A Crash-course on statistics

Likelihood based inference

Master in Modelling for Science and Engineering

Introduction

According to Efron[1] the story of statistics as a scientific discipline can be divided into three eras:

1. The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?
2. The classical period of Pearson, Fisher, Neyman ... who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple: *Is treatment A better than treatment B?* but the new methods were suited to the kinds of small data sets individual scientists might collect.
3. The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce massive data sets . But now the food of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind.

---

[1] *Large-scale inference: empirical Bayes methods for estimation, testing and Prediction* (2010)

The *Second Era* problems

The following data give the number of flying bomb hits recorded in each of 576 small areas of $\frac{1}{4}km^2$ in the south of London during World War II.

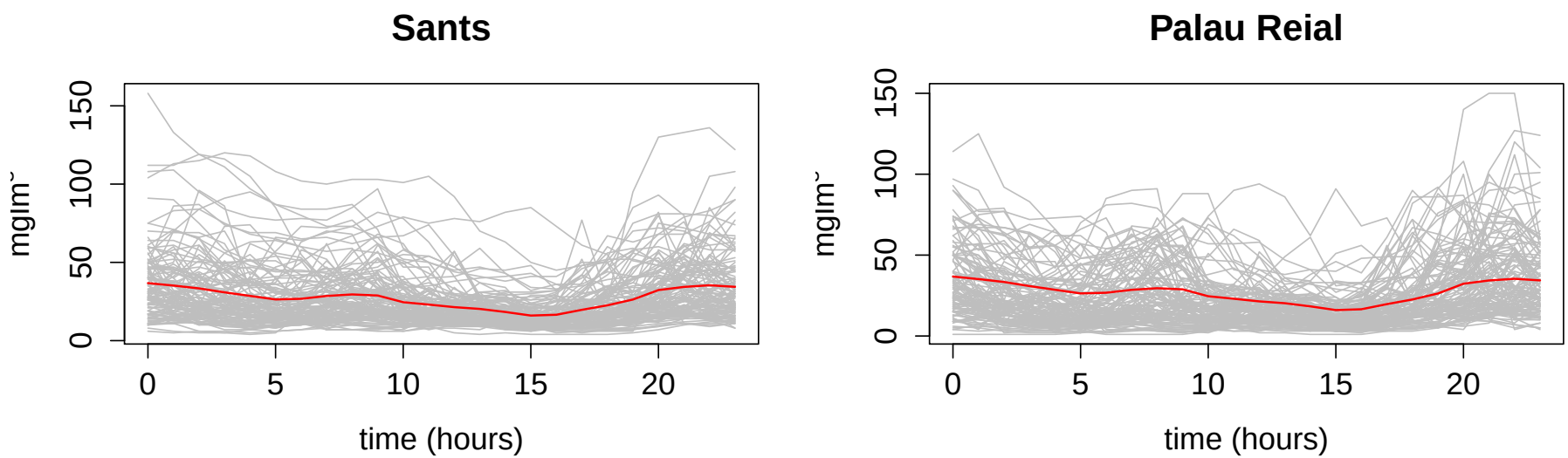
Example of some of the data used in the Bomb Sight project
1940s bomb census images reproduced by permission of The National Archives, London, England

Flying bomb hits on London

Number of impacts	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be randomly distributed over the area and should therefore be fitted by a Poisson distribution. Is this the case?

Levels of pollutants in non-working days in Sants and Palau Reial in 2014 in gray, ant the respective pointwise mean functions in red[2].



Were these neighbourhoods equally polluted?

[2] A. Cabaña, A. Estrada, J. Peña and A. Quiroz (2016) *Permutation tests in the two-sample problem for functional data*

And the *Third Era* knocks the door

The prostate cancer data[3]

Consider the genetic expression levels for $N = 6033$ genes obtained for $n = 102$ men, $n_1 = 50$ healthy control subjects and $n_2 = 52$ prostate cancer patients.

The problem we are interested in is to find out genes that are expressed in a different way in both groups, because this would be an indication of a relationship between these genes and prostate cancer.

We would like to know wether the 2 populations (healthy and ill patients) have, for each of the 6033 genes, the same mean.

The data are available in `R` package `sda` with the instruction `data(singh2002)`.

