Statistics
in Particle Physics

Tanguy Marsault

August 12, 2024

Contents

1 Modeling **1**

1.1 Probability density functions 1

1.2 Likelihood 2

1.3 Model 2

2 Generalities on statistics **3**

2.1 Probabilities 3

2.2 Estimators 6

2.2.1 Generalities 6

2.2.2 Cramér-Rao bound 7

2.2.3 Asymptotic behaviour 7

2.3 Confidence intervals 8

2.4 Hypothesis testing 10

2.5 Bayesian statistics 11

3 Likelihood **11**

3.1 Maximum likelihood estimator (MVE) 11

3.2 Hypothesis testing 12

4 Uncertainty propagation **12**

1 Modeling

Modeling consists in finding relationship between observed experimental data and the parameters of a model. The model is a mathematical representation of the data.

1.1 Probability density functions

In the context of particle physics, one has a number of channels labelled by c, and a number of events n_c in each channel. Achannel is defined by some cuts on the data. The number of events in each channel is a random variable. In each channel, there are various observables x_c which are also random variables.

By looking at an observable x one can build up a model for its probability density function *pdf* (in the large sense, with respect to continuous or discrete measures). We denote this *pdf* by f and it verifies,

$$\int f(x)\mathrm{d}x = 1 \tag{1}$$

To do so, one just needs to collect data and build a histogram and fit it to some regular function (one could also keep a discretized probability distribution).

In general, the *pdf* is parametrized by some parameters $\boldsymbol{\alpha}$ that can be parameters of the physicla theory (*e.g.* the mass of a particle) or some unknown detector behaviours (*e.g.* the energy resolution). The *pdf* is then written as $f(x|\boldsymbol{\alpha})$.

Usually not all parameters are interesting and one refers to relevant parameters as *parameters of interest* (POI) and irrelevant parameters as *nuisance parameters*.

Let us suppose we work in a signel channel for now, and we have many events for an observable x that shall be denoted as $\mathcal{D} = \{x_i\}_{i\in[\![1,n_c]\!]}$. Then one is interested in the *pdf* of $\mathcal{D}$. We emphasized that n_c is a random variable as well. In general it follows a Poisson distribution $\mathcal{P}(n_c|\nu_c)$ where ν_c is the expected number of events in the channel. Then, the *pdf* for $\mathcal{D}$, parameterized by $\boldsymbol{\alpha}$ is

$$f(\mathcal{D}|\boldsymbol{\alpha}) = P(n_c|\nu_c(\boldsymbol{\alpha}))\prod_{i=1}^{n_c} f(x_i|\boldsymbol{\alpha}) \tag{2}$$

where we mae clear that in general, ν_c depends on $\boldsymbol{\alpha}$.

Now if one combines the channels, the *pdf* for the whole data set is

$$f(\mathcal{D}_{tot}|\boldsymbol{\alpha}) = \prod_c P(n_c|\nu_c(\boldsymbol{\alpha}))\prod_{i=1}^{n_c} f_c(x_i^c|\boldsymbol{\alpha}) \tag{3}$$

Note that, in general, the observables x^c are not the same in different channels and so the *pdf* are different.

1.2 Likelihood

The likelihood is a measure of how well the model fits the data. It is defined as

$$\mathcal{L}(\boldsymbol{\alpha}) = f(\mathcal{D}_{tot}|\boldsymbol{\alpha}) \tag{4}$$

Obviously it is not integrated to one anymore. The likelihood is a function of the parameters $\boldsymbol{\alpha}$.

$$\int \mathcal{L}(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\alpha} \neq 1$$

1.3 Model

In general we call the model the full representation of the data, that is $f(\cdot|\cdot)$ considered as a function of the observables and the parameters. To summarized everything,

- The model is the full representation of the data $f(\cdot|\cdot)$

- The likelihood is the model evaluated at the data $\mathcal{L}(\boldsymbol{\alpha}) = f(\mathcal{D}_{tot}|\boldsymbol{\alpha})$, that is a function of the parameters $\boldsymbol{\alpha}$
- The *pdf* of the data is the model evaluated at the data with the parameters given

2 Generalities on statistics

In this section, we recall some formal definition on statistics that shall be useful.

