A glimpse of the Statistician's toolbox

Christophe Pouzat

MAP5, Paris-Descartes University and CNRS

christophe.pouzat@parisdescartes.fr

LASCON, January 22 2018, Lecture 2

Outline

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

Where are we ?

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

What are we going to talk about?

- Descriptive statistics that are robust: `median`, `median absolute deviation`, `five-number summary`.
- `Cumulative Distribution Functions` (CDF) and their observed or empirical versions (ECDF).
- The `Likelihood` function.
- The `Maximum Likelihood Estimator` (MLE) and its properties.

Where are we ?

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

What makes a statistic "robust"?

- Robust statistics are "well behaved" even when something goes wrong.
- We will illustrate what that means with robust versions of the classical *location* and *scale* parameters:
  - the median instead of the *mean*,
  - the median absolute deviation (MAD) instead of the *standard deviation*.

An example

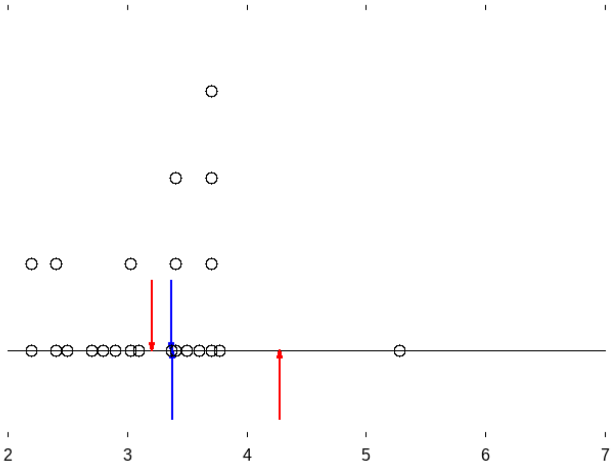
24 determinations of the copper content in the wholemeal floor (in parts per million) sorted in ascending order (example 1.1 of Maronna, Martin & Yohai, *Robust Statistics. Theory and Methods.* 2006 J. Wiley):

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90
3.03	3.03	3.10	3.37	3.40	3.40	3.40	3.50
3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

- The sample *mean* is 4.28 while the *median* is 3.38.

If we remove the outlier, 28.95, we get:

- A *mean* of 3.21 and a *median* of 3.37.



The data with the outlier out of scale. Bottom: mean (red) and median (blue) computed with the outlier. Top: mean (red) and median (blue) computed *without* the outlier.

What is the MAD?

- We have just seen that using a median instead of a mean "stabilizes" the results (understand: make the result look the same when some observations are removed).
- We want to adapt this idea to the statistics characterizing the *scale* or *spread* of the data for which the *standard deviation* (SD) is usually used.
- The SD is moreover obtained from the square root of a mean of squared differences (the difference between each individual observation and the sample mean). If one observation is a genuine outlier *it will dominate the estimate*.

- The *Median Absolute Deviation* addresses both issues.
- It is proportional to the median of the absolute deviations with respect to the median:

$$\texttt{MAD} = \frac{1}{0.67449}\,\texttt{median}\left(|X_i - \texttt{median}(X)|\right) ,$$

  where $X = \{X_1, \ldots, X_n\}$ is the sample.
- The division by 0.67449 makes the MAD equal to the SD (on average) when the sample is drawn from a Gaussian.
- For the copper data, the SD is 5.30 with the complete sample and becomes 0.69 when the outlier is removed.
- For the same sample, the MAD is 0.53 with the complete sample and becomes 0.50 when the outlier is removed.

- When you work with real data use the median instead of the mean and the MAD instead of the SD unless you are pretty sure that your sample contains no "pathological" observations.
- We will see that at work on neurophysiological data when we will discuss spike sorting.

The five-numbers summary

This is a set of statistics that turns out to be very useful to summarize large data set. It is:

- The *minimum* of the sample.
- The *first quartile*.
- The *median* (second *quartile*).
- The *third quartile*
- The *maximum* of the sample.