---

[3] D. Singh et al. 2002. *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell 1:203–209

In order to answer the previous questions, we are in the following situation: given the sample $X_1, \ldots, X_n \sim F$,

> *Can we admit that F satisfies a certain property "$\mathcal{P}$", or the sample is so unlikely under the premise that "$\mathcal{P}$" holds, that we would rather reject this hypothesis?*

If we are in a parametric context, we asume that $F \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ a certain family depending on a finite number of parameters, the previous question reduces to

> *Can we admit the null hipothesis $\mathcal{H}_0 : \theta \in \Theta_0$, where $\Theta_0 \subset \Theta$ or the sample leads us to reject it?*

We call $\mathcal{H}_1 : \theta \in \Theta \setminus \Theta_0$ alternative hipothesis.

In order to answer the previous questions, we are in the following situation: given the sample $X_1, \ldots, X_n \sim F$,

> *Can we admit that F satisfies a certain property "$\mathcal{P}$", or the sample is so unlikely under the premise that "$\mathcal{P}$" holds, that we would rather reject this hypothesis?*

If we are in a parametric context, we asume that $F \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ a certain family depending on a finite number of parameters, the previous question reduces to

> *Can we admit the null hipothesis $\mathcal{H}_0 : \theta \in \Theta_0$, where $\Theta_0 \subset \Theta$ or the sample leads us to reject it?*

We call $\mathcal{H}_1 : \theta \in \Theta \setminus \Theta_0$ alternative hipothesis.
If Θ_0 consists of only one point , $\mathcal{H}_0 = \theta_0$ is simple.

So, we have a basic premise that we are willing to accept as true $\mathcal{H}_0 : \theta \in \Theta_0$ (null hypothesis) unless the data $X_1, \ldots, X_n$ tell us to reject $\mathcal{H}_0$ in favour of an alternative hypothesis $\mathcal{H}_1$.

The testing procedure asumes that to each result of the experiment we have to assign one of two possibilities

reject or **not reject**

the null hypothesis.

This originates a partition of the sample space.

The sample points to which we assign the decision of rejecting the null are the *critical region*.

We do not make the partition directly from the sample, but we make a reduction through a test statistic.

This statistic will give us a measure of the discrepancy between the sample data and $\mathcal{H}_0$.

We must know (or be able to approximate) the distribution of the test statistic, in order to judge whether the discrepancy is big or not in probabilistic terms.

Level and Power.

We call *level* of the test, α, to the supremum of the probability of rejecting the null when it is true:

$$\alpha = \sup \mathbf{P}(\text{reject} \mathcal{H}_0 | \mathcal{H}_0 \text{ holds})$$

and *power* $\Pi(\theta)$ corresponding to a given value of θ in $\Theta_1 = \Theta \setminus \Theta_0$ to the probability of rejecting the null when the true probability is F_θ:

$$\Pi(\theta) = \mathbf{P}(\text{reject} \mathcal{H}_0 | \theta)$$

The mapping $\theta \mapsto \Pi(\theta)$ defined in Θ_1 is usually calledr *power function* of the test.

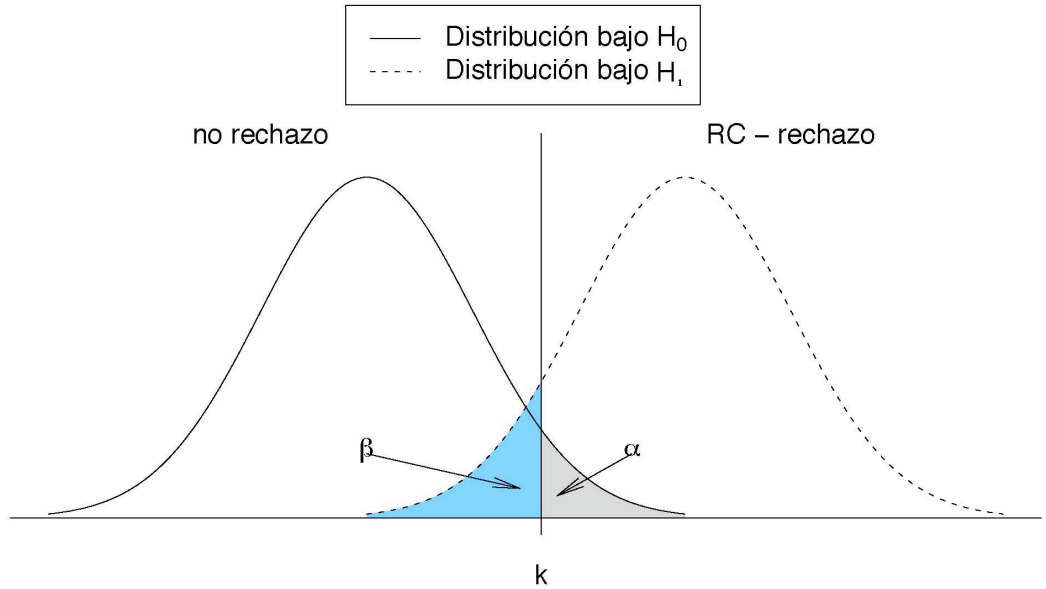
This function provides us with a measurement of the ability of the test in detecting alternatives.

