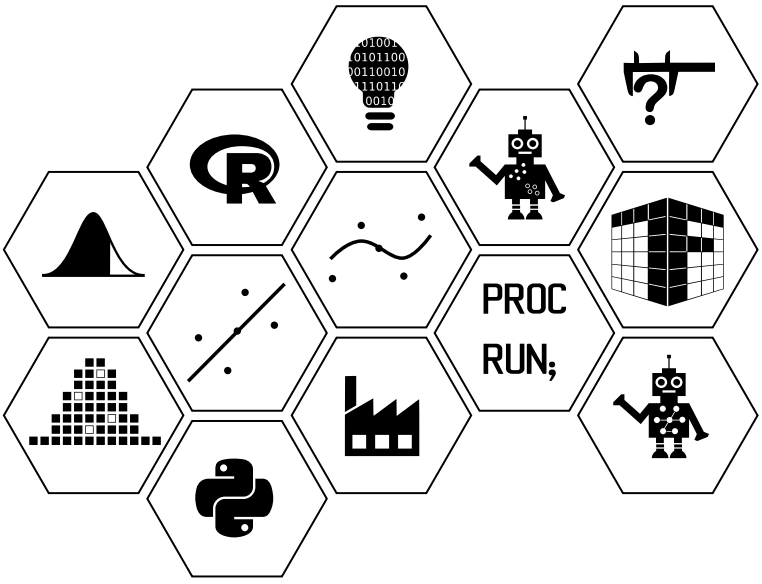
Learning from Data/Data Science Foundations

Week 5: Point estimation and likelihood



DATA ANALYTICS
GLASGOW

Point estimation and Likelihood

Introduction

In the material for week 4 we introduced the idea of testing and interval estimation for population parameters in contexts where we assume that our continuous (or binary) data have arisen from (or are well approximated by) a normal distribution. For each context, we established a *best guess* for the values of our unknown population parameters of interest, e.g. population mean, population proportion, (we refer to these as *point estimates*) and we account for variability and hence uncertainty in the estimation through testing and/or producing confidence intervals as part of statistical inference. In the contexts that we have considered so far there are natural estimators for the unknown parameters of our distribution based on our sample mean (and sample proportion). However, it is not always the case that the estimators for distributions of interest arise naturally. To help us generalise to estimating the parameters for other distributions (which as you've seen in probability are useful depending on the context), in this week we will formalise the idea of point estimators and their properties, and generalise how to determine point estimators (and hence point estimates) for the population parameters of discrete probability distributions through the method of **Maximum Likelihood**.

If we return to our initial diagram for the course, shown below in Figure 1, we can see that we've suggested three alternative approaches which use a probability model to describe our data.

> Estimator: a rule for calculating an estimate of a given quantity based on observed data: thus the rule (the estimator), the quantity of interest (the estimand) and its result (the estimate) are distinguished

