I.1 The Role of Statistical Inference

I.1.0 Loredo's Lead-in

I.1.0.1 What is probability

$p : x \mapsto p(x)$ is a map that
takes a value of a random variable to its image $p(x)$
we are asking the **interpretation of** $p(x)$

Frequentists' viewpoint:
$p(x)$ is the frequency of x in the ensemble
from frequentists' perspective, the values of random variable is distributed

Bayesian's viewpoint:
$p(x)$ is the probability that the random variable's value is x
from Bayesian's perspective, the "random variable" has a single value, probability is distributed

I.1.0.2 Essentials from logic

Construct arguments with **propositions** and **logic connectives**.

proposition

A proposition is a statement that is either true or false

logic connective

argument

$H|\mathcal{P}$: premise $\mathcal{P}$ implies hypothesis H
notice both H and $\mathcal{P}$ are propositions

Validity and Soundness of An Argument

Validity：
An argument is said to be **valid**, if: H is true given $\mathcal{P}$ is true.
The validity of an argument only concerns its form rather then content

Factually Correct
An argument is said to be **factually correct** if its premise $\mathcal{P}$ is true.
This concerns only the content of an argument rather than form.

Soundness:
An argument is sad to be **sound** if it's both **factually correct** and **valid**.

Integer Representation of Deduction

Denote V a map such that sends valid argument to 1 and invalid argument to 0.
Then we have:
$V(A \vee B \mid \mathcal{P}) = V(A \mid \mathcal{P}) + V(B \mid \mathcal{P}) - V(A \wedge B \mid \mathcal{P})$
$V(A \wedge B \mid \mathcal{P}) = V(A \mid \mathcal{P})V(B \mid \mathcal{P})$

Extend IR of Deductions to RR of Inductions

To measure the strength of argument, we want to construct a map $P(H|\mathcal{P})$ such that

Let $P(H \mid \mathcal{P})$ be a map that assigns the value 1 to valid arguments and 0 to invalid arguments. We aim to construct a map that assigns a real number between 0 and 1 to inductive arguments, where the value assigned to an inductive argument reflects its degree of reliability.

To construct such a map, we draw inspiration from the properties of a mapping in the previous "deductive arguments" context, which assigns integer values:

1. For an argument leading to an "and" proposition, its validity/strength is equal to the product of:
    - The validity/strength of the argument "the same premises derive A";
    - The validity/strength of the argument "premises $+A$ can derive B".
2. For an argument leading to an "or" proposition, its validity is equal to:
    - The validity/strength of the argument "the same premises can derive the sub-proposition A";
    - Plus the validity/strength of the argument "the same premises can derive the sub-proposition B";
    - Minus the validity/strength of the argument "the same premises derive the 'or' proposition".

The only thing we need to modify is the product rule:
$P(A \wedge B \mid \mathcal{P}) = P(A \mid \mathcal{P})P(B \mid A, \mathcal{P})$

Why make this modification?
Suppose that A implies B, namely $P(B|A) = P(B|A, \mathcal{P}) = 1$, then we don't expect $P(A, B|\mathcal{P})$ to differ from $P(A|\mathcal{P})$; but if we use the product rule for validity, the RHS of $P(A, B|\mathcal{P})$ would be $P(A|\mathcal{P}P(B|\mathcal{P}))$ which differs from $P(A|\mathcal{P})$ by a scaling factor $P(B|\mathcal{P})$ which is not generally 1.

Surprisingly, we find that the two dominating rules (product rule (AND), sum rule (OR)), coincide with those of **probability theory**. We thus steal everything from probability theory.

I.1.1 Goal and Methodology of Science

Roughly speaking, the ultimate goal of (the majority) of physicists is to **find the rules that rule everything of our universe**, from these rules we can describe the real world by a **mathematical model** such that explain or predict measurement/experiments...
This can be concluded as: physicists make arguments.

