b) very general features of the underlying distribution of response values, e.g. that two distributions differ only in their location (median), not in their shape or spread.

Sometimes, it is plausible to assume that we actually know the distributional form of the response variable (in the population). For example, we might assume that a sample of data comes from a Normally-distributed population. In other words, we might adopt a Normal probability model for the population. This means that, implicitly, we are claiming to know everything about the population apart from the values of the model's two unknown parameters: the population mean (μ) and variance (σ^2).

Parametric inference can lead to firmer conclusions than nonparametric inference, sometimes much firmer. The more we believe we know about the population to begin with, the stronger the inferences we should be able to draw on the basis of the data. On the other hand, the inferences we draw could well be invalid if the assumptions we choose to make are invalid.

Parametric statistical inference naturally focuses on the unknown parameters of the model. In particular, we might wish to:

- find a point estimate ('best guess') of a parameter's value;
- find an interval estimate ('range of plausible values') for a parameter;
- test hypotheses about a parameter.

With nonparametric tests, we implicitly drew inferences about unknown parameters, e.g. population median(s) in the Wilcoxon Signed-Ranks Test and the Mann-Whitney U Test. The difference is that, in parametric inference, the parameters are assumed to be the only unknowns. Also, we are able more easily to examine the theoretical properties of our estimation and testing procedures.

3.2 Point Estimation

A point estimate is a 'best guess' at the value of an unknown parameter. It is a single number, based on a sample of data from the model. There are general procedures for determining point estimates in any model, e.g. the method of maximum likelihood that we will discuss later in this chapter. For many common statistical models (e.g. Binomial, Poisson, Normal), though, a little thought about the meaning of a parameter suggests a 'natural' way to estimate it.

1. Binomial

In the **Binomial model** with parameter $\theta\,(0 \leq \theta \leq 1)$, θ is the probability of a 'success', i.e. the proportion of successes in the population (using the relative frequency definition of probability). If a sample of n trials is conducted, in which x 'successes' and $n - x$ 'failures' are recorded, then an obvious estimate of θ is the proportion of successes in the sample, i.e.

$$\hat{\theta} = \frac{x}{n}$$

Example 1 - Shareholders

A study was recently carried out to investigate the effect of a continuous decrease in the share price on the number of shareholders for a small company. Of 10,318 shareholders, in a period of six months, 600 people had sold their shares.

$$\hat{\theta} = \frac{600}{10318} = 0.058$$

The proportion of people to have sold their shares was 0.058.

2. Poisson

In the **Poisson model** with parameter λ $(0 \leq \lambda)$, λ is the expected value, or population mean number, of events in a specified time interval. If independent counts $x_1, x_2, \ldots, x_n$ are observed in different time intervals, then an obvious estimate of λ is the sample mean number of events, i.e.

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

Example 2 - Car accidents on a motorway

The numbers of car accidents at a fixed point on a motorway were recorded for 20 consecutive months. The results are as shown below, with x_i being the number of accidents at the fixed point in the i'th month.

Month	1	2	3	4	5	6	7	8	9	10
No. of accidents	2	2	1	1	0	4	2	1	2	1
Month	11	12	13	14	15	16	17	18	19	20
No. of accidents	1	1	3	1	2	2	3	2	3	4

$$\hat{\lambda} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}(2 + 2 + 1 + 1 + \ldots + 4) = 38/20 = 1.9$$

On average, 1.9 accidents occurred each month.

3. Normal

In the **Normal model** with parameters μ and $\sigma^2 (0 \leq \sigma)$, μ is the population mean and σ^2 is the population variance. If response values $x_1, x_2, \ldots, x_n$ are observed, then it is natural to estimate the population mean μ by the sample mean,

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

A natural estimate of the population variance σ^2 is the sample variance,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example 3 - Heights of Primary School children

The heights of a sample of 25 children in the West of Scotland were measured just after they began Primary School (aged 4 - 5 years). The values (cm) are shown below.

```
100.1   103.7   105.3   108.3   108.4   108.4   108.7   109.1   109.1
109.6   109.8   110.2   110.4   110.4   111.4   112.2   115.3   116.0
116.4   116.7   117.6   118.0   118.1   118.6   120.5
```

Many studies have shown that height is Normally distributed in populations of specified race, sex and age. So, assuming that the heights of West of Scotland children aged about 5 years are Normally distributed, we can estimate the population mean height and standard deviation using the point estimators defined above, $\hat{\mu} = 111.7$, $\hat{\sigma}^2 = 26.2$.