The *inter quartile range* (IQR), the difference between the third and first quartile is another robust estimator of the spread of the data.

When working with large datasets my recommendation is to compute systematically the five-numbers summary and the MAD. These statistics should appear in your lab-book.

The Empirical Cumulative Distribution Function (ECDF)

- The *Cumulative Distribution Function* (CDF) of a random variable X is by definition:

  $$F_X(x) \equiv \mathbb{P}(X \leq x)\,,$$

  where $\mathbb{P}(X \leq x)$ stands for "the probability of the event $X \leq x$".
- Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F_X$, the *Empirical Cumulative Distribution Function* (ECDF) of the sample $\{X_1, \ldots, X_n\}$ is (by definition):

  $$\widehat{\mathrm{F}} \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)\,,$$

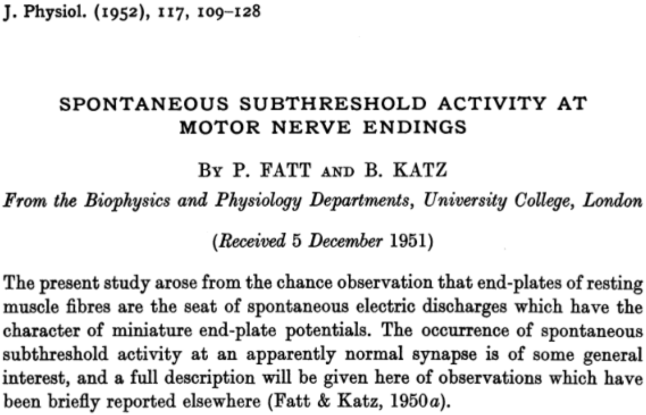
  where

  $$\mathbb{1}(X_i \leq x) = \begin{cases} 1 & \text{if} \quad X_i \leq x \\ 0 & \text{if} \quad X_i > x. \end{cases}$$

- The ECDF is a function that makes a "jump" of size 1/n at each observation X_i.
- If we write $X_{(i)}$ the *order statistics*, that is:
  - $X_{(1)} = \min\{X_1, \dots, X_n\}$
  - $X_{(2)} = \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}$
  - $X_{(k)} = \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(k-1)}\}$
  - $X_{(n)} = \max\{X_1, \dots, X_n\}$

  the ECDF graph is piecewise constant, continuous on the right side with a limit of the left side (the function's value at a jump site, $X_{(k)}$, is the staircase's height on the right side of $X_{(k)}$).

An example with historical data



J. Physiol. (1952), 117, 109–128

SPONTANEOUS SUBTHRESHOLD ACTIVITY AT MOTOR NERVE ENDINGS

BY P. FATT AND B. KATZ

From the Biophysics and Physiology Departments, University College, London

(*Received* 5 *December* 1951)

The present study arose from the chance observation that end-plates of resting muscle fibres are the seat of spontaneous electric discharges which have the character of miniature end-plate potentials. The occurrence of spontaneous subthreshold activity at an apparently normal synapse is of some general interest, and a full description will be given here of observations which have been briefly reported elsewhere (Fatt & Katz, 1950*a*).

In 1952, Fatt and Katz reported the first observation at the frog neuro-muscular junction of miniature end-plate potentials (mEPPs).