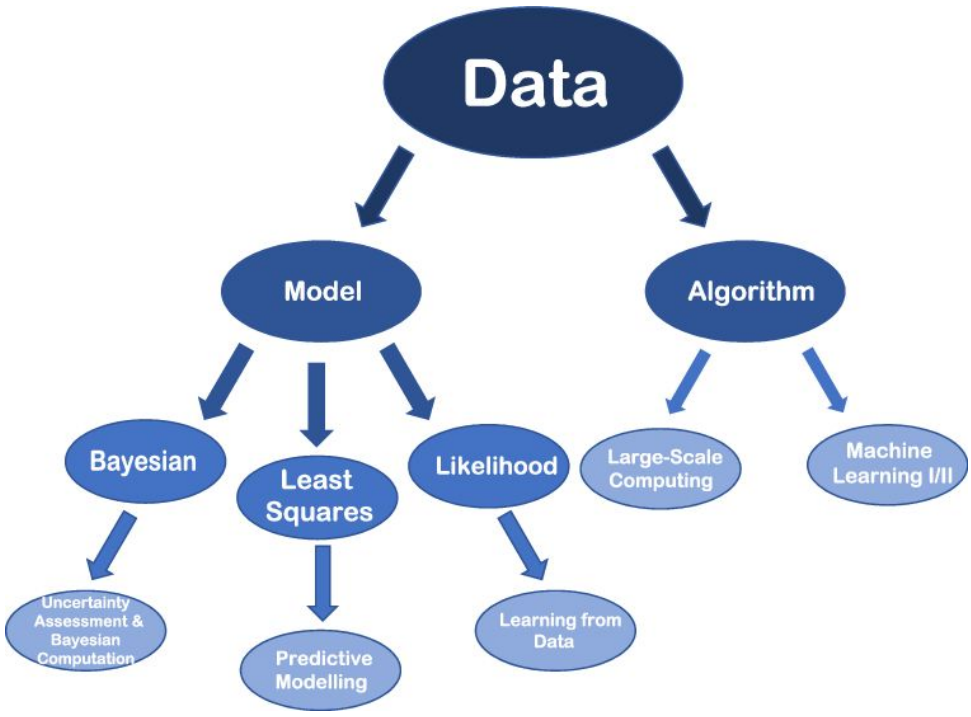


Figure 1: *Approaches to Learning from Data*

In this course we will consider **likelihood**. This is going to allow us to introduce a general framework for point estimation, interval estimation and hypothesis testing for estimating parameters for an assumed probability distribution. In simplest form, this allows us to make conclusions, for example, about the mean, proportion, variance, rate etc. for particular contexts. However, additionally the course **predictive modelling** introduces that we can obtain more information by investigating relationships between variables, and estimating a response variable using information from other (potential) explanatory variables, through writing down an appropriate statistical model. **Predictive modelling** will introduce the idea of **least squares** to do this. Likelihood can also be used, and provides more flexibility when extending probability distribution assumptions beyond the normal distribution. We'll consider the connection here in a little more detail when we introduce likelihood for the normal distribution in week 8.

We'll introduce the ideas of Bayesian inference through supplementary material for this course (from week 7 onwards) and these ideas will be extended upon in the course **Uncertainty assessment and Bayesian computation**. Likelihood is also a key component of the Bayesian framework where we use likelihood in order to describe our data and we update our data using prior beliefs about our population parameters.

Week 5 learning material aims

The material in week 5 covers:

- properties of point estimators;
- the idea and concept of maximum likelihood;
- maximum likelihood estimation for discrete distributions.

Point Estimation

As we have already seen, a point estimate is a *best guess* at the value of an unknown population parameter. It is a single number, based on a sample of data and for common statistical models such as the binomial and normal, seen in week 4, a little thought about the meaning of a parameter suggests a *natural* way to estimate it.

Binomial model

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.

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

In week 4 example 4, we used the sample proportion of people that could taste the chemical PTC to estimate the proportion of people that could taste PTC. For this we had:

$$\frac{156}{213} = 0.732 \quad \text{or} \quad 73.2\%.$$

Normal model

In a similar way, in the **normal model** with parameters μ and $\sigma^2 (\sigma \geq 0)$, μ is the population mean and σ^2 is the population variance. If 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,

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

In example 1 from week 4, we used the sample mean IQ of 92.2 from a sample of 39 people to estimate the population mean IQ.

A further example could be for the Poisson distribution.

Poisson model

In the **Poisson model** with parameter λ ($\lambda \geq 0$), λ is the expected value, or population mean number, of events in a specified 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 1

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**
No. of accidents	2	2	1	1	0	4	2

Month	**8**	**9**	**10**	**11**	**12**
No. of accidents	1	2	1	1	1

Month	**13**	**14**	**15**	**16**
No. of accidents	3	1	2	2

Month	**17**	**18**	**19**	**20**
No. of accidents	3	2	3	4

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

On average, 1.9 accidents occurred each month.

Definition and Properties of Point Estimators

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. In this notation, we are referring to θ as any unknown parameter for any probability distribution.

Definition 1

A **point estimator** of θ is an algebraic function, $t(\mathbf{X})$, of the data.

Definition 2

A **point estimate** is a particular numeric value of the function, $t(\mathbf{x})$, obtained from a particular set of data, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Example 2

Poisson model - estimating λ

In the Poisson model:

- the point estimator for the population parameter is $\hat{\lambda} = \sum_{i=1}^{n} X_i/n$, where X_i is the number of events at time or point i and n is the sample size.
- the point estimate is the numerical example found from our sample of data.

For example, in example 1 here on car accidents the point estimator is $\sum_{i=1}^{n} X_i/n$, where X_i is the number of events at time i and the point estimate is $\hat{\lambda}$ = 1.9 accidents each month.

Task 1

State the point estimators and point estimates for:

- θ, the population proportion of tasters in the binomial model for the PTC example, given at the beginning of this section on *Point Estimation*, where the number of tasters was 156 and the total sample size was 213;

- μ, the population mean IQ in the normal model for the IQ example, given at the beginning of this section on *Point Estimation*, where the sample mean was 92.2.

In any situation there will be a variety of possible estimators and we need some way of choosing between them.

In order to estimate our true population parameter(s), we would like:

1. our estimator to have values that cover the same range;

2. an estimator which, on average, provides the true population parameter when calculated over all possible samples;

3. the variance of our point estimates obtained over all possible samples to decrease towards zero as the sample size increases towards the size of the population.

More formally, we would like any point estimator $t(\mathbf{X})$ to have the following properties, for a general population parameter θ:

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. (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,

$$\mathbb{E}\{t(\mathbf{X})\} = \theta.$$

3. 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.

$t(\mathbf{X})$ 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

$$\mathrm{Var}\{t(\mathbf{X})\} \to 0 \quad \text{as} \quad n \to \infty,$$

(when $t(\mathbf{X})$ is unbiased).

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

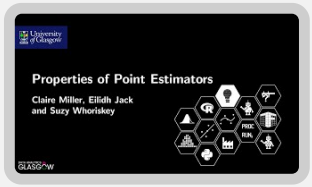
There are additional properties which will be mentioned in subsequent weeks.