Point Estimation

In general, the estimate of a parameter will vary from sample to sample. This leads to the following definition(s):

Suppose we intend to observe data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a probability model with an unknown parameter θ. A point estimator of θ is an algebraic function, $t(\mathbf{X})$, of the data. A point estimate is a particular value of the function, $t(\mathbf{x})$, obtained from a particular set of data, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

In any situation there will be a variety of possible estimators and we need some way of choosing between them. We would like any point estimator to have the following properties.

1. The range of $t(\mathbf{X})$ should be the same as the range of θ.

2. $t(\mathbf{X})$ should be unbiased, i.e. $\mathrm{E}\{t(\mathbf{X})\} = \theta$ (the bias for an estimator is zero). Although we cannot require $t(\mathbf{x})$ to be equal to θ for every possible sample, $\mathbf{x}$, we do require that, on average over all possible samples, $t(\mathbf{X})$ is equal to θ.

   An estimator can be unbiased but can vary so much that it is not useful. It is useful, therefore, to consider the variance of an estimator as a further property of its reliability.

3. $t(\mathbf{X})$ should be consistent. As the sample size (n) gets bigger, then the probability distribution of the estimator should become more concentrated around the true value of θ. This means that $\mathrm{var}\{t(\mathbf{X})\} \to 0$ as $n \to \infty$ (when $t(\mathbf{X})$ is unbiased).

   This means that as the sample size increases the procedure delivers more and more reliable results.

There are additional properties which will be mentioned at the end of this chapter.

Example 4 - Binomial model with parameter $\theta\,(0 \leq \theta \leq 1)$

It is assumed that $X \sim \text{Bi}(n, \theta)$. The proposed estimator of θ is $t(X) = X/n$.

Range :

Unbiased :

Consistent :

See the Probability (Level M) course for definitions of expectation and variance and associated rules. See the probability formula sheet for expectations and variances of standard distributions.

3.3 Maximum Likelihood Estimation

In the examples discussed in Section 3.2, we relied on adhoc arguments to propose possible estimators of unknown parameters in some simple models. The method of Maximum Likelihood offers us a general approach to obtaining estimators of the parameters in any model.

Maximum likelihood estimation is the best known and most widely used method of estimation (ordinary least squares is also widely used, see regression modelling). In this approach we are selecting the value of θ, our parameters, for a chosen probability distribution, for which our given set of observations has maximum probability.

The maximum likelihood estimator (MLE) is the value of $\hat{\theta}$ which maximises $L(\theta; \mathbf{x})$, where $L(\theta; \mathbf{x})$ is the likelihood function.

Assuming that a sample of data, $x_1, x_2, \ldots, x_n$ arise from independent replicates of an experiment, then probabilities associated with the individual observations are multiplied together to give an overall probability associated with the data.

3.3.1 Discrete Distributions

Suppose that the discrete random variable X_i is observed on the ith replicate of an experiment and that X_i has probability mass function $p_i(x_i)$. Then assuming independence, the overall probability of the data is:

$$\begin{aligned} & P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \ldots \text{and } X_n = x_n) \\ & \qquad = \quad P(X_1 = x_1) \times P(X_2 = x_2) \times \ldots P(X_n = x_n) \\ & \qquad = \quad p_1(x_1) \times p_2(x_2) \times \ldots \times p_n(x_n) \end{aligned}$$

The likelihood function of the unknown parameter θ, given a sample of data, $x_1, x_2, \ldots, x_n$ is defined as:

$$L(\theta; x_1, x_2, \ldots, x_n) = p_1(x_1) \times p_2(x_2) \times \ldots \times p_n(x_n) = \prod_{i=1}^{n} p_i(x_i)$$

It is often easier to work with the (natural) logarithm of the likelihood function. This is known as the log-likelihood function, and is denoted

$$\ell(\theta) = \log_e\{L(\theta; \mathbf{x})\}$$

Since $\log_e(.)$ is a monotonic function, $\ell(\theta)$ reaches its maximum at the same value of θ as $L(\theta; \mathbf{x})$. This means that the maximum likelihood estimate can be found by maximising either $L(\theta; \mathbf{x})$ or $\ell(\theta)$. We will shorten $L(\theta; \mathbf{x})$ to $L(\theta)$.

