Lecture 1: Probabilities
Licia Verde

Goals: Not a rigorous introduction with proofs etc, but
a) a "practical manual", to give you enough knowledge to be able understand cosmological data analysis
b) a "bag of tricks" (hopefully) useful for your future work. We'll start talking about probability and statistics from a Cosmologist point of view, then we go on with description of random fields (ubiquitous in Cosmology). We then introduce Monte Carlo methods both Monte-Carlo error estimate and Markov Chain Monte Carlo. We conclude with forecasting the performance of future experiments via Fisher matrix. Depending on time I will cover issues and questions that may arise during the tutorials.

1 What's probability: Bayesian vs Frequentist

Probability can be interpreted as a **frequency**

$$P = \frac{n}{N} \tag{1}$$

where n stands for the successes and N for the total number of trials.

Or it can be interpreted as a lack of information: if I knew everything , i know that an event is a sure event, then $P = 1$, if I know nothing then $P = 0$ and in between I can use my judgment and or information from frequencies to estimate P.

The world is divided in Frequentists and Bayesians. In general Cosmologists are Bayesians and High Energy Physicists are Frequentists.

For Frequentists events are just frequencies of occurrence: probabilities are only defined as the quantities obtained in the limit after the number of independent trials tends to infinity.

Bayesians interpret probabilities as the degree of belief in a hypothesis: they use judgment, prior information, probability theory etc...

As we do cosmology we will be Bayesian.

2 Dealing with probabilities

In probability theory probability, distributions are fundamental concepts. They are used to calculate confidence intervals, for modeling purposes etc. We first need to introduce the concept of random variable in statistics (and in Cosmology). Depending on the problem at hand, the random variable may be the face of a dice, the number of galaxies in a volume δV of the Universe, the CMB temperature in a given pixel of a CMB map, the measured value of the power spectrum $P(k)$ etc. The probability that x (your random variable) can take a specific value is $P(x)$ where P denotes the probability distribution.
The properties of P are:
1) $P(x)$ is a non negative, real number for all real values of x.

2) $P(x)$ is normalized so that [1] $\int dx P(x) = 1$
3) For mutually exclusive events x_1 and x_2, $P(x_1 + x_2) = P(x_1) + P(x_2)$ the probability of x_1 or x_2 to happen is the sum of the individual probabilities. $P(x_1 + x_2)$ is also written as $P(x_1 U x_2)$ or $P(x_1.OR.x_2)$.
4) In general:

$$P(a, b) = P(a)P(b|a) = P(b)P(a|b) \tag{2}$$

The probability of a and b to happen is the probability of a times the conditional probability of b given a. Here we have also made the (apparently tautological) identification $P(a, b) = P(b, a)$. [see figure] For independent events then $P(a, b) = P(a)P(b)$.

Exercise: "Will it be sunny tomorrow?" answer in the frequentist way and in the Bayesian way.
Exercise: Produce some examples for rule n 4)

While Frequentists only consider distributions of events, Bayesians consider hypotheses as "events", giving us Bayes theorem:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \tag{3}$$

where H stands for hypothesis (generally the set of parameters specifying your model, although many cosmologists now also consider model themselves) and D stands for data. Note that this is nothing but equation 2 with substitutions: $b \longrightarrow H$ and $a \longrightarrow D$.

$P(H|D)$ is called the **posterior** distribution. $P(H)$ is called the **prior** and $P(D|H)$ is called **likelihood**. Eq.3 is a really important equation!!!

The usual points of heated discussion follow: How do you chose $P(H)$? Does the choice affects your final results? (yes, in general it will). Isn't this a bit subjective?

Exercise: Consider a positive definite quantity (like for example the tensor to scalar ratio r or the optical depth to the last scattering surface τ). What prior should one use? a flat prior in the variable? or a logaritmic prior (i.e. flat prior in the log of the quantity)? for example CMB analysis may use a flat prior in $\ln r$, and in $Z = \exp(-2\tau)$. How is this related to using a flat pror in r or in τ?