I.1.2 Parameterized Hypotheses

Assume that hypotheses can be parameterized by a set of (finite or infinite number of) parameters $\vec{\lambda} = (\lambda^1, \lambda^2, \dots)$, thus we may consider a hypothesis as a vector in some multi-dimensional vector space.
Assuming our universe allow only one unique set of rules, thus only one hypothesis could be true. Then, given we know enough facts of our universe, then a hypothesis can either be true or false, namely the probability density $P(\vec{\lambda}_0|D)$ for any specific $\vec{\lambda}_0$ would be either 0 or $+\infty$, where D is a data set sufficiently abundant. Now that we only have access to a limit range of facts (by using the word facts, we assume there is no bias), we may expect the **probability density** $P(\vec{\lambda}|D)$ to be spread in some subspaces of the space of $\vec{\lambda}$, instead of being a Dirac delta function at some unique point.
And we expect that:

$$A \int P(\vec{\lambda}|D) d\vec{\lambda} = 1$$

Then, given a set of observational facts D, $P(\vec{\lambda}|D)$, from probability theory, is the **probability density** that **hypothesis labeled $\vec{\lambda}$ is true, given D is true**.
Now our question become how to calculate this quantity, the answer is through **Baye's theorem**.

I.1.3 Bayes Theorem

By making one single presumption that **physical hypothesis can be parameterized**, we can now transplant **all frequentists' theorems from probability theory** to Bayesian inference. Among which the most important one is **Bayes's Theorem**:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Now let's substitute A by parameters $\vec{\lambda}$ of hypotheses and B by observational/experimental facts D:

$$P(\vec{\lambda}|D) = P(\vec{\lambda}) \frac{P(D|\vec{\lambda})}{P(D)} =: \pi(\vec{\lambda}) \frac{\mathcal{L}(D|\vec{\lambda})}{E(D)}$$

the quantity on LHS is called **posterior distribution** of parameters; and on RHS we:
take $P(\vec{\lambda})$ out and denote it by $\pi(\vec{\lambda})$ (prior) because this distribution can be derived from nowhere, this distribution has to be chosen manually **prior** to our observation (where we get D), in a sense this quantity is dependent to the parameterization of the multi-dimensional vector space of hypotheses
we denote $P(D|\vec{\lambda})$ by $\mathcal{L}(D|\vec{\lambda})$ (likelihood) because this function represents that **likelihood** (probability density) that our observation give output valued D (consider D some specific value) when hypothesis $\vec{\lambda}$ is true.

Thus the essential question lies on the computation of **likelihood** $\mathcal{L}(D|\vec{\lambda})$. For event-level inference, this is rather simple:
For example, we want to estimate the **mass** m of an astrophysical object from its **observational data** D.
Which means our hypotheses are 'the mass of the object is $m = m_0$' where m_0 varies for different hypotheses, and thus the hypothesis is parameterized by a single parameter m.
For any given input value $m = m_0$, the output of our observational instrument is **not a unique value** D_0, but is spread in some range $\mathbb{D}$, and the probability that our instrument pop up with $D \in \mathbb{D}$ is characterized by some distribution $f_{m_0}(D)$, i.e. the probability density that our instrument pop up D given the true input is m_0. From this definition, we realize that this function - characterizing the intrinsic property of our instrument - is the **likelihood function** we are looking for.

This can be easily generalized to multi-parameter inference.
But what if we want to estimate the hypotheses on the distributions of these event-level parameters?

I.2 Hierarchical Inference

For simplicity, we denote the parameters of event-level hypotheses by θ, I'm omitting the vector symbol but it should always be recognized as a vector.

What happens if we want to make a hypothesis on the distribution of these parameters?
For examples, the event-level parameters of LIGO's BBH merger detections includes component masses m_1, m_2 and effective spin χ_{eff}, one may wonder if the shape of the distribution of m_1 among the BBH population, and make hypothesis like 'the primary mass distribution of BBH merger follows a power law whose index is $\alpha = \alpha_0$', and thus these **population-level** hypotheses can be parameterized by α, and we call α the **hyper-parameter** of these hypotheses. A specific type of hyper-parameterized population-level hypotheses on BBH population is called a **population model**. (for instance, power-law model; Gaussian-peak model, ...)

To distinguish **population-level** hyper-parameters from parameters of specific events, we denote:
Hyper-parameters of a population model by Λ or $\vec{\Lambda}$
Parameters of an event by θ or $\vec{\theta}$
To determine the **supportiveness** our question become **how to calculate the posterior of hyper-parameters**.