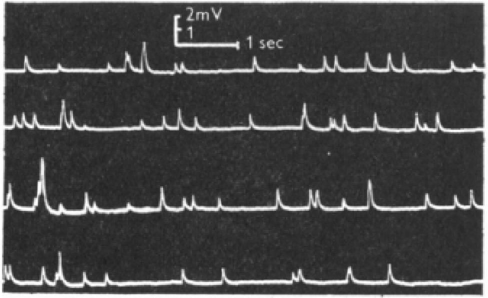
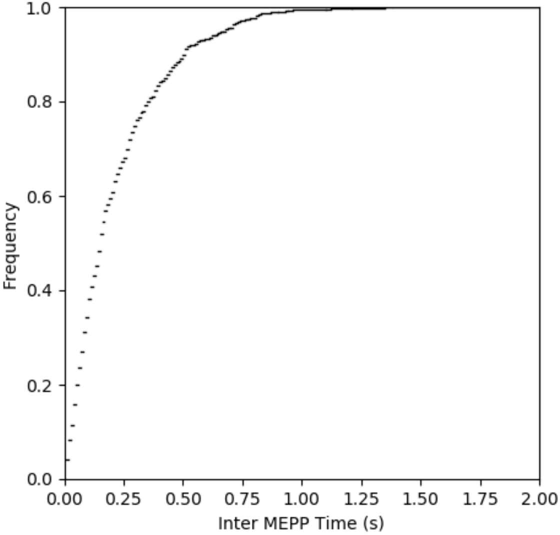


Fig. 2. Example of miniature e.p.p.'s in a muscle treated with 10^{-6} prostigmine bromide.

These observations quickly lead to the quantal release model of synaptic transmission. Katz got the Nobel Prize for that (Paul Fatt was forgotten. . .).

The data

- In their 1952 paper, Fatt and Katz study the time intervals between two successive mEPPs (the results are shown on their Figs. 11 & 12).
- The data can be found in the appendix of a book by David Cox and Peter Lewis (1966) *The Statistical Analysis of Series of Events* who thank Katz and Miledi for providing the data and who wrongly describe them as measurements between *nerve impulses* (this is at least what I concluded when I tried to figure out how intervals between nerve impulses–from a single axon–could follow so perfectly a Poisson distribution).
- The data reappear in two recent and excellent books by Larry Wasserman (*All of Statistics*, 2004 and *All of Nonparametric Statistics*, 2006).
- They can be downloaded from L. Wasserman website: `http://www.stat.cmu.edu/~larry/all-of-nonpar/`.



ECDF of the inter MEPP times from Fatt and Katz (1952).

Adding confidence bands

- It is possible and even straightforward to add a confidence band to the graph of $\widehat{\mathrm{F}}$.
- A confidence band is a domain that contains the complete graph of the true CDF with a probability set by us.
- The Kolmogorov distribution can be used, but a simpler to compute distribution–leading to almost as tight bands–results from the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality:

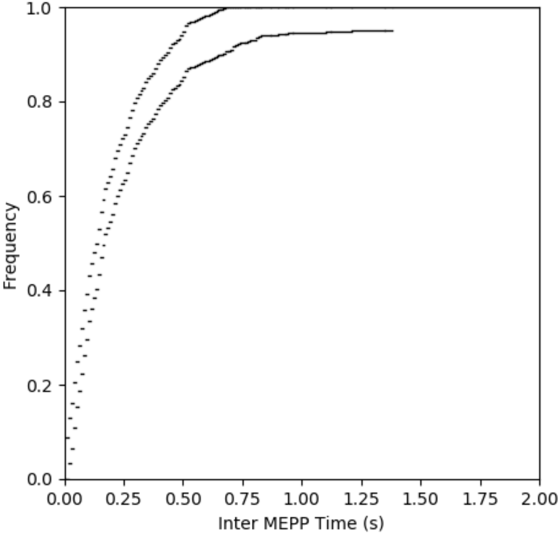
  $$\mathbb{P}\left(\sup_x \mid \mathrm{F}(x) - \widehat{\mathrm{F}}_n(x) \mid > \epsilon\right) \le 2\, e^{-2n\epsilon^2} \, ,$$

  where n is the sample size.
- So if we want ϵ such that:

  $$\mathbb{P}\left(\sup_x \mid \mathrm{F}(x) - \widehat{\mathrm{F}}_n(x) \mid > \epsilon\right) \le 1 - \alpha$$

  we find:

  $$\epsilon(\alpha) = \sqrt{\frac{1}{2n}}\sqrt{\log\left(\frac{2}{1-\alpha}\right)} \, .$$



ECDF with 95% confidence band of the inter MEPP times from Fatt and Katz (1952).

Why use the ECDF?