2.1 Probabilities

In this section we give a few results on probabilities that shall be useful. We denote by F_Y the *cumulative distribution function* of a random variable Y. $F_Y(y) = P(Y \leq y)$.

First let us introduce few convergences mode of random variables.

Définition 2.1. Different convergences
Let X_n be a sequence of random variables and X a random variable. Denote by

- We say that X_n converges to X in probability if for all $\epsilon > 0$,

  $$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0 \tag{5}$$

  We write $X_n \xrightarrow{P} X$.

- We say that X_n converges to X in distribution if for all x such that F_X is continuous at x, F_{X_n} converges to F_X pointwise, *i.e*

  $$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{6}$$

  We write $X_n \xrightarrow{D} X$.

- We say that X_n converges to X almost surely if

  $$P(\lim_{n\to\infty} X_n = X) = 1 \tag{7}$$

  We write $X_n \xrightarrow{a.s.} X$.

- We say that X_n converges to X in L^p if

  $$\lim_{n\to\infty} E(|X_n - X|^p) = 0 \tag{8}$$

  We write $X_n \xrightarrow{L^p} X$.

These convergences are related in the following fashion.

Theorem 2.2. Let X_n be a sequence of random variables and X a random variable. Then

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X \tag{9}$$

and

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{P} X \tag{10}$$

A very important result is the Law of Large Numbers,

Theorem 2.3. Law of Large Numbers
Let X_n be a sequence of i.i.d. random variables with mean μ, such that $\sigma(X_1) < \infty$. Then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s} \mu \tag{11}$$

and

$$\bar{X}_n \xrightarrow{P} \mu \tag{12}$$

and

$$\bar{X}_n \xrightarrow{L^2} \mu \tag{13}$$

This is very useful in many cases and an extremly strong result. It is also interesting to understand how different convergences can combine.

Remarque 2.4. We shall use the notation for the mean of a sequence of random variables as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{14}$$

in the rest of this document.

A similar result that can also be useful is called Wald's theorem.

Theorem 2.5. Wald's theorem
Let X_n be a sequence of i.i.d. random variables and N be an integer valued random variable independent of X_n. If $E(X_1) < \infty$ et $E(n) < \infty$, then

$$E\left(\sum_{i=1}^{N} X_i\right) = E(N)E(X_1) \tag{15}$$

Theorem 2.6. Mann-Wald's theorem
Let X_n be a sequence of random variables and X a random variable. Then for any measurable function g,

- If $X_n \xrightarrow{P} X$ then $g(X_n) \xrightarrow{P} g(X)$
- If $X_n \xrightarrow{D} X$ then $g(X_n) \xrightarrow{D} g(X)$
- If $X_n \xrightarrow{a.s.} X$ then $g(X_n) \xrightarrow{a.s.} g(X)$

Note that this theorem applies to random variables with values in $\mathbb{R}^d$, *i.e* random vectors.

This shall be very useful, especially when we combine it with the following major result.

Theorem 2.7. Slutsky's theorem
Let X_n and Y_n be two sequences of random variables such that

- $X_n \xrightarrow{D} X$
- $Y_n \xrightarrow{P} c$

where c is a constant. Then

$$(X_n, Y_n) \xrightarrow{D} (X, c) \tag{16}$$

Remarque 2.8. The two previous theorems yield the following result that can be very interesting in practice.Let X_n and Y_n be two sequences of random variables such that

- $X_n \xrightarrow{D} X$
- $Y_n \xrightarrow{P} c \in \mathbb{R}$

then

- $X_n + Y_n \xrightarrow{D} X + c$
- $X_n Y_n \xrightarrow{D} Xc$
- $X_n / Y_n \xrightarrow{D} X/c$

It follows from the continuity of a wisely chosen g in each case.

Finally, let us state two interesting results that tells us about the rate of convergence of the sample mean. The first one in the Central Limit Theorem.