The video below provides illustrations and visualisations for the ideas of unbiased and consistent estimators:

Video

Unbiased and consistent estimators

Duration 8:42



The example below, investigates these three properties (range, unbiased and consistent) for a point estimator for the binomial distribution.

Example 3

Binomial model with parameter θ, $0 \leq \theta \leq 1$

It is assumed that $X \sim \mathrm{Bi}(n, \theta)$, where X is the number of success, n is the sample size and θ is the probability of success. The proposed point estimator of θ is $t(X) = X/n$.

Show that $t(X) = X/n$ has the same range as θ and that it is an unbiased and consistent point estimator.

Range :

Since $0 \leq X \leq n$, X/n has the same range as θ.

When $X = 0$, $X/n = 0$ and when $X = n$, $X/n = 1$. Therefore, $(0 \leq t(X) \leq 1)$.

Unbiased :

$\mathbb{E}(X/n) = 1/n\{\mathbb{E}(X)\} = \frac{1}{n}(n\theta)^* = \theta$

Ex: (50/20) = (1/20)*50 = (1/20)*(20*2.5) = 2.5

Consistent :

$\mathrm{Var}(X/n) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}(n\theta(1-\theta))^* = \frac{1}{n}(\theta(1-\theta)).$

and so since $0 \leq \theta \leq 1$, $\mathrm{Var}(X/n) \to 0$ as $n \to \infty$

Probability theory - a reminder

For any random variable X and any real constant k:

$$\mathbb{E}(kX) = k\mathbb{E}(X)$$

$$\mathrm{Var}(kX) = k^2\mathrm{Var}(X).$$

See the **Probability and Stochastic Models** or the **Probability and Sampling Fundamentals/Sampling Fundamentals** courses for definitions of expectation and variance and associated rules.

*See the probability formula sheet (a copy is included on Moodle with this week's material) for expectations and variances of standard distributions.

Task 2

Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables from a $\mathrm{Poi}(\lambda)$ distribution (where $\lambda \geq 0$ is unknown). It is proposed to estimate λ using the sample mean,

$t(\mathbf{X}) = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$.

Show that $t(\mathbf{X})$ has the same range as λ, and that it is an unbiased and consistent estimator of λ.

Moodle cheat sheet:
https://moodle.gla.ac.uk/pluginfile.php/5664309/mod_resource/content/1/formulae.pdf

Perhaps helpful?
https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php

Maximum Likelihood Estimation

Motivating ideas

In week 4 we stated that if we adopt a probability model for the variable of interest in the population, this means that, implicitly, we are claiming to know everything about the population apart from the values of the model's unknown parameters. We referred to this as *parametric inference*, and in parametric inference the problem is reduced to then estimating the model parameters.

In the examples used in week 4, and for defining point estimators above, the estimators were intuitive based on the nature of the data and the question of interest, and checking the properties of range, unbiased and consistent, defined in the *Definition and Properties of Point Estimators*' section, that they were appropriate point estimators for our population parameters.

However, for a given scenario it is not always obvious what an appropriate estimator for a population parameter may be in order to use a particular probability distribution to describe the variability in the data.

For example, let's consider the contexts below:

Suppose for a business model we were interested in predicting the:

- number of customers you would have to call before finding a customer that has experienced a faulty product;
- number of people you would have to poll before finding someone who would vote for an independent candidate;
- number of tweets you would have to assess before finding one from a fabricated source.

From probability theory, we know that a geometric distribution is a possibility to describe the properties of a random variable of the number of trials until (and including) the first failure, with parameter θ the probability of success. However, unlike with the binomial or Poisson distribution, it's not obvious what a sensible estimator is for the parameter θ of the geometric distribution i.e. what is an appropriate way to summarise the 'no. of trials until first failure' information?

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[1]. 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.

To motivate this let's consider the following example for the binomial model where we already know that $\hat{\theta} = X/n$ is a sensible extimator:

Example 4

Suppose we are interested in the number of dentists in the UK that often feel stressed at work. We take a random sample of the number of dentists in the UK, say $n = 30$, and ask them whether or not they often feel stressed at work. From this sample we get the result that 60% often feel stressed at work.

Take X to be the number of dentists who often feel stressed at work (which in this sample is 18), then $X \sim \text{Bi}(30, \theta)$.

For this example, let's consider the probability that we would have seen a value of $X = 18$ for different values of θ using the probability mass function of a binomial distribution.

$\theta = 0.1$

$$p_{0.1}(18) = \binom{30}{18}(0.1)^{18}(0.9)^{12} =< 0.00001$$

$\theta = 0.3$

$$p_{0.3}(18) = \binom{30}{18}(0.3)^{18}(0.7)^{12} = 0.0005$$

These are all very small probabilities. It is unlikely that a sample of 30 dentists would include as many as 18 that often feel stressed if only 10 or 30% of the population of dentists often felt stressed. So 0.1 and 0.3 are implausible values for θ, given the evidence in the data.