- The only "data manipulation" involved is sorting (no bin width setting).
- It is easy to get confidence bands making the ECDF a quantitative tool.
- Histograms lead too easily to baseless conclusions.

Where are we ?

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

The setting

- We consider a situation where a sample (or a set of observations) $x = (x_i)_{i=1,\dots,n}$ is available.
- These observations are modeled as a draw from a probability distribution $\mathcal{M}$, we say that the sample x is the realization of the random variable X whose distribution is $\mathcal{M}$ ($X \sim \mathcal{M}$).
- Our model $\mathcal{M}$ is in fact partly unknown, otherwise we would not need the experiment that gave us x.
- So we really have in mind a collection of models that we write $\mathcal{M}(\theta)$, where $\theta \in \mathbb{R}^p$ and $p < \infty$.

As an example, we could assume:

- The data were generated by measurements along a decaying mono-exponential that we will call "our signal", s:

  $$s(t;b,\Delta,\tau) \equiv b + \Delta \; \exp - t/\tau \, ,$$

  where b is the baseline, Δ, the jump at zero and τ the decay time constant.
- These three quantities constitute our model parameter: $\theta \equiv (b,\Delta,\tau)$.
- The measurements were done at some specific (positive) times $(t_i)_{i=1,\ldots,n}$.
- The measurements were corrupted by an independent Gaussian noise with a know variance σ^2 and a null mean, leading to the following expression for the *probability density*:

  $$p(X_i = x; t_i, \theta) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(-\frac{(x - s(t_i;\theta))^2}{2\sigma^2}\right) \, .$$

- Since we assume that the measurement noise is independent of the signal value and of the time, the probability (density) of our sample can be written:

$$p\left((x_i)_{i=1,\ldots,n}; (t_i)_{i=1,\ldots,n}, \theta\right) = \prod_{i=1}^{n} p(x_i; t_i, \theta)\,.$$

- Our collection of models, $\mathcal{M}(\theta)$, is then made of all the functions $\mathbb{R}^n \mapsto \mathbb{R}$ of the form:

$$\prod_{i=1}^{n} p(X_i; t_i, \theta) \quad \text{with} \quad (t_i)_{i=1,\ldots,n} \geq 0 \quad \text{fixed}\,.$$

Our problem

- We want to find the member of our collection that "explains best" the data.
- Stated differently, we want to find $\hat{\theta}$ such that $\mathcal{M}(\hat{\theta})$ "explains best" the data.
- If the data are put to use that means that $\hat{\theta}$ will depend on them, $\hat{\theta}$ must be a function of (x_i).
- We are going to be optimistic and assume that the data were actually generated by one member of our collection and we will write θ_0 the index of this member.
- $\hat{\theta}(x_1, \ldots, x_n)$ is then called an estimator of θ_0.
- Finding the "best explanation" amounts then to finding $\hat{\theta}(x_1, \ldots, x_n)$ as close as possible to θ_0.

Consequences

- Since all members of our collection are probability densities, we are explicitly considering that if we repeat our experiment in the exact same conditions we will get $(y_i)_{1,\dots,n} \neq (x_i)_{1,\dots,n}$.
- We therefore expect (in general) that $\hat{\theta}(y_1,\dots,y_n) \neq \hat{\theta}(x_1,\dots,x_n)$.
- Stated differently, since $\hat{\theta}$ depends on the observations, it is a random variable.
- The notion of "finding $\hat{\theta}(x_1,\dots,x_n)$ as close as possible to θ_0" must then be made more precise.
- Doing as if we could perform as many experiments as we wish, we will look for estimators whose mean value satisfy $\lim_{n\to\infty} \mathrm{E}\hat{\theta}(x_i) = \theta_0$. Such estimators are called asymptotically unbiased or consistent.

