Statistical Methods
Lecture notes

Luca Amendola
University of Heidelberg
l.amendola@thphys.uni-heidelberg.de
11/2017
http://www.thphys.uni-heidelberg.de/~amendola/teaching.html
v. 2.8

May 7, 2018

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Contents

Chapter 1

Random variables

1.1 Notation

Wherever not otherwise specified, we will use the following notation:

- data (or observations, measurements etc) will be denoted by x_i or d_i (or n_i for discrete values), where the Latin index i goes from 1 to the number of data points
- random variables will be denoted as x, y etc
- parameters will be denoted as θ_α, where the Greek index runs over the number of parameters (unless of course we use a more specific name as μ, σ etc)
- estimators of parameters, seen as random variables, will be denoted as $\hat{\theta}_\alpha$. If they are obtained as maximum likelihood estimators we can use a ML superscript. Specific values for the estimators (i.e., sample estimates) will be denoted as $\bar{\theta}_\alpha$ or similar notation like for instance the sample mean $\bar{x}$ (or $\bar{n}$ if the data are discrete values). Sometimes however we do not need to distinguish between estimators and estimates and use a bar for estimators as well.
- expected values will be denoted as $E[x]$ or $\langle x \rangle$; we use also $Var[x] \equiv E[(x - E[x])^2]$.
- we always assume the data can be ordered on a discrete or continuous line or space, i.e. they are numbers, and not categories. Categorical data are values that have no obvious numerical ordering and therefore can be arbitrarily ordered. For instance, collecting in a table the US states where people was born gives a list of states that can be represented by an arbitrary number sequence, eg ordering the states alphabetically or by size or population etc.

1.2 Some intuitive concepts

Let us consider a group of $N = 120$ people and let's record how many of these people were born in a given month. Suppose we obtain:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
12	12	11	11	7	13	5	9	7	10	8	15

The number of people in every i-th month, n_i, is a random variable, i.e. a variable that can assume random (unpredictable) values in a certain range. The theory of probability describes quantitatively the behavior of random variables.

The first intuitive concept to quantify the behavior of n_i is to define an average:

$$\bar{n} \equiv \frac{\sum_i^s n_i}{s} \tag{1.1}$$

where $s = 12$ is the number of months in the year. If we don't have any reason to suspect that some month is somehow special (we neglect here the slight difference in month lengths), we expect that every month contains N/s

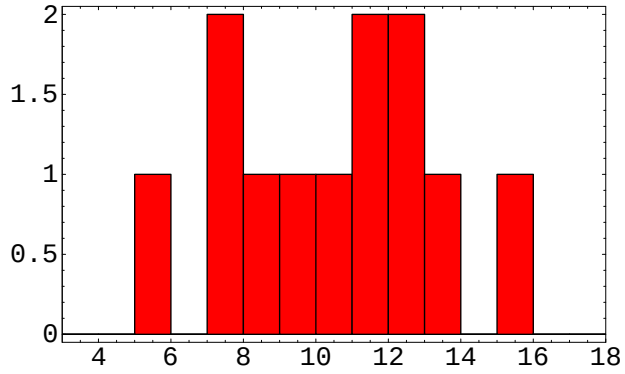

Figure 1.1: Experiment histogram.

people and since $N = \sum n_i$ we can define the average as above. The average itself may or may not be a random variable. In the present case, $\bar{n} = 10$ is not a random variable, since we fixed the total number of persons in our experiment. In most cases, eg in physical measurements, the average of a sample of measures is itself a random variable.

We ask now ourselves how to define the deviation from the mean, i.e. how to quantify the fact that generally $n_i \neq \bar{n}$. Perhaps we are interested in discovering anomalies in the distribution. If we are now not interested in the sign of the deviation but only on its amplitude we could use something proportional to $|n_i - \bar{n}|$. We could define then something like

$$\frac{\sum |n_i - \bar{n}|}{s} \tag{1.2}$$

However, for reasons that will be clear later one, usually we define a related but different quantity:

$$\bar{\sigma} \equiv \sqrt{\frac{\sum_i^s (n_i - \bar{n})^2}{s}} \tag{1.3}$$

where σ is called the root mean square or standard deviation (or more exactly an estimate of the standard deviation). Note that $\bar{\sigma}^2$ is itself an average:

$$\bar{\sigma}^2 \equiv \frac{\sum_i^s (n_i - \bar{n})}{s} \tag{1.4}$$

In our present example we have then

$$\bar{n} = 10 \quad \bar{\sigma} = 3.41 \tag{1.5}$$

We expect then that most data n_i will not deviate more than a few $\bar{\sigma}$ from average and this is in fact what we observe. Then $\bar{n}$ and $\bar{\sigma}$ give us important information on the behavior of n_i; since this experiment can be repeated many time, these averages describe fundamental properties of the data distribution. It is intuitively clear that these averages will be more and more precise when we average over more and more experiments. These ideas will be more precisely defined in the following.

Every function of n_i is itself a random variable (although it could be a trivial random variable with probability 1). For instance, the number of people born in the first M months; the number of months with more than P people and so on.

1.3 Probability and frequency

Let us build now a data histogram, ie bars with height equal to the number p_i of months with n_j people, as in Fig. 1.1.

The bar height gives an estimate of the probability of obtaining that number n_i of people. If we repeat the experiment N_{exp} times the estimate will be more precise, in the sense that it will converge to a limiting case. For instance, in Figs 1.2,1.3 we have the cases $N_{exp} = 10$ and 50.

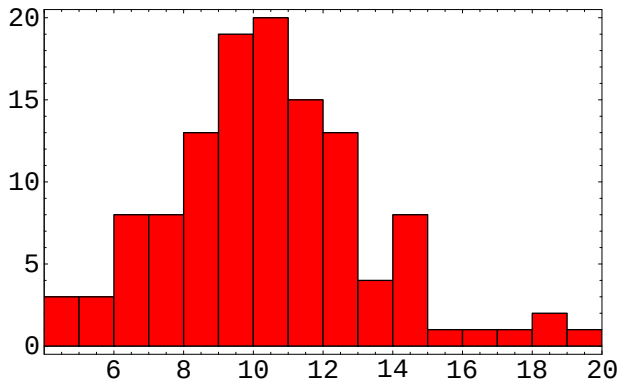

Figure 1.2: Repeating the experiment 10 times...

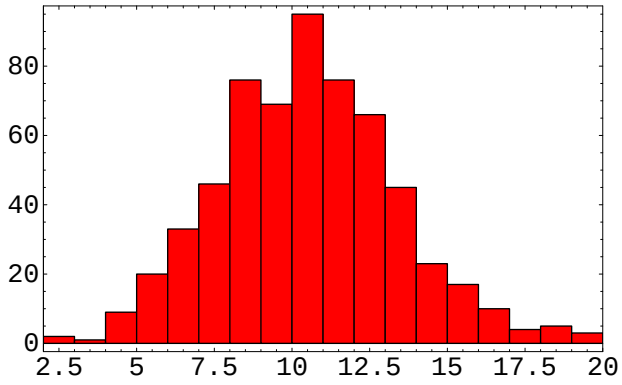

Figure 1.3: ...and 50 times.

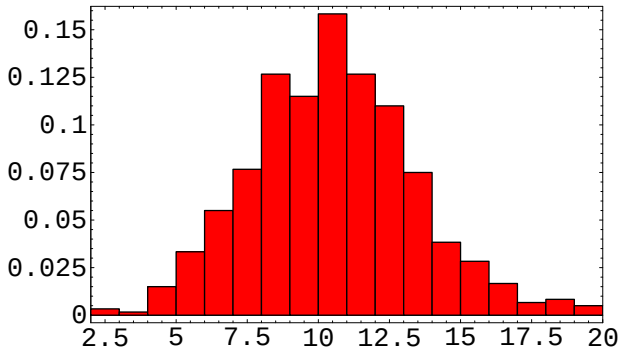

Figure 1.4: Same histogram, but normalized to unity.

For $N_{exp} \to \infty$ we can assume we have the "true probability" of having n_i. We *define* then

$$P(n_i) \equiv \lim_{N_{exp}->\infty} \frac{\text{number of occurrences of } n}{sN_{exp}} \tag{1.6}$$

that is, the frequency of events. Now since $\sum$number of occurrences of $n = sN_{exp}$, we have

$$\sum_{i}^{N_{tot}} P(n_i) = 1 \tag{1.7}$$

The sum of all possible probabilities of any given experiments equals 1. Fig. 1.3 becomes then as in Fig. 1.4. The histogram approximates the probability distribution of n_i.

In the limit in which the random variable n_i becomes a continuous variable (eg a temperature, a magnitude etc), we define a probability density or *probability distribution function* (PDF) $f(x)$

$$f(x)dx = P(x) \tag{1.8}$$

and we have, within the domain of x(i.e. all its possible values)

$$\int f(x)dx = 1 \tag{1.9}$$

In the same experiment we can identify other random variables and therefore other PDFs. For instance, if we ask the birth month we will get an answer in the range $m_i \in 1-12$ which itself is a random variable. Then we can create an histogram as in Fig. 1.5 in which we plot the month frequency n_i/N_{tot} versus the months. If we increase N_{tot} to 1200 we obtain Fig. 1.6. Here clearly the distribution tends to a *uniform distribution* with $P(m_i) = 1/12$.

1.4 Properties of the PDFs

The two most fundamental properties of probability distributions are

$$\int f(x)dx = 1 \tag{1.10}$$

$$f(x) \geq 0 \tag{1.11}$$

We can easily extend the idea to joint events, for instance the probability of obtaining at the same time (non necessarily in the chronological sense) the measurement x in dx (eg a galaxy magnitude) and y in dy (eg the galaxy redshift). Then we have

$$f(x,y)dxdy = P(x,y) \tag{1.12}$$

$$f(x,y) \geq 0 \tag{1.13}$$

$$\int f(x,y)dxdy = 1 \tag{1.14}$$

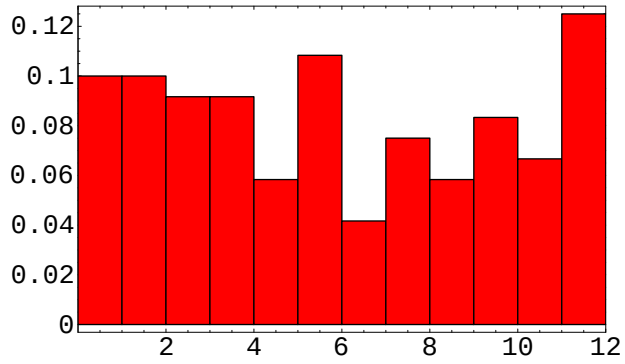

Figure 1.5: Histogram for the frequency of months for $N = 120$...