Example 5 - Binomial model (for one data point x)

Model: $X \sim \text{Bi}(n, \theta)$

Data: x

Likelihood:

Log-likelihood:

Assuming that $1 \leq x \leq n-1$, then these functions are maximised at an interior point of the interval $[0, 1]$. The **maximum** can be found by differentiation:

To find the **turning point** set $\frac{\partial \ell}{\partial \theta} = \ell'(\theta) = 0$.

This means that the log likelihood has a turning point at $\theta = \frac{x}{n}$. We need to check that this is indeed a local **maximum**:

This second derivative must be less than 0 at all values of θ in the range $(0, 1)$, since θ^2 and $(1 - \theta)^2$ are always non-negative:

This means that the turning point at $\theta = \frac{x}{n}$ is indeed a local maximum. So assuming that $1 \leq x \leq n - 1$, the maximum likelihood estimate of θ in this model is $\hat{\theta}_{MLE} = \frac{x}{n}$. Notice that this is the natural estimator we discussed in Section 3.2.

Usually, more than one piece of data is used to estimate the parameter.

Example 6 - Poisson model

Model: $X_1, X_2, \ldots, X_n$ independent with each $x_i \sim \text{Poi}(\lambda)$
Data: $x_1, x_2, \ldots, x_n$

Likelihood:

Log-likelihood:

Differentiate wrt λ:

Solve for λ:

Second derivative:

If we apply the above results to Example 2 on car accidents we see that $\hat{\lambda}_{MLE} = \bar{x} = 38/20 = 1.9$. This result is also displayed in Figure 1 where it can be seen that the maximum of the likelihood function occurs at $\lambda = 1.9$.

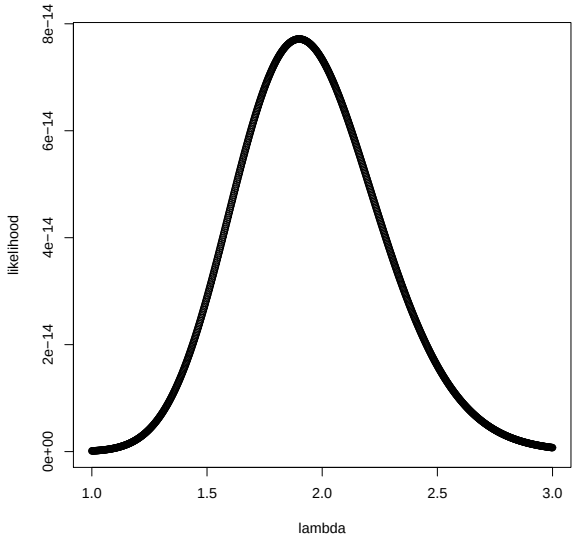

Figure 1: Likelihood function for Poisson distribution, examples 2 and 6.

3.3.2 Continuous Distributions

When X_i is a continuous random variable, then the probability density function of X_i evaluated at x_i does not directly represent the probability of the data. However, $f_i(x_i)$ is (approximately) proportional to the model probability that X_i lies in a small interval around the value x_i. So, it is reasonable to take the likelihood of the unknown parameter, θ, to be

$$f_1(x_1) \times f_2(x_2) \times \ldots \times f_n(x_n)$$

i.e.

$$L(\theta; x_1, \ldots, x_n) \propto \prod_{i=1}^{n} f_i(x_i)$$

As in the discrete case, the maximum of this function can usually be determined by differentiating the log-likelihood. Again, we will shorten $L(\theta; x_1, \ldots, x_n)$ to $L(\theta)$.

Example 7 - Exponential model

Air-conditioning failures data

Data were recorded on the time (intervals in service-hours) between the failures of the air-conditioning equipment in a Boeing 720 aircraft. The observations were:

```
50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88,
46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14.
```

What can we say about the mean time between failures?

Model: $X_1, X_2, \ldots, X_n$ independent, with each $X_i \sim \text{Expo}(\theta)$ $\theta > 0$

Data: $x_1, x_2, \ldots, x_n$

Likelihood:

Log-likelihood:

Using the data for this example, it can be seen that $\hat{\theta}_{MLE} = 0.016$. The likelihood is displayed in Figure 2.