It will be useful to consider the following: Effectively we are comparing $P(x)$ with $P(f(x))$, where f denotes a function of x. for example x is τ and $f(x)$ is $\exp(-2\tau)$. Recall that: $P(f) = P(x(f)) \left|\frac{df}{dx}\right|^{-1}$. The Jacobian of the transformation appears to conserve probabilities.

Exercise: Under which conditions the choice of prior does not matter?

---

[1] for discrete distribution $\int \longrightarrow \sum$

3 Moments and cumulants

In the language of probability distribution **averages** are defined as follows:

$$\langle f(x)\rangle = \int dx f(x) P(x) \tag{4}$$

These can then be related to "expectation values" (see later). For now let's just introduce the moments: $\hat{\mu}_m = \langle x^m \rangle$ and, of special interest the central moments: $\mu_m = \langle (x - \langle x \rangle)^m \rangle$.

Exercise: show that $\hat{\mu}_0 = 1$ and that the average $\langle x \rangle = \hat{\mu}_1$. Also show that $\mu_2 = \langle x^2 \rangle - \langle x \rangle^2$

Here, μ_2 is the variance, μ_3 is called the skewness, μ_4 is related to the kurtosis. If you deal with the statistical nature of initial conditions (i.e. primordial non gaussianity) or non-linear evolution of gaussian initial conditions, you will encounter these quantities again (and again..). [fig here]

Up to the skewness central moments and cumulants coincide. For higher order terms things become more complicated. To keep things a simple as possible let's just take the gaussian distribution (see below) as reference. While moments of order higher than 4 are non-zero for both Gaussian and non-Gaussian distribution, the **cumulants** of higher orders are zero for a Gaussian distribution. For a Gaussian distribution all moments of order higher than 2 are specified by μ_1 and μ_2, for non-Gaussian distribution, yhe relation between central moments and cumulants κ for the first 6 orders is reported below.

$$\begin{aligned}
\mu_1 &= 0 && (5)\\
\mu_2 &= \kappa_2 && (6)\\
\mu_3 &= \kappa_3 && (7)\\
\mu_4 &= \kappa_4 + 3(\kappa_2)^2 && (8)\\
\mu_5 &= \kappa_5 + 10\kappa_3\kappa_2 && (9)\\
\mu_6 &= \kappa_6 + 15\kappa_4\kappa_2 + 10(\kappa_3)^2 + 15(\kappa_2)^3 && (10)\\
& && (11)
\end{aligned}$$

3.1 Useful trick: the generating function

Define the generating function as

$$Z(k) = \langle \exp(ikx) \rangle = \int dx \exp(ikx) P(x) \tag{12}$$

WHich may sound familiar as it is a sort of Fourier transform... Note that this can be written as an infinite series (by expanding the exponential) giving (exercise)

$$Z(k) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \hat{\mu}_n \tag{13}$$

So far nothing special, but now the neat trick is that the moments are obtained as:

$$\hat{\mu}_n = (-i^n)\frac{d^n}{dk^n}Z(k)|_{k=0} \tag{14}$$

and the **cumulants** are obtained by doing the same operation on $\ln Z$.

4 Two useful distributions

4.1 The Poisson distribution

The Poisson distribution describes an independent point process: photon noise, radioactive decay, galaxy distribution for very few galaxies, point sources .... It is an example of a discrete probability distribution. For cosmological applications it is useful to think of a Poisson process as follows: Consider a random process (example a random distribution of galaxies in space) of average density ρ. Divide the space in infinitesimal cells, of volume δV so small that their occupation can only be 0 or 1 and the probability of having more than one object per cell is 0. Then the probability of having one object in a given cell is $P_1 = \rho\delta V$ and the probability of getting no object in the cell is therefore $P_0 = 1 - \rho\delta V$. Thus for one cell the generating function is $Z(k) = \sum_n P_n \exp(ikn) = 1 + \rho\delta V(\exp(ik) - 1)$ and for a volume V with $V/\delta V$ cells, we have $Z(k) = (1 + \rho\delta V(\exp(ik) - 1))^{V/\delta V} \sim \exp[\rho V(\exp(ik) - 1)]$.