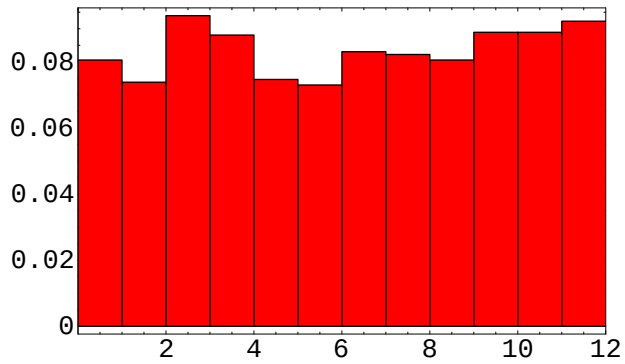

Figure 1.6: ...and for $N = 1200$.

Immediate consequence of the first law is that if $F(<X) = \int_{-\infty}^{X} f(x)dx$ is the probability of obtaining a result less than X , then the probability of obtaining a result greater than or equal to X is $F(\geq X) = 1 - F(<X)$. So in general if $P(A)$ is the probability of A, the probability of non-A (ie anything but A) is simply $1 - P(A)$, to be denoted as $P(\bar{A})$.

Other examples of probability:

$$P(x) = \lim_{N\to\infty} \frac{\text{number of people voting for party X}}{\text{number of interviewed}} \tag{1.15}$$

$$P(x) = \lim_{N\to\infty} \frac{\text{number of measures (distances, temp., etc) that give X}}{\text{number of experiments}} \tag{1.16}$$

Clearly if x is a continuous variable we have

$$f(x)dx = \lim_{N\to\infty} \frac{\text{number of measures in x,x+dx}}{\text{number of experiments}} \tag{1.17}$$

1.5 Probability of independent events

Suppose we throw two dice; the joint probability of obtaining 1 in a throw and 2 in the other one is the product of the single-throw probabilities, $P_{12} = P_1 P_2$. This is true only because the two throws are independent, that is, do not influence each other. Then we have $P_1 = P_2 = 1/6$ and $P_{12} = 1/36$, as of course one could see by the number of occurrences over the number of experiments. If we have the PDF $f(x, y)$, the event x can be said to be independent of event y if the probability of x, $p(x)dx$, does not depend on y for any range of y. This requires

$$P(X, \Delta X)P(Y, \Delta Y) = \int_{Y,Y+\Delta Y} dy \int_{X,X+\Delta X} dx f(x, y) \tag{1.18}$$

for any range $\Delta X, \Delta Y$. But this can be true if and only if $f(x, y)$ is separable, i.e. $f(x, y) = f_1(x)f_2(y)$ so that

$$P(X, \Delta X)P(Y, \Delta Y) = \int_{Y,Y+\Delta Y} dy \int_{X,X+\Delta X} dx f(x, y) = \int_{X,X+\Delta X} dx f_1(x) \int_{Y,Y+\Delta Y} dy f_2(y) \tag{1.19}$$

So, two events are independent if and only if the joint PDF $f(x, y)$ is separable. This extends obviously to N events.

1.6 Problem: the birthday "paradox"

We can now use several of these concepts in an exercise. Let us estimate the probability that in N random people there are at least two with the same birthday.

A person B has the same birthday of person A only once in 365. Then $P(coinc., N = 2) = 1/365$ and the probability of non-coincidence is $P(non - coinc., N = 2) = 1 - 1/365 = 364/365$.

Let's add a third person. His/her birthday will not coincide with the other two 363 times over 365. The joint probability that the 3 birthdays do *not* coincide is then

$$P(non - coinc., N = 3) = \frac{364}{365}\frac{363}{365} \tag{1.20}$$

It is clear then that for N persons we have

$$P(\text{non} - \text{coinc.}, N) = \frac{365}{365}\frac{364}{365}\frac{363}{365}...\frac{365 - N + 1}{365} \tag{1.21}$$

We can now use

$$e^{-x} \approx 1 - x \tag{1.22}$$

to write

$$\frac{365-N+1}{365} = 1 - \frac{N-1}{365} \approx e^{-(N-1)/365}$$

and therefore

$$P(\text{non}-\text{coinc}, N) = e^{-1/365}e^{-2/365}e^{-3/365}...e^{-(N-1)/365} = e^{-\frac{N(N-1)}{2}\frac{1}{365}} \tag{1.23}$$

Finally, the probability of having at least one coincidence must be the complement to unity to this, i.e.

$$P(\text{coinc}, N) = 1 - e^{-\frac{N(N-1)}{2}\frac{1}{365}} \approx 1 - e^{-\frac{N^2}{730}} \tag{1.24}$$

For $N = 20$ one has, perhaps surprisingly (this is the "paradox") $P(N) = 0.5$ i.e. almost 50%.

1.7 Joint, disjoint, conditional probability

When we combine several events, we can define three kinds of probabilities according to how we want to combine the events.

Joint P. If P_A and P_B are the probabilities of the independent events A and B, the probability of having both A *and* B is $P_A P_B$, as we have seen already. Then ($\cap$=AND; $\cup$=OR)

$$P(A \cap B) = P(B \cap A) = P(A)P(B) \tag{1.25}$$

For instance, the prob. of having 1 in a dice throw and 2 in another one is $(1/6)^2 = 1/36$. If the events are not independent, we cannot write $P(A \cap B) = P(A)P(B)$ but we can still write

$$P(A \cap B) = P(B \cap A) \tag{1.26}$$

Disjoint P.. If P_A and P_B are the prob. of events A and B (not necessarily independent) that are mutually exclusive (i.e. $A\texttt{AND}B = A \cap B = 0$), the prob. of A *or* B is $P_A + P_B$. Therefore

$$P(A \cup B) = P(A) + P(B) \tag{1.27}$$

We have already seen an example of disjoint prob. when we have seen that $P(A) = 1 - P(\overline{A})$. Since A and $\overline{A}$ are mutually exclusive, we can write

$$P(A \cup \overline{A}) = 1 = P(A) + P(\overline{A}) \tag{1.28}$$

So for instance the prob. of having 1 or 2 in a dice roll is 1/6+1/6=1/3. Considering continuous variables we have

$$p(x \in A \cup B) = \int_A f(x)dx + \int_B f(x)dx \tag{1.29}$$

only if the ranges A and B do not overlap. If they overlap, the events are not mutually exclusive (there will be some event that is both A and B so $A \cap B \neq 0$), and we have:

$$p(x \in A \cup B) = \int_A f(x)dx + \int_B f(x)dx - \int_{A\cap B} f(x)dx \tag{1.30}$$

In general therefore

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.31}$$

So for instance if A is the prob. of having one "1" in the first die, whatever the second is, and B the prob. of "1" in the second die, whatever the first is, and we consider the prob. of having at least a "1" in two throws, the event "11" is both A and B. So we have $P(A \cup B) = 1/6 + 1/6 - 1/36 = 11/36$, as we can verify easily since the winning combinations are (11,12,13,14,15,16,21,31,41,51,61) are 11 over 36.

Let's make another example. Suppose we have a number of galaxies and we measure their redshift and their apparent magnitude. I win a bet if the next galaxy has magnitude higher than (i.e. dimmer than) 24 *or* if its redshift is higher than 2. Based on previous data, I know that 10% of galaxies in my survey have that high magnitude and 20% that high redshift. What is the probability I win my bet? If I am sure that no high redshift galaxy in my sample has that high magnitude, then my chances are 10+15=25%. But if, on the contrary, I discover that all the

galaxies with that high magnitudes are indeed high redshift ones, then of course those 15 high-redshift galaxies out of 100 already includes the 10 out of 100 that are very dim, so my chances are just 15%, that is 10+15-10=15%. Note that we did not assume that magnitude and redshift are independent; in fact, in the second case they are highly correlated.

Since the probability that y is in separated intervals dy_i, dy_2 etc is the sum of disjoint elements, the probability that x is in dx and y is in dy_1 or dy_2 or dy_3 etc, is

$$p(x)dx = dx \sum_i f(x,y)dy_i = dx \int f(x,y)dy \tag{1.32}$$

that is, $p(x) = \int f(x,y)dy$. So we identify the probability density of having x in a range dx regardless of y as

$$p(x) = \int f(x,y)dy \tag{1.33}$$

This operation is often called *marginalization*, and is immediately extended to N variables. If, and only if, x, y are independent, the marginalization is independent of y in any range.

Conditional P.. If the events are not independent, we can define the conditional probability (prob. of A given B):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{number of cases that are both A and B}}{\text{number of cases that are B}} \tag{1.34}$$

So for instance, the probability of the combination 1-2 *after* obtaining 1 in the first roll equals $(1/36)/(1/6) = 1/6$.

This extends obviously to continuous variables. The prob. of x in the range $I = (-1, 1)$ given that $x < 0$ is $P(x \in I | x < 0)$. The prob. of having $x < 0$ is

$$P(x < 0) = \int_{<0} f(x)dx \tag{1.35}$$

and the prob. of having $x \in I$ and at the same time $x < 0$ is

$$P(x < 0, x \in I) = \int_{-1}^{0} f(x)dx$$

Now, the fraction of cases (or area) such that $P(x \in I | x < 0)$ is clearly the fraction $P(x < 0, x \in I)/P(x < 0)$, which agrees with the rule above. In other words, if in 100 measures there are 50 with $x < 0$ and 20 with $-1 < x < 0$ it is clear that the fraction of measures with $x \in I$ among those with $x < 0$ is 20/50=2/5.

Another example. The prob. of obtaining ≥ 9 in two dice rolls is 10/36: there are in fact 10 successful events: $36, 45, 46, 55, 56, 66, 63, 54, 64, 65$ in 36 possibilities. Which is the prob. of obtaining a score ≥ 9 given that in the first roll the result is 6 ? We have

$$P(x + y \geq 9 | x = 6) = P(x + y \geq 9, x = 6)/P(x = 6) = \frac{4}{36}\frac{6}{1} = \frac{2}{3} \tag{1.36}$$

which indeed is true since if the first die has a 6, then it is sufficient that the second result is 3,4,5,6 to win, i.e. 4 cases out of 6.