- We would also like the distribution of $\hat{\theta}$ to be as concentrated as possible around θ_0, that is, to have a small variance.
- Comparing the results of two independent experiments means comparing $\hat{\theta}(y_1, \ldots, y_n)$ and $\hat{\theta}(x_1, \ldots, x_n)$ and we would like a *yardstick* allowing us to judge how different are two estimated values.
- Stated more formally we want a procedure that providence confidence intervals on the estimated parameters.
- Implementing reproducible research is just impossible without confidence intervals (as soon as the observations are variable at least).

Where are we ?

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

The likelihood function

- In the previous section we defined a "model" as a probability (density) function depending on some parameters $\mathcal{M}(\hat{\theta})$ like

$$\prod_{i=1}^{n} p(X_i; t_i, \theta) \quad \text{with} \quad (t_i)_{i=1,\ldots,n} \geq 0 \quad \text{fixed} \,.$$

- Once data have been observed and "plugged-in" the probability density function:

$$\prod_{i=1}^{n} p(x_i; t_i, \theta) \quad \text{with} \quad (t_i, x_i)_{i=1,\ldots,n} \geq 0 \quad \text{fixed} \,,$$

  we can view this object as a function of θ.
- The likelihood function is "just" that: the probability density applied to "fixed" data, viewed as a function of the model parameters, $L(\theta)$.

Variations

- We will not focus on the precise values taken by the likelihood function but on each shape.
- We will therefore (in general) drop the factors that do not depend on the model parameters.
- Since the likelihood function originates from a probability (density), it is positive and we can take its logarithm. Knowing the log-likelihood is like knowing the likelihood.
- For theoretical reasons that will be explained later and for numerical reason (the likelihood can be very very small) we will work with the log-likelihood, $l(\theta) = \log L(\theta)$, most of the time.

An empirical study

Let us almost go back to the previous example, the mono-exponential decay, with the following modification:

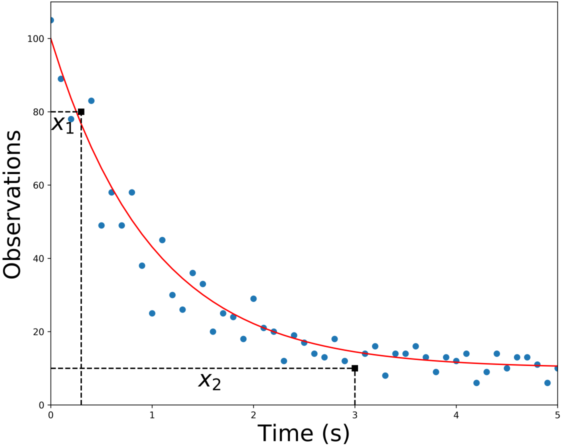
- The observations are made a regularly separated time points with a time δ between two successive observations: $(t_i)_{i=1,\ldots,n} = (\delta i)_{i=1,\ldots,n}$.
- The measurement noise depends on the signal $s(t;\theta)$,
- The observation at time t_i is the realization of a Poisson random variable with parameter $s(t_i;\theta)$:

$$\mathbb{P}\{X_i = n\} = \frac{(s_i)^n}{n!} \exp(-s_i) \, , \quad \text{for} \quad n = 0, 1, 2, \ldots$$

and

$$s_i = s(\delta i; \theta) = b + \Delta \exp(-\delta i / \tau) \, .$$

This simple scheme contains all the key ingredients to work with fluorescence measurements.



Example of simulated data with $b = 10$, $\Delta = 90$, $\tau = 1$.

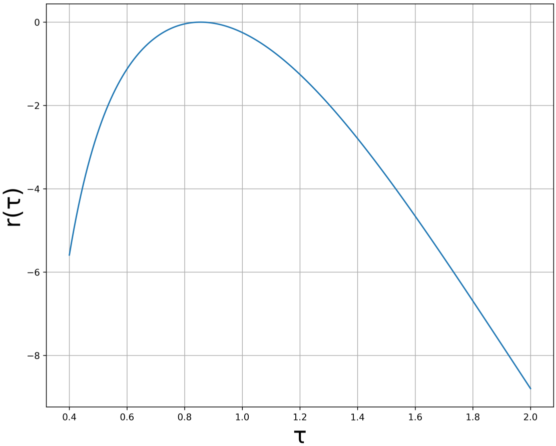
- In order to have simpler displays, we will start with a setting where both parameters b and Δ are known.
- The log-likelihood is then a function of a single variable τ.
- We will also do as if only two times add been used, t_1 and t_2 leading to observations x_1 and x_2 on the previous figure.
- The log-likelihood is then:

$$l(\tau) = x_1 \log s(t_1, \tau) - s(t_1, \tau) + x_2 \log s(t_2, \tau) - s(t_2, \tau)\,.$$

