STAT0008 Lecture 1
Introduction and Likelihood-based Inference

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

1st October 2018

STAT0008 – Basic Information

Lecturer

- Dr. Aidan O'Keeffe
- Office hour: Tuesday 3pm–4pm, Office 237, 1-19 Torrington Place, Department of Statistical Science (or please e-mail the lecturer to arrange an alternative time).

Moodle

- Lecture slides and tutorial sheets etc. can be found on the STAT0008 Moodle page.

STAT0008 – Basic Information

Timetable

- LECTURES: Monday 0900–1100: Medical Sciences 131 AV Hill Lecture Theatre

- TUTORIALS: Weeks 7–16 only (excl. reading week)
  - Thursday 0900–1000 - Groups 1 and 2
  - Friday 1100–1200 - Groups 3 and 4

- WORKSHOPS: Weeks 10 and 14
  - Friday 1400–1600, Bentham House SB31, Denys Holland Lecture Theatre

Assessment

Examination: A $2\frac{1}{2}$ hour written examination (for Level 6 students) or a 2 hour written examination (for Level 7 students) in Term 3.

In course assessment: A closed book in-class test (45 minutes) will take place on Monday 19th November at 09:00, in place of the first hour of the lecture on that day.

Final mark for STAT0008 is weighted as follows:

- 90% Written examination
- 10% In-course assessment

Course Objectives

Statistical inference concerns the use of data to make conclusions/learn about a population. Typically, we take data from a sample where the sample size is far smaller than the population size. In making inference we could be taking data and asking questions such as:

- What are these data telling me?
- Based on these data, what can we believe about our population?

Statistical inference is used in many applied fields. For example: finance, medicine, education, agriculture, commerce, politics, psychology, sport, marketing. . .

In summary, any piece of data analysis = 'Statistical inference'!

Types of Statistical Inference

Statistical inference refers to a wide range of methods/approaches to data analysis. Broadly speaking, we may define our analyses as either '**descriptive**' or '**inferential**'.

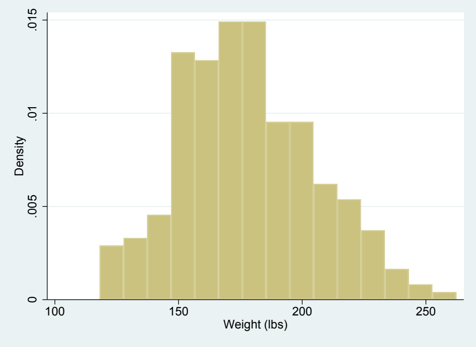
Descriptive statistics: These are basic summaries of observed data and, typically, the data are not assumed to have any particular probability distribution. We concentrate on the observed data and do not necessarily attempt to infer anything about a wider population using descriptive statistics. Common examples of descriptive statistics include:

- Mean, standard deviation
- Graphs: bar charts, histograms, scatter plots
- Summary tables (categorical data)

Descriptive statistics: Example

251 adult males participate in a medical study. The sample mean weight of these subjects is calculated as 178lbs.

A histogram of the weights of the subjects is shown below.



Just information – no distributional assumptions!

Types of Statistical Inference

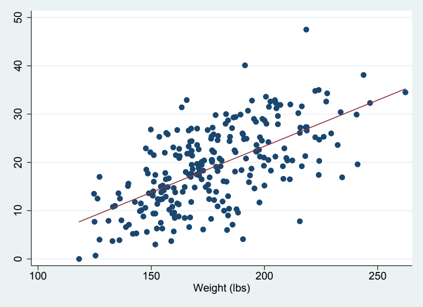
Inferential statistics: Statistical approaches where the main aim is to infer something about a population.

Typically, a probability distribution may be assumed for some outcome of interest and data from a sample are modelled to draw conclusions about the behaviour of a the same outcome at the population level. Common examples of inferential statistics include:

- 95% confidence intervals
- Hypothesis testing (e.g. t-tests, Pearson's Chi-square test)
- Linear models and analysis of variance

Inferential statistics: Example

Using the same group of 251 males, we fit a **normal linear model** to examine the relationship between % body fat and weight.



We obtain the following regression line equation:

$$\text{\% Body Fat} = -14.89 + 0.19(\text{Body Weight (lbs)})$$

A 95% confidence interval for the coefficient of Body Weight is (0.16, 0.22).

Statistical Inference: Process

Traditionally, we take the following steps when performing statistical inference.