Yet another example. Suppose we know that on average one person out of 1000 randomly chosen persons is a physics student and plays piano; suppose we also know that in general one out of 100 people plays piano; then the probability that a person, among those that play piano, is also a physics student is $(1/1000)/(1/100) = 1/10$. In other words, out of 1000 people, we know that 10 (on average) play piano; we also know that among those 1000 people there is one that both plays piano and study physics and of course this person has to be among the 10 that play piano. Then this person is indeed 10% of those that play piano.

Consequently, the fraction of people that play piano and study physics , $P(A \cap B)$, is equal to the fraction that play piano, $P(B)$, times the fraction of people that study physics among those that play piano, $P(A|B)$.

1.8 Bayes' Theorem

Now, since, $A \cap B = B \cap A$ we have Bayes' Theorem

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{1.37}$$

or

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.38}$$

Note that if A and B are independent, ie if

$$P(A \cap B) = P(A)P(B) \tag{1.39}$$

it follows that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. For instance, if the fraction of people that study physics is $1/100$ and the fraction that play piano is $1/100$, if we find out that the fraction of people that both study physics and play piano is not $1/10000$ then we can conclude that playing piano and studying physics are *not* independent events. The prob. $P(A)$ in this case is called *prior probability*; the prob. $P(A|B)$ is called *posterior probability.*

The prob. that $B \cup \overline{B}$ occurs is of course always 1, even in the case of conditional prob. We have therefore

$$P(B \cup \overline{B}|A) = 1 = \frac{P(A, B \cup \overline{B})}{P(A)} \tag{1.40}$$

or

$$P(A, B \cup \overline{B}) = P(A) \tag{1.41}$$

In terms of PDF this rule says that integrating a PDF of two variables over the whole domain of one of the two: this is again the *marginalization.*

Bayes' theorem applies immediately to PDFs. Considering the probabilities in infinitesimal elements dx, dy we have

$$P(x|y)dx = \frac{L(y|x)dyp(x)dx}{E(y)dy} \tag{1.42}$$

that is

$$P(x|y) = \frac{L(y|x)p(x)}{E(y)} \tag{1.43}$$

(notice again that the conditional probability $P(x|y)$ gives the probability of x being in dx, while $P(y|x)$ gives the probability of $y$being in dy). Beside the prior p and the posterior P, we denote the denominator $E(y)$ as *evidence*, and the conditional probability $L(y|x)$ as *likelihood*.

Problem.

1% of people has the tropical disease Z. There exist a test that gives positive 80% of the times if the disease is present (true positive), but also 10% of the times when the disease is absent (false positive). If a person tests positive, which is the prob. that he/she has the Z disease?

Answer. We have:

prob. of having Z: $P(Z) = 1\%$

cond. prob. of being positive (event labeled p) having Z : $P(p|Z) = 80\%$

cond. prob. of being positive while not having Z: $P(p|\overline{Z}) = 10\%$.

From the first we deduce that $P(\overline{Z}) = 99\%$. We need now the prob. of having Z being positive; from Bayes' theorem we have:

$$P(Z|p) = \frac{P(p|Z)P(Z)}{P(p)} \tag{1.44}$$

We should then evaluate $P(p)$. The prob. of testing positive and having Z is:

$$P(p, Z) = P(p|Z)P(Z) = 0.8 \cdot 0.01 = 0.008 \tag{1.45}$$

The prob. of testing positive being healthy is instead:

$$P(p, \overline{Z}) = P(p|\overline{Z})P(\overline{Z}) = 0.1 \cdot 0.99 \approx 0.1 \tag{1.46}$$

Moreover

$$P(p) \quad = \quad P(p,Z) + P(p,\overline{Z}) = P(p|Z)P(Z) + P(p|\overline{Z})P(\overline{Z}) = 0.108$$

It follows finally

$$P(Z|p) = \frac{P(p|Z)P(Z)}{P(p)} = \frac{0.8 \cdot 0.01}{0.108} = 0.075 \tag{1.47}$$

ie there is only a prob. of 7.5% of having Z. The reason of this perhaps surprising result is that $P(Z)$ (the absolute prob. of being infected with Z) is much smaller than $P(p)$, the absolute prob. of testing positive.

Exchanging the conditional probabilities is a very common logical error: the probability that one is a great artist because he/she is "misunderstood" (ie, nobody likes his/her paintings) is not equal to the probability of being misunderstood being a great artist. In fact, the probability of being great artists is much less than the probability of making bad paintings.

Exercise: what is the probability of throwing two dice and obtain a score between 1 and 3 in the first roll and 6 in the second? what is the probability of having an overall score less than 8 in the launch of two dice? And if we had already obtained 5 in the first die?

Chapter 2

Probability distributions

In this Chapter we review the main properties of the probability distributions (expected values, moments etc) and present the most one-dimensional and multi-dimensional important PDFs.

2.1 Expected values

Let's briefly introduce two examples of PDFs.

Uniform distribution.

$f(x) = const.$ in the range $x \in (a, b)$. We have

$$\int_a^b f(x)dx = cost. \times (b-a) \tag{2.1}$$

and the normalization requires $const = 1/(b-a)$.

Gauss distribution.

$$f(x) = Ae^{-\frac{(x-x_0)^2}{2\sigma^2}} \tag{2.2}$$

Normalization

$$\int f(x)dx = A\int_{-\infty}^{+\infty} \exp(-\frac{(x-x_0)^2}{2\sigma^2})dx = A\sqrt{2\pi\sigma^2} \tag{2.3}$$

from which $A = (2\pi\sigma^2)^{-1/2}$.

Indeed:

$$\begin{aligned}
\int e^{-x^2/2}dx &= \sqrt{(\int e^{-x^2/2}dx)(\int e^{-y^2/2}dy)} \\
&= \sqrt{\int e^{-\frac{(x^2+y^2)}{2}}dxdy} \\
&= \sqrt{\int e^{-r^2/2}rdr\int_{-\pi}^{+\pi}d\theta} \\
&= \sqrt{2\pi\int_0^{\infty}e^{-z}dz} = \sqrt{2\pi(-e^{-\infty}+e^0)} = \sqrt{2\pi}
\end{aligned}$$

Finally since

$$\int e^{-x^2/2\sigma^2}dx = \sigma\int e^{-z^2/2}dz = \sigma\sqrt{2\pi}$$

we obtain the result. The parameters x_0 and σ^2 are called mean and variance.

PDFs can be synthetically characterized by several quantities.

Quantile α:

value of x such that

$$\int_{-\infty}^{x} f(x')dx' = \alpha \tag{2.4}$$

($0 \leq \alpha \leq 1$). If $\alpha = 0.5$ the quantile is called median.

Mode.

The value of x such that $P(x)$ is maximal.

Moments or expected values.

The expected value of a quantity $g(x)$ is defined as

$$E[g] =< g >\equiv \int g(x)f(x)dx \tag{2.5}$$

The mean is therefore the expectation value of x :

$$E[x] = \int xf(x)dx \tag{2.6}$$

For discrete variables we have

$$E[n] = \sum_{1}^{N} n_i P(n_i) \tag{2.7}$$

Since $P(n_i)$ is defined as the number of events n_i divided by the total number of cases, we retrieve the intuitive definition of mean of a variable as the sum of all the values divided by the number of cases.

The variance (or central moment of second order) is defined as

$$E[(x - \langle x \rangle)^2] = \int (x - \langle x \rangle)^2 f(x)dx = \int x^2 f(x)dx - \hat{x}^2 \tag{2.8}$$

(sometimes also $Var(x)$). For a Gaussian one has

$$E[x] = x_0 \tag{2.9}$$

$$E[(x - \langle x \rangle)^2] = \sigma^2 \tag{2.10}$$

Note that $E[x - \langle x \rangle] = 0$ and $E[y^2] \neq E[y]^2$. For a uniform variable, one has

$$E[x] = \frac{b+a}{2} \tag{2.11}$$

$$E[(x - \langle x \rangle)^2] = \frac{(b-a)^2}{12} \tag{2.12}$$

The variance has great importance in scientific measures. Conventionally in fact the error associated to each measure is given by the square root of the variance, or standard deviation, and is denoted generally with σ also for non-Gaussian distributions. Note that $E^{1/2}[(x - \langle x \rangle)^2]$ coincides indeed with the definition (eq. 1.4) in the limit of an infinite number of observations.

The n-th order moment is

$$E[x^n] = \int x^n f(x)dx \tag{2.13}$$

$$E[(x - \langle x \rangle)^n] = \int (x - \langle x \rangle)^n f(x)dx \tag{2.14}$$

Exercises:

Evaluate $E[x]$ and $E[(x - \langle x \rangle)^2]$ for a uniform distribution in the range $(a - b)$ and for a Gaussian. Invent a PDF, normalize it, and evaluate mean and variance.

Prove that

$$E[ax] = aE[x] \tag{2.15}$$
$$E[x+a] = E[x] + a \tag{2.16}$$

that is, the mean is a linear operation.

More in general

$$\langle f(x) + g(x) \rangle = \langle f(x) \rangle + \langle g(x) \rangle \tag{2.17}$$

2.2 Population and sampling

Suppose in an experiment we find the results $x_1, x_2,$...etc. Expected values can be seen as averages over an infinite number of trials:

$$E[x] = \lim_{N\to\infty} \frac{\sum_i x_i}{N} \tag{2.18}$$

Let us show this for a variable x that can assume a number of discrete values $x_\alpha = x_{1,2,3,...}$ with probability $P_\alpha = P_{1,2,3,...}$. Notice that here $x_{1,2,3..}$ represent the possible values of x_α, while in Eq. (2.18) they represented the individual occurrences of the measurements: one can have several times the same outcome $x_\alpha = 3$; to distinguish these two interpretations of the subscript, I use now Greek letters. In the limit of infinite trials, we should obtain by definition a fraction P_α of values x_α, so a number NP_α of them (for instance, NP_1 =5 times the value x_1, NP_2 =10 times the value x_2 and so on), and the average will be

$$\frac{\sum_\alpha x_\alpha N P_\alpha}{N} = \sum_\alpha x_\alpha P_\alpha \tag{2.19}$$

which is exactly Eq. (2.7), i.e. $E[x]$. So the expected value of a function $g(x)$ is the average of that quantity taken over an infinite number of trials.