Theorem 2.9. Central Limit Theorem
Let X_n be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{17}$$

This can be generalized to any kind of random vectors, in the following way.

Theorem 2.10. Multivariate Central Limit Theorem
Let X_n be a sequence of i.i.d. random vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Then

$$\sqrt{n}\,(\bar{X}_n - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}(0, \mathbf{\Sigma}) \tag{18}$$

This can be further generalized when one applies a function g to the sequence of random variables.

Theorem 2.11. Delta theorem
Let X_n be a sequence of random variables such that

$$\sqrt{n}\,(X_n - \mu) \xrightarrow{D} Z \tag{19}$$

where Z is a random variable and μ a constant. Then for any function g differentiable at μ with Jacobian matrix $J_g(m)$. Then

$$\sqrt{n}\,(g(X_n) - g(\mu)) \xrightarrow{D} J_g(m) Z \tag{20}$$

Remarque 2.12. In the case of a scalar function and a single random variable, the Jacobian is just the derivative of the function. For instance, if $\sqrt{n}(X_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} \mathcal{N}(0, g'(m)^2 \sigma^2) \tag{21}$$

2.2 Estimators

In this section we setup the framework for statistics and parametric estimation.

2.2.1 Generalities

This short paragraph recalls the general definitions about estimators.

Définition 2.13. Parametric model
In everything that follows, we shall call a parametric model a set of probability measures $\mathcal{P}$ on a measurable space $(\mathcal{X}, \mathcal{A})$ that is indexed by a parameter $\boldsymbol{\alpha}$ (that might be multidimensional). In our case we shall focus on models that are finite dimensional, *i.e.* $\boldsymbol{\alpha} \in \mathbb{R}^d$.
Furthermore we shall assume that the model is dominated by a σ-finite measure μ, so that any probability measure $\mathbb{P}_{\boldsymbol{\alpha}}$ is determined by a density $f(x|\boldsymbol{\alpha})$ with respect to μ. In our case we shall take the Lebesgue measure as the dominating measure.
Finally we shall assume that the model is identifiable, *i.e.* the map,

$$\boldsymbol{\alpha} \mapsto f(x|\boldsymbol{\alpha}) \tag{22}$$

is injective.

Définition 2.14. Statistic
A statistic is a measurable function of the data, *i.e.* a function $T : \mathcal{X}^n \to \mathcal{Y}$ where $\mathcal{Y}$ is a measurable space.

Définition 2.15. Estimator
An estimator is a statistic that is used to estimate a parameter of the model. It is a function of the data and is denoted by $\hat{\eta}$. In general it estimates a function of the parameters, $\eta = g(\boldsymbol{\alpha})$. η can be a scalar or a vector.

Remarque 2.16. In general, the estimator is a random variable, as it depends on the data that is a random variable. One can denote the estimator as a function of the data,

$$\hat{\eta} = \hat{\eta}(\mathcal{D}) \tag{23}$$

A realisation of the data is denoted by $\mathcal{D}^n = (x_i)_{i\in[\![1,n]\!]}$, while the date as a random variable is denoted by $\underline{\mathcal{D}}^n = (X_i)_{i\in[\![1,n]\!]}$.

Définition 2.17. Risk, bias and variance
The risk of an estimator is defined as

$$R(\hat{\eta}) = E\left(\|\hat{\eta} - g(\boldsymbol{\alpha})\|^2\right) \tag{24}$$

The bias of an estimator is defined as

$$B(\hat{\eta}) = E(\hat{\eta}) - g(\boldsymbol{\alpha}) \tag{25}$$

The variance (or covariance matrix) of an estimator is defined as

$$V(\hat{\eta}) = E\left((\hat{\eta} - E(\hat{\eta}))^2\right) = E((\hat{\eta} - E(\hat{\eta}))(\hat{\eta} - E(\hat{\eta}))^T) \tag{26}$$

Note how these quantities are defined a function of the parameters of the model. To make this precise, one could add a subscript $\boldsymbol{\alpha}$, but this shall make thing more complicated and blurred.