1. Choose some outcome of interest (the variable about which you'd like to find out something). Assuming that this variable is measured in a population, call this variable X_i (i denotes the i^{th} member of the population).

2. Usually (but not always!) choose a probability distribution/model for the outcomes X_i. That is, set

$$X_i \sim \mathcal{D}(\theta)$$

where $\mathcal{D}(\theta)$ is some probability distribution with (unknown) parameter(s) θ.

Statistical Inference: Process

3. Assume that a sample of size n is to be taken from the population. The sample is denoted

$$\{X_1, \ldots, X_n\}$$

   and we construct the **joint distribution** of $\{X_1, \ldots, X_n\}$, conditional on parameters θ and our earlier choice of $X_i \sim \mathcal{D}(\theta)$.

4. Take **sampled values** from the population (denoted $\{x_1, \ldots, x_n\}$) and use these, together with the assumptions made in Steps 1–3, to **estimate** the **parameter(s)** θ.

Then, we might assess the fit of the model/reflect on the sampled data and perhaps change our assumptions - an iterative procedure!

Statistical Inference: Process - Example

1. Suppose that we want to learn about the weight of babies born in London during 2017. Let X_i be the weight of the i^{th} baby born in London within some population of size N ($N =$ total number of babies born in London in 2017).

2. Now, weight would naturally have a continuous probability distribution, so let us assume that

   $$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

   (so we've chosen a Normal distribution for $\mathcal{D}(\theta)$ and our parameters of interest are $\theta = (\mu, \sigma)$).

Statistical Inference: Process - Example

3. We shall take a sample of n babies and denote the, as yet unknown, weights from the sample $\{X_1, \ldots, X_n\}$.

   To consider a joint distribution for $\{X_1, \ldots, X_n\}$, let us assume that each baby's weight is independent of all other baby weights. We say that $X_1, \ldots, X_n$ are **independent and identically distributed** and this specifies the joint distribution of $\{X_1, \ldots, X_n\}$.

Statistical Inference: Process - Example

4. Then we take measured the weights of the n sampled babies, denoting these $\{x_1, \ldots, x_n\}$. These are realised values of $\{X_1, \ldots, X_n\}$ and we use them to **make inference** about the distribution of $\{X_1, \ldots, X_N\}$.

   In this case, making inference would probably mean using $x_1, \ldots, x_n$ to **estimate** the **parameters** μ and σ^2. For example, the sample mean

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

   as an estimate of μ.

Random Sample

The random variables $X_1, \ldots, X_n$ are a **random sample** of size n from a population with probability distribution $\mathcal{D}(\theta)$ if $X_1, \ldots, X_n$ are independent of oneanother (mutually independent) and each X_i has the same probability distribution $\mathcal{D}(\theta)$

Then we say that $X_1, \ldots, X_n$ are independent and identically distributed random variables, each having the probability distribution $\mathcal{D}(\theta)$.

Parametric Model

A **parametric statistical model** occurs where a random variable of interest is assumed have some **probability distribution**, characterised by an unknown parameter (or set of parameters) θ.

A parametric model consists of a set of probability distributions, characterised by θ. The set of possible values of θ is called the **parameter space**. Often, we write

$$\theta \in \Theta \subseteq \mathbb{R}^p.$$

where θ is the parameter or set of parameters and Θ denotes the parameter space. Here, p denotes the dimension of the parameter/set of parameters θ.

Parametric Model - Example

In a factory that produces electric kettles, let $X_1, \ldots, X_n$ be a random sample of lifetimes of n kettles produced by the factory. We shall assume that the lifetimes of kettles are independent and that each kettle's lifetime has an Exponential distribution with parameter λ. That is,

$$X_i \sim \mathsf{Exp}(\lambda).$$

So, we have assumed a parametric model for the kettle lifetime, in that each lifetime has probability density function

$$f_{X_i}(x_i \mid \lambda) = \lambda \exp(-\lambda x_i)$$

with unknown parameter $\lambda \in (0, \infty) \leftarrow$ parameter space.

Parametric Model - Example

Calculate an expression for the probability that all n kettles in the sample have a lifetime longer than 2 time units.

Parametric Model - Statistics

When forming a parametric model, our aim is often to estimate the model parameters. For parameter estimation we typically consider a **statistic** formed from our sample.

Definition: Statistic

Given a random sample $X_1, \ldots, X_n$, a **statistic** is any real-valued function $T = T(X_1, \ldots, X_n)$ of the random variables $X_1, \ldots, X_n$.