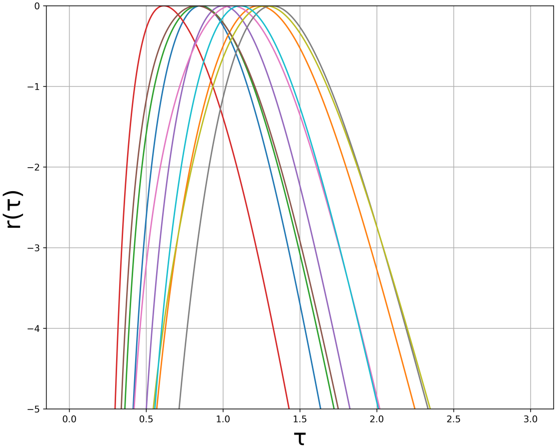
- To make comparison with subsequent simulations in the same setting we will show the graph of:

$$r(\tau) = l(\hat{\tau}) - l(\tau)\,,$$

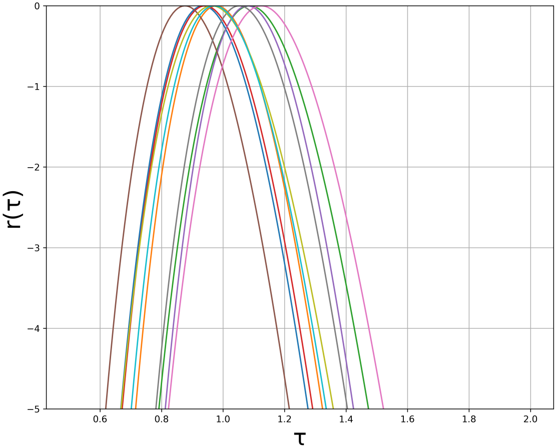
where $\hat{\tau}$ is the location of the maximum of $l(\tau)$.



On the next slide we repeat what we just did with 9 additional samples.



On the next slide we repeat what we just did but we now use 4 times more observations (that is 8) per sample.



What do you see?

What you should have seen

As the sample size increase:

- The log-likelihood becomes more symmetric.
- The curvature increases.
- The location of the peak gets closer to the true value.

Where are we ?

Introduction

Descriptive statistics

Models and estimators

The likelihood function

The Maximum Likelihood Estimator

The MLE

- In 1922 Fischer proposed to take as an estimator $\hat{\theta}$ the θ that maximizes $l(\theta)$.
- In that he was essentially following and generalizing Daniel Bernoulli, Lambert and Gauss.
- But he went much farther claiming that when the maximum was a smooth maximum, obtained by taking the derivative / gradient with respect to θ and setting it equal to 0, then:
    - The accuracy (standard error of the estimate) can be found to a good approximation from the curvature of $l(\theta)$ at its maximum.
    - $\hat{\theta}$ expresses all the relevant information available in the data.
    - This estimate is the best of all the consistent ones.

Some remarks

- The MLE is just the value of the parameter that makes the observations most probable *a posteriori*.
- Some technical precautions are required in order to fulfill all of Fischer's promises; they are referred to as "the appropriate smoothness conditions" in the literature.
- They are heavy to state and a real pain to check (that's why no one checks them)!
- My recommendation is to go ahead and after the MLE, $\hat{\theta}$, has been found, do a parametric bootstrap:
  - take $\hat{\theta}$ has the "true" value and simulate 500 to 1000 samples from $\mathcal{M}(\hat{\theta})$,
  - for each simulated sample, repeat the estimation procedure and get a sample of 500 to 1000 $\hat{\theta}$ values,
  - check that this sample has the properties expected from MLE theory.