The possibility of achieving high power without rising the level passes through the choice of an adequate sample size.

Decisions basedon the critical region S
taken from the sample $\mathbf{X} = (X_1, \ldots, X_n)$.

Nature ↓	Reject $\mathcal{H}_0$: $\mathbf{X} \in S$	Do not reject $\mathcal{H}_0$: $\mathbf{X} \notin S$
$\mathcal{H}_0$ true	(Type I error) Level: $\alpha = \sup_{\theta \in \Theta_0} \mathbf{P}_\theta\{\mathbf{X} \in S\}$	
$\mathcal{H}_0$ false	Power: $\Pi(\theta) = \mathbf{P}_\theta\{\mathbf{X} \in S\}$	(Type II error)

Probability of Type II error is usually denoted by β.



Probability of errors I and II in the particular case of mean comparisons .

The 5% significance level (or p-value ≤ 0.05)

The question arises in each example considered: *what is the critical level for the p-value? Is there some generally accepted level at which null hypotheses are automatically rejected?*

A significance level of $p < 0.05$ is often taken to be of interest, because it is below the magic level of 0.05.

For example suppose that we had tested a new drug versus standard drug, which under the null hypothesis of no difference between the two drugs, gave $p = 0.04$.
This says that the apparent difference between the two drugs being due to chance is less than 1 in 20. The p-value of 0.05 is the watershed used by the American control board (the FDA) which licences new drugs from pharmaceutical companies.

As a result it has been almost universally accepted right across the board in all walks of life.

However this level can be, to say the least, inappropriate and possibly even catastrophic. Suppose, for example, we were considering test data for safety critical software for a nuclear power station, N representing the number of faults detected in the first 10 years. Would we be happy with a p-value on trials which suggests that

$$P(N \geq 1) = 0.05?$$

We might be more comfortable if $p = 0.0001$, but even then, given the number of power stations (over 1000 in Europe alone) we would be justified in worrying.

The significance level which should be used in deciding whether or not to reject a null hypothesis ought to depend entirely on the question being asked; it quite properly should depend upon the consequences of being wrong.

At the very least we should qualify our rejection with something like the following.

$0.05 < p \leq 0.06$	"Weak evidence for rejection"
$0.03 < p \leq 0.05$	"Reasonable evidence for rejection"
$0.01 < p \leq 0.03$	"Good evidence for rejection"
$0.005 < p \leq 0.01$	"Strong evidence for rejection"
$0.001 < p \leq 0.005$	"Very strong evidence for rejection"
$0.0005 < p \leq 0.001$	"Extremely strong evidence for rejection"
$p \leq 0.0005$	"Overwhelming evidence for rejection"

The likelihood function

The basic idea starts with the joint distribution of $X = X_1, X_2, \ldots, X_n$ depending upon a parameter θ,

$$f(\mathbf{x}; \theta) = f(x_1, x_2, \ldots, x_n; \theta).$$

For fixed θ, probability statements can be made about X. If we have observations, but θ is unknown, we regard information about θ as being contained in the likelihood

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta),$$

where L is regarded as a function of θ with $\mathbf{x}$ fixed.

