INFORMATION

• 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 ~ 2000 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 values of $\vec{x}$ is $\Omega_{\vec{\theta}}$, 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 $\mathcal{L}$:

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

• Sufficiency

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

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

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

⇒ No other function of the data can give any further information about $\vec{\theta}$.

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 $\vec{\theta}$, the likelihood factorizes as:

$$\mathcal{L}(\vec{x}|\vec{\theta}) = g(t, \vec{\theta})\, h(\vec{x}) \quad \text{and vice versa}$$

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

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

~~Therefore: $A(t|\vec{\theta}) = \int_R \mathcal{L}(\vec{x}|\vec{\theta})\, d\vec{x}$~~

In general, for any statistic t:

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

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.

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~~ 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 of θ?

⇒ Interval estimation

~~Based on the~~ m

→ 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 new signal over a known background, are the measured data described better by the background-only or by the signal+background model?

⇒ Hypothesis Testing

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~~ and 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 ~~the~~ 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)?