Theorem 2.18. Bias-variance decomposition
The risk of an estimator can be decomposed as

$$R(\hat{\eta}) = \mathrm{Tr}(V(\hat{\eta})) + B(\hat{\eta})^T B(\hat{\eta}) \tag{27}$$

Définition 2.19. Preferred estimator
An estimator is said to be preferred if it has a lower risk than any other estimator.

2.2.2 Cramér-Rao bound

In this paragraph we derive a lower for the risk. This is called the Cramér-Rao bound.

Définition 2.20. Regularity conditions
Let $\mathcal{P}$ be a parametric model and $\boldsymbol{\alpha}$ be a parameter of the model. We call the regularity conditions the following properties,

1. The *pdfs* $f(\cdot|\boldsymbol{\alpha})$ have the same support $\mathcal{S}$
2. The three conditions are satisfied (assuming also that the 1. is satisfied):
   (i) The space of the parameters is open in $\mathbb{R}^d$
   (ii) $\boldsymbol{\alpha} \mapsto f(x|\boldsymbol{\alpha})$ is differentiable for almost every x
   (iii) $\int_{\mathcal{S}} \nabla_{\boldsymbol{\alpha}} f(x|\boldsymbol{\alpha}) \mathrm{d}x = 0$ for any $\boldsymbol{\alpha}$
3. $\boldsymbol{\alpha} \mapsto f(x|\boldsymbol{\alpha})$ is C^2 for almost every x
4. For any $\boldsymbol{\alpha}$,
$$\int_{\mathcal{S}} \frac{\partial^2 f(x|\boldsymbol{\alpha})}{\partial \alpha_i \partial \alpha_j} \mathrm{d}x = \frac{\partial}{\partial \alpha_i} \int_{\mathcal{S}} \frac{\partial f(x|\boldsymbol{\alpha})}{\partial \alpha_j} \mathrm{d}x$$

Définition 2.21. Score
When the regularity conditions 1 and 2 are verified, the score of the model is defined as

$$S(x|\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \log f(x|\boldsymbol{\alpha}) \tag{28}$$

It can also be considered as a random variable, $S(\underline{\mathcal{D}}|\boldsymbol{\alpha})$.

Définition 2.22. Regular estimator
We suppose that regularity conditions 1. and 2. are verified. An estimator is said to be regular if the map $\boldsymbol{\alpha} \mapsto E(\hat{\eta}(\mathcal{D}))$ is differentiable.

Theorem 2.23. Cramér-Rao bound
Let $\hat{\eta}$ be a regular and unbiased estimator of $\eta = g(\boldsymbol{\alpha})$ in a parametric model. Then for any $\boldsymbol{\alpha}$,

$$R(\hat{\eta}) \geq \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})^T V(S(\underline{\mathcal{D}}|\boldsymbol{\alpha}))^{-1} \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) \tag{29}$$

where the inequality is to be understaood as the difference of the two matrices is semi-definite positive.

2.2.3 Asymptotic behaviour

It is usually hard to build perfect estimators, that is to say unbiased ones with a low variance. In fact, if it were easy, one could ususally just analytically derive the expression of the parameters. Instead, estimators usually rely on the fact that they approximate the true value of the parameter when the dataset becomes large. Then in general, one builds up a sequence of estimators defined by,

$$\hat{\eta}_n = \hat{\eta}(\mathcal{D}^n) \tag{30}$$

and one studies the asymptotic behaviour of this sequence. This is usually done using the probabilistic results we stated previously.

Définition 2.24. Consistency
A sequence $\hat{\eta}_n$ of estimators of $\eta = g(\boldsymbol{\alpha})$ is said to be

- Consistent if
$$\hat{\eta}_n \xrightarrow{P} g(\boldsymbol{\alpha}) \tag{31}$$
- Strongly consistent if
$$\hat{\eta}_n \xrightarrow{a.s.} g(\boldsymbol{\alpha}) \tag{32}$$
- Quadradically consistent if
$$\hat{\eta}_n \xrightarrow{L^2} g(\boldsymbol{\alpha}) \tag{33}$$

Définition 2.25. Asymptotic bias
An estimator is asymptotically unbiased if

$$\lim_{n\to\infty} B(\hat{\eta}_n) = 0 \tag{34}$$

One can use the following result for the qudratic consistency.