Let's consider some larger values for θ:

$\theta = 0.5$

$$p_{0.5}(18) = \binom{30}{18}(0.5)^{18}(0.5)^{12} = 0.081$$

R SYNTAX: (factorial(30)/(factorial(18)*factorial(30-18)))*.5^18*(1-.5)^12

$\theta = 0.6$

$$p_{0.6}(18) = \binom{30}{18}(0.6)^{18}(0.4)^{12} = 0.15$$

R SYNTAX: (factorial(30)/(factorial(18)*factorial(30-18)))*.6^18*(1-.6)^12

$\theta = 0.7$

$$p_{0.7}(18) = \binom{30}{18}(0.7)^{18}(0.3)^{12} = 0.075$$

R SYNTAX: (factorial(30)/(factorial(18)*factorial(30-18)))*.7^18*(1-.7)^12

These are much larger probabilities. It is much more likely that a sample of 30 dentists would include 18 that often suffer from stress if 50-70% of the population of dentists often suffer from stress. So these are all plausible values of θ given the evidence in the data.

We can therefore define our likelihood function $L(\theta; \mathbf{x})$ based on the probability mass function for our data,

$$L(\theta; \mathbf{x}) = p(x).$$

For the example above if we plot the likelihood function by calculating $p(x)$ over a range of θ we get:

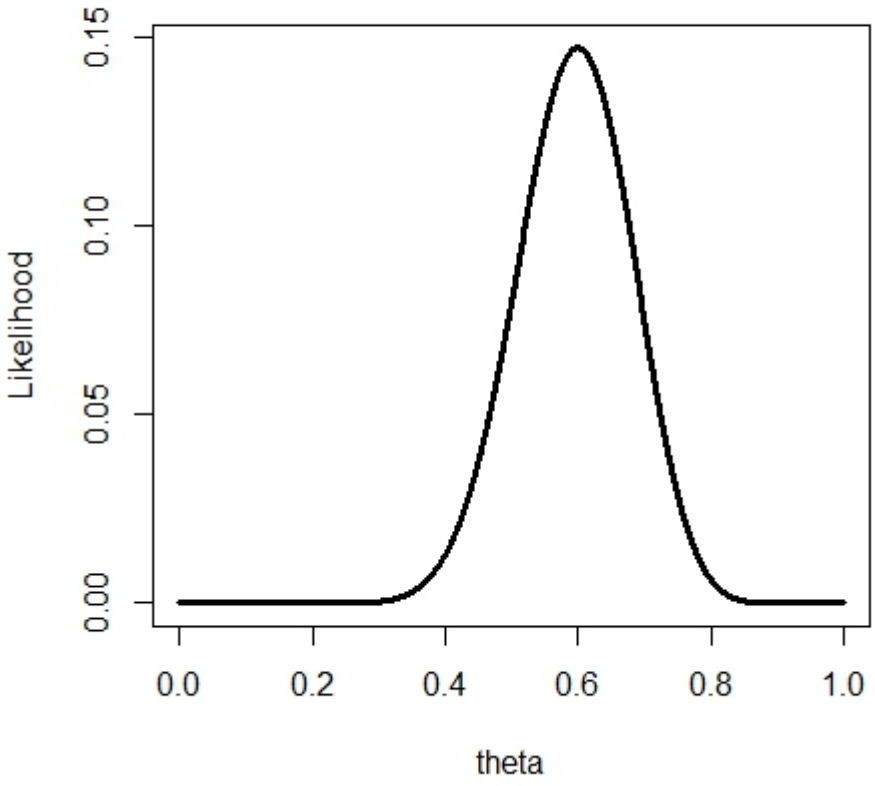


Figure 2

In this example, the likelihood function reaches its maximum value at θ = 0.6. On the basis of the sample data, this is the most plausible value of θ, i.e. $\hat{\theta} = X/n = 18/30 = 0.6$. The value of θ at which the likelihood function reaches its maximum is known as the maximum likelihood estimate of θ. In most models the likelihood is based on a product of several terms and this will be formally derived below.

Summary of motivational ideas

In maximum likelihood estimation, the key idea is that, given observed data and an assumed probability model, we want to find estimates for the population paramaters that maximise the likelihood that our distribution fits our data,

i.e. if we observe x, assuming a geometric distribution, what is the best estimate of the parameter θ for the geometric distribution to fit the data?

The method of maximum likelihood

The maximum likelihood estimate (MLE) is the value of $\hat{\theta}$, where θ is the population parameter for any probability distribution, which maximises the likelihood function $L(\theta; \mathbf{x})$.

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[2] to give an overall probability associated with the data.

In order to find the maximum value of a simple function:

- We first need to find the turning points of the function and this can be done by computing the gradient of a function at each point through differentiation with respect to the unknown parameters of a function.
- Turning points are found where the first derivative is equal to zero.
- We then need to establish which of the turning points are a maximum. A maximum turning point can be identified where there is evidence of negative curvature at a turning point i.e. where the second derivative is less than zero.