Some examples:

$$T(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{(sample mean)}$$

$$T(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{(sample variance)}$$

$$T(\mathbf{X}) = \max(X_1, \ldots, X_n) \quad \text{(sample maximum)}$$

Parametric Model - Statistics

We note that $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ is a random variable, because $T(\mathbf{X})$ is a function of the random variables $X_1, \ldots, X_n$.

In addition, the distribution of $T(\mathbf{X})$ can be derived from the joint distribution of $X_1, \ldots, X_n$ (and this distribution will depend on θ if the joint distribution of $X_1, \ldots, X_n$ also depends on θ).

Formally, the probability distribution of $T(\mathbf{X})$ is called the **sampling distribution** of the statistic $T(\mathbf{X})$.

Sampling Distribution - Definition

Definition: Sampling Distribution of a Statistic

Suppose that $T = T(X_1, \ldots, X_n)$ is a statistic formed from the sample $X_1, \ldots, X_n$. If all possible samples of size n were taken from the population and the statistic T calculated for each sample, then the distribution of these values is the sampling distribution of the statistic T.

In practice, the sampling distribution of a statistic can be obtained mathematically or through simulations.

Estimator, Estimate and Estimand

As previously, suppose that $X_1, \ldots, X_n$ are an iid sample such that $X_i \sim \mathcal{D}(\theta)$. Then

An **estimator** of the parameter θ is a statistic $T(\mathbf{X}) = \phi(X_1, \ldots, X_n)$ (some function of $X_1, \ldots, X_n$) used to estimate θ

The **estimand** is the parameter that we would like to estimate. In this case, the estimand is θ.

Once our sample of data is actually observed (denoted $x_1, \ldots, x_n$) then we can calculate a realised value of T, written

$$t = t(x_1, \ldots, x_n).$$

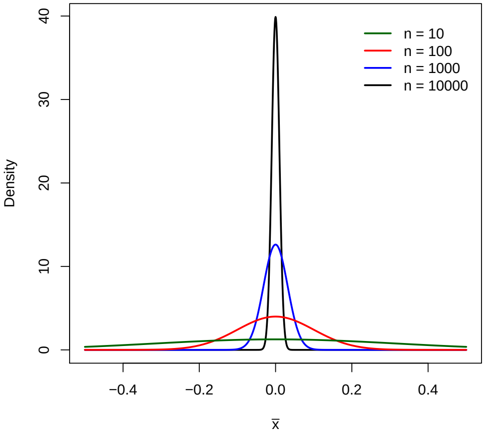
This value is called the **estimate** of θ and is typically written $\hat{\theta}$.

Example

In a London borough, 1000 eligible voters are surveyed in the run-up to an election. Of these 1000 responses, 400 say that they intend to vote *Conservative*.

Sampling Distribution - Example

Suppose that $X_i \sim \mathcal{N}(0,1)$ independently for each i. Below are simulation plots of the distribution of the sample mean for samples of different sizes ($n = 10, 100, 1000, 10000$).



Sampling Distribution - Example

We see that, as n gets larger, the variability of the sampling distribution of $\bar{X}$ decreases (i.e. the **estimator**, $\bar{X}$ becomes more **precise**.

Here, we took $\bar{X}$ as an *intuitive* estimator for μ.

However, for now, this choice was just a 'guess' with no theoretical justification. We shall now consider how to choose estimators for parameters of interest.

First, we shall consider and define an important function that is used commonly in parametric inference, known as a **likelihood function**.

Likelihood Function - Single Observation

Definition: Likelihood Function

Suppose that X is a random variable with probability density (or mass) function $f_X(x;\theta)$, where θ denotes the parameter(s) of the distribution function of X. Given some observed value x of X, the **likelihood function** for θ is defined as

$$\mathcal{L}(\theta \mid x) = f_X(x;\theta).$$

Thus, the likelihood function is simply the density function but written as a function of the parameter(s) θ for a fixed x.

Likelihood Function - Single Observation: Example

The number of accidents per year at a particular road junction is assumed to have a Poisson distribution with mean λ. Suppose that x accidents are observed in a year. Write down the likelihood function for θ.

Likelihood Function - Multiple Observations

If we have a case of multiple observations, where $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector of observed values of the random variables $X_1, \ldots, X_n$ then the likelihood function for θ is written

$$\mathcal{L}(\theta \mid \mathbf{x}) = f_X(\mathbf{x}; \theta).$$

That is, the joint probability density (or mass) function of $\mathbf{x}$ written as a function of θ.