What we normally deal with, however, is a finite number of data, eg measurements. We can construct averages like

$$\hat{x} = \frac{\sum x_i}{N} \tag{2.20}$$
$$s^2 = \frac{\sum (x-\hat{x})^2}{N} \tag{2.21}$$

for finite Ns, that will be called *sample* mean and *sample* variance. Only in the limit $N \to \infty$ will these quantities become identical with the expected values of x and of $(x - E[x])^2$. So the sample quantities are random variables, since if I take a new set I will get in general a different value, while the expected values are parameters, not random variables, that are supposed to measure an intrinsic properties of the underlying "population" (as often called in sociology) of measurements.

As will soon show, we should think of the sample quantities as *estimators* of the distribution parameters: for instance, the sample average of x should be seen as an estimator of the espected value of x. The same parameter can have several estimators but clearly an estimator, in order to be a good one, should be a good approximation, that is, obey a number of properties we will discus later on.

2.3 Transformation of variables

Given a random variable x and its PDF $f(x)$, we could be interested to derive a PDF of a variable function of x, for instance x^2 or $1/x$ or $y(x)$. For instance, if we know the PDF of the absolute magnitude M of a galaxy , we could be interested in the PDF of its distance r

$$M = m - 25 - 5\log_{10} r \tag{2.22}$$

assuming we know the apparent magnitude m. If $dy = y'dx$ is the infinitesimal interval of the new variable y as a function of the old one, it is clear that the prob. of having x in $x, x+dx$ must be equal to the one of having y in $y, y+dy$ (we assume a monotonic relation $y(x)$)

$$f(x)dx = g(y)dy \tag{2.23}$$

and therefore the new PDF $g(y)$ is

$$g(y) = f(x)|\frac{dx}{dy}| \tag{2.24}$$

where the absolute value ensures the positivity of the new PDF. So if the PDF of M is a Gaussian, the PDF of $r(M)$ is

$$\begin{aligned} f(r) &= A|\frac{dM}{dr}| \exp -\frac{(M(r)-M_0)^2}{2\sigma^2} = A(\frac{5\log_e 10}{r}) \exp -\frac{(m-25-5\log r - M_0)^2}{2\sigma^2} \\ &= \frac{A'}{r} \exp -\frac{(m-25-5\log r - M_0)^2}{2\sigma^2} \end{aligned}$$

and defining r_0 such that $M_0 = m - 25 - 5\log r_0$ one can write

$$f(r) = \frac{A'}{r} \exp -\frac{(5\log r/r_0)^2}{2\sigma^2} \tag{2.25}$$

called a log-normal distribution. Notice that in general

$$E[g(x)] \neq g(E[x]) \tag{2.26}$$

We can also consider the transformation of variables in the case of many random variables. The transformation from $x_1, x_2, ...$ to $y_1, y_2, ..$ can be performed introducing the Jacobian of the transformation

$$f(x_i)d^n x = g(y_i)d^n y \tag{2.27}$$

from which

$$g(y_i) = f(x_i)|J| \tag{2.28}$$

where $J_{ij} = \partial x_i/\partial y_j$ and $||$ denotes the determinant.

Exercise.

If the variable x is distributed in a uniform manner in (a, b) (both > 0), which is the distribution of $y = x^2$?

2.4 Error propagation

We can now use these formulae to find the error (standard deviation) associated to a function of a random variable x in the limit of small deviations from the mean.

Suppose we have a variable x, eg the side of a square, distributed as $f(x)$ with mean μ and variance σ_x^2 and we are interested in the PDF of the area $y = x^2$. We can expand y around μ:

$$y(x) = y(\mu) + \frac{dy}{dx}|_\mu (x-\mu) \tag{2.29}$$

We can then evaluate the mean and variance in the limit of small deviations from μ:

$$E[y] = \int y(\mu)f(x)dx + \int \frac{dy}{dx}|_\mu (x-\mu)f(x)dx = \int y(\mu)f(x)dx + \frac{dy}{dx}|_\mu \int (x-\mu)f(x)dx = y(\mu)\int f(x)dx = y(\mu) \tag{2.30}$$

(because $\int (x-\mu)f(x)dx = \int xf(x)x - \mu\int f(x)dx = \mu - \mu = 0$) and

$$\begin{aligned} E[y^2] &= \int [y^2(\mu) + y'(\mu)^2(x-\mu)^2 + 2y(\mu)y'(\mu)(x-\mu)]f(x)dx & (2.31) \\ &= y^2(\mu) + y'^2\sigma_x^2 & (2.32) \end{aligned}$$

(where $y' \equiv \frac{dy}{dx}|_\mu$). It follows that the variance of y for small deviations of x from μ is

$$\sigma_y^2 = E[(y - y(\mu))^2] = E[y^2] - y(\mu)^2 = y'^2\sigma_x^2 \tag{2.33}$$

In the case of area $y = x^2$ we have then $\sigma_y^2 = 4\mu^2\sigma_x^2$. This is the fundamental rule of error propagation.

We can easily extend this rule to several variables. Suppose for instance that $y(x_1, x_2)$ depends on two variables, for instance y is the sum of the sides of two squares measured independently. Because of independence $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Then we have

$$y(x_1, x_2) = y(\mu_1, \mu_2) + \sum_i \frac{\partial y}{\partial x}|_\mu (x_i - \mu_i) \tag{2.34}$$

from which

$$\begin{aligned} E[y^2] &= \int [y^2(\mu_1, \mu_2) + (\sum_i y_i'(\mu)(x_i - \mu))^2 + 2y(\mu_1, \mu_2) \sum_i y_i'(\mu)(x_i - \mu)] f_1(x_1) f_2(x_2) dx_1 dx_2 \\ &= y^2(\mu_1, \mu_2) + \sum_i y'^2 \sigma_{x_i}^2 \end{aligned}$$

(the first step depends on the assumption of independency) and finally

$$\sigma_y^2 = E[(y - y(\mu_1, \mu_2))^2] = \sum_i y_i'^2 \sigma_{x_i}^2 \tag{2.35}$$

This rules extends obviously to any number of independent variables.

2.5 Sum and products of variables. Variance of the sample mean.

In the case $y = x_1 + x_2 + .. + x_n$ the above rule gives

$$\sigma_y^2 = \sum_i \sigma_{x_i}^2 \tag{2.36}$$

i.e., the variance of a sum of random variables is the sum of the variances. The error in y is therefore

$$\sigma_y = \sqrt{\sum_i \sigma_{x_i}^2} \tag{2.37}$$

i.e. *the errors add in quadrature*.

In the important case of the sample mean, i.e. assuming a number of data x_i from the same distribution with variance σ^2,

$$\bar{x} = \frac{\sum_i x_i}{N} \tag{2.38}$$

we see immediately that

$$\sigma_{\bar{x}}^2 = \frac{1}{N^2} \sum \sigma^2 = \frac{\sigma^2}{N} \tag{2.39}$$

This is the very important result that states that the variance of the mean is smaller by a factor of $1/N$ than the variance of each individual data point.

E**xercise**: generalize to $y = a_1x_1 + a_2x_2 + ... + a_nx_n$ (not necessarily with the same variance).

In the case of a product, $y = x_1x_2...x_n$ we have instead

$$\frac{\sigma_y^2}{\mu_y^2} = \sum_i \frac{\sigma_{x_i}^2}{\mu_{x,i}^2} \tag{2.40}$$

where $\mu_y = \mu_1\mu_2...\mu_n$ e $\mu_{x,i} \equiv \mu_i$. The quantity

$$\frac{\sigma_x}{\mu_x} \tag{2.41}$$

is the relative error. For a product of variables, then, the square of the relative error is the sum of the squares of the individual relative errors. That is, for a product of variables the *relative errors add in quadrature*.

Exercise: generalize to $y = x_1^{m_1} x_2^{m_2} ... x_n^{m_n}$.

2.6 The main PDFs

2.6.1 Binomial PDF

Let us consider N independent events, eg the scores $1-3-2-6$ etc in a series of dice rolls, or the sequence $TTHHTH$ of heads/tails in coin tosses. We want to evaluate the probability that a joint event, eg 8 heads out of 10 tosses, or three times a 1 out of 4 dice rolls, regardless of the order in the sequence, i.e. considering the events as indistinguishable. This is exactly the same kind of statistics we need in eg the statistics of a gas, which depends on the probability for indistinguishable particles to be in a given region of phase space.

We need first of all to evaluate the number of possible sequences. If we have N different elements, eg a, b, c, we can permute the N elements $N!$ times. For instance, $N = 3$ elements can be combined $3! = 6$ times: $abc, acb, cab, cba, bac, bca$. Then $N!$ is the number of permutations of distinguishable elements.

Suppose now we have only two elements, eg head or tail, or event A and any other event $\bar{A}$. Then many permutations are identical, for instance $HHTTT$ remains the same by exchanging the two Hs and the three Ts. Suppose we have n times one of the two elements and, therefore, $N-n$ the number of the other. Then, among the total $N!$ permutations, a fraction $n!$ is identical because we permute the same identical n element, and a fraction $(N-n)!$ will also be identical for the same reason. How many indistinguishable combinations will we obtain? Clearly

$$\frac{\text{total permutations}}{(\text{permutations among } n)(\text{permutations among } N-n)} = \frac{N!}{n!(N-n)!} \equiv \left(\begin{array}{c} N \\ n \end{array} \right) \tag{2.42}$$

For instance, if $N = 4$ and $n = 2$ (as in $TTHH$) I will have $4!/2!/2! = 6$ equivalent combinations ($HHTT, HTHT$, $TTHH$,$THTH, THHT, HTTH$). Notice that for $n = 0$ we define $n! = 1$.