Theorem 2.26.
Let $\hat{\eta}_n$ be a sequence of estimators of $\eta = g(\boldsymbol{\alpha})$. This sequence is quadratically consistent if,

- It is asymtotically unbiased
- The variance of the estimator goes to zero,
$$\lim_{n\to\infty} \mathrm{Tr}(V(\hat{\eta}_n)) = 0 \tag{35}$$

2.3 Confidence intervals

In this paragraph we introduce the concept of confidence intervals. This is a very important concept that shall allow use to make statements about the value of the parameters of the model. It shall indicate us how sure we are about the value of the parameters.

Définition 2.27. Confidence interval
Let $\hat{\eta}$ be an estimator of $\eta = g(\boldsymbol{\alpha})$. A confidence interval of level $1-\alpha$ is an interval $I(\underline{\mathcal{D}})$ such that

$$P(\eta \in I(\underline{\mathcal{D}}|\boldsymbol{\alpha}) \geq 1-\alpha \tag{36}$$

for any $\boldsymbol{\alpha}$.
The confidence interval is a statistic, *i.e.* a function of the data. It should be seen as a random variable.

Remarque 2.28. The notation α for the level of the confidence interval is completely unrelated to the notation for the parameters of the model. It is purely conventional.

Theorem 2.29. Construction of confidence intervals
Let $\hat{\eta}$ be an estimator of $\eta = g(\boldsymbol{\alpha})$. Let $T(\underline{\mathcal{D}}, \eta)$ be random variable whose distribution does not depend on $\boldsymbol{\alpha}$. Denote by F its cumulative distribution function and suppose it is continuous and increasing. We call quantile of level r the value q_r such that

$$F(q_r) = r \tag{37}$$

With the above setup, $q_r = F^{-1}(r)$. One can then construct confidence intervals of level $1-\alpha$ using $\gamma \in (0, \alpha)$ by taking

$$I_\gamma(\underline{\mathcal{D}}) = T^{-1}(\underline{\mathcal{D}}, [q_\gamma, q_{1+\gamma-\alpha}]) \tag{38}$$

When the distribution of is symmetric, it is customary to take $\gamma = \alpha/2$ to make the interval symmetric, see figure 1.

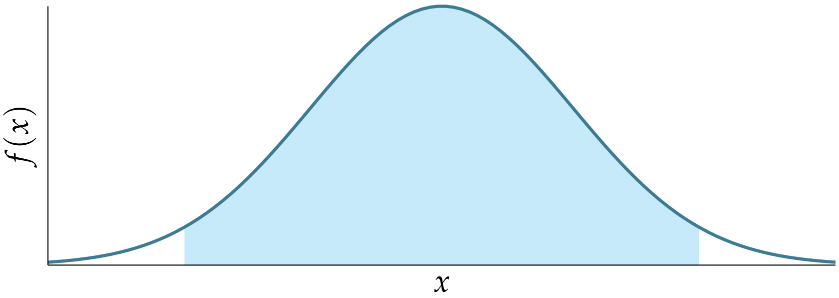

Figure 1: Confidence interval for a Gaussian distribution

Usually, one doesn't have access to such a confidence interval because the distribution of T is unknown or depends on the parameters. In this case, one can use asymptotic confidence intervals.

Définition 2.30. Asymptotic confidence interval
Let $I_n(\underline{\mathcal{D}}^n)$ be a sequence of intervals. It is said to be an asymptotic confidence interval of level $1-\alpha$ if

$$\lim_{n\to\infty} P(\eta \in I_n(\underline{\mathcal{D}}^n)|\boldsymbol{\alpha}) \geq 1-\alpha \tag{39}$$

It is usually much easier to work with asymptotic intervals since as we saw previously, one have many results on the asymptotic behaviour of the estimators. Let us give an example on how to do that.