Initially, we will introduce likelihood using discrete distributions of binomial and Poisson to illustrate how the process works, and confirm the intuitive point estimators that we already know from previous examples.

Discrete Distributions

Suppose that the discrete random variable X_i is observed on the $_i$th 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:

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

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

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

It is often easier, algebraically, 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)

We considered an example of the binomial distribution in the motivational example of dentists that often feel stressed at work, and visualised the maximum likelihood estimate from the plot of the likelihood. Now let's work through the full derivation of this. So we have,

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

Data: x

and we want to find the estimator and hence estimate of θ which maximises the probability that the binomial distribution fits our data x.

Likelihood:

The idea which lies behind the principle of likelihood is that $p(x)$ can be used to indicate the relative plausibility of any particular value of θ to be the true value of θ.

In this example there is only one data point x and hence the likelihood is just equal to the probability that $X = x$:

$$L(\theta) = p(x),$$

$p(x)$ can be found from the probability distribution sheet for the binomial distribution, and hence:

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad 0 \leq \theta \leq 1.$$

Log-likelihood:

Algebraically, it is then more convenient to take the natural log of the likelihood expression:

$$\ell(\theta) = \log_{\mathrm{e}} \left(\binom{n}{x} \theta^x (1-\theta)^{n-x} \right), \qquad 0 < \theta < 1$$

$$\ell(\theta) = \log_{\mathrm{e}} \binom{n}{x} + \log_{\mathrm{e}}(\theta^x) + \log_{\mathrm{e}} \left((1-\theta)^{n-x} \right), \qquad 0 < \theta < 1$$

$$\ell(\theta) = K + x \log_e \theta + (n-x) \log_e (1-\theta), \qquad 0 < \theta < 1$$

where K is some constant.

The quantity $\binom{n}{x}$ and hence $\log_e \binom{n}{x}$ is a known numeric value based on the sample size and observed data value for x. Therefore, it is often more convenient to simplify this by simply referring to it as a constant K.

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:

$$\frac{d\ell}{d\theta} = \ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}.$$

K+xlog(o)+(n-x)*log(1-o)

https://www.derivative-calculator.net/

Since we are differentiating with respect to θ and K is a known constant (that does not depend on θ), then it disappears after differentiation.

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

$$\begin{aligned}
\frac{d\ell}{d\theta} = \ell'(\theta) &= 0 \\
\frac{x}{\theta} - \frac{n-x}{1-\theta} &= 0 \\
\frac{x}{\theta} &= \frac{n-x}{1-\theta} \\
x(1-\theta) &= \theta(n-x) \\
x - \theta x &= \theta n - \theta x \\
x &= \theta n \\
\theta &= \frac{x}{n}.
\end{aligned}$$

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**:

$$\frac{d^2\ell}{d\theta^2} = \ell''(\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.$$

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:

$$\ell''(\theta) = -\left(\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right) < 0.$$

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 earlier.

Mathematical reminders

You will see in the example above that there are a few mathematical rules that are very useful. Here is a reminder of the key results that we'll use throughout the course:

log rules:

- $\log_e(AB) = \log_e(A) + \log_e(B)$
- $\log_e(A/B) = \log_e(A) - \log_e(B)$
- $\log_e(A^n) = n\log_e(A)$

differentiation:

$$\frac{d}{d\theta}\log_e \theta = \frac{1}{\theta}$$

$$\frac{d}{d\theta}\log_e f(\theta) = \frac{f'(\theta)}{f(\theta)}$$

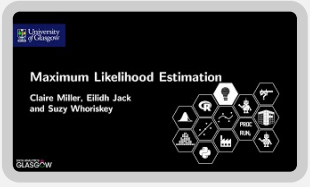
Now let's find the maximum likelihood estimator for the Poisson distribution and in particular apply the results for example 1, to obtain the maximum likelihood estimate.

The video talks through the general approach for maximum likelihood and derives the steps for obtaining the maximum likelihood estimator for the Poisson distribution in detail:

Video

Maximum likelihood estimation with a Poisson example

Duration 12:02



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:

$$L(\lambda) = \prod_{i=1}^{n} p_i(x_i)$$

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i}\, e^{-n\lambda}}{\prod_{i=1}^n x_i!}, \quad \lambda \geq 0$$

Log-likelihood:

$$\ell(\lambda) = \sum_{i=1}^{n} x_i \log_e \lambda - n\lambda + K, \quad \lambda \geq 0$$

for a constant K.

Differentiate wrt λ:

$$\ell'(\lambda) = \frac{\sum_{i=1}^n x_i}{\lambda} - n,$$

Solve for λ:

$$\ell'(\lambda) = 0 \text{ when } \lambda = \frac{\sum_{i=1}^n x_i}{n} = \bar{x},$$

Second derivative:

$$\ell''(\lambda) = -\frac{\sum_{i=1}^n x_i}{\lambda^2},$$

and this is < 0 for all $\lambda > 0$.