Notably (and often encountered in this course) if $X_1, \ldots, X_n$ are iid with probability density (mass) function $f(x_i; \theta)$, then the likelihood function of θ is written

$$\mathcal{L}(\theta \mid \mathbf{x}) = f(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Likelihood Function - Multiple Observations: Example

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables such that each $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Given that a sample $\mathbf{x}$ of $\mathbf{X}$ is observed, write down the likelihood function for $(\mu, \sigma^2)^T$.

Maximum Likelihood Estimation

Typically, we are interested in estimating the parameter(s) θ. Since the likelihood function reflects the probability density of the observed data, it makes intuitive sense to use the value of θ that **maximises** the likelihood function as an estimate of θ.

This is known as the **maximum likelihood estimate** of θ and is often written as $\hat{\theta}$.

Note that we can also consider the likelihood function $\mathcal{L}(\theta \mid \mathbf{x})$ as a function of random variables $X_1, \ldots, X_n$ and, in this case we would write the likelihood function as

$$\mathcal{L}(\theta \mid \mathbf{X})$$

Now, on maximising this function we obtain the **maximum likelihood estimator** for θ and this is written as $\hat{\Theta}$.

Log-likelihood Function

Typically, we consider a **log-likelihood function** which is defined as the natural logarithm of a likelihood function and denoted

$$\ell(\theta \mid \mathbf{x}) = \log[\mathcal{L}(\theta \mid \mathbf{x})].$$

Since the natural logarithm is a strictly increasing function, the value of θ that maximises $\mathcal{L}(\theta \mid \mathbf{x})$ is also the value of θ that maximises $\ell(\theta \mid \mathbf{x})$. That is, a maximum likelihood estimate of θ is such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta \mid \mathbf{x}) = \arg\max_{\theta \in \Theta} \ell(\theta \mid \mathbf{x}).$$

In many situations, it is easier to work with the log-likelihood function and obtain a maximum likelihood estimate by maximising the log-likelihood function.

Obtaining a Maximum Likelihood Estimate

Finding a maximum likelihood estimate involves determining $\hat{\theta}$ where

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta \mid \mathbf{x}).$$

Typically, this supremum is calculated by differentiating the log-likelihood function with respect to θ, setting equal to zero and solving for θ.

That is, the solution to the equation (or set of equations if θ has more than one dimension)

$$\left.\frac{\partial \ell(\theta \mid \mathbf{x})}{\partial \theta}\right|_{\hat{\theta}} = 0.$$

This equation called a **score equation**.

Maximum Likelihood Estimation - Example

Suppose that we observe r successes in a sequence of n independent Bernoulli trials where the probability of a success is p. Write down the likelihood function of p and compute the corresponding maximum likelihood estimate.

Maximum Likelihood Estimation - Example

Obtaining a Maximum Likelihood Estimate

We should remember that maximising a likelihood function by solving a score equation provides only a local maximum (if it exists).

A global maximum may lie elsewhere (possibly at the boundary of the parameter space).

Invariance Property of the MLE

If $\hat{\theta}$ is a maximum likelihood estimate of θ and g is a function then $g(\hat{\theta})$ is a maximum likelihood estimate of $g(\theta)$.

Proof - see course moodle page (Non-examinable)

This is known as the **invariance property** of maximum likelihood estimates.

The Likelihood Principle

Given a sample of data and an assumed statistical model, then the **likelihood principle** states that all information from the sample that is relevant to the model parameters is contained within the likelihood function.

Moreover, two likelihood functions are considered to be equivalent if one is a scalar multiple of the other.

More formally, suppose that $\mathbf{x}$ and $\mathbf{y}$ are two sets of observations that are modelled to depend on a parameter θ, with likelihood functions $\mathcal{L}_1(\theta \mid \mathbf{x})$ and $\mathcal{L}_2(\theta \mid \mathbf{y})$ respectively. Then if there exists some constant c such that

$$\mathcal{L}_1(\theta \mid \mathbf{x}) = c\mathcal{L}_2(\theta \mid \mathbf{y})$$

is satisfied for every θ then these samples contain the same information about θ and will lead to identical inference on θ.

The Likelihood Principle - An Example

In a factory that produces electrical components, the management are interested in estimating the proportion of defective components, θ produced by the factory.