- With this bootstrap procedure we also make sure that the "asymptotic" regime is reached (the theorems are valid when the sample size goes to infinity).

Functions associated with the Likelihood

- The *score function* is defined by: $S(\theta) \equiv \frac{\partial l(\theta)}{\partial \theta}$.
- The *observed information* is defined by: $\mathcal{J}(\theta) \equiv -\nabla \nabla^T l(\theta)$.
- the *Fischer information* is defined by: $\mathcal{I}(\theta) \equiv \mathtt{E}\mathcal{J}(\theta)/n$.

Asymptotic properties of the MLE

Under the "appropriate smoothness conditions" (see the Wikipedia page for a full statement), we have:

- $\hat{\theta}$ converges in probability to θ_0.
- $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{\text{dist.}}{\to} \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta_0)\right)$.
- $\left(\hat{\theta}-\theta_0\right) \sim \mathcal{N}\left(0, \mathcal{J}^{-1}(\hat{\theta})\right)$.
- $2\left(l(\hat{\theta})-l(\theta_0)\right) \overset{\text{dist.}}{\to} \chi^2_p$.

The last two statements can be used to get confidence intervals.

Illustration: confidence intervals

We therefore have:

- $2\left(l(\hat{\theta}) - l(\theta_0)\right) \overset{\text{dist.}}{\to} \chi^2_p$.
- In our previous figures we showed the graph of $r(\theta) = l(\theta) - l(\hat{\theta})$.
- If we want a 95 % confidence interval we should therefore select the segment of θ values for which $r(\theta) \geq -1.92$ since the 0.95 quantile of a χ^2_1 distribution is 3.84.

Illustration: comparison of two estimators

We reconsider our first simulation, 2 observations, b and Δ known and we compare two estimators for τ:

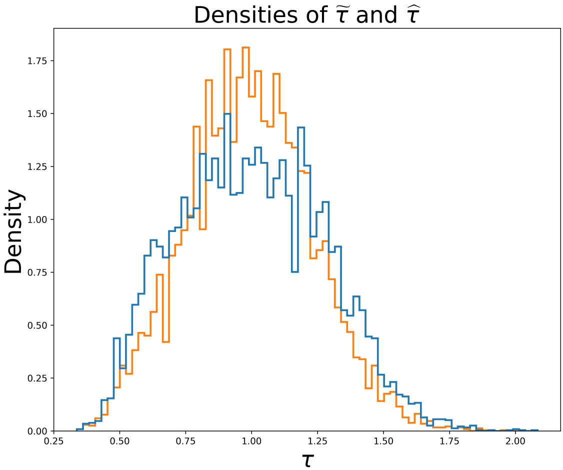
- The MLE, $\hat{\tau}$, that maximizes:

$$l(\tau) = x_1 \, \log s(t_1,\tau) - s(t_1,\tau) + x_2 \, \log s(t_2,\tau) - s(t_2,\tau) \, .$$

- The least squares estimator, $\tilde{\tau}$, that minimizes:

$$rss(\tau) = \left(x_1 - s(t_1,\tau)\right)^2 + \left(x_2 - s(t_2,\tau)\right)^2 \, .$$

We simulate 10000 samples and get both estimators for each.



$\tilde{\tau}$ is blue and $\hat{\tau}$ is orange.

Some remarks

- Maximum likelihood estimation is very general: as soon as we have an expression for the probability (density) of our data, we can use it.
- It assumes that the true model is in the family we consider, this is a very strong assumption.
- We must therefore always do some goodness of fit test, otherwise the confidence intervals we will get will be meaningless.
- Doing maximum likelihood means that we do optimization all the time. The users of this method should make a minimal effort to master the numerical optimization routines they are going to use.
- Be careful with the optimization routines of `Scipy`.