INFORMATION

- ~~Fisher~~ Information

When we perform an experiment, we typically collect a huge amount of data that we need to clean and reduce in order to make a statement on whatever quantity we are interested on.

Examples: → In CUORE, we have ~ 200 TB of raw data, but our publications just report the result on the halflife of an isotope.

→ CMS or ATLAS have PB of data, but just measured the Higgs mass and cross section.

We need to define a method to select the useful information.
But first we need to define the requirements for what we call information:

→ The information should increase with the number of observations

→ The information should be conditional on what we want to learn from the experiment.
Data which are irrelevant to the hypothesis under test should contain no information.

→ The greater the information, the better should be the precision of the experiment.

- Likelihood

Let's take a real random variable $\vec{x}$ with PDF $f(\vec{x}|\vec{\theta})$,
where $\vec{\theta}$ is a set of real parameters.
The set of allowed ~~par~~ values of $\vec{x}$ is Ω_θ, which might depend on $\vec{\theta}$.

Suppose we make a set of n observations of $\vec{x}$: $\vec{x}_1, \dots, \vec{x}_n$

The joint PDF of $\vec{x}$ is: $P(\vec{x}|\vec{\theta}) = P(\vec{x}_1, \dots, \vec{x}_n|\vec{\theta}) = \prod_{i=1}^{n} f(\vec{x}_i|\vec{\theta})$

Since the values $\vec{x}_i$ are fixed (they are measured!), P is no longer a PDF, but only a function of $\vec{\theta}$, and we denote it as $\mathcal{L}$:

$$\boxed{\mathcal{L}(\vec{\theta}) = \mathcal{L}(\vec{x}|\vec{\theta}) = \prod_{i=1}^{n} f(\vec{x}_i|\vec{\theta})}$$

- Statistic

A statistic is any new random variable $t = t(\vec{x}_1, \ldots, \vec{x}_n)$.

For example, the average $\langle \vec{x} \rangle$ is a statistic.

- Fisher information

Assume that 1) Ω_θ is independent of $\vec{\theta}$

2) $\mathcal{L}(\vec{x}|\vec{\theta})$ is regular enough so that the operators $\frac{\partial^2}{\partial\theta_i \partial\theta_j}$ and $\int d\vec{x}$ commute.

The Fisher information given by an observation x about the parameter θ is defined as:

$$I_x(\theta) = E\left[\left(\frac{\partial \ln \mathcal{L}(x|\theta)}{\partial\theta}\right)^2\right]$$

$$= \int_\Omega \left(\frac{\partial \ln \mathcal{L}(x|\theta)}{\partial\theta}\right)^2 \mathcal{L}(x|\theta)\, dx$$

If $\vec{\theta}$ has k dimensions, $I_x(\theta)$ is a $k \times k$ matrix:

$$\left[I_x(\vec{\theta})\right]_{ij} = \int_\Omega \frac{\partial \ln \mathcal{L}}{\partial\theta_i} \cdot \frac{\partial \ln \mathcal{L}}{\partial\theta_j} \cdot \mathcal{L} \cdot dx$$

Equivalently, one can prove that:

$$\left[I_x(\vec{\theta})\right]_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln \mathcal{L}\right]$$

The Fisher information is additive: $I_N(\theta) = N\, I_1(\theta)$

- Sufficiency

A statistic $t = t(\vec{x})$ is sufficient for ϑ if the conditional density function of $\vec{x}$ given t, $f(\vec{x}|t)$ is independent of ϑ.

If t is a sufficient statistic, any strictly monotonic function of t is also a sufficient statistic.

$\Rightarrow$ There is as much information about ϑ in T as there is in the original data $\vec{x}$.

$\Rightarrow$ No other function of the data can give any further information about ϑ.

Example: The set ~~___~~ $t = \vec{x}$ is sufficient, since it carries all the initial information. However, it provides no data reduction, so it is useless.

If $t(\vec{x})$ is a sufficient statistic for ϑ, the likelihood factorises as:

$$\mathcal{L}(\vec{x}|\vec{\vartheta}) = g(t, \vec{\vartheta})\, h(\vec{x}) \quad \text{and viceversa}$$

where: $h(\vec{x})$ does not depend on $\vec{\vartheta}$

$g(t, \vec{\vartheta}) \propto A(t|\vartheta)$, the conditional probability density for t given ϑ.

~~Therefore~~ $\cancel{A(t|\vec{\vartheta}) = \ldots \int_{\Omega} \mathcal{L}(\vec{x}|\vec{\vartheta})\, d\vec{x}}$

In general, ~~___~~ for any statistic t:

$$I_t(\vec{\vartheta}) \leq I_x(\vec{\vartheta})$$

with the equality if and only if t is a sufficient statistic.

In other words, the information provided by a sufficient statistic is the same as that of the original sample $\vec{x}$.