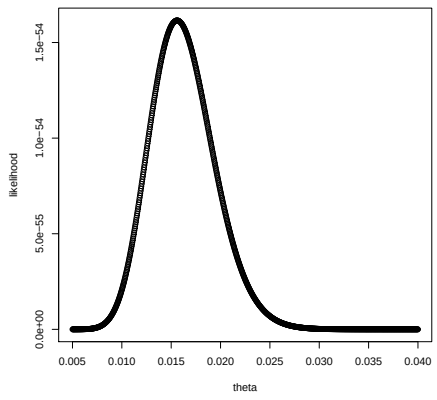

Figure 2: Likelihood function for Exponential distribution, example 7.

It is always important to bear in mind, though, that the MLE of θ might be found on the boundary of the range of θ. Example 8 demonstrates this, in a very important special case.

Example 8 - A Uniform model

Model: $X_1, X_2, \ldots, X_n$ independent, with each $X_i \sim \mathrm{U}(0, \theta)$ $\quad 0 \leq x_i \leq \theta$

Data: $x_1, x_2, \ldots, x_n$

Likelihood:

Strangely, this doesn't depend on the x_i's!

Actually, it does, because we require $\theta \geq x_i$ for each x_i, otherwise the probability mass function, and therefore the contribution to the likelihood function is 0. So we must have $\theta \geq \max x_i$.

Log-likelihood:

We will consider the normal distribution in later lectures, when we consider multi-parameter models.

3.3.3 Important Concepts

The following concepts will be very important for us in future lectures.

Combining Likelihoods

For two independent samples X_1, X_2, we have

$$P(X_1, X_2; \theta) = P(X_1; \theta)P(X_2; \theta)$$

The likelihoods therefore combine in a very simple way.

$$L(\theta) = L_1(\theta)L_2(\theta)$$

where L_1 and L_2 denote the likelihood functions of the two separate samples.

Similarly, the log likelihood functions combine as

$$\ell(\theta) = \ell_1(\theta) + \ell_2(\theta)$$

Relative likelihood

Suppose that the maximum likelihood estimate for θ is $\hat{\theta}$. Relative plausibilities of other θ values may be found by comparing the likelihood of those other values with the likelihood of $\hat{\theta}$.

The relative likelihood function ($R(\theta)$) and the relative log-likelihood function ($r(\theta)$) are defined as

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta}_{MLE})} \\ r(\theta) &= \log_e \frac{L(\theta)}{L(\hat{\theta}_{MLE})} = \ell(\theta) - \ell(\hat{\theta}_{MLE}) \end{aligned}$$

Sample Information

Since we know that the second derivative will be negative at the maximum, we define the **sample information**, denoted by $k(\mathbf{x})$, as

$$k(\mathbf{x}) = -\ell''(\hat{\theta}_{MLE}).$$

3.3.4 Computation of MLE's

We have found closed-form expressions for $\hat{\theta}$ for simple cases. However, to find the maximum of $L(\theta; \mathbf{x})$ can be an optimisation problem.

Complications arise when it is not possible to find the root of the equation $\ell'(\theta) = 0$ in closed form. In such cases, it is often possible to find the root numerically, by an iterative procedure known as the Newton-Raphson method.

This method will be illustrated below for a general function $g(x)$. For $\ell'(\theta) = 0$ replace $g(x)$ by $\ell'(\theta)$ below.

This is based on the idea that, if $x^{(0)}$ is a first approximation to a root of the general equation

$$g(x) = 0,$$

then a better approximation to the root is given by

$$x^{(1)} = x^{(0)} - \frac{g(x^{(0)})}{g'(x^{(0)})}$$

A series of better and better approximations may then be obtained, using the iterative formula

$$x^{(j+1)} = x^{(j)} - \frac{g(x^{(j)})}{g'(x^{(j)})}$$

This iterative procedure is stopped when the numerical approximation has converged to the correct value of the root of the equation (to a pre-determined level of accuracy), e.g. when

$$-0.001 \leq \frac{g(x^{(j)})}{g'(x^{(j)})} \leq 0.001$$

In order to start the Newton-Raphson method for finding the root of the equation, and hence the MLE of θ, we need a first approximation. This is often estimated for $\hat{\theta}$ using a plot of the likelihood or log-likelihood function. In order to find the MLE of θ, $g(x)$ is taken to be $\ell'(\theta)$ in the above.