The binomial PDF generalizes this calculation to the case in which I have a series of n independent events A each with the same probability p (eg for "head" the prob. is $1/2$, for a 2 in a dice roll is $1/6$ etc). In this case, the occurrence of n events A or prob. p out of N implies the occurrence of $N-n$ events $\bar{A}$ with prob. $1-p$. All this implies a joint prob. of

$$p^n(1-p)^{N-n} \tag{2.43}$$

But clearly we have $\left(\begin{array}{c} N \\ n \end{array} \right)$ of such combinations and therefore the binomial prob. will be

$$P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \tag{2.44}$$

where n is the discrete random variable $0 \leq n \leq N$ (number of events A) while N, p are the distribution parameters. Notice that by employing the rules of the binomial we have, as indeed we should have expected:

$$\sum_{n=0}^{N} P(n; N, p) = (p + (1-p))^N = 1 \tag{2.45}$$

It is also intuitive that the mean of events A of prob. (frequency) p out of N events should be the fraction p of N and indeed

$$E[n] = Np \tag{2.46}$$

$$\sigma^2 = E[(n - Np)^2] = Np(1-p) \tag{2.47}$$

Let's demonstrate the first one:

$$E[n] = \sum_n n P(n; N, p) = \sum_{n=0}^{N} \frac{nN!}{n!(N-n)!} p^n (1-p)^{N-n} \tag{2.48}$$

$$= \sum_{n=1}^{N} \frac{N(N-1)!}{(n-1)!(N-n)!} p p^{n-1} (1-p)^{N-n} \tag{2.49}$$

$$= Np \sum_{n'=0}^{N'} \frac{(N')!}{(n')!(N'-n')!} p^{n'} (1-p)^{N'-n'} = Np \tag{2.50}$$

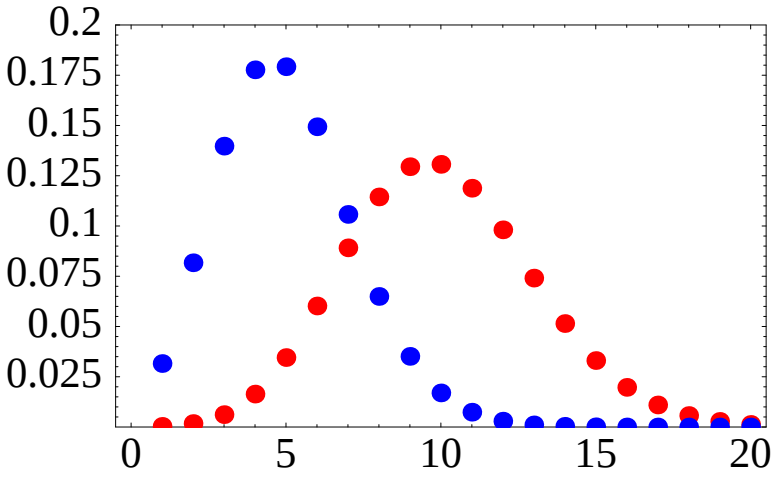

Figure 2.1: Binomial for $N = 120$ and $p = 1/12$ (red dots) e $p = 1/24$ (blu dots).

(where we defined $n' = n-1$ and $N' = N-1$). The binomial distribution for $N = 1$ is called Bernoulli distribution:

$$P(n;1) = p^n(1-p)^{1-n} \tag{2.51}$$

for $n = 0, 1$. It applies for instance to a single toss of a coin and gives the probability that an event, e.g. tail, happens ($n = 1$) with probability p or does not happen ($n = 0$), with probability $1 - p$.

Exercises:

1) Which is the probability of obtaining two heads out of 4 throws ?

The prob. of having exactly $n = 2$ two heads, each with prob. $p = 0.5$, out of $N = 4$ events is

$$P(2;4,0.5) = 3/8 \tag{2.52}$$

2) In the birthday experiment we have obtained 15 persons in December. Which is the prob. of obtaining 15 or more in a given month?

The prob. that the birthday of a person is in December is p =1/12. The total number of events is N =120 and $n = 15$ is the number of events A =December . The statistics is then a Binomial $P(15;120,1/12)$. The prob. of having more than 15 events A is therefore

$$F(>15) = \sum_{n>15}^{120} \frac{N!}{n!(N-n)!} p^n(1-p)^{N-n} = 0.074 \tag{2.53}$$

that is, only 7.4%.

2.6.2 Poissonian PDF

Let us consider now the limit of the Binomial for $N \to \infty$ and $p \to 0$ (rare events), but keeping $Np = \nu$ finite. We can approximate $N!/(N-n)! \approx N^n$ and (since $\lim_{n\to\infty}(1+\frac{a}{n})^n = e^a$)) $(1-p)^{N-n} = (1-\frac{\nu}{N})^{N-n} \approx (1-\frac{\nu}{N})^N \approx e^{-\nu}$ so that

$$P(n;\nu) = \frac{N^n}{n!} p^n e^{-\nu} = \frac{\nu^n}{n!} e^{-\nu} \tag{2.54}$$

and we obtain the Poissonian PDF.

The moments are

$$E[n] = e^{-\nu} \sum_0^\infty n \frac{\nu^n}{n!} = \nu e^{-\nu} \sum_1^\infty \frac{\nu^{n-1}}{n-1!} = \nu e^{-\nu} \sum_0^\infty \frac{\nu^{n'}}{n'!} = \nu \tag{2.55}$$

$$E[(n-\nu)^2] = \nu \tag{2.56}$$

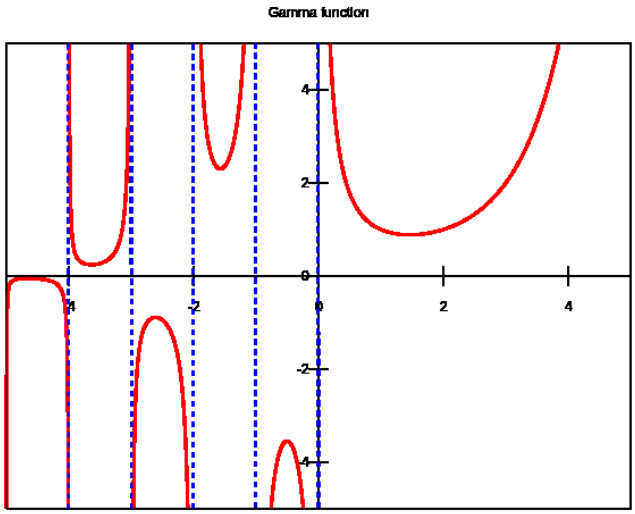

Figure 2.2: Gamma function (from Wikipedia, by Alessio Damato - CC BY-SA 3.0, commons.wikimedia.org/w/index.php?curid=365942)

For large n, we can assume that n is a continuous variable. In this case we generalize to

$$P(x;\nu) = \frac{\nu^x}{\Gamma(x+1)} e^{-\nu} \tag{2.57}$$

where $\Gamma(x)$ (equal to $(x-1)!$ for x integer) is the gamma function

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \tag{2.58}$$

2.6.3 Gaussian PDF

For large ν, the Poissonian is well approximated by the Gaussian with mean and variance ν. The Gaussian is defined as:

$$G(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.59}$$

and has mean μ and variance σ^2. Defining the new variable $z = (x-\mu)/\sigma$ the Gaussian becomes the Normal distribution:

$$N(x) = G(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.60}$$

We can define the *error function*

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{2.61}$$

so that the so-called *cumulative function* $F(x) = \int_{-\infty}^x G(x;0,1)dx$, which increases monotonically from 0 to 1, becomes

$$F(x) = \frac{1}{2}[1 + \mathrm{erf}(\frac{x}{\sqrt{2}})] \tag{2.62}$$

The prob. that the gaussian variable x distributed as $G(x;\mu,\sigma)$ is in the range $(\mu-a, \mu+a)$ is

$$P(x \in (-a,a)) = \mathrm{erf}(\frac{a}{\sigma\sqrt{2}}) \tag{2.63}$$

The Gaussian PDF is of such great importance not only because is the large-ν limit of the Poissonian but also because of the Central Limit Theorem (to be discussed later on):

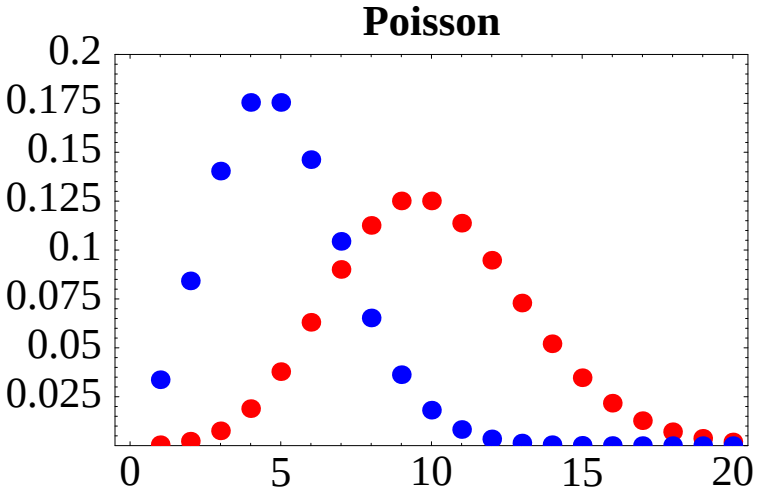

Figure 2.3: Poissonian for $\nu = 10$ (red) and $\nu = 5$ (blue). Note the similarity to the Binomial.

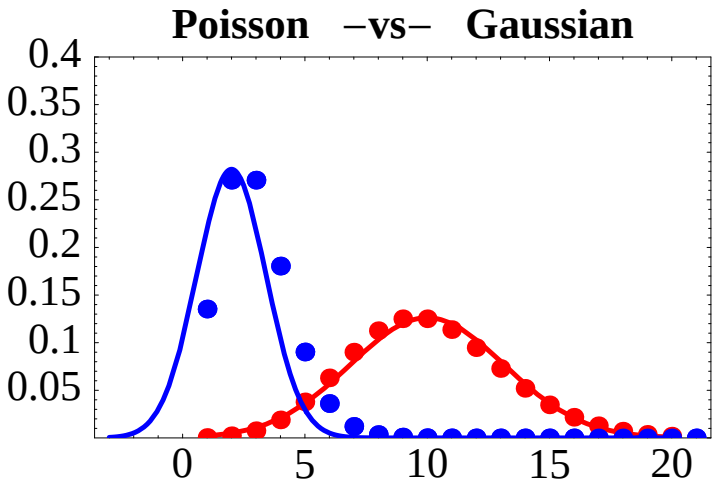

Figure 2.4: Comparing Poissonian and Gaussian PDFs for $\nu = 2$ (blue) and $\nu = 10$ (red).

Every random variable X sum (or linear combination) of many independent variables x_i (i.e. $X = \sum_i x_i$) is distributed approximately as a Gaussian of mean $\sum_i \mu_i$ and variance $\sigma_X^2 = \sum_i \sigma_i^2$ in the limit $n \to \infty$ independently of the individual PDFs .

In practice, the CLT can be applied in many experimental situations in which the error is the sum of many independent causes: reading errors, instrumental noise, contaminations etc. In these cases, the measure can be assumed to be gaussian distributed.

