Statistical estimation

Statistical modelling: theory and practice

Gilles Guillot

gigu@dtu.dk

September 3, 2013

1. Introductory example
2. Principles of estimation
3. Likelihood theory
4. Reading
5. Exercises

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 $y_1, ..., y_n$ of temperatures at dates $t_1, .., t_n$ 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 a 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 te 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 $U\{1, ..., N\}$

Estimator, estimate

Estimator

Denoting generically θ the unknown parameter value, an estimator is a rule (or algorithm) allowing us to “guess” θ, from the data.
From a mathematical point of view, it is a function

$$\begin{array}{rcl} \mathbb{R}^n & \longrightarrow & \mathbb{R}^d \\ (x_1, ..., x_n) & \longrightarrow & \hat{\theta} \end{array}$$

- d is the dimension of the parameter space (often $d = 1$ for us)
- Since we assume that data are random, we will often stress this by denoting them $(X_1, ..., X_n)$.
- The number $\hat{\theta}$ is an estimate of θ. It is a random variable, denoted sometimes $\hat{\theta}(X_1, ..., X_n)$

Bias of an estimator

Definition: bias

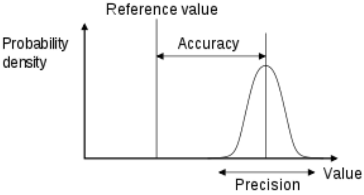
The bias of an estimator is the average discrepancy between the estimate and the true parameter value:

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

An estimator is said to be 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 scaterred around its mean (which may differ from the true value θ).

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 scaterred around the true value θ.

When $\hat{\theta}$ is unbiased, $E[\hat{\theta}] = \theta$ hence $V[\hat{\theta}] = MSE[\hat{\theta}]$.

Confidence interval

Definition: confidence interval

A confidence interval at level $(1 - \alpha) \in [0, 1]$ is an interval that contains the true unknown parameter value θ with probbaility $1 - \alpha$.

Example: estimation of a proportion

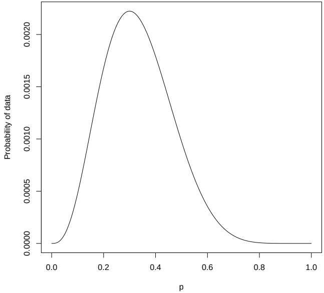
We have a sample of n objects of which x are faulty. We estimate the unknown proportion p by $\hat{p} = x/n$.
Exercise: give the bias and variance of $\hat{p}$.

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 $L(p)$.



The maximum likelihood principle

- The above suggests a method to estimate the unknown parameter p:

$$\hat{p} = Argmax_p \, p^x (1-p)^{n-x}$$

- $\hat{p}$ 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 $Argmax_p \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$ 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 θ.

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 λ, i.e. $P(X_i = x) = \exp(-\lambda)\lambda^x/x!$
- 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 "

$$
\begin{aligned}
L(x_1, ..., x_n; \lambda) &= \prod_{i=1}^{n} \exp(-\lambda)\lambda^{x_i}/x_i! \\
&= \exp(-n\lambda)\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!} \\
&\propto \exp(-n\lambda)\lambda^{\sum_i x_i}
\end{aligned}
$$

Likelihood for i.i.d Normal observations

"Likelihood = probability of data for a given parameter value "
Parameter: $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$

$$\begin{aligned} L(x_1, ..., x_n; \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}] \\ &\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}] \end{aligned}$$

General maximum likelihood principle

Maximum likelihood estimator

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

The maximum likelihood estimator of θ is defined as

$$\hat{\theta}_{ML}(x_1, ..., x_n) = Argmax_\theta L(x_1, ..., x_n; \theta)$$

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 $l(p) = \ln L(p)$.

$$\begin{aligned} l(p) &= \ln[p^x(1-p)^{n-x}] = x\ln p + (n-x)\ln(1-p) \\ l'(p) &= x/p - (n-x)/(1-p) \end{aligned} \quad (1)$$

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

- $\hat{p} = x/n$ is the estimate of p
- for a generic sample with random outcome X, $\hat{p} = X/n$ is the estimator or p, it is a random variable

Maximum likelihood estimator for i.i.d Poisson observations

Omitting the term that does not depend on λ, we have

$$\begin{aligned} l(x_1, ..., x_n; \lambda) &= \ln L(x_1, ..., x_n; \lambda) = \ln[\exp(-n\lambda)\lambda^{\sum_i x_i}] \\ &= -n\lambda + \sum_i x_i \ln \lambda \end{aligned}$$

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

And $\dfrac{d}{d\lambda} l(x_1, ..., x_n; \lambda) = 0 \Longleftrightarrow \lambda = \sum_i x_i / n$

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

Likelihood for i.i.d Normal observations

"Likelihood = probability of data for a given parameter value "
Parameter: $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$

$$\begin{aligned} L(x_1, ..., x_n; \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}] \\ &\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}] \end{aligned}$$

$$\begin{aligned} l(x_1, ..., x_n; \mu, \sigma) &= \sum_{i=1}^{n} \left[-\ln\sigma - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} \right] \\ &= -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

$$\begin{aligned}
\frac{d}{d\mu} l(x_1, ..., x_n; \mu, \sigma) &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\
\frac{d}{d\sigma} l(x_1, ..., x_n; \mu, \sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}$$

$\frac{d}{d\mu} l = 0$ and $\frac{d}{d\sigma} l = 0$ give

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \mu) &= 0 \\
\text{and } -n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 &= 0
\end{aligned}$$

hence

$$\hat{\theta}_{ML} = \widehat{(\mu, \sigma)}_{ML} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)$$

Remarks on the MLE

- The rule “likelihod = 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 θ belongs to a discrete set, differentiating $l(\theta)$ is meaningless. One has to resort to discrete optimization methods.
- The likelihood $L(x_1, ..., x_n; \theta)$ is sometimes denoted $L(\theta | x_1, ..., x_n)$.

  This is misleading and mathematically completely wrong since in the likelihood theory, θ is not a random variable.

Reading

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]

Exercises

1. We assume that we have recorded the life duration of n light bulbs denoted $x_1, ..., x_n$. We assume that they are n iid replicates of an exponential $\mathcal{E}(\alpha)$ distribution.
   Derive analytically the expression of the MLE of α.
2. Derive analytically the MLE of a for a dataset consisting of n iid replicates of a $\mathcal{U}[0, a]$ distribution. Evaluate the bias of this estimator. What is the limit of the bias when n tends to $+\infty$?
3. For a distribution f_θ, the expectation of X under f_θ 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?