We will also be interested in considering some related functions (random variables):

- Log-Likelihood:

$$l(\theta; \mathbf{x}) = \log L((\theta; \mathbf{x}))$$

- Score:

$$U(\theta) = \frac{\partial l(\theta; \mathbf{x})}{\partial \theta}$$

- Observed Information:

$$J(\theta) = -\frac{\partial^2 l(\theta; \mathbf{x})}{\partial \theta^2}$$

Example: Suppose $X = X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with parameter $\theta \in [0, 1]$.

$$i.e. \quad P(X_i = 1) = \theta, \; P(X_i = 0) = 1 - \theta.$$

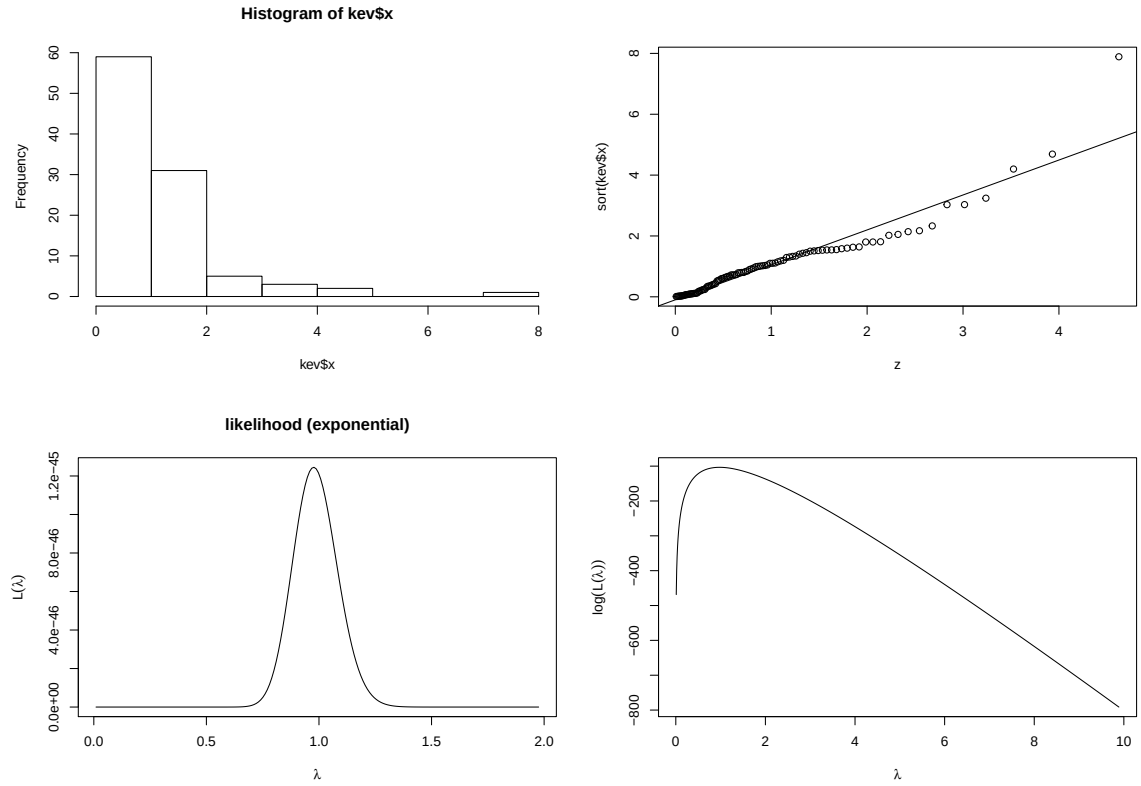
Observations are $x = (1, 0, 0, 1, 0, 1, 1)$ and

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{7} f(x_i; \theta) = \prod_{i=1}^{7} \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^4 (1 - \theta)^3 .$$

In general, for a sample size n,

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} .$$

Example: The data in `http://mat.uab.cat/~acabana/data/kevlar.txt` correspond to the time-to-failure (in hours) of 101 independent blocks of Kevlar49/epoxy, under a stress of 90% [4]



[4] Barlow et al. (1984)

Example: Structured data. Consider the data on O-ring thermal distress in a series of launches of the Challenger collected before the accident that occurred in 1986.

`http://mat.uab.cat/~acabana/data/challenger.txt`

We have records of

```
y = number of O-rings with t d (0,1,..., 6)
x1= temperature (F)
x2= pressure (psi)
```

for 23 independent launches between April 21th, 1981 to January 21st, 1986. The temperature on January 28, 1986 was 31, and the pressure undertaken by the rocket 200 psi.

$$\mathbf{P}(Y = r) = \binom{m}{r} p^r (1-p)^{m-r} \quad m = 6$$

and one possible relation between the probability of failure p and temperature and pressure is $p = \dfrac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$.

