Statistical estimation
Statistical modelling: theory and practice

Gilles Guillot

gigu@dtu.dk

September 3, 2013
Introductory example

- A batch of 1000 electronic components contains some faulty items. One takes a sample of size 100 with replacement, of which 3 are faulty. What is the proportion of faulty items in the batch?
- Arriving in a new city, you see a tram passing in the street with the number 16. How many tram lines are there in this city?
- For a set of measurements $y_1, \ldots, y_n$ of temperatures at dates $t_1, \ldots, t_n$ observed at a certain location, we want to fit a line $y = at + b$. What are the values $a$ and $b$ that best fit the data?
A common set up

- We have some data
- There is a “mechanism” that generates the data
- This mechanism depends on an unknown parameter that we want to estimate
The Statistics way

We relate the unknown parameter to the data by mean of a probability distribution.

- Proportion of faulty items: the number of faulty items in the sample can be assumed to be follow a binomial distribution $B(n, p)$
- Tram lines: the number observed can be assumed to follow a uniform distribution $U\{1, ..., N\}$
Denoting generically $\theta$ the unknown parameter value, an estimator is a rule (or algorithm) allowing us to “guess” $\theta$, from the data. From a mathematical point of view, it is a function

$$\mathbb{R}^n \rightarrow \mathbb{R}^d$$

$$(x_1, \ldots, x_n) \rightarrow \hat{\theta}$$
\begin{itemize}
  \item $d$ is the dimension of the parameter space (often $d = 1$ for us)
  \item Since we assume that data are random, we will often stress this by denoting them $(X_1, \ldots, X_n)$.
  \item The number $\hat{\theta}$ is an estimate of $\theta$. It is a random variable, denoted sometimes $\hat{\theta}(X_1, \ldots, X_n)$
\end{itemize}
Bias of an estimator

Definition: bias

The bias of an estimator is the average discrepancy between the estimate and the true parameter value:

\[
\text{Bias}(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta
\]

An estimator is said to be \textit{unbiased} if \(\text{Bias}(\hat{\theta}) = 0\)
Precision and accuracy

Precision and accuracy are two concepts that belong to science and engineering best explained by the figure below:

In statistics, we have two related concepts: variance and mean square error.
Definition: variance of an estimator

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

\( V[\hat{\theta}] \) is a measure of how much \( \hat{\theta} \) is scattered around its mean (which may differ from the true value \( \theta \)).

Definition: mean square error of an estimator

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

is a measure of how much \( \hat{\theta} \) is scattered around the true value \( \theta \).

When \( \hat{\theta} \) is unbiased, \( E[\hat{\theta}] = \theta \) hence \( V[\hat{\theta}] = MSE[\hat{\theta}] \).
Definition: confidence interval

A confidence interval at level $(1 - \alpha) \in [0, 1]$ is an interval that contains the true unknown parameter value $\theta$ with probability $1 - \alpha$. 
Example: estimation of a proportion

We have a sample of \( n \) objects of which \( x \) are faulty. We estimate the unknown proportion \( p \) by \( \hat{p} = x/n \).

Exercise: give the bias and variance of \( \hat{p} \).
A new look at the probability of the data

We consider again the problem of estimating a proportion with binomial sampling.

- $P(X = x) = p^x (1 - p)^{n-x}$ is the probability to obtain $x$ faulty objects.
- If we consider $p^x (1 - p)^{n-x}$ as a function of $p$, it can be interpreted as the likelihood of the unknown parameter $p$.

To acknowledge the dependence on $p$, we denote $L(x; p) = p^x (1 - p)^{n-x}$ or for short $L(p)$. 

![Graph showing the probability of data against $p$.]
The maximum likelihood principle

- The above suggests a method to estimate the unknown parameter $p$:
  \[ \hat{p} = \text{Argmax}_p \ p^x (1 - p)^{n-x} \]

- $\hat{p}$ is the parameter value that makes our data most probable.
- It is known as the **Maximum Likelihood Estimate** of $p$. and denoted $\hat{p}_{ML}$

Note that defining $\hat{p}$ as $\text{Argmax}_p \binom{n}{x} p^x (1 - p)^{n-x}$ would lead to the same result as $\binom{n}{x}$ does not depend on $p$. 
Likelihood in a general statistical model

Definition: likelihood function

- We consider a dataset consisting of $n$ observations $(x_1, ..., x_n)$
- We assume that the probability density function or probability mass function of $(x_1, ..., x_n)$ denoted by $f_\theta(x_1, ..., x_n)$ is known up to an unknown parameter $\theta$.

The likelihood function $L$ is defined as

$$L(x_1, ..., x_n; \theta) = f_\theta(x_1, ..., x_n)$$
Examples of likelihood functions

Poisson counts

- We observe the number of phone calls at various calling centres over a given period and denote them by \((x_1, ..., x_n)\).
- We assume that the \(x_i\) are independent realizations of a Poisson random variable \(X_i\) with parameter \(\lambda\), i.e.
  \[
P(X_i = x) = \exp(-\lambda)\frac{\lambda^x}{x!}
  \]
- NB: \(x \in \mathbb{N}\) and \(\lambda \in \mathbb{R}_+\)
- \(E[X_i] = \lambda\) and \(V[X_i] = \lambda\)
Likelihood for i.i.d Poisson observations