With the substitution $\rho V \longrightarrow \lambda$ we obtain
$Z(k) = \exp[\lambda(\exp(ik) - 1)] = \sum_{n=0}^{\infty} \lambda^n/n! \exp(-\lambda)\exp(ikn)$. Thus the Poission probability distribution we recover is:

$$P_n = \frac{\lambda^n}{n!}\exp[-\lambda] \tag{15}$$

Exercise: show that for the Poisson distribution $\langle n\rangle = \lambda$ and that $\sigma^2 = \lambda$.

4.2 The Gaussian distribution

The Gaussian distribution is extremely useful because of the "Central Limit theorem". The Central Limit theorem states that the sum of many independent and identically distributed random variables will be approximately Gaussianly distributed. The conditions for this to happen are quite mild; that is the variance of the distribution one starts off with has to be finite. The proof is remarkably simple. Let's take n events with probability distributions $P(x_i)$ and $< x_i >= 0$ for simplicity and let's Y be their sum. What is the $P(Y)$? Well the generating function for Y is the product of the generating functions for the x_i:

$$Z_Y(k) = \sum_{m=0}^{m=\infty}\left[\frac{(ik)_m}{m!}\mu^m\right]^n \simeq \left(1 - \frac{1}{2}\frac{k^2 < x^2 >}{n} + ...\right)^n \tag{16}$$

for $n \longrightarrow \infty$ then $Z_Y(k) \longrightarrow \exp[-1/2k^2 < x^2 >]$. By recalling the definition of generating function (eq. 12 remember it is a sort of Fourier transform of the Probability distribution)

we can see that the probability distribution which generated this Z is

$$P(Y) = \frac{1}{\sqrt{2\pi < x^2 >}} \exp\left[-\frac{1}{2}\frac{Y^2}{< x^2 >}\right] \tag{17}$$

that is a Gaussian!

Exercise: Verify that higher order cumulants are zero for the Gaussian distribution.

Exercise: show that the Central limit theorem holds for the Poisson distribution.

Beyond the Central Limit theorem, the Gaussian distribution is very important in cosmology as we believe that the initial conditions, the primordial perturbations generated from inflation, had a distribution very very close to Gaussian. (although it is crucial to test this experimentally)

We should also remember that thanks to the Central Limit theorem, when we estimate parameters in Cosmology in many cases we approximate our data as having a Gaussian distribution, even if we know that each data point is NOT drawn from a Gaussian distribution. The Central Limit theorem simplifies our lives every day...

There are exceptions though. Let us for example consider N independent data points drawn from a Cauchy distribution: $P(x) = [\pi\sigma(1 + [(x - \bar{x})/\sigma]^2]^{-1}$. This is a proper probability distribution as it integrates to unity, but moments diverge. One can show that the numerical mean of a finite number N of observations is finite but the "population mean" (the one defined through the integral of equation (4) with $f(x) = x$) is not. Note also that the scatter in the average of N data points drawn from this distribution is the same as the scatter in 1 point: the scatter never diminishes regardless of the sample size....

5 Modeling of data and statistical inference

If you have an urn with N red balls and M blue balls and you draw from the urn, probability theory can tell you what are the chances of you to pick a red ball given that you has so far drawn x blue and y red ones... However in practice what you want to do is to use probability to tell you what is the distribution of balls in the urn having made a few drawn from it! In other words, if you knew everything about the Universe probability theory could tell you what are the probabilities to get a given outcome for an observation. However, especially in cosmology, you want to make few observations and draw conclusions about the Universe! WIth the added complication that experiments in Cosmology are not quite like experiments in the lab: you can't poke the Universe and see how it reacts, and in many cases you can't repeat the observation, and you can only see a small part of the Universe! keeping this caveat in mind let's push ahead.