Three important values of the cumulative function are

$$F(\mu + j\sigma) - F(\mu - j\sigma) = \text{erf}(\frac{j}{\sqrt{2}}) = 0.68, 0.95, 0.997 \tag{2.64}$$

for $j = 1, 2, 3$: these give the prob. of finding x at $j = 1, 2, 3\sigma$ from the mean μ. Conventionally, errors are quoted at 1σ even for non-Gaussian distributions.

2.6.4 χ^2 distribution.

Let us now consider the Normal PDF and let's transform the variable x through the function $y = x^2$. Then we have

$$(2\pi)^{-1/2} \exp(-\frac{x^2}{2})dx = f(y)dy \tag{2.65}$$

from which, since $dx/dy = y^{-1/2}/2$, we obtain

$$f(y) = 2e^{-y/2}2^{-3/2}(\pi y)^{-1/2} \tag{2.66}$$

the factor of 2 must be inserted in order to normalize the resulting function $f(y)$. In this case in fact there are two x's for every y. The probability to obtain y in dy equals the sum of the two disjoint probabilities of having x in $x, x + dx$ *and* x in $-x, -x - dx$. This means effectively that $f(y)$ has to be multiplied by 2.

If we have instead n independent Gaussian variables, we can define

$$z = \sum_i \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{2.67}$$

In order to find the PDF of z we need to transform from $x_1, x_2, x_3..$ to n variables analog to spherical coordinates in n dimensions: a radius $r^2 = z$ and $n - 1$ angles θ_i. In 3D, this would mean replacing $dxdydz$ with $r^2 \sin\theta d\theta d\phi$. In general we find

$$\int f_x(x_1, x_2, ...)dx_1dx_2dx_3... = \int f_r(r)r^{n-1}dr|J|d\theta_1d\theta_2d\theta_3... = \int f_z(z)z^{(n-1)/2}|\frac{dr}{dz}|dz|J|d\Omega \tag{2.68}$$

$$= \frac{1}{2}\int f_z(z)z^{n/2-1}dz|J|d\Omega \tag{2.69}$$

where $|J|$ is a Jacobian factor that depends only on angles. For instance, in 3D, it amounts to $\sin\theta$. The integrals over the angles $d\Omega$ produces a normalization factor N and therefore the z PDF assumes the form $Ne^{-z/2}z^{n/2-1}$. Then we define the χ^2distribution (Fig. 2.5)

$$f(z \equiv \chi^2; n) = \frac{1}{2^{n/2}\Gamma(n/2)}z^{n/2-1}e^{-z/2} \tag{2.70}$$

where n is denoted the "number of degrees of freedom" and the Gamma function is defined as

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt \tag{2.71}$$

For integer numbers we have $\Gamma(n) = (n-1)!$ (and also $\Gamma(1/2) = \sqrt{\pi}$). The normalization is obviously such that $\int f(z; n)dz = 1$ for every n. For the χ^2 PDF we have

$$E(x) = n \tag{2.72}$$

$$\mathrm{Var}(x) = 2n \tag{2.73}$$

2.7 Moment generating function

The moments are the most important descriptors of a PdF. It is therefore useful to be able to calculate them easily. To this scope, one introduces the moment generating function (MGF), defined for a single variable as

$$m_x(t) \equiv \langle e^{tx} \rangle = \int e^{tx} f(x)dx \tag{2.74}$$

By expanding $e^{tx} = \sum_0^\infty \frac{(tx)^n}{n!}$, it it easy to show that

$$\frac{d^r m_x(t)}{dt^r}|_{t=0} = \langle x^r \rangle \tag{2.75}$$

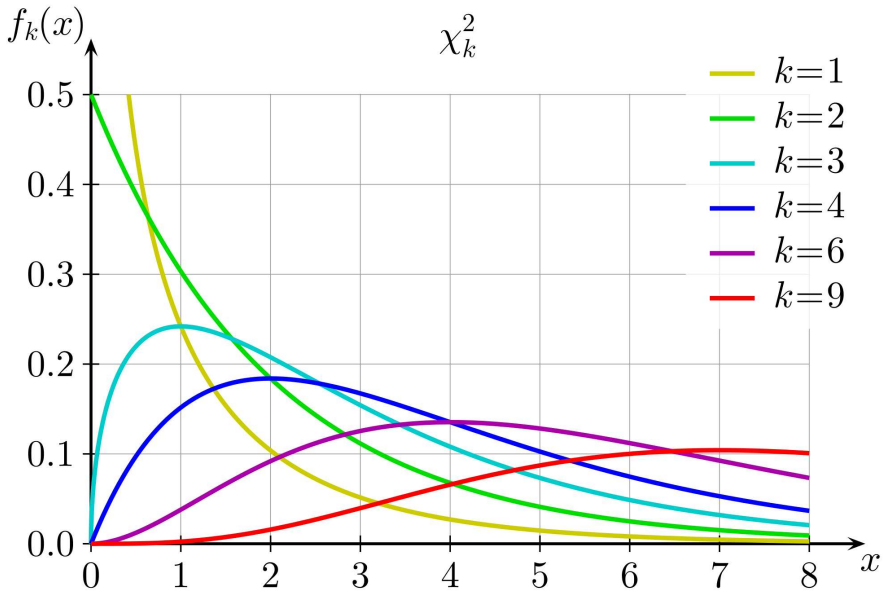

Figure 2.5: χ^2_k distribution for various values of degrees-of-freedom k. (By Geek3 - Own work, CC BY 3.0, commons.wikimedia.org/w/index.php?curid=9884213)

One has in fact

$$\frac{d^r m_x(t)}{dt^r}|_{t=0} = \frac{d^r \langle e^{tx}\rangle}{dt^r}|_{t=0} = \langle \frac{d^r e^{tx}}{dt^r}|_{t=0}\rangle \tag{2.76}$$

$$= \langle x^r e^{tx}\rangle_{t=0} = \langle x^r\rangle \tag{2.77}$$

Analogously, one can define the MGF for central moments:

$$m_{x-\mu}(t) \equiv \langle e^{t(x-\mu)}\rangle = \int e^{t(x-\mu)} f(x)dx \tag{2.78}$$

Suppose now we have two independent variables x, y distributed as $f(x), g(y)$. Let us find the MGF of the sum $s = x + y$. We can write directly

$$m_s = \int e^{t(x+y)} f(x)g(y)dxdy = \int e^{tx} f(x)dx \int e^{ty} g(y)dy = m_x m_y \tag{2.79}$$

i.e. the MGF of the sum of two independent variables is the product of the two MGFs. This extends obviously to the sum of any number of independent variables.

If the MGF exists, then two PFDs with the same MGF are identical; in other words, the MGF characterizes completely the PDF.

Exercise.

Show that the PDF of s can be written as the convolution of the individual PDFs.

Exercise.

Show that for a Gaussian with parameters μ, σ,

$$m_x(t) = e^{\frac{1}{2}t^2\sigma^2 + \mu t} \tag{2.80}$$

2.8 Central limit theorem

The MGF helps us to demonstrate the Central Limit Theorem, according to which the Gaussian is the asymptotic distribution of the sum of n independent identically distributed (IID) random variables in the limit of $n \to \infty$.

Let x_i with $i = 1, .., n$ be n IID random variables with mean μ and variance σ^2, with an unknown PDF. The CLT states that the variable

$$Y \equiv \frac{\hat{x} - \mu}{\sigma/\sqrt{n}} \tag{2.81}$$

where $\hat{x} = \sum_i x_i/n$ tends to a Gaussian variable for $n \to \infty$. Notice that $\sigma/\sqrt{n}$ is the variance of $\hat{x}$. Let us define the normal variables

$$z_i = \frac{x_i - \mu}{\sigma} \tag{2.82}$$

with $\langle z_i \rangle = 0$ and $\langle z_i^2 \rangle = 1$. Clearly

$$Y = \frac{1}{\sqrt{n}} \sum_i z_i \tag{2.83}$$

Let us find the MGF of Y. By the property of additive variables we have

$$m_Y(t) = \langle e^{Yt} \rangle = \langle e^{z_i t/\sqrt{n}} \rangle^n \tag{2.84}$$

Now

$$\langle e^{z_i t/\sqrt{n}} \rangle^n = \langle 1 + \frac{z_i t}{\sqrt{n}} + \frac{z_i^2 t^2}{2!n} + \frac{z_i^3 t^3}{3! n^{3/2}} + ... \rangle^n \tag{2.85}$$

$$= (1 + \frac{\langle z_i \rangle t}{\sqrt{n}} + \frac{\langle z_i^2 \rangle t^2}{2!n} + \frac{\langle z_i^3 \rangle t^3}{3! n^{3/2}} + ...)^n \tag{2.86}$$

Since $\langle z_i \rangle = 0$ and $\langle z_i^2 \rangle = 1$ we obtain for $n \gg 1$

$$\langle e^{z_i t/\sqrt{n}} \rangle^n = (1 + \frac{t^2}{2!n} + \frac{\langle z_i^3 \rangle t^3}{3! n^{3/2}} + ...)^n \tag{2.87}$$

$$\approx (1 + \frac{t^2}{2n})^n \approx 1 + \frac{t^2}{2} \approx e^{\frac{1}{2}t^2} \tag{2.88}$$

where in the last step we assume $t \ll 1$, i.e. we approximate only near the origin, which is where the moments obtained from the MGF are evaluated. We obtain then the MGF of a Normal Gaussian variable, QED.

The importance of this theorem is that it guarantees that if the errors in a measure are the results of many independent errors due to various parts of the experiment, then they are expected to be distributed in a Gaussian way. It can be extended to the case of independent variables with different mean and variances but in this case the condition $n \gg 1$ is not a sufficient condition for normality.

Exercise. Find the MGF for a uniform distribution P between in the range $-1, 1$. Find the MGF for the sum of n IID variables distributed as P

$$z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i}{\sigma} \tag{2.89}$$

where σ^2 is the variance of the x_i's. Notice that x_i/σ are normalized variable, $E[x_i] = 0, V[x_i] = 1$, and also z is. Plot the MGF of z for increasing values of n and confirm that it tends to the MGF of the Normal distribution $e^{t^2/2}$ for $n \gg 1$, as in Fig. 2.6.