If the data for the j-th flight consists on r_j O-rings with thermal distress at launch temperature $x_{1,j}$, (disregard pressure for a while) $j = 1, \ldots, n = 23$, $m = 6$, and we consider all the launches as independent events, then the likelihood of the model is

$$L(\beta_0, \beta_1) = \prod_{j=1}^{n} \binom{m}{r_j} \frac{\exp(\beta_0 \sum_{j=1}^{n} r_j + \beta_1 x_{1,j})}{\prod_{j=1}^{n} (1 + \exp(\beta_0 + \beta_1 x_{1,j})^m}$$

It might be convenient to make a change of parameters: in general, if the law of Y depends on θ, and there is a 1-1 transformation such that $\theta = \theta(\psi)$, the likelihood for ψ, $L^*(\psi) = L(\theta(\psi))$. In this case,

$$\psi = \frac{\exp(\beta_0 + x_1\beta_1)}{1 + \exp(\beta_0 + x_1\beta_1)} \quad \lambda = \beta_1 \Rightarrow \beta_0(\psi, \lambda) = \log\left(\frac{\psi}{1-\psi}\right) - x_1\lambda$$

Dependent data

The dependence structure of the data may be very complex, making it hard to write down $f(y;\theta)$ explicitly.

When the data come from a stochastic process indexed by an order set (such as time), we still can manage to work out the likelihood function:

$$L(\theta;\mathbf{y}) = f(y_T,\ldots,y_1;\theta) = \prod_{j=p+1}^{T} f(y_j|y_1,\ldots,y_{j-1};\theta)\times f(y_p,\ldots y_1;\theta)$$

hence the log-likelihood is

$$l(\theta;\mathbf{y}) = \sum_{j=p+1}^{T} \log f(y_j|y_1,\ldots,y_{j-1};\theta) + \log f(y_p,\ldots y_1;\theta)$$

Maximum likelihood estimation

The value of θ which maximises $L(\theta)$ (or equivalently $l(\theta) = \log(L(\theta))$ is called a (*maximum likelihood estimate* (MLE) of θ.

- Under regularity conditions, θ is a solution of the score equation $U(\theta) = 0$, and $J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2} > 0$, but there are important examples where the MLE exists but there is no score equation.

- Generally $J(\theta) > 0$, and $J(\widehat{\theta})$ measures the concentration of $l(\theta)$ around $\widehat{\theta}$. Close to $\widehat{\theta}$, we can write

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}).$$

- Likelihood concentrates around the maximum, and as n increases, values far from the maximum become less and less likely.

Before performing the experiment we don't have data, hence we can't compute $J(\theta)$. But we can compute its expected value

$$I(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right).$$

This quantity is known as *expected information* or *Fisher information*. If the observations come from an i.i.d. sample, the information is

$$I(\theta) = ni(\theta) \quad \text{where} \quad i(\theta) = E\left(-\frac{\partial^2}{\partial \theta^2} \log f\left(X_i; \theta\right)\right),$$

is the Fisher Information corresponding to a single data point.

Example

Suppose $X = X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$

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

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

and

$$l\left(\mu, \sigma^2\right) = \log L\left(\mu, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

In this case thes score U is a vector with components,

$$U_\mu = \frac{\partial l\left(\mu,\sigma^2\right)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right),$$

$$U_{\sigma^2} = \frac{\partial l\left(\mu,\sigma^2\right)}{\partial \sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

Equating these derivatives to zero results in

$$\widehat{\mu}=\overline{X}, \quad \widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\widehat{\mu}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2.$$

In order to see that these are the unique solutions, we need to see that l is concave and that the Hessian is positive-definite.

