AstroStatistics and Cosmology

Jacopo Tissino

2020-10-06

Contents

Introduction

Monday 2020-9-28, compiled 2020-10-06

In this course we will discuss

1. Bayesian statistics for parameter estimation and model comparison;
2. Application of these to practical data analysis problems in cosmology.

The goal of the first part will *not* be on mathematical proofs, but on *applications*. We are "customers" of statistics. We need a good understanding of the theory, but not necessarily a very *formal* one.

However, we will not take the "cookbook" approach: we need to understand the statistics in depth, blindly applying a technique is bad.

Lectures from the fifth of October will be in room P1C in the Paolotti building. We have the handwritten notes from the professor, and LaTeX notes from students of the earlier years.

Email: `michele.liguori@unipd.it`, or `liguori.unipd@gmail.com` (it's the same).

There will be **homework**. Some exercises to solve and other things. We should hand it in within 3 weeks of the assignment, this is not strict, but we should let him know if we cannot do it in time. The final homework will require some coding, making some Monte Carlo Markov chains, and it will not have the deadline since coding takes time. We should hand it in before the exam.

We can choose whatever programming language we want (Python is good, C++ is slightly worse since the professor is not so familiar with it, but it's fine). We should have a summary of the results with plots, and show the source code.

The exam is an oral, to do whenever we want. When we contact him we will be given a journal paper to read. At the exam, we will do a blackboard presentation of the paper and be asked questions about it, like in journal club.

A book which does things similarly to this course is "Data Analysis: a Bayesian tutorial" by Silvia and Skilling [SS06].

Usual COVID safety procedure if we come to class physically. There are 23 seats available. Of course, we can also follow the lectures online.

Chapter 1

Bayesian statistics

1.1 Inference

We need to apply inductive reasoning, since deductive reasoning cannot work in real life, since we cannot know anything with certainty.

We apply reasoning in the form: "if A, then B is more plausible" and "we see evidence for B": so, A is more plausible. If "we see evidence for $\neg B$", instead, then A is less plausible.

We then need to establish clear mathematical rules for this plausible reasoning.

1.1.1 Cox's theorem

This theorem gives constraint on our system of 'probability', by which we mean a function which associates a real number to each 'proposition/hypothesis', by which we mean a possible state of the world. We want the probability of a certain event to be 1, and the probability of an impossible event to be 0. The probability is denoted as $\mathbb{P}$, and we interpret it as a **degree of belief**.

We want the rules we use to be self-consistent:

1. if $\mathbb{P}(A) > \mathbb{P}(B)$ and $\mathbb{P}(B) > \mathbb{P}(C)$, then $\mathbb{P}(A) > \mathbb{P}(C)$;
2. for any events, $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B)$, where the notation $\mathbb{P}(A|B)$ denotes the probability of A, *given that* B has happened;
3. starting from the same information, we must arrive at the same conclusions.

Find examples of inconsistency if these are not verified.

Maybe write point 2 in a more verbose way…

If these hold, then we have the **Kolmogorov axioms**:

$$\mathbb{P}(X|I) + \mathbb{P}(\neg X|I) = 1 \qquad \text{and} \qquad \mathbb{P}(AB|I) = \mathbb{P}(A|B, I)\mathbb{P}(B|I)\,. \tag{1.1.1}$$

We write the probabilities as conditioned on *preexisting knowledge* I. This quickly becomes annoying in the notation, so I will stop writing it, but it is always implied: we never discuss probabilities in a vacuum.

A corollary of the second rue is **Bayes' theorem**:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}\,. \tag{1.1.2}$$

Tuesday 2020-9-29, compiled 2020-10-06

The procedure for hypothesis testing will look like

$$\mathbb{P}(\text{hypothesis}|\text{data}) = \frac{\mathbb{P}(\text{data}|\text{hypothesis})\mathbb{P}(\text{hypothesis})}{\mathbb{P}(\text{data})}\,. \tag{1.1.3}$$

The key difference from the frequentist approach is that, while there the parameters have certain fixed values, here we can describe our *belief* about their values through a probability distribution.

The things we will want to do can be classified into

1. **hypothesis testing**, "are CMB data consistent with gaussianity?";
2. **parameter estimation**, "what is the value of the mass of the Sun?";
3. **model selection**, "is GR the correct theory of gravity?".

1.1.2 Parameter estimation

We start with an example: the toss of a coin. The question is: we toss it N times and get R heads. Is it a fair coin?

If $H \in [0, 1]$ is the probability of getting heads in a single coin flip, then

$$\mathbb{P}(R\text{ heads}|H, I) \propto H^R(1-H)^{N-R}\,. \tag{1.1.4}$$

Here, I is the other information we have about the coin: the fact that every throw is independent, the fact that there are no outcomes beyond heads or tails.