Given a set of observations often you want to fit a model to the data, where the model is described by a set of parameters $\vec{\alpha}$. Sometimes the model is physically motivated (say CMB angular power spectra etc.) or a convenient function (e.g. initial studies of large scale

structure were fitting galaxies correlation functions with power laws). Then you want to define a merit function, that measures the agreement between the data and the model: by adjusting the parameters to maximize the agreement one obtains the *best fit parameters*. Of course because of measurement errors there will be errors associated to the parameter determination. To be useful a fitting procedure should provide a) best fit parameters b) error estimates on the parameters c) possibly a statistical measure of the goodness of fit. When c) suggests that the model is a bad description of the data then a) and b) make no sense.

Remember at this point Bayes theorem: while you may want to ask:" what is the probability that a particular set of parameters is correct?" what you can ask to a "figure of merit" is "given a set of parameters what is the probability that that this data set could have occurred?". This is the likelihood. You may want to estimate parameters by maximizing the likelihood and identify the likelihood (probability of the data given the parameters) with the likelihood of the model parameters.

6 Chisquare, goodness of fit and confidence regions

Following Numerical recipes (press et al 1992, Chapter 15) it is easier to introduce model fitting and parameter estimation using the least-squares example. D_i are our data points and $y(\vec{x_i}|\vec{\alpha})$ a model with parameters $\vec{\alpha}$. For example if the model is a straight line then $\vec{\alpha}$ denotes the slope and intercept of the line. [figure]

The least squares is given by:

$$\chi^2 = \sum_i w_i (D_i - y(x_i|\vec{\alpha})^2 \tag{18}$$

and you can show that the minimum variance weights are $w_i = 1/\sigma_1^2$.

Exercise: if the points are correlated how does this equation changes?

Best fit value parameters are the parameters that minimize the χ^2. Note that by solving $\partial\chi^2/\partial\alpha_i \equiv 0$ you can find the best fit parameters.

6.1 Goodness of fit

In particular, if the measurement errors are Gaussianly distributed, and (as in this example) the model is a linear function of the parameters, then the probability distribution of for different values of χ^2 at the minimum is the χ^2 distribution for $\nu \equiv n - m$ degrees of freedom (where m is the number of parameters and n is the number of data points. The probability that the observed χ^2 even for a correct model is less than a value $\hat{\chi}^2$ is $P(\chi^2 < \hat{\chi}^2, \nu) = P(\nu/2, \hat{\chi}^2/2) = \Gamma(\nu/2, \hat{\chi}^2/2)$ where Γ stands for the incomplete Gamma function. Its complement, $Q = 1 - P(\nu/2, \hat{\chi}^2/2)$ is the probability that the observed χ^2 exceed by chance $\hat{\chi^2}$ even for a correct model. See numerical recipes (Press et at) for more details.

It is common that the chi-square distribution holds even for models that are non linear in the parameters and even in more general cases (see an example later).

The computed probability Q gives a quantitative measure of the goodness of fit when evaluated at the best fit parameters (i.e. at χ^2_{min}. If Q is a very small probability then

a) the model is wrong and can be rejected

b) the errors are really larger than stated or

c) the measurement errors were not Gaussianly distributed.

If you know the actuall error distribution you may want to **Monte Carlo simulate** synthetic data stets, subject them to your actual fitting procedure, and determine both the probability distribution of your χ^2 statistic and the accuracy with which model parameters are recovered by the fit.

On the other hand Q may be too large, if it is too near 1 then also something's up: a) errors may have been overestimated b) the data are correlated and correlations were ignored in the fit. In principle it may be that the distribution you are dealing with is more compact than a gaussian but this is almost never the case.