Observe that since $\theta = (\mu, \sigma^2)$ is a vector, $\mathbf{I}(\theta)$ and $\mathbf{J}(\theta)$ are symmetric $(p \times p)$ matrices.

$$\{\mathbf{J}(\theta)\}_{rs} = -\frac{\partial^2 l(\theta)}{\partial\theta_r \partial\theta_s} \quad \text{y} \quad \{\mathbf{I}(\theta)\}_{rs} = E\left(-\frac{\partial^2 l(\theta)}{\partial\theta_r \partial\theta_s}\right).$$

In our examples with the Gaussian sample

$$\mathbf{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_i (x_i - \mu) \\ \frac{1}{\sigma^4}\sum_i (x_i - \mu) & \frac{1}{\sigma^6}\sum_i (x_i - \mu)^2 - \frac{n}{2\sigma^4} \end{pmatrix}$$

$$\mathbf{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Asymptotic distribution of the MLE $\widehat{\theta}$

In many situations, under some regularity conditions

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}),$$

We can take advantage of this in order to obtain asymptotic confidence intervals for the parameters. A 95% confidence interval for θ is

$$\widehat{\theta} \pm 1.96 I\left(\widehat{\theta}\right)^{-1/2}.$$

If the parameter is multi-dimensional,

$$\widehat{\theta} \xrightarrow{D} N(\theta, \mathbf{I}(\theta)^{-1}).$$

The precise statement of this result is due to Harald Cramér[5]. The required regularity conditions are

(i) The true value of θ is interior to the parameter space.

(ii) Differentiation under the integral is valid, so that $E\left[U(\theta)\right] = 0$ and $V\left[U(\theta)\right] = ni(\theta)$. This allows a CLT to apply to $U(\theta)$.

(iii) Taylor expansions are valid for the derivatives of the log-likelihood, so that higher order terms may be neglected

(iv) A week law of large numbers applies to $J(\theta)$.

[5] see Chap. 18, Ferguson (1996) *A Course in Large Sample Theory* Chapman & Hall for details

MLE for dependent data

The full log-likelihood function is called the exact log-likelihood

$$l(\theta;\mathbf{y}) = \sum_{j=p+1}^{T} \log f(y_j|y_1,\ldots,y_{j-1};\theta) + \log f(y_p,\ldots y_1;\theta)$$

The first term is called the conditional log-likelihood, and the second term is called the marginal log- likelihood for the initial values.

In the maximum likelihood estimation of time series models, two types MLEs may be computed:

The exact MLEs $\widehat{\theta}$ and the conditional MLEs $\widehat{\theta}_{cond}$ obtained maximising just the conditional log-likelihood.

They are both consistent estimators, but in general, for finite time series, they do not coincide.

Neyman-Pearson Lemma (simple hypotheses)

With the notation

$$S_k = \{(x_i, \ldots, x_n) : L(x_1, \ldots, x_n; \theta_1) > kL(x_1, \ldots, x_n; \theta_0)\},$$

$$\bar{S}_k = \{(x_i, \ldots, x_n) : L(x_1, \ldots, x_n; \theta_1) \geq kL(x_1, \ldots, x_n; \theta_0)\}.$$

If R and S are two critical regions of equal level and S is such that $S_k \subset S \subset \bar{S}_k$ for some $k \in [0, +\infty]$ then the power of R is less than or equal to that of S.

Critical regions with this property are called *Neyman-Pearson critical regions*.

Proof The level α of R is $\int_R L(x;\theta_0)d\mu(x) = \int_{R\setminus S} L(x;\theta_0)d\mu(x)$
$+ \int_{R\cap S} L(x;\theta_0)d\mu(x)$ this value is equal to the level of S, that can be written as $\int_S L(x;\theta_0)d\mu(x) = \int_{S\setminus R} L(x;\theta_0)d\mu(x) + \int_{R\cap S} L(x;\theta_0)d\mu(x)$.
Since levels are the same $\int_{R\setminus S} L(x;\theta_0)d\mu(x) = \int_{S\setminus R} L(x;\theta_0)d\mu(x)$, and this allows writing

$$\Pi_R = \int_R L(x;\theta_1)d\mu(x) = \int_{R\setminus S} L(x;\theta_1)d\mu(x) + \int_{R\cap S} L(x;\theta_1)d\mu(x)$$

$$\leq k\int_{R\setminus S} L(x;\theta_0)d\mu(x) + \int_{R\cap S} L(x;\theta_1)d\mu(x)$$

$$= k\int_{S\setminus R} L(x;\theta_0)d\mu(x) + \int_{R\cap S} L(x;\theta_1)d\mu(x)$$

$$\leq \int_{S\setminus R} L(x;\theta_1)d\mu(x) + \int_{R\cap S} L(x;\theta_1)d\mu(x) = \int_S L(x;\theta_1)d\mu(x) = \Pi_S.$$

□

The likelihood ratio

We often want to test in situations where the adopted probability model involves several unknown parameters. Thus we may denote an element of the parameter space by

$$\theta = (\theta_1, \theta_2, \ldots \theta_k)$$

Some of these parameters may be *nuisance* parameters, (*e.g.* testing hypotheses on the unknown mean of a normal distribution with unknown variance, where the variance is regarded as a nuisance parameter).