This is the **likelihood**, what we want to do is to invert the relation, finding a probability density function for H given the data. The probability $\mathbb{P}(\text{data})$, also called the **evidence**, is not something we need to calculate when doing parameter estimation: we can just write

$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters})\mathbb{P}(\text{parameters})\,, \tag{1.1.5}$$

since we are computing probability density functions, which need to be normalized in order to make sense. This is a useful parameter estimation toy problem, since we only have one parameter to estimate.

In order to use the formula we need a prior, $\mathbb{P}(\text{hypothesis})$. This is hard in general, and it depends on the problem. We might want a prior which is peaked around 0.5 for a regular coin, if we have no reason to think that it is unfair. Let us suppose we have doubts about the honesty of who gave us the coin: then, we might want a noninformative prior, like a flat one. Let us suppose we are in this case: $\mathbb{P}(\text{parameters}) = \text{const}$.

Since the prior is flat, the posterior is proportional to the likelihood:

$$\mathbb{P}(H|R\text{ heads}, I) \propto \mathbb{P}(R\text{ heads}|H, I) \propto H^R(1-H)^{N-R}\,. \tag{1.1.6}$$

We can simulate this experiment! We need a binomial random number generator.

Do the simulation!

As N increases, the posterior "zeroes in" onto the correct value. If we split the N simulated throws in two, and use the posterior from the first batch as a prior for the second, we get the same result!

Monday 2020-10-5, compiled 2020-10-06

A completely agnostic prior in the coin-flip example would be flat.

As we toss many times, the prior becomes less and less relevant.

As long as we never set it to zero!

Where is objectivity, if we include our beliefs in the analysis? There are fundamentalist bayesians and frequentists. In cosmology the Bayesian approach is best since we only have one realization of the universe.

Also, we want to include previous experience in our analysis. We can take an objective approach to priors, by selecting a *noninformative* one. These are *objective*, in the sense that they express the *a priori* ignorance about the process.

After our analysis in parameter estimation we find a PDF, which contains everything we need, but we want to give a number in our paper in order to summarize the result. This does not give us *more* information than the PDF.

A common approach, which works as long as the posterior is unimodal, is to give the maximum of the posterior: if the parameter is x, then the central value x_0 is calculated by setting

$$\left.\frac{\mathrm{d}P}{\mathrm{d}x}\right|_{x_0} = 0\,. \tag{1.1.7}$$

This is called *Maximum A-Posteriori parameter estimation*, MAP.

How do we provide error bars? The precision becomes higher as the peak becomes narrower. Let us consider the log of the posterior, $L = \log P$. This is typical, since the log is monotonic, and it is convenient since P is typically very "peaky".

Around the maximum, L is given by

$$\begin{aligned} L &= \log \mathbb{P}(x|\text{data}, I) \qquad &(1.1.8)\\ &= L(x_0) + \cancel{\left.\frac{\mathrm{d}L}{\mathrm{d}x}\right|_{x_0}}(x-x_0) + \frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}x^2}\right|_{x_0}(x-x_0)^2 + \mathcal{O}\Big((x-x_0)^3\Big)\,. \qquad &(1.1.9)\end{aligned}$$

If the peak is well-behaved, then we can approximate it at second order in a large enough region around the maximum: this is

$$P = \exp(L) \approx \underbrace{P_0}_{e^{L(x_0)}} \exp\left(\frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}x^2}\right|_{x_0}(x-x_0)^2\right)\,, \tag{1.1.10}$$

which is a Gaussian whose variance is

$$\sigma^2 = -\left(\left.\frac{\mathrm{d}^2L}{\mathrm{d}x^2}\right|_{x_0}\right)^{-1}\,, \tag{1.1.11}$$

which will be positive: if x_0 is a maximum the second derivative is negative.

So the expansion in L is justified a posteriori through the Central Limit Theorem?

There will be a theorem telling us that the posterior converges to a Gaussian, and the estimate will converge to the maximum likelihood estimate.

Typically we have many parameters, not just one. Even if the theory only has a few, the experiment will also have several.

Integration is difficult in multidimensional contexts, we want to have something better than exponential time in the parameter number.

Suppose that our posterior is very asymmetric. Then, it might be better to give the mean of the posterior instead of the maximum:

$$\langle x \rangle = \int x \mathbb{P}(x|\text{data}, I) \, \mathrm{d}x \,. \tag{1.1.12}$$

If there is symmetry, this is similar to the maximum. Quoting both if there is asymmetry might be good.

We then want to build a **credible interval**, which we define as the *shortest* interval $[x_1, x_2]$ containing the representative value we choose, say $\langle x \rangle$, and which integrates to a certain chosen value, often chosen to be 95 %:

$$\int_{x_1}^{x_2} \mathbb{P}(x|\text{data}, I) = 0.95 \,. \tag{1.1.13}$$

Let us apply this procedure to the coin toss problem. Recall that the posterior PDF, with a flat prior, was given by

$$P = \mathbb{P}(H|\text{data}, I) \propto H^R (1-H)^{N-R} \,. \tag{1.1.14}$$

The derivative is given by

$$\begin{aligned}
\frac{\mathrm{d}L}{\mathrm{d}H} = \frac{\mathrm{d}\log P}{\mathrm{d}H} &= \frac{\mathrm{d}}{\mathrm{d}H}\left[R \log H + (N-R)\log(1-H)\right] &\quad (1.1.15)\\
&= \frac{R}{H} - \frac{N-R}{1-H} \,, &\quad (1.1.16)
\end{aligned}$$

which we set to zero: this yields

$$\frac{R}{H_0} = \frac{N-R}{1-H_0} \implies R - H_0 N = 0 \,, \tag{1.1.17}$$

so $H_0 = R/N$. This is what we get in the end, when we have many data.

The errorbar can be found by differentiating again:

$$\begin{aligned}
\frac{\mathrm{d}^2 L}{\mathrm{d}H^2} &= -\frac{R}{H^2} - \frac{N-R}{(1-H)^2} &\quad (1.1.18)\\
&= \frac{R(2H-1) - NH^2}{H^2(1-H)^2} \,, &\quad (1.1.19)
\end{aligned}$$

so we can find the errorbar by computing it in $H = R/N$: skipping a few steps, it is

$$
\begin{aligned}
\sigma^2 = \frac{(R/N)^2(1-R/N)^2}{N(R/N)^2 - R(2R/N-1)} = \frac{H_0(1-H_0)}{N} \qquad (1.1.20) \\
\sigma = \sqrt{\frac{H_0(1-H_0)}{N}}\,. \qquad (1.1.21)
\end{aligned}
$$

As is expected, this scales like $1/\sqrt{N}$. The reason we're doing this with the full Bayesian machinery is that the procedure will not yield these simple results in general.

Let us solve a probability problem using Bayesian statistics: the Monty Hall problem.

There are three doors, one of which is desirable, the other two are not. We choose one door; the host knows which the desirable door is, he excludes a door as undesirable and asks us whether we want to change our choice.

Tuesday 2020-10-6, compiled 2020-10-06

Coming back to the Monty Hall problem. We picked A, the host picked C.

Let us denote the presence of the desirable object with 1, 0 its absence. So, we have three options: $(A, B, C) = (1, 0, 0)$, or $(0, 1, 0)$, or $(0, 0, 1)$. *A priori*, we assign a probability of 1/3 to each: this is, then, our prior. Let us assume that WLOG A is the door we picked. This has probability 1, since we can make it true in any case by relabeling the doors.

We want to compute

$$
\mathbb{P}(B = 1 | \text{host picked } C, \text{ we picked } A)\,. \qquad (1.1.22)
$$

We will write the condition as $[C]$ for compactness. The complement to this probability will be

$$
\mathbb{P}(A = 1 | [C])\,. \qquad (1.1.23)
$$

We can then apply Bayes' theorem:

$$
\mathbb{P}(B = 1 | [C]) = \frac{\mathbb{P}([C] | B = 1)\mathbb{P}(B = 1)}{\mathbb{P}([C])}\,. \qquad (1.1.24)
$$

$\mathbb{P}(B = 1) = 1/3$ is our prior. Now, if $B = 1$ and we chose A, then the host is *forced* to pick C since otherwise he will uncover the prize: $\mathbb{P}([C] | B = 1) = 1$.

Then, what we are left with is the computation of $\mathbb{P}([C])$.

We can write this through *marginalization*, integrating over all the possible events which can happen:[1]

$$
\mathbb{P}([C]) = \underbrace{\mathbb{P}([C] | A = 1)}_{=1/2} \mathbb{P}(A = 1) + \underbrace{\mathbb{P}([C] | B = 1)}_{=1} \mathbb{P}(B = 1) + \underbrace{\mathbb{P}([C] | C = 1)}_{=0} \mathbb{P}(C = 1) \qquad (1.1.25)
$$

[1] We assume that, if we selected the good door, the host chooses uniformly between the two: that is, $\mathbb{P}([C] | A = 1) = x = 1/2$. A host can actually be biased towards C or B, maybe he will choose the nearest door to him or something like that. In any case, if we leave x as a variable the final probability to find the prize by switching is found to be $1/(1 + x)$, which is always larger than 1/2: we are always better off switching. An interesting fact, however, is that by changing x the probability can move from 1/2 to 1.

$$= \frac{3}{2} \times \frac{1}{3} = \frac{1}{2}\,. \tag{1.1.26}$$

Another example. This is an investigation. The probability of anybody in a neighborhood to die of overdose is $\mathbb{P}(O) = 1/2$. Also, 30 % of murder victims are drug addicts:

$$\mathbb{P}(\text{addict}|\text{murder victim}) = 0.3\,. \tag{1.1.27}$$

We want to compute the probability that someone who was an addict was indeed murdered, and did not die of overdose: $\mathbb{P}(O|A)$: this will be given by

$$\mathbb{P}(O|A) = \frac{\mathbb{P}(A|O)\mathbb{P}(O)}{\mathbb{P}(A)}\,. \tag{1.1.28}$$

We do not have $\mathbb{P}(A|O)$, but we can estimate it, and then try to understand how much it affects the final result. Let us start out by estimating it as 0.9, since overdoses will likely most often happen to addicts.

We can calculate the probability of being an addict by marginalizing over the cause of a drug-induced death:

$$\mathbb{P}(A) = \mathbb{P}(A|M)\mathbb{P}(M) + \mathbb{P}(A|O)\mathbb{P}(O)\,, \tag{1.1.29}$$

and since $\mathbb{P}(O) = 0.5$, we can also have $\mathbb{P}(M) = 0.5$. This then means that $\mathbb{P}(A) = 0.6$. If there were other possible events, we would need to sum over them.

With all of this, we have

$$\mathbb{P}(O|A) = 0.75\,. \tag{1.1.30}$$

If we were to change $\mathbb{P}(A|O)$ this would not change much, so it is ok to estimate this roughly.

Let us discuss marginalization in some more detail. We typically do it for all the "nuisance parameters", which we must account for but do not really care about in the end: typically, parameters connected to experimental noise.

Suppose we have a PDF like

$$\mathbb{P}(X, Y)\,, \tag{1.1.31}$$

where Y can take values in the set of *exhaustive* and *mutually exclusive* events Y_k. Marginalization is the process of computing $\mathbb{P}(X)$ through

$$\mathbb{P}(X) = \sum_{k=1}^{N} \mathbb{P}(X, Y_k) = \sum_{k=1}^{N} \mathbb{P}(X|Y_k)\mathbb{P}(Y_k) \tag{1.1.32}$$

$$= \underbrace{\sum_{k=1}^{N} \mathbb{P}(Y_K|X)}_{=1} \mathbb{P}(X)\,. \tag{1.1.33}$$

Thus, we have shown that $\mathbb{P}(X)$ is indeed given by the expression above. It is crucial to assume that the Y_k are exhaustive and mutually exclusive in order for the sum of $\mathbb{P}(Y_k|X)$ to equal 1.

Typically, the events are not discrete but continuous. Exhaustivity and exclusivity can still apply, however we need to turn the sum into an integral.

We are considering probability density functions of these continuous parameters: they are defined as

$$\frac{\mathrm{d}p}{\mathrm{d}y} = \lim_{\delta y \to 0} \frac{\mathbb{P}(X, y \leq Y \leq y + \delta y)}{\delta y} \,. \tag{1.1.34}$$

Then, we can compute the finite probability of X as

$$\mathbb{P}(X) = \int_{\mathbb{R}} \mathrm{d}Y \, \frac{\mathrm{d}p}{\mathrm{d}y} \,. \tag{1.1.35}$$

If we integrate keeping Y in a certain region we find the probability of finding a value for it in that range.

We now set up the problem for tomorrow: we do parameter estimation. We have a set of measurements of the same quantity: $\vec{d} = \{x_i\}$. Each of the measurements is affected by error, and by the "experimentalist's Central Limit Theorem" their sum will resemble a Gaussian.

Each measurement x_i will then be given by $x_i = \mu + n_i$, the mean value plus a noise term. In this example, we assume that n_i is Gaussian. We want to write a posterior distribution for μ: it will be given in terms of the likelihood of the data given the true value, assuming that σ^2 is a fixed and known quantity,

$$\mathscr{L}(x_i|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\right) . \tag{1.1.36}$$

This is the likelihood from a single value x_i, while if we want to compute the joint likelihood of all the data $\{x_i\}$ it is harder. We can assume, in this specific case, that the errors are independent: therefore, the probability factors and we can write

$$\mathbb{P}(\{x_i\} \,|\mu) = \prod_i \mathbb{P}(x_i|\mu) = \frac{1}{\sigma^N \sqrt{2\pi}^N} \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right) . \tag{1.1.37}$$

Bibliography

[SS06] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, June 2006. 259 pp. ISBN: 978-0-19-856831-5. Google Books: lYMSDAAAQBAJ.