In many examples the Newton-Raphson method converges to a solution in very few steps as long as a sensible first approximation is available.

There are other optimisation methods which have been found useful in computing MLEs. These are: the method of scoring, the simplex method and the EM algorithm. Please see the references for this course for more information on these approaches.

We will re-visit the Newton-Raphson method when we consider interval estimation.

3.4 Additional Properties of Point Estimators

Earlier in this chapter, the properties of unbiasedness and consistency were discussed and checked for different estimators. Using these two properties may still leave a number of candidate estimators. A further property is to look for minimum variance unbiased estimators (MVUEs). The words 'efficient' and 'efficiency' when applied to estimators refer to the variances of the estimators. The lower the variance of an unbiased estimator, the more efficient it is.

Mean Square Error

One way of considering bias and variance together is via a combined measure called the mean square error.

The quality of a point estimator is sometimes assessed by the mean square error (MSE). The MSE of an estimator $\hat{\theta}$ of a parameter θ is defined to be:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

This can be written as:

$$MSE(\hat{\theta}) = \text{var}(\hat{\theta}) + \left[\text{bias}(\hat{\theta})\right]^2$$

Efficiency

An unbiased estimator is said to be efficient if it has the minimum possible variance; the efficiency of an unbiased estimator is the ratio of the minimum possible variance to the variance of the estimator.

Minimum possible variance

It is usual to take the following well-known lower bound to the variance of unbiased estimators as the 'minimum variance'.

The Cramér-Rao inequality (and lower bound)

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from the distribution with p.d.f $f(x;\theta)$. Subject to certain regularity conditions on $f(x;\theta)$, we have that for any unbiased estimator $\hat{\theta}$ for θ,

$$\text{var}\left[\hat{\theta}\right] \geq I_\theta^{-1}$$

where

$$I_\theta = E\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right]$$

$L(\theta;\mathbf{x})$ is the likelihood function defined earlier and $\ell = \log_e(L(\theta))$. I_θ^{-1} is known as the Cramér-Rao lower bound, and the corresponding inequality is the Cramér-Rao inequality. I_θ is sometimes known as the Fisher information about θ in the observations and $\frac{\partial l(\theta)}{\partial \theta}$ is sometimes known as the score function $s(\mathrm{x};\theta)$.

Sufficiency

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from $f(x;\theta)$. Suppose further that $t(x_1, x_2, \ldots, x_n)$ is a function of the observations $x_1, x_2, \ldots, x_n$ and not of θ and that $T(X_1, X_2, \ldots, X_n)$ is the corresponding random variable.

T is then a statistic, and T is sufficient for θ (a sufficient statistic for θ) if the conditional distribution of $X_1, X_2, \ldots, X_n$ given the value T, does not depend on θ.

What this means is that if T is sufficient for θ, then it contains all the information about θ which is contained in the X_i; once the value of T is known we can squeeze no more information out of the X_i regarding θ.

This definition of sufficiency does not indicate how to go about finding a sufficient statistic. The following 'factorisation theorem' can help with this.

For $T(X_1, X_2 \ldots X_n)$ to be sufficient for a parameter θ, the joint probability function factors in the form:

$$f(\mathbf{x} \mid \theta) = g[T(\mathbf{x}), \theta]h(\mathbf{x})$$

The Rao-Blackwell Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample of observations from a distribution with pdf $f(x; \theta)$. Suppose that T is a sufficient statistic for θ and that $\hat{\theta}$ is any unbiased estimator for θ. Define $\hat{\theta}_T = E[\hat{\theta} \mid T]$. Then,

- $\hat{\theta}_T$ is a function of T alone.
- $E(\hat{\theta}_T) = \theta$
- $\text{var}(\hat{\theta}_T) \leq \text{var}(\hat{\theta})$.

The Rao-Blackwell theorem gives a quantitative rationale for basing an estimator of a parameter θ on a sufficient statistic if one exists.

We will revisit many of these properties later when we consider properties of Maximum Likelihood Estimators.