We use the *likelihood ratio*, $\lambda(\mathbf{x})$, defined as

$$\lambda(\mathbf{x}) = \frac{\sup\{L(\theta;\mathbf{x}) : \theta \in \Theta_0\}}{\sup\{L(\theta;\mathbf{x}) : \theta \in \Theta\}}, \quad \mathbf{x} \in \mathbb{R}_X^n.$$

The informal argument for this is as follows.

For a realisation x, determine its best chance of occurrence under H_0 and also its best chance overall. The ratio of these two chances can never exceed unity, but, if small, would constitute evidence for rejection of the null hypothesis.

A *likelihood ratio test* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is a test with critical region of the form

$$C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\},$$

where k is a real number between 0 and 1.[6]

[6] In case both the null and the alternative are simple, this is the Neymann-Pearson critical region, hence, it is optimal.

Clearly the test will be at significance level α if k can be chosen to satisfy

$$\sup\{P(\lambda(\mathbf{X}) \leq k;\ \theta \in \Theta_0)\} = \alpha.$$

If H_0 is a simple hypothesis with $\Theta_0 = \{\theta_0\}$, we have the simpler form

$$P(\lambda(\mathbf{X}) \leq k;\ \theta_0) = \alpha.$$

To determine k, we must look at the c.d.f. of the random variable $\lambda(\mathbf{X})$, where the random sample $\mathbf{X}$ has joint p.d.f. $f_{\mathbf{X}}(\mathbf{x}; \theta_0)$.

This can be done in relatively few cases, including

- exponencial rates
- inference on the mean of Gaussian samples (yielding the t-test)
- comparing nested linear models (the F-test)

The likelihood ratio as a statistic

Since the function $-2\log\lambda(\mathbf{x})$ is a decreasing function, it follows that the critical region of the likelihood ratio test can also be expressed in the form

$$C_1 = \{\mathbf{x} : -2\log\lambda(x) \geq c\}\,.$$

Writing

$$\Lambda(\mathbf{x}) = -2\log\lambda(\mathbf{x}) = 2\left[l(\widehat{\theta} : \mathbf{x}) - l(\theta_0 : \mathbf{x})\right]$$

the critical region may be written as

$$C_1 = \{\mathbf{x} : \Lambda(\mathbf{x}) \geq c\}$$

and $\Lambda(\mathbf{X})$ is called the likelihood ratio statistic.

The asymptotic distribution of the LRS

We have been using the idea that values of θ close to $\widehat{\theta}$ are well supported by the data so, if θ_0 is a possible value of θ, then it turns out that, for large samples,

$$\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2_p$$

where $p = \dim(\theta)$.

Let us see why (for scalar θ). Write

$$l(\theta_0) = l(\widehat{\theta}) + (\widehat{\theta} - \theta_0)l'(\widehat{\theta}) + \frac{1}{2}(\widehat{\theta} - \theta_0)^2 l''(\widehat{\theta}) + \dots$$

and, remembering that $l'(\widehat{\theta}) = 0$, we have

$$\begin{aligned} \Lambda &\simeq (\widehat{\theta} - \theta_0)^2 \left[-l''(\widehat{\theta})\right] \\ &= (\widehat{\theta} - \theta_0)^2 J(\widehat{\theta}) \\ &= (\widehat{\theta} - \theta_0)^2 I(\theta_0) \frac{J(\widehat{\theta})}{I(\theta_0)}. \end{aligned}$$