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

Likelihood in `R`

Let's look at plots of the likelihood and log-likelihood function in `R` for this example to visualise the results:

```
## create a plausible sequence of values for the population parameter
lambda
lambda = seq(1,3,length=1000)

## the data - number of accidents in each month
x <- c(2,2,1,1,0,4,2,1,2,1,1,1,3,1,2,2,3,2,3,4)

## the sample size
n <- length(x)

## the likelihood function
lik <- (lambda^(sum(x))*exp(-n*lambda))/prod(factorial(x))

## plotting the likelihood function, with a line added at the MLE

par(mfrow=c(1,2))
plot(lambda, lik, ylab="likelihood", xlab="lambda", type="l", lwd=3,
cex.lab=1.5)
abline(v=mean(x), col="yellow", lwd=3)

## the log-likelihood function
loglik <- sum(x)*log(lambda)-n*lambda

## plotting the log-likelihood function, with a line added at the MLE
plot(lambda,loglik, ylab="log-likelihood", type="l", lwd=3, cex.lab=1.5)
abline(v=mean(x), col="yellow", lwd=3)
```

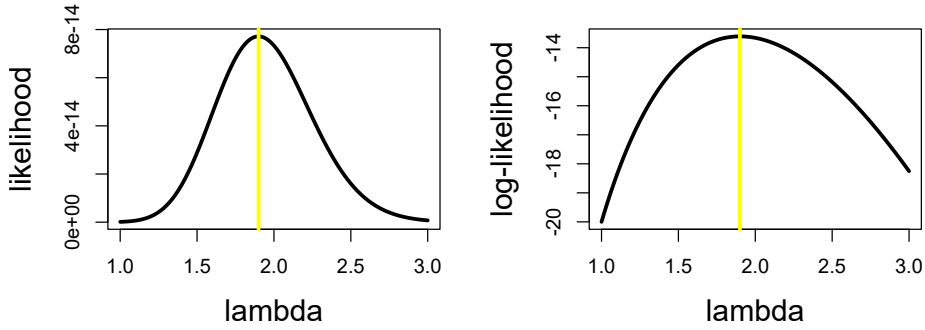


Figure 3

Task 3

Now let's return to our example for the geometric distribution. Suppose that you are interested in predicting the number of customers you would have to call before finding a customer that has experienced a faulty product, and to assess this data $x_1, \ldots, x_n$ have been collected for each of n weeks, where x_i is the number of customers called before finding a customer with a faulty product in week i.

Suppose that $x_1, x_2 \ldots, x_n$ are observations of the independent random variables $X_1, X_2, \ldots, X_n$ respectively. Obtain the maximum likelihood estimator of the unknown parameter θ in $X_i \sim \text{Geo}(\theta)$ where $0 < \theta < 1$.

Learning outcomes for week 5

By the end of week 5, you should be able to:

- write down and justify criteria required of 'good' point estimators;
- check whether or not a proposed estimator within a stated statistical model has the same range as the parameter of interest, is unbiased and consistent;
- apply the principle of maximum likelihood to obtain point estimators (and hence point estimates) of parameters in one-parameter discrete statistical models.

Review exercises, selected video solutions and written answers to all tasks/review exercises are provided overleaf.

Review exercises

Task 4

Suppose that $X_1, X_2, \ldots, X_n$ are independent $\mathsf{U}(0, \theta)$ random variables, for unknown $\theta > 0$. It is proposed to estimate θ by $t(\mathbf{X}) = 2\bar{X}$. By using the fact that a $\mathsf{U}(0, \theta)$ random variable has $\mathbb{E}(X_i) = \frac{1}{2}\theta$ and $\mathrm{Var}(X_i) = \frac{1}{12}\theta^2$. Show that $t(\mathbf{X})$ is an unbiased and consistent point estimator of θ.

Task 5

Suppose that $X_1, X_2, \ldots, X_n$ are independent observations from a $\mathsf{N}(\mu, \sigma^2)$ distribution (where μ is unknown and σ^2 is known). It is proposed to estimate μ using the sample mean, $t(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Show that $t(\mathbf{X})$ has the same range as μ, and that it is an unbiased and consistent estimator of μ.

Task 6

Suppose that $x_1, x_2, \ldots, x_n$ are observations of the independent random variables $X_1, X_2, \ldots, X_n$ respectively. Obtain the maximum likelihood estimator of the unknown parameter θ for $X_1, X_2, \ldots, X_n$, where $X_i \sim \mathsf{Bi}(m_i; \theta)$, $0 < \theta < 1$ and $m_1, m_2, \ldots, m_n$ are known positive integers. (Hint: $L(\theta) = \prod_{i=1}^{n} p_i(x_i)$).

Task 7

A Geiger counter is a device which is used to measure radioactivity. In order to check that the device is calibrated correctly, a measurement can be taken from a source of known

radioactive strength. The counts recorded by the Geiger counter over 200 one second intervals were recorded and these can be represented by the random variables $X_1, \ldots, X_{200}$.