Example 2.31. Asymptotic confidence interval for the mean
Let X_n be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\bar{X}_n$ be the sample mean. Then, by the Central Limit Theorem,

$$\sqrt{n}\,(\bar{X}_n-\mu) \xrightarrow{D} \mathcal{N}(0,\sigma^2) \tag{40}$$

Then, one can construct an asymptotic confidence interval for the mean by using the quantiles of the normal distribution. One can take

$$I_n(\underline{\mathcal{D}}^n) = \left[\bar{X}_n - \frac{q_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X}_n + \frac{q_{\alpha/2}\sigma}{\sqrt{n}}\right] \tag{41}$$

where $q_{\alpha/2}$ is the quantile of the normal distribution of level $1-\alpha/2$.
This example explains why it is often said that uncertainties on parameters scale as $1/\sqrt{n}$.

Définition 2.32. Coverage
The coverage of a confidence interval is defined as

$$\tau(I) = P(\eta \in I(\underline{\mathcal{D}})|\boldsymbol{\alpha}) \tag{42}$$

Remarque 2.33. The coverage is a measure of how well the confidence interval is doing. Usually one computes exact confidence intervals when the *cdf* of T is known *i.e* the coverage is exactly the level of the confidence interval $1-\alpha$. However, for asymptotic intervals it is interesting to see how the coverage scales with n, to undersatnd how well the confidence interval is converging towards an exact confidence interval.

2.4 Hypothesis testing

In this paragraph we focus on Hypothesis testing. This statistical tool will help us choose some values of the parameters compared to others. hypothesis testing consists in comparing two models, the null model and the alternative model. Let us denote them as,

$$\mathcal{H}_0 : \boldsymbol{\alpha} \in \Theta_0 \quad \text{and} \quad \mathcal{H}_1 : \boldsymbol{\alpha} \in \Theta_1 \tag{43}$$

One chooses Θ_0 and Θ_1 such that $\Theta_0 \cap \Theta_1 = \emptyset$.

What we call a test is a statistic that is a function of the data that shall help us choose between the two models. It is thus defined by,

$$\delta(\underline{\mathcal{D}}) = \begin{cases} 0 & \text{if} \quad \mathcal{H}_0 \quad \text{is accepted} \\ 1 & \text{if} \quad \mathcal{H}_1 \quad \text{is accepted} \end{cases} \tag{44}$$

We shall ususally work with a setup such that $\delta(\mathcal{D}) = 1$ is equivalent to the fact that some statistics $T(\mathcal{D})$ is greater than a threshold c.

Let us define a few important objects,

Définition 2.34. Type I and Type II errors
Let δ be a test of the hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$. Then

- The Type I error is defined as

$$\alpha = P(\delta = 1 | \boldsymbol{\alpha} \in \Theta_0) \tag{45}$$

- The Type II error is defined as

$$\beta = P(\delta = 0 | \boldsymbol{\alpha} \in \Theta_1) \tag{46}$$

Tyoe I error is the probability to reject the null hypothesis when it is true. Type II error is the probability to accept the null hypothesis when it is false.
One ususally calls type I error the significance level of the test and type II error the power of the test.

In general, what is done is to fix the significance level of the test, *i.e.* the probability of making a type I error. Then one tries to minimize the probability of making a type II error. This is done by choosing the test that maximizes the power of the test.

Définition 2.35. Comparison of tests
Let δ_1 and δ_2 be two tests of the hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$. We say that δ_1 is more powerful than δ_2 if

$$\beta_1(\boldsymbol{\alpha}) \leq \beta_2(\boldsymbol{\alpha}) \quad \forall \boldsymbol{\alpha} \in \Theta_1 \tag{47}$$

Once we work with a test, that we chose to be as powerful as possible, one ususally wnats to understand to which level some hypothesis is rejected. This is done by computing the p-value of the test. Suppose for now that we work with simple hypothesis, *i.e.* $\Theta_0 = \{\boldsymbol{\alpha}_0\}$ and $\Theta_1 = \{\boldsymbol{\alpha}_1\}$.