Twelve components were randomly selected from the factory's production line, of which two were found to be defective.

Now suppose that two statisticians (Statistician A and Statistician B) are asked to use these data to estimate θ.

The Likelihood Principle - An Example

Statistician A decides to use a **Binomial distribution** to model the data and defines:

$$X = \text{ Number of failed components in a sample of 12;}$$

so $X \sim \text{Bin}(12, \theta)$.

The likelihood function of θ is

$$\begin{aligned}\mathcal{L}_1(\theta \mid x = 2) = \mathbb{P}(X = 2 \mid \theta) &= \binom{12}{2}\theta^2(1-\theta)^{10} \\ &= 66\theta^2(1-\theta)^{10}.\end{aligned}$$

The Likelihood Principle - An Example

Statistician B decides to use a **Negative Binomial distribution** to model the data. We briefly recap the negative binomial distribution before outlining Statistician B's model.

Let Y be the number of independent Bernoulli trials required to observe r 'successes' where the probability of a success is p. Then we say that Y has a **Negative Binomial** distribution and the probability mass function of Y is

$$\mathbb{P}(Y = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

where $r \in \{0, 1, 2, \ldots\}$ and $k \in \{r, r+1, \ldots\}$.

The Likelihood Principle - An Example

Statistician B uses a **Negative Binomial distribution** to model the data and defines:

$$Y = \text{ Number of components sampled until two defective components are identified;}$$

so $Y \sim \text{NegBin}(2, \theta)$.

The likelihood function of θ is then

$$\begin{aligned}\mathcal{L}_2(\theta \mid y = 12) = \mathbb{P}(Y = 12 \mid \theta) &= \binom{12-1}{2-1}\theta^2(1-\theta)^{10} \\ &= 11\theta^2(1-\theta)^{10}.\end{aligned}$$

The Likelihood Principle - An Example

Hence, we have two likelihood functions of θ:

$$\mathcal{L}_1(\theta \mid x = 2) = 66\theta^2(1-\theta)^{10}$$
$$\mathcal{L}_2(\theta \mid y = 12) = 11\theta^2(1-\theta)^{10}$$

and

$$\mathcal{L}_1(\theta \mid x = 2) = 6\mathcal{L}_2(\theta \mid y = 12) \quad (\textit{proportionality}).$$

Therefore, according to the likelihood principle, both likelihood functions contain the same information on θ and each will lead to the same inference on θ. (Check this as an easy exercise!)

The Likelihood Principle - An Example

In both cases (Statisticians A and B) we needed to know only the total number of trials (samples) and the total number of 'successes' (defective items).

Since, in each case, the individual samples (Bernoulli trials) are assumed to be independent, then knowing the actual result of each trial (the order in which the defective items occurred) was of no benefit when seeking to estimate θ.

More formally, assuming that there are n independent Bernoulli trials and r 'successes', then we define the random variables

$$\begin{aligned} R &= \text{ Total number of successes} \\ X_i &= \text{ Result of the } i^{\text{th}} \text{ trial } (1 = \text{success}, 0 = \text{failure}) \end{aligned}$$

where $i \in \{1, \ldots, n\}$.

The Likelihood Principle - An Example

The joint probability mass function of $X_1, \ldots, X_n$, conditional on $R = r$, is written

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid R = r) &= \frac{\mathbb{P}(X_1 = x_1 \cap X_2 = x_2 \ldots \cap X_n = x_n)}{\mathbb{P}(R = r)} \\
&= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} \\
&= \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} \\
&= \frac{\theta^r(1-\theta)^{n-r}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} \\
&= \frac{1}{\binom{n}{r}}.
\end{aligned}$$

where $x_1 + \ldots + x_n = r$.

The Likelihood Principle - An Example

We see that the probability mass function for $X_1, \ldots, X_n$ conditional on R does not depend on θ.

As as result, we do not need to know about the outcomes of the *individual* Bernoulli trials to gain information about θ. Instead, we need only know the total number of successes, r, from the n trials in order to learn about θ.

We say that the number of successes is **sufficient** for θ. When we consider the number of successes as a random variable, R, we say that R is a **sufficient statistic** for θ. We shall consider sufficiency in more detail during Lecture 2.

Summary and Learning Outcomes

- Types of statistical inference
  - Descriptive
  - Inferential
- Parametric models and inference
- A **statistic** and its **sampling distribution**
- **Estimators**, **estimates** and **estimands**
- The likelihood function and maximum likelihood estimation