For this particular radioactive source, if the Geiger counter is functioning correctly then the mean count in a one second interval should be 20.

Assume that the counts $X_1, \ldots, X_{200}$ are independent and each follow a $\mathrm{Poi}(\theta)$ distribution, and that the sum of the observed counts was $\sum_{i=1}^{200} x_i = 3654$.

Write down the log-likelihood function for θ. Use this to find the maximum likelihood estimate of θ in terms of the observed values $x_1, \ldots, x_n$, making sure to check that you have, indeed, found a maximum likelihood estimate.

Answer 1

PTC example (example 4 from week 4):

Point estimator = $\hat{\theta} = X/n$, where X is the number of people that could taste PTC in the sample and n is the sample size. The point estimate is $\hat{\theta}$ = 0.732.

IQ example (example 1 from week 4):

Point estimator = $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, the sample mean of the IQ data. The point estimate is the numerical value of this which was $\hat{\mu}$ = 92.2.

Answer 2

It is assumed that $X_1, X_2, \ldots, X_n$ are all independently distributed, with each $X_i \sim \mathrm{Poi}(\lambda)$. This means, in particular, that $\mathbb{E}(X_i) = \lambda$ and $\mathrm{Var}(X_i) = \lambda$. The parameter λ is restricted to be non-negative. The proposed estimator of λ is $t(\mathbf{X}) = \bar{X}$.

Range : Each X_i is a count (0,1,$\cdots$) and so $\bar{X}$ must be non-negative, like λ itself.

Unbiased :
$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mathbb{E}(\sum_{i=1}^{n} \frac{X_i}{n}) = \sum_{i=1}^{n}(\frac{1}{n}\mathbb{E}(X_i))^* = \sum_{i=1}^{n} \frac{1}{n}\lambda = n.\frac{1}{n}\lambda = \lambda$

Consistent :
$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mathrm{Var}(\sum_{i=1}^{n} \frac{X_i}{n}) = \sum_{i=1}^{n}(\frac{1}{n^2}\mathrm{Var}(X_i))^* = \sum_{i=1}^{n} \frac{1}{n^2}\lambda = n.\frac{1}{n^2}\lambda = \frac{\lambda}{n}$

and so $\mathrm{Var}(\bar{X}) \to 0$ as $n \to \infty$.

*See the probability formula sheet (a copy is included on Moodle with this week's material) for expectations and variances of standard distributions.

Answer 3

$$p(x_i) = \theta^{x_i - 1}(1-\theta), \qquad x_i = 1, 2, \ldots$$

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=}^{n} \theta^{x_i - 1}(1-\theta), \quad 0 < \theta < 1$$

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} (x_i - 1)\log_e(\theta) + \log_e(1-\theta)^n \\ &= \log_e(\theta)(\sum_{i=1}^{n} x_i - n) + n\log_e(1-\theta) \end{aligned}$$

$$\ell'(\theta) = \frac{1}{\theta}(\sum_{i=1}^{n} x_i - n) - \frac{n}{1-\theta}$$

So $\ell'(\theta) = 0$ when,

$$\begin{aligned} (1-\theta)(\sum_{i=1}^{n} x_i - n) &= n\theta \\ \sum_{i=1}^{n} x_i - n &= \theta \sum_{i=1}^{n} x_i \\ \theta &= \frac{\sum_{i=1}^{n} x_i - n}{\sum_{i=1}^{n} x_i} \end{aligned}$$

So $\ell(\theta)$ has a turning point at $\theta = \frac{\sum_{i=1}^{n} x_i - n}{\sum_{i=1}^{n} x_i}$.

Now,

$$\ell''(\theta) = -\frac{1}{\theta^2}(\sum_{i=1}^{n} x_i - n) - \frac{n}{(1-\theta)^2} < 0$$

for all θ in the range 0 to 1.

So the turning point at $\theta = \frac{\sum_{i=1}^{n} x_i - n}{\sum_{i=1}^{n} x_i}$ is a local maximum and $\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} x_i - n}{\sum_{i=1}^{n} x_i}$.

Answer 4

Illustrating properties of unbiased and consistent:

Unbiased :
$\mathbb{E}(2\bar{X}) = \mathbb{E}(\frac{2}{n}\sum_{i=1}^{n} X_i) = \mathbb{E}(\sum_{i=1}^{n} \frac{2X_i}{n}) = \sum_{i=1}^{n}(\frac{2}{n}\mathbb{E}(X_i)) = \sum_{i=1}^{n} \frac{2}{n}\frac{\theta}{2} = n.\frac{1}{n}\theta = \theta$

Consistent :
$\mathrm{Var}(2\bar{X}) = \mathrm{Var}(\frac{2}{n}\sum_{i=1}^{n} X_i) = \mathrm{Var}(\sum_{i=1}^{n} \frac{2X_i}{n}) = \sum_{i=1}^{n}(\frac{4}{n^2}\mathrm{Var}(X_i)) = \sum_{i=1}^{n} \frac{4}{n^2}\frac{\theta^2}{12} = \frac{\theta^2}{3n}$

and so $\mathrm{var}(\bar{X}) \to 0$ as $n \to \infty$

Answer 5

Illustrating properties of unbiased and consistent:

It is assumed that $X_1, X_2, \ldots, X_n$ are all independently distributed, with each $X_i \sim \mathrm{N}(\mu, \sigma^2)$. This means, in particular, that $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma$. The proposed estimator of μ is $t(\mathbf{X}) = \bar{X}$.