MEASUREMENT THEORY

In general, whenever we perform a measurement, we need to convey the result in a clear and synthetic way. Often times our result is a number (or a set of numbers) that will/should be used by other in the future, so we need to minimize the possible ambiguity on the underlying meaning of the quantity we quote.

~~When~~ Suppose we collect some data $\vec{x}$ distributed with a PDF $f(\vec{x}|\vec{\theta})$, and want to make a statement on ~~some of the par~~ one parameter θ (out of the vector $\vec{\theta}$).

We can ask the following questions:

→ Based on the measured data $\vec{x}$, what is the single value $\hat{\theta}$ that is closest to the true (unknown) value of θ?

⇒ Point estimation

→ Based on the measured data $\vec{x}$, what is the range of values that is most likely to include the true (unknown) value ~~s~~ of θ?

⇒ Interval estimation

~~→ Based on the~~

→ Is our model $f(\vec{x}|\vec{\theta})$ good enough to describe the measured data?

⇒ Goodness of fit

→ In the case we want to test the existence of new physics, e.g. the presence of a ~~sig~~ new signal over a known background, are the measured data described better by the background-only or by the signal+background model?

⇒ Hypothesis testing

- Addressed question vs required method

Each of the 4 questions listed above requires the use of dedicated statistical and computational methods.

Understanding the relation between addressed questions and required methods is fundamental, and will save you a lot of time in the future (trust me)!

~~Point estimation → Maximize like~~ ~~Find parameter values $\hat{\theta}$ that maximize the likelihood (or minimize the χ^2)~~

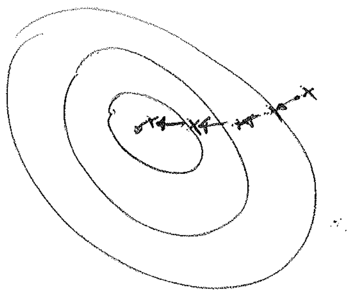
Point estimation → Find the parameter values $\hat{\theta}$ that maximize $\mathcal{L}$ → Minimizer algorithms (gradient follower)

Interval estimation → Study the tails of $\mathcal{L}$ or of the posterior $P(\theta|\vec{x})$ → Study all possible combinations of $\vec{\theta}$ giving $\vec{x}$
↳ Toy MC
→ Map the posterior $P(\theta|\vec{x})$
↳ Markov Chain MC

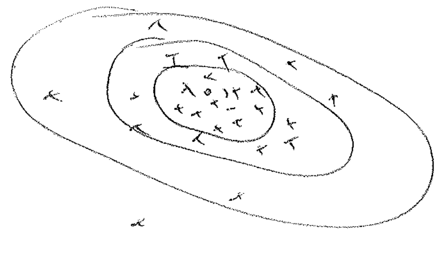
Goodness of fit → Quantify the probability of a random fluctuation to give a worse fit → ~~test~~ Analytical method (e.g. χ^2) plus lookup tables
→ Toy MC

Hypothesis testing → Compare the signal strength with the probability of a similarly strong background fluctuation → Toy MC + some method to compare validity of alternative hypotheses.

Minimizer:



Mapper:



So far, we've used a very vague language on purpose. To be more specific, we need to choose either the frequentist or the Bayesian approach, and specify the questions addressed by each of them.

- Frequentist approach

~~Point estimation~~

Assumptions: The true value of the parameter ϑ is fixed but unknown.
We cannot associate a PDF to ϑ, but just to the data $\vec{x}$.

Point estimation: ~~What~~ Based on the measured data, what's our best "estimate" for the fixed unknown parameter?
What's the estimate that is closer to the true value?

Interval estimation: Based on the measured data, what interval contains the true value with a predefined amount of probability (e.g. 68%)?
~~→ This has to be true also if we repeat~~
→ If we repeat the measurement 100 times, we will have 100 different intervals, and ~~but~~ the true value will be contained in them 68 times

Goodness of fit: ~~Is~~ Does my model provide a suitable description of the data, or is there any indication that it should be modified somehow?

Hypothesis testing: Based on the data, which among two (or more) alternative hypotheses is true?
~~What is the probability that~~
→ Assuming H_0 is true, what is the probability that the data will fake H_1 (and viceversa)?

• Bayesian approach

In the Bayesian approach, the probability is interpreted as a "degree of belief" and can be therefore applied to a wider range of ~~...~~ elements, including:

- random variables
- (true) parameters of a model
- hypotheses

Point estimation: based on the measured data, what is the most probable value for the parameter θ?

Interval estimation: based on the measured data, what is the interval ~~values that contains 68% of P(θ)~~ of the PDF of θ, $f(\theta)$, that contains a given amount of probability (e.g. 68%)?

Goodness of fit: This question makes no sense in the Bayesian approach, because we cannot compare one hypothesis with N unknown ones.

Hypothesis testing: based on the data, what is the ratio of the probabilities of hypotheses H_0 and H_1?

Plan 4