Remember: "Likelihood = probability of data for a given parameter value"

\[
L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \exp(-\lambda) \frac{\lambda^{x_i}}{x_i!} \\
= \exp(-n\lambda) \frac{\lambda^{\sum x_i}}{\prod x_i!} \\
\propto \exp(-n\lambda) \lambda^{\sum x_i}
\]
Likelihood for i.i.d Normal observations

"Likelihood = probability of data for a given parameter value"

Parameter: $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$

$$L(x_1, \ldots, x_n; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right]$$

$$\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left[-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right]$$
General maximum likelihood principle

Maximum likelihood estimator

- We consider a dataset consisting of $n$ observations $(x_1, \ldots, x_n)$
- We assume that we know the likelihood function
  $$L(x_1, \ldots, x_n; \theta) = f_\theta(x_1, \ldots, x_n)$$

The maximum likelihood estimator of $\theta$ is defined as

$$\hat{\theta}_{ML}(x_1, \ldots, x_n) = \text{Argmax}_\theta L(x_1, \ldots, x_n; \theta)$$
Deriving $\hat{p}$ explicitly for the previous binomial sampling

We want to maximize $L(p) = p^x (1 - p)^{n-x}$

We could work on $L(p)$ directly in this case but let us denote $l(p) = \ln L(p)$.

$$l(p) = \ln[p^x (1 - p)^{n-x}] = x \ln p + (n - x) \ln(1 - p)$$

$$l'(p) = \frac{x}{p} - \frac{(n - x)}{(1 - p)}$$

(1)

$$l'(p) = 0 \text{ if } p = \frac{x}{n}$$

- $\hat{p} = \frac{x}{n}$ is the estimate of $p$
- for a generic sample with random outcome $X$, $\hat{p} = \frac{X}{n}$ is the estimator or $p$, it is a random variable
Omitting the term that does not depend on $\lambda$, we have

$$l(x_1, \ldots, x_n; \lambda) = \ln L(x_1, \ldots, x_n; \lambda) = \ln[\exp(-n\lambda)\lambda\sum_i x_i]$$

$$= -n\lambda + \sum_i x_i \ln \lambda$$

Hence

$$\frac{d}{d\lambda} l(x_1, \ldots, x_n; \lambda) = -n + \sum_i x_i / \lambda$$

And

$$\frac{d}{d\lambda} l(x_1, \ldots, x_n; \lambda) = 0 \iff \lambda = \sum_i x_i / n$$

The MLE of $\lambda$ is $\hat{\lambda}_{ML} = \sum_i x_i / n = \bar{x}$
Likelihood theory

Likelihood for i.i.d Normal observations

"Likelihood = probability of data for a given parameter value"

Parameter: \( \theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+ \)

\[
L(x_1, \ldots, x_n; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right]
\]

\[
\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left[-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right]
\]

\[
l(x_1, \ldots, x_n; \mu, \sigma) = \sum_{i=1}^{n} \left[-\ln \sigma - \frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right]
\]

\[
= -n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\]
\[
\frac{d}{d\mu} l(x_1, \ldots , x_n; \mu, \sigma) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)
\]

\[
\frac{d}{d\sigma} l(x_1, \ldots , x_n; \mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
\frac{d}{d\mu} l = 0 \quad \text{and} \quad \frac{d}{d\sigma} l = 0 \quad \text{give}
\]

\[
\sum_{i=1}^{n} (x_i - \mu) = 0
\]

\[
\text{and} \quad -n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0
\]

hence

\[
\hat{\theta}_{ML} = (\hat{\mu}, \hat{\sigma})_{ML} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)
\]
Remarks on the MLE

- The rule “likelihood = product of marginal densities” applies only when the observations are independent.
- Taking the log
  - linearizes the product into a sum
  - simplifies greatly the math expressions in the case density proportional to $\exp(ax^b)x^c$
  - avoids numerical instabilities when using numerical computation.
Remarks on the MLE (cont’)

- Deriving the MLE in closed form is often impossible in real-life problems. One has to resort to numerical optimization. Hence the importance of optimization methods in statistics.
- If the parameter $\theta$ belongs to a discrete set, differentiating $l(\theta)$ is meaningless. One has to resort to discrete optimization methods.
- The likelihood $L(x_1, \ldots, x_n; \theta)$ is sometimes denoted $L(\theta|x_1, \ldots, x_n)$.

This is misleading and mathematically completely wrong since in the likelihood theory, $\theta$ is not a random variable.
To go beyond these slides, you can read the first two chapters of *In All Likelihood*, Yudi Pawitan, Oxford Science Publications, 2001. This book is not in DTU digital library but almost completely on [Google books](https://books.google.com)
1. We assume that we have recorded the life duration of $n$ light bulbs denoted $x_1, \ldots, x_n$. We assume that they are $n$ iid replicates of an exponential $E(\alpha)$ distribution. Derive analytically the expression of the MLE of $\alpha$.

2. Derive analytically the MLE of $a$ for a dataset consisting of $n$ iid replicates of a $U[0, a]$ distribution. Evaluate the bias of this estimator. What is the limit of the bias when $n$ tends to $+\infty$?

3. For a distribution $f_\theta$, the expectation of $X$ under $f_\theta$ can be expressed as a function $\phi(\theta)$. The moment method consists in identifying $\phi(\theta)$ to the empirical mean. Apply this principle to the case above and discuss the estimator in terms of bias, variance, other remarks?