Range : Clearly $-\infty < \bar{X} < \infty$

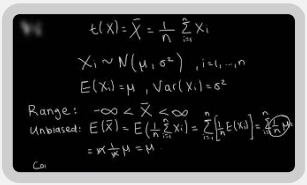
Unbiased :
$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mathbb{E}(\sum_{i=1}^{n} \frac{X_i}{n}) = \sum_{i=1}^{n}(\frac{1}{n}\mathbb{E}(X_i)) = \sum_{i=1}^{n} \frac{1}{n}\mu = n.\frac{1}{n}\mu = \mu$

Consistent :
$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mathrm{Var}(\sum_{i=1}^{n} \frac{X_i}{n}) = \sum_{i=1}^{n}(\frac{1}{n^2}\mathrm{Var}(X_i)) = \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2 = n.\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$
and so $\mathrm{Var}(\bar{X}) \to 0$ as $n \to \infty$

Video

Video model answers for task 5

Duration 4:55



Answer 6

$$p(x_i) = \binom{m_i}{x_i}\theta^{x_i}(1-\theta)^{m_i-x_i}, \qquad x_i = 0, 1, \ldots, m_i$$

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n}\binom{m_i}{x_i}\theta^{x_i}(1-\theta)^{m_i-x_i}, \quad 0 < \theta < 1$$

$$\begin{aligned}\ell(\theta) &= K + \sum_{i=1}^{n} x_i \log_e(\theta) + \sum_{i=1}^{n}(m_i - x_i)\log_e(1-\theta) \\ &= K + \log_e(\theta)\sum_{i=1}^{n} x_i + \log_e(1-\theta).\left(M - \sum_{i=1}^{n} x_i\right)\end{aligned}$$

where $M = \sum_{i=1}^{n} m_i$ is the total number of 'trials' and $\sum_{i=1}^{n} x_i$ is the total number of 'successes'.

$$\ell'(\theta) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\left(M - \sum_{i=1}^{n} x_i\right) = 0$$

$$\begin{aligned}(1-\theta)\sum_{i=1}^{n} x_i &= \theta\left(M - \sum_{i=1}^{n} x_i\right) \\ \sum_{i=1}^{n} x_i &= \theta M \\ \theta &= \frac{1}{M}\sum_{i=1}^{n} x_i\end{aligned}$$

So $\ell(\theta)$ has a turning point at $\theta = \frac{1}{M}\sum_{i=1}^{n} x_i$.

Now,

$$\ell''(\theta) = -\frac{1}{\theta^2}\sum_{i=1}^{n} x_i - \frac{1}{(1-\theta)^2}\left(M - \sum_{i=1}^{n} x_i\right) < 0$$

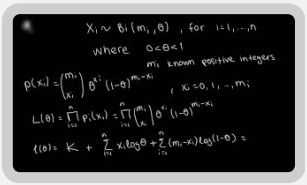
for all θ in the range 0 to 1.

So the turning point at $\theta = \frac{1}{M}\sum_{i=1}^{n} x_i$ is a local maximum and $\hat{\theta}_{MLE} = \frac{1}{M}\sum_{i=1}^{n} x_i$.

Video

Video model answers for task 6

Duration 9:25



Answer 7

Geiger counter

The likelihood function is

$L(\theta) = \prod_{i=1}^{n} \theta^{x_i} e^{-\theta} / x_i! = \theta^{\sum_i x_i} e^{-n\theta} / \prod_i x_i!$

The log-likelihood function is

$\ell(\theta) = (\sum_{i=1}^{n} x_i) \log_e \theta - n\theta + K$

The derivatives are

$\ell'(\theta) = (\sum_{i=1}^{n} x_i)/\theta - n$

$\ell'(\theta) = (\sum_{i=1}^{n} x_i)/\theta - n = 0, \quad \frac{\sum x_i}{\theta} = n, \ \theta = \frac{\sum x_i}{n} = \bar{x}$

$\ell''(\theta) = -(\sum_{i=1}^{n} x_i)/\theta^2$

The first derivative equals 0 when $\theta = \bar{x}$.

The fact that the second derivative is negative everywhere establishes that this is the MLE.

$$\hat{\theta}_{MLE} = 3654/200 = 18.27$$

Footnotes

1. ordinary least squares is also widely used, see the Predictive Modelling course, but in the first instance is restricted to normally distributed data ↩

2. See the Probability and Stochastic Models or the Probability and Sampling Fundamentals/Sampling Fundamentals course for details here ↩