Définition 2.36. p-value
Let δ be a test of the hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$. The p-value of the test is defined as

$$p(\mathcal{D}) = P(T(\underline{\mathcal{D}}) \geq T(\mathcal{D})|\boldsymbol{\alpha} = \boldsymbol{\alpha}_0) \tag{48}$$

where $\mathcal{D}$ is the observed data.
The p-value is the probability to observe a test statistic as extreme as the one observed with the data, given that the null hypothesis is true.

Note how the p-value is a random variable, as it depends on the data (or it can be seen as a function of the data, meaning that it will chnage between two different experiments).

Remarque 2.37. If we know the distribution of the test statistic, we can compute the p-value exactly as,

$$p(\mathcal{D}) = 1 - F(T(\mathcal{D})) \tag{49}$$

In case we work with composite hypothesis, *i.e.* Θ_0 and Θ_1 are not singletons, one can still compute the p-value by taking the supremum of the p-values of the simple hypothesis.

Définition 2.38. Composite p-value
Let δ be a test of the hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$. The composite p-value of the test is defined as

$$p(\mathcal{D}) = \sup_{\boldsymbol{\alpha}_0 \in \Theta_0} P(T(\underline{\mathcal{D}}) \geq T(\mathcal{D})|\boldsymbol{\alpha} = \boldsymbol{\alpha}_0) \tag{50}$$

where $\mathcal{D}$ is the observed data.

2.5 Bayesian statistics

3 Likelihood

3.1 Maximum likelihood estimator (MVE)

As explained in the introduction of this document, the likelihood is a measure of how probable it is to observe the data given the parameters of the model. Hence it is natural to use the likelihood to estimate the parameters of the model. Namely, one chooses as an estimator of the parameters the value that maximizes the likelihood. *These are the most probable parameters given the data.*

The likelihood is defined as,

$$\mathcal{L}(\boldsymbol{\alpha}) = f(\mathcal{D}_{tot}|\boldsymbol{\alpha}) \tag{51}$$

and the estimator of maximum likelihood is defined as

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}) \tag{52}$$

It is not obvious that it exists but we shall assume that the likelihood is sufficiently well-behaved so that it exists. As explained previously, the likelihhod is typically the product of the different *pdf*s of the data in each channel. It is usually easier to work with the log-likelihood, defined as

$$-\log \mathcal{L}(\boldsymbol{\alpha}) \tag{53}$$

So that we minimize a sum instead of maximizing a product. Mathematically, this is equivalent.

From now on we suppose the model to be regular, *i.e.* the regularity conditions are satisfied and the Fisher information matrix, defined by,

$$I(\boldsymbol{\alpha}) = -E\left(\nabla^2_{\boldsymbol{\alpha}} \log f(\mathcal{D}_{tot}|\boldsymbol{\alpha})\right) \tag{54}$$

is positive definite.

One already has a necessary condition for $\hat{\boldsymbol{\alpha}}$ to be a maximum of the likelihood.

Theorem 3.1. Necessary condition for the MVE
Let $\hat{\boldsymbol{\alpha}}$ be the MLE. Then

$$\nabla_{\boldsymbol{\alpha}} \log L(\mathcal{D}_{tot}|\hat{\boldsymbol{\alpha}}) = 0 \tag{55}$$

and

$$\nabla^2_{\boldsymbol{\alpha}} \log L(\mathcal{D}_{tot}|\hat{\boldsymbol{\alpha}}) < 0 \tag{56}$$

Why is the MVE so important. It is because it is asymptotically efficient.

Theorem 3.2. MVE asymptotic efficiency
Let us denote by $\boldsymbol{\alpha}_0$ the true value of the parameters and by $\hat{\boldsymbol{\alpha}}_n(\underline{\mathcal{D}}^n)$ the MVE, then

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{D} \mathcal{N}(0, I(\boldsymbol{\alpha}_0)^{-1}) \tag{57}$$

This is a very strong result, that shall be compared with Cramér-Rao bound.

3.2 Hypothesis testing

4 Uncertainty propagation