But

$$(\widehat{\theta} - \theta_0) I(\theta_0)^{1/2} \xrightarrow{D} N(0,1) \quad \text{and} \quad \frac{J(\widehat{\theta})}{I(\theta_0)} \xrightarrow{P} 1$$

so

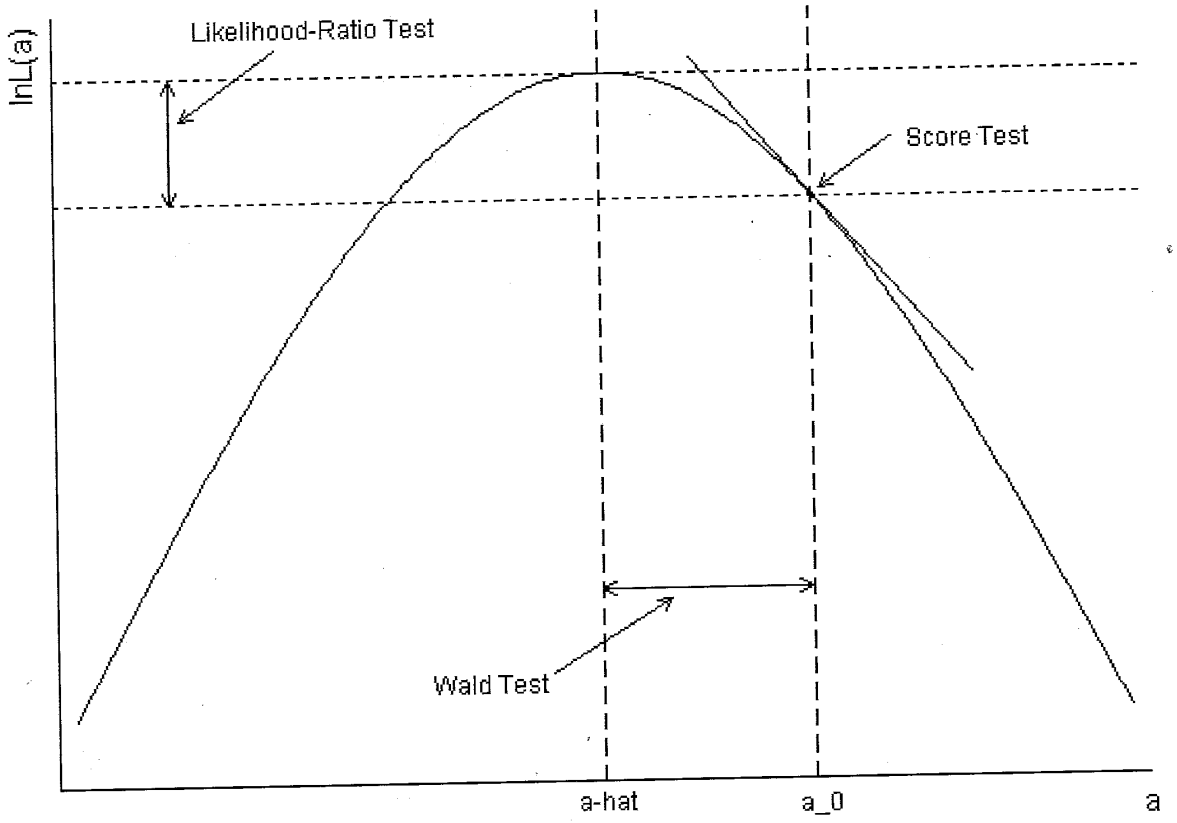
$$(\widehat{\theta} - \theta_0)^2 I(\theta_0) \xrightarrow{D} \chi^2_1$$

or

$$\Lambda \xrightarrow{D} \chi^2_1$$

provided θ_0 is the true value of θ.

Related asymptotic tests



Wald's test $W = (\hat{\theta} - \theta_0)^{tr} I(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi^2_p$

Rao's test (Score test) $S = U(\theta_0)^{tr} I(\theta_0) U(\theta_0) \sim \chi^2_p$. If $H_0 : \theta \in \Theta_0$, replace θ_0 by MLE restricted to Θ_0.

Examples of LRTs

It is not hard to derive LRTs for different situations, such as

- Goodness-of-fit tests for discrete data (Flying bomb hits in London)
- Contingency tables
- t-test (equivalent)
- F test for linear models
- the usual test for proportions is a score test
- $\vdots$

The comparison of the two samples of curves corresponds to another setting: the (nonparametric) permutation tests.

And the prostate cancer example is "just" performing the 6033 t-tests. How can we do that and pretend to obtain a decent p-value?