2.9 Multivariate distributions

So far we have seen mostly distributions of single variables. We have however already defined the joint probability

$$f(x, y)dxdy \tag{2.90}$$

of having x, y in the area $dxdy$. The definition of probability requires now that

$$\int f(x, y)dxdy = 1 \tag{2.91}$$

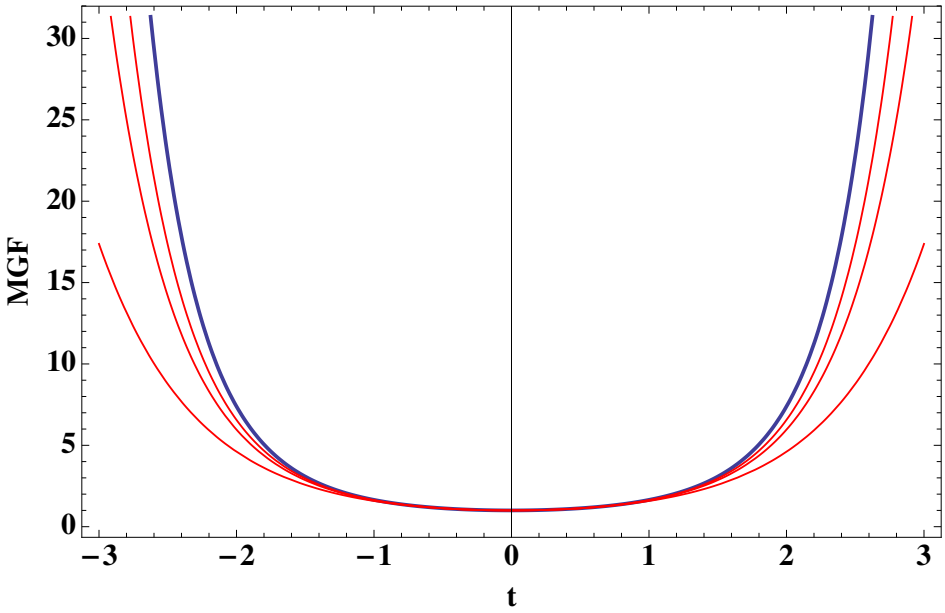

Figure 2.6: MGF for a Normal Gaussian distribution (blue thick line) and for a normalized sum of independent uniform variables (red thin lines) for $n = 1, 3, 5$, from bottom up.

It is clear that we can extend this definition to many variables x_i in the volume $d^n x$. For independent variables we know that $f(x, y) = f_1(x)f_2(y)$.

Analogously to the 1D case, we define the means

$$E[x] = \int x f(x, y) dx dy = \mu_x \tag{2.92}$$

$$E[y] = \int y f(x, y) dx dy = \mu_y \tag{2.93}$$

and the covariance matrix

$$C_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \tag{2.94}$$

$$= \int (x_i - \mu_i)(x_j - \mu_j) f(x_1, x_2) dx_1 dx_2 \tag{2.95}$$

where x_i is the vector of random variables and μ_i the mean vector. In case of more than two variables, the covariance integral is assumed to have been already marginalized over all the variables except i, j . The elements along the diagonal are the variances σ_i^2 of the individual random variables . If x_1, x_2 are independent then

$$C_{12} = \int (x_1 - \mu_x)(x_2 - \mu_y) f_1(x_1) f_2(x_2) dx_1 dx_2 \tag{2.96}$$

separates out and by definition of mean $C_{12} = 0$: then the covariance matrix of independent variables is diagonal (however in general $C_{12} = 0$ does not imply independent variables, but only uncorrelated variables). The covariance matrix is of course symmetric and also positive-definite, since

$$q_i C_{ij} q_j = \int [q_i(x_i - \mu_i)][q_j(x_j - \mu_j)] f(x_1, x_2) dx_1 dx_2 \tag{2.97}$$

$$= \int [q_i(x_i - \mu_i)]^2 f(x_1, x_2) dx_1 dx_2 > 0 \tag{2.98}$$

(sum over repeated indexes) for any vector q_i. The eigenvalues of C_{ij} are then real and positive.

The degree of correlation is indicated by the weight of the off-diagonal elements. For any two variables we define

$$\rho_{xy} \equiv \frac{C_{xy}}{\sigma_x \sigma_y} \tag{2.99}$$

and to maintain positive-definiteness $|\rho_{xy}| < 1$.

If we have the multivariate PDF $f(x, y)$ we can obtain the PDF of the individual variables by integrating out the other one:

$$g(x) = \int f(x, y) dy \tag{2.100}$$

This new PDF (*marginalized* over y) gives the probability of having x in dx whatever is y. We realize immediately that

$$E[x] = \int x g(x) dx \tag{2.101}$$

and similarly for all the other moments of x . All these definitions extend immediately to n dimensions, e.g.

$$C_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \tag{2.102}$$

$$= \int (x_i - \mu_i)(x_j - \mu_j) f(x_1, x_2, x_{3,...}) d^n x \tag{2.103}$$

2.9.1 Multinomial distribution

The binomial distribution can be generalized to the case in which there are not just two possible outcomes with probability p and $1 - p$ but k possible outcomes each with probability p_i, $i = 1, .., k$, with the constraint that the outcomes exhaust all the possibilities, so $\sum_i p_i = 1$. Now the probability of having a particular sequence of independent outcomes formed by x_1 outcomes of type 1, x_2 of type 2, etc will be

$$p_1^{x_1} p_2^{x_2} p_3^{x_3} ... p_k^{x_x} \tag{2.104}$$

Just as for the binomial distribution, accounting for all the possible internal permutations leads to the multinomial distribution, i.e. the probability that in a sequence of N trial one finds x_1 items of type 1, x_2 of type 2 etc . This is given by

$$P(x_1, x_2, ..x_k) = \frac{N!}{x_1! x_2! ... x_k!} p_1^{x_1} p_2^{x_2} p_3^{x_3} ... p_k^{x_k} \tag{2.105}$$

(provided $\sum_i p_i = 1$ and $\sum_i x_i = N$). The expected values and variances are

$$E[x_i] = N p_i \tag{2.106}$$

$$Var[x_i] = N p_i (1 - p_i) \tag{2.107}$$

Here we have however also a non-zero covariance

$$Cov[x_i x_j] = -n p_i p_j \tag{2.108}$$

The negative value reflects the fact that if x_i is large (i.e. several items of type i are extracted), then is more likely to have fewer items j, since the total number of outcomes is fixed to N.

2.9.2 Multivariate gaussian

The most interesting case of multivariate PDF is the multivariate Gaussian. Let us consider the most general Gaussian of two variables x_1, x_2 (with zero mean for simplicity)

$$G(x_1, x_2) = N \exp[-\frac{1}{2(1 - \rho^2)} (\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} - 2\frac{\rho x_1 x_2}{\sigma_1 \sigma_2})] \tag{2.109}$$

where $N = 1/(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})$. The covariance matrix is

$$C_{ij} = \int x_i x_j f(x_1, x_2) dx_1 dx_2 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{2.110}$$

and so $\rho = C_{xy}/\sigma_x\sigma_y$ is indeed the correlation coefficient. For $\rho = 1$ the distribution is degenerate, i.e. $\det C = 0$.

This PDF can be written as

$$G(x_1, x_2) = N \exp -\frac{1}{2}(X_i C_{ij}^{-1} X_j) \tag{2.111}$$

where we defined the vector $X_i \equiv (x_i - \mu_i)$ (we put back non-zero means μ_i) and $N = 1/2\pi\sqrt{\det C}$. This can be immediately generalized to n variables:

$$G(x_i, i = 1...n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp -\frac{1}{2}(X_i C_{ij}^{-1} X_j) \tag{2.112}$$

The contours of equiprobability $P = const$ are ellipsoids with principal axes oriented along the eigenvectors and semiaxes proportional to the square root of the eigenvalues of **C**.

Exercise:

Show that

$$\int x_1^2 G(x_1, x_2) dx_1 dx_2 = \sigma_1^2 \tag{2.113}$$

where $G(x_1, x_2)$ is given in (2.109).

We have

$$N \int x_1^2 \exp[-\frac{1}{2(1-\rho^2)}(\frac{x_1^2}{\sigma_1^2})] dx_1 \int \exp[-\frac{1}{2(1-\rho^2)}(\frac{x_2^2}{\sigma_2^2} - 2\frac{\rho x_1 x_2}{\sigma_1\sigma_2})] dx_2$$

where

$$N = [2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}]^{-1} \tag{2.114}$$

Adding and subtracting $-\frac{1}{2}\frac{x_1^2\rho^2}{\sigma_1^2(1-\rho^2)}$ within the exponent we obtain

$$N \int x_1^2 \exp[-\frac{1}{2(1-\rho^2)}(\frac{x_1^2}{\sigma_1^2} - \frac{x_1^2\rho^2}{\sigma_1^2})] dx_1 \int \exp[-\frac{1}{2(1-\rho^2)}(\frac{x_2}{\sigma_2} - \frac{\rho x_1}{\sigma_1})^2] dx_2$$

and the second integral becomes

$$\exp[-\frac{1}{2(1-\rho^2)}(\frac{x_2}{\sigma_2} - \frac{\rho x_1}{\sigma_1})^2] = \exp -\frac{1}{2}\frac{(x_2 - x_0)^2}{\hat{\sigma}_2^2}$$

where $x_0 = \rho x_1 \sigma_2/\sigma_1$ and $\hat{\sigma_2}^2 = \sigma_2^2(1-\rho^2)$. The second integral gives then the norm of the Gaussian, $\sqrt{2\pi\hat{\sigma}_2} = \sqrt{2\pi\sigma_2^2(1-\rho^2)}$. The first integral becomes then

$$\int x_1^2 \exp[-\frac{1}{2(1-\rho^2)}(\frac{x_1^2}{\sigma_1^2} - \frac{x_1^2\rho^2}{\sigma_1^2})] dx_1 = \int x_1^2 \exp[-\frac{1}{2}\frac{x_1^2}{\sigma_1^2}] dx_1 = \sigma_1^2\sqrt{2\pi\sigma_1^2}$$

Multiplying the various factors we obtain

$$\frac{\sigma_1^2\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2(1-\rho^2)}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = \sigma_1^2 \tag{2.115}$$

2.10 Gaussian integrals

Useful Gaussian integrals (integration always from $-\infty$ to $+\infty$):

$$\int e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx = \sigma\sqrt{2\pi} \tag{2.116}$$

$$\int x^{2n} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx = \sqrt{2\pi}\sigma^{2n+1}(2n-1)!! \tag{2.117}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int x^{2n} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx = (2n-1)!!\sigma^{2n} \tag{2.118}$$

(where $n!! = n(n-2)(n-4)...$). Notice that all odd moments of x vanish identically.