Thanks to our effective notation, this can be easily done by transplanting formulae from probability theory:

$$P(\Lambda|D) = p(\Lambda)\frac{P(D|\Lambda)}{p(D)}$$

Again, $p(\Lambda)$ is **prior** which we have the 'freedom' to choose, and the **evidence** $P(D)$ is a constant for fixed D in practice, so the **likelihood** $P(D|\Lambda)$ is the only distribution of interest. And this can be easily expressed as marginalization:

$$P(D|\Lambda) = \int P(D|\theta,\Lambda)P(\theta|\Lambda)d\theta = \int P(D|\theta)P(\theta|\Lambda)d\theta$$

where the first equality is a general identity from probability theory, and the second stands at the probability distribution of getting D is determined by given θ and irrelevant of Λ once θ are given.
Now what left on RHS undetermined are $P(D|\theta)$ (which represents instrumental properties) and $P(\theta|\Lambda)$ (which represents population model properties), thus all knowledge used to determine the posterior distribution are accessible to us.

I.3 Including Selection Effect

What indeed is the (values) of posterior $P(\Lambda|D)$ stand for in the last section? According to our interpretation introduced in I.1.2 Parameterized Hypotheses, it is the **probability density** of 'hypothesis labeled Λ_0 be true' **given** 'the observations give $D^{\text{obs}} = D_0$'.
But According to our interpretation, the **more background knowledge we have** the **better we can constrain the hypotheses**. What additional background knowledge we have ignored in the calculation above? That is, in our dataset $D = \{D_i\}$ in practice, we **don't include outputs of all detections**, actually the data set D is **sampled** from a larger data set, according to some rule of sampling which we denote S. (In practice, this S often means excluding data such that SNR lower than some threshold value.)
So, to include as more background information as possible, we shall either target at $P(\Lambda|D^*)$ where D^* is the complete data set, or $P(\Lambda|D, S)$ where S is our knowledge on sampling procedure.

First Approach

Let's first try $P(\Lambda|D^*)$, whose calculation is in essence calculation of likelihood $P(D^*|\Lambda)$. to solve for which we simply replace D by D^* in the last equation given in I.2 Hierarchical Inference:

$$P(D^*|\Lambda) = \int P(D^*|\theta,\Lambda)P(\theta|\Lambda)d\theta = \int P(D^*|\theta)P(\theta|\Lambda)d\theta$$

But are these all background knowledge we could use? No, we have **additional knowledge** that $D_i^* \in D^*$ such that its SNR lower than some value are meaningless, which means the second approach actually constrain the hypotheses better.

Second Approach

We thus move to the second approach: targeting at $P(\Lambda|D,S)$. Again what we are actually solving for is the **likelihood**:

$$P(D|\Lambda,S)$$

which stands for the probability density distribution of **sampled data**, given Λ, S. In other words, the values of this function at each D_0 is equal to $P(D^{\text{sampled}} = D_0|\Lambda,S)$, but this is equal to $P(D^{\text{det}} = D_0|\Lambda,S,\text{detection be sampled})$. we calculate this by inverse the product rule:

$$P(D^{\text{det}} = D_0|\Lambda,S,\text{detection be sampled}) = \frac{P(D^{\text{det}} = D_0,\text{detection be sampled}|\Lambda,S)}{P(\text{detection be sampled}|\Lambda,S)}$$

Let's first look at the **numerator**, by product rule:

$$P(D^{\text{det}} = D_0,\text{detection be sampled}|\Lambda,S) = P(\text{dbs}|D^{\text{det}} = D_0,\Lambda,S)P(D^{\text{det}} = D_0|\Lambda,S)$$

where the first term is always equal to 1 for any D_0 in our sample set, while the second term can be calculated by calculated by **marginalization**, thus we conclude:

$$\text{numerator} = P(D^{det} = D_0|\Lambda,S) = P(D^{\text{det}} = D_0|S)$$

which can be easily calculated by marginalization over θ.

Now let's move to the **denominator**, which should again be calculated via marginalization:

$$P(\text{dbs}|\Lambda,S) = \int d\theta \cdot P(\text{dbs}|\Lambda,S,\theta)P(\theta|\Lambda,S)$$

where $P(\text{dbs}|\Lambda,S,thets) = P(\text{dbs}|S,\theta)$ is the probability that the detection of an event of parameter θ be sampled, given sampling rule S; and $P(\theta|\Lambda,S) = P(\theta|\Lambda)$.
Thus we conclude:

$$\text{denominator} = \int d\theta \cdot P(\text{dbs}|\theta, S)P(\theta|\Lambda)$$