Postscript: the Chi-by eye rule is that the minimum χ^2 should be roughly equal to the numebr of parameters. Can you justify this statement?

6.2 Confidence region

Rather than presenting the full probability distribution of errors it is useful to present confidence limits or confidence regions: a region in the m-dimensional space (m being the number of parameters), that contain a certain percentage of the total probability distribution. Obviously you want a suitably compact region around the best fit value. It is customary to choose 68.3%, 95.4%, 99.7%... Ellipsoidal regions have connections with the normal (Gaussian) distribution but in general things may be very different... A natural choice for the shape of confidence intervals is given by constant χ^2 boundaries. For the observed data set the value of parameters $\vec{\alpha}_0$ minimize the χ^2, denoted by χ^2_{min}. If we perturb $\vec{\alpha}$ away from $\vec{\alpha}_0$ the χ^2 will increase. From the properties of the χ^2 distribution it is possible to show that there is a well defined relation between confidence intervals, formal standard errors, and $\Delta\chi^2$. We report here the $\Delta\chi^2$ for the conventionals $1, 2$,and $3-\sigma$ as a function of the number of parameters for the joint confidence levels:

p	1	2	3
68.3%	1.00	2.30	3.53
95.4%	2.71	4.61	6.25
99.73%	9.00	11.8	14.2

In general, let's spell out the following prescription. If μ is the number of fitted parameters of which you want to plot the join confidence region and p is the confidence limit desired
a) find the $\Delta\chi^2$ such that the probability of a chi-square variable with μ degrees of freedom being less than $\Delta\chi^2$ is p: For general values of p this is given by Q described above (for the standard 1,2,3$-\sigma$ see table above).

P.S. Frequentists use χ^2 a lot.

7 Likelihoods

One can be more sophysticated than χ^2, if $P(D)$ (D is data) is known. Remember from the Bayes theorem (eq.3) the probability of the data given the model (Hypothesis) is the likelihood. If we set $P(D) = 1$ (after all we got the data) and ignore the prior by maximizing the likelihood we find the most likely Hypothesis, or , often, the most likely parameters of a given model.

Note that we have ignored $P(D)$ and the prior so in general this technique does not give you a goodness of fit and not an absolute probability of the model, only relative probabilities. Frequentists rely on χ^2 analyses where a goodness of fit can be established.

In many cases (thanks to the central limit theorem) the likelihood can be well approximated by a multi-variate Gaussian:

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2}|detC|^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}(D-y)_i C_{ij}^{-1}(D-y)_j\right] \tag{19}$$

where $C_{ij} = \langle (D_i - y_i)(D_j - y_j)\rangle$ is the covariance matrix.

Exercise: when are likelihood analyses and χ^2 analyses the same?

7.1 Confidence levels for likelihood

For Bayesian statistics, confidence regions are found as regions R in *model space* such that $\int_R P(\vec{\alpha}|D)d\vec{\alpha}$ is ,say, 0.68 for 68% confidence level and 0.95 for 95% confidence. Note that this encloses the prior information. To report results independently of the prior the likelihood ratio is used. In this case compare the likelihood at a particular point in model space $\mathcal{L}(\vec{\alpha})$ with the value of the maximum likelihood $\mathcal{L}_{max}$. Then a model is said acceptable if

$$-2\ln\left[\frac{\mathcal{L}(\vec{\alpha})}{\mathcal{L}_{max}}\right] \leq \text{threshold} \tag{20}$$

Then the threshold should be calibrated by calculating the distribution of the likelihood ratio in the case where a particular model is the true model. There are some cases however when the value of the threshold is the corresponding confidence limit for a χ^2 with m degrees of freedom, for m the number of parameters. (The data must have Gaussian errors, the model must depend linearly on the parameters, the gradients of the model wrt the parameters are not degenerate, the parameters do not affect the covariance).