For n variables we have (sum over repeated indexes)

$$\int \exp[-\frac{1}{2} x_i C^{-1}_{ij} x_j] d^n x = (2\pi)^{n/2} (\det C)^{1/2} \tag{2.119}$$

$$\int \exp[-\frac{1}{2} x_i C^{-1}_{ij} x_j + J_i x_i] d^n x = (2\pi)^{n/2} (\det C)^{1/2} \exp[\frac{1}{2} J_i C_{ij} J_j] \tag{2.120}$$

Higher order moments of a multivariate Gaussian integrals can be evaluated as

$$E[x_1 x_2 ... x_{2n}] = \sum \prod E[x_i x_j] \tag{2.121}$$

where the sum and productory means summing over all possible distinct ways of partitioning $x_1 x_2 ... x_{2n}$into pairs $x_i x_j$ where each term in the sum is the product of n pairs. All odd moments vanish. Here, in the product $x_1 x_2 ... x_{2n}$, one can also repeat the variables, for instance if one has three variables x, y, z, moments like $E[x^2 y^2] = E[xxyy]$ or $E[x^2 y^4 z^2] = E[xxyyyyzz]$ are also included. For instance if $n = 2$ one has

$$E[x_1 x_2 x_3 x_4] = E[x_1 x_2] E[x_3 x_4] + E[x_1 x_3] E[x_2 x_4] + E[x_1 x_4] E[x_2 x_3] \tag{2.122}$$

and

$$E[x_1^2 x_2^2] = E[x_1 x_1 x_2 x_2] = E[x_1 x_1] E[x_2 x_2] + E[x_1 x_2] E[x_1 x_2] + E[x_1 x_2] E[x_1 x_2] = E[x_1^2] E[x_2^2] + 2E^2[x_1 x_2] \tag{2.123}$$

and

$$E[x_1^4] = E[x_1 x_1 x_1 x_1] = 3E[x_1 x_1] E[x_1 x_1] = 3E^2[x_1^2] = 3\sigma_1^4 \tag{2.124}$$

which agrees with (2.118). This result is variously known as Wick's theorem, Isserlis' theorem, etc.

Problem. As an application of Gaussian integral, find the PDF of the sum $z = x + y$ of two Gaussian random variables x, y with zero mean and the same variance σ^2. Also, confirm that the variance of z is $2\sigma^2$.

We have seen in Sect. (2.4) that the variance of the sum of random variables is simply $\sum_i \sigma_i^2$, so one would get indeed $z = 2\sigma^2$. However the error propagation formula is supposed to be an approximation, while here the result is exact because the function $z = z(x, y)$ is a linear function.

To find the distribution of z we need to transform variables from x, y to a new pair of variables, z and, say, $X = x$ (we could have used any other linearly independent combination of x, y). Calculating the Jacobian of the transformation we see that $|J| = 1$. Then we have

$$f(z, X) = G(x, y; \sigma^2)|J| = G(x, y; \sigma^2) \mid_{y=z-X; x=X} \tag{2.125}$$

and

$$f(z)dz = \int G(x, y; \sigma^2) \mid_{y=z-X; x=X} dX \tag{2.126}$$

Now we write

$$G(x, y; \sigma^2) \mid_{y=z-X; x=X} = N \exp -\frac{1}{2}\frac{x^2 + y^2}{\sigma^2} \mid_{y=z-X; x=X} = N \exp -\frac{1}{2\sigma^2}(z^2 - 2xy) \mid_{y=z-X; x=X} \tag{2.127}$$

$$= N \exp -\frac{1}{2\sigma^2}[z^2 - 2X(z - X)] = N \exp -\frac{1}{2\sigma^2}[z^2 + 2X^2 - 2Xz] \tag{2.128}$$

$$= N \exp[-\frac{1}{2}\frac{z^2}{\sigma^2} + \frac{1}{4}\frac{z^2}{\sigma^2}] \exp -\frac{1}{2\sigma^2}[2X^2 - 2Xz + \frac{z^2}{2}] \tag{2.129}$$

Then we have

$$f(z)dz = \int G(x,y;\sigma^2)\mid_{y=z-X;x=X} dX \tag{2.130}$$

$$= N\exp[-\frac{1}{2}\frac{z^2}{\sigma^2}+\frac{1}{4}\frac{z^2}{\sigma^2}]\int \exp -\frac{1}{2\sigma^2}[2X^2-2Xz+\frac{z^2}{2}]dX \tag{2.131}$$

$$= N\sqrt{\pi}(\sigma^2)\exp[-\frac{1}{4}\frac{z^2}{\sigma^2}] = N'\exp[-\frac{1}{2}\frac{z^2}{(2\sigma^2)}] \tag{2.132}$$

which shows that indeed z has variance $2\sigma^2$ and that z is itself a Gaussian variable. This result can be easily generalized to N Gaussian variables with variances σ_i.

2.11 Parameter estimation: Statistics, sample, bias

So far we have analyzed the theoretical properties of the distributions. However, what we really normally have is a number of measurements $x_1, x_2, x_3...x_n$. If the measures are independent, we can assume that the joint PDF of the full set x_i is

$$f_{sample}(x_i) = f(x_1)f(x_2)f(x_3)...f(x_n) \tag{2.133}$$

Our problem now is to derive from the n measures the estimates of the population parameters, that is, the parameters that characterize $f(x)$, for instance the mean μ and the variance σ^2 . We need to find then functions $\theta(x_i)$ of the data x_i (and *only* of the data, not of unknown parameters), generally called *statistics*, such that they approximate the parameters of the $f(x)$. Since x_i are random variables, $\hat{\theta}(x_i)$ is also a random variable. The central problem of statistics, called *inference*, is to obtain estimates of unknown parameters from a collection of data, assuming that we know the type of distribution each single datapoint has, eg whether Gaussian or Binomial etc, but not the value of some of their parameters, e.g the mean or the variance.

We have already seen an example of estimator: the mean

$$\hat{x} = \frac{\sum_i x_i}{n} \tag{2.134}$$

(now we use a hat to denote the estimator as a random variable, rather than any specific estimate) is in fact an estimator of $\mu = E[x]$. We can certainly have several estimators for any given parameter; here we see now which are the main properties that "good" estimator should possess.

Let θ be the parameter of $f(x)$ to be estimated and $\hat{\theta}$ the estimator, function of the n measures x_i. If $\hat{\theta}$ approximates θ in the limit of large n, the estimator is said to be *consistent* :

$$\lim_{n\to\infty} P(|\hat{\theta}-\theta| > \epsilon) = 0 \tag{2.135}$$

for every $\epsilon > 0$.

The expected value of $\hat{\theta} = \hat{\theta}(x_1, x_2, ...)$ is by definition

$$E[\hat{\theta}] = \int \hat{\theta} f(x_1, x_2, ...x_n;\theta)...dx_1...dx_n \tag{2.136}$$

If the *bias*

$$b = E[\hat{\theta}] - \theta \tag{2.137}$$

is zero for every n, the estimator $\hat{\theta}$ is *unbiased.* If $b \to 0$ only for large n, the estimator is said to be asymptotically *unbiased*. The bias is a *systematic* error because it does not depend on how good the measures are but on our choice of the estimator. At least in principle, one can always choose a better estimator or a unbiased one.

We define also the variance of the estimator:

$$V[\hat{\theta}] = E[(\hat{\theta}-E[\hat{\theta}])^2] = \int (\hat{\theta}-E[\hat{\theta}])^2\hat{\theta} f(x_1, x_2, ...x_n;\theta)...dx_1...dx_n \tag{2.138}$$

The variance of $\hat{\theta}$ is a *statistical* error because is unavoidable (although it can be minimized), since it depends ultimately on the fact that $\hat{\theta}$ is a random variable.

We define also the mean square error

$$MSE = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] = V[\hat{\theta}] + b^2 \tag{2.139}$$

which can be indeed interpreted as the sum of the statistical and systematic errors. An estimator can be chosen that minimizes the bias (the estimator is said to be *accurate*), the variance (the estimator is said to be *precise*), or the MSE. Normally, you can't have all of them.

Please notice: θ is a parameter, not a random variable; $\hat{\theta}$ is a function of the data and therefore a random variable; $E(\hat{\theta})$ is no longer a function of the data since the data have been integrated over using the PDF (as if we really had performed an infinity of experiments) and therefore is not a random variable.

2.12 Estimators of mean and variance.

In this section we assume always independent and identically distributed variables. The sample mean

$$\hat{x} \equiv \frac{\sum_i x_i}{n} \tag{2.140}$$

is an unbiased estimator of $\mu = E[x]$. In fact

$$E[\hat{x}] = \frac{1}{n}\sum_i \int x_i f(x) dx = \frac{\sum \mu}{n} = \mu \tag{2.141}$$

Notice also that even a single measure, eg x_1, is an unbiased estimator of the mean: $E[x_1] = \mu$. A good choice should then be the estimator of minimal variance. As we have seen already and check again below, the variance of the mean goes like $1/n$; the mean is therefore a better estimator than a single measure, or any mean of a subset of measures.

The sample variance

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \hat{x})^2 \tag{2.142}$$

is an unbiased estimator of $\sigma^2 = E[(x-\mu)^2]$. Notice that both x_i and $\hat{x}$ are random variables; if the mean is known in advance, the denominator of the unbiased estimator would be n instead of $n-1$.

Analogously, the unbiased estimator of the covariance is

$$\hat{V}_{ij} = \frac{1}{n-1}\sum_k (x_{i,k} - \hat{x}_i)(x_{j,k} - \hat{x}_j) \tag{2.143}$$

For two variables this is

$$\hat{V}_{xy} = \frac{n}{n-1}(\hat{xy} - \hat{x}\hat{y}) \tag{2.144}$$

where in this specific instance we use the $\hat{}$ to denote the sample average, for instance

$$\hat{xy} = \frac{1}{n}\sum_i x_i y_i \tag{2.145}$$

Finally, an estimator of the correlation coefficient is

$$r = \frac{\hat{V}_{xy}}{s_x s_y} = \frac{\hat{xy} - \hat{x}\hat{y}}{\sqrt{(\hat{x^2} - \hat{x}^2)(\hat{y^2} - \hat{y}^2)}} \tag{2.146}$$

(note $\hat{x^2} \neq \hat{x}^2$). This is only asymptotically unbiased, even if $\hat{V}_{xy}, s_x, s_y$ are unbiased quantities; however is often used because of its simplicity.

We can now estimate the variance of the mean:

$$Var[\hat{x}] = E[