7.2 Marginalization, combining different experiments

Of all the model parameters α_i some of them may be uninteresting. Typical examples of nuisance parameters are calibration factors, galaxy bias parameter etc, but also it may be that we are interested on constraints on only one cosmological parameter at the time rather

than on the *joint* constraints on 2 or more parameters simultaneously. One then marginalizes over the uninteresting parameters by integrating the posterior distribution:

$$P(\alpha_1..\alpha_j|D) = \int d\alpha_{j+1}, ...\alpha_m P(\vec{\alpha}|D) \tag{21}$$

If there are in total m parameters and we are interested in j of them ($j < m$).

Note that if you have two independent experiments the combined likelihood of the two experiments is just the product of the two likelihoods. (of course if the two experiments are non independent then one would have to include their covariance). In many cases one of the two experiments can be used as a prior. A word of caution is on order here. We can always combine independent experiments by multiplying their likelihoods, and if the experiments are good and sound and the model used is a good and complete description of the data all is well. However it is always important to a) think about the priors one is using and to quantify their effects. b) to make sure that results from independent experiments are consistent: by multiplying likelihood from inconsistent experiments you can always get some sort of results but it does not mean that the result actually makes sense....

Sometimes you may be interested in placing an prior on the uninteresting parameters before marginalization. The prior may come from a previous measurement or from your "belief".

Typical examples of this are: marginalization over calibration uncertainty, over point sources amplitude or over beam errors for CMB studies. It is useful to know of the following trick for Gaussian likelihoods:

$$\begin{aligned} P(\alpha_1..\alpha_{m-1}|D) &= \int \frac{dA}{(2\pi)^{m/2}||C||^{1/2}} e^{\left[-\frac{1}{2}(C_i-(\hat{C}_i+AP_i))\Sigma_{ij}^{-1}(C_j-(\hat{C}_j+AP_j))\right]} \\ &\times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(A-\hat{A})^2}{\sigma^2}\right] \end{aligned} \tag{22}$$

repeated indices are summed over. A the amplitude of , say, a point source constribution ot the C_ℓ angular power spectrum is the $m-th$ parameter which we want to marginalize over with a Gaussian prior with variance σ^2. The trick is to recognize that this integral can be written as:

$$P(\alpha_1..\alpha_{m-1}|D) = C_0 \exp\left[-\frac{1}{2}C_1 - 2C_2A + C_3A^2\right] dA \tag{23}$$

(where $C_{0...3}$ denote constants) and that this kind of integral is evaluated by using the substitution $A \longrightarrow A - C_2/C_3$ giving something $\propto \exp[-1/2(C_1 - C_2^2/C_3)]$.

It is left as an exercise to write the constants explicitly.

7.3 An example

Let's say you want to constrain cosmology by studying clusters number counts as a function of redshift. The observation of a discrete number N of clusters is a Poisson process, the probability of which is given by the product

$$P = \Pi_{i=1}^{N}[e_i^{n_i} \exp(-e_i)/n_i!] \tag{24}$$

where n_i is the number of clusters observed in the $i-th$ experimental bin and e_i is the expected number in that bin in a given model:$e_i = I(x)\delta x_i$ with i being the proportional to the probability distribution. Here δx_i can represent an interval in clusters mass and/or redshift. Note: this is a product of Poisson distributions, thus one is assuming that these are independent processes. Clusters may be clustered, so when can this be used?

For unbinned data (or for small bins so that bins have only 0 and 1 counts) we define the quantity:

$$C \equiv -2 \ln P = 2(E - \sum_{i=1}^{N} \ln I_i) \tag{25}$$

where E is the total expected number of clusters in a given model. The quantity ΔC between two models with different parameters has a χ^2 distribution! (so all that was said in the χ^2 section applies, even though we started from a highly non-Gaussian distribution.

(This is from the paper of Cash 1979)

Thanks

I am indebited to: Andy Taylor crash course on statistics at ROE in 1997, Ned Wright Journal club on statistics, and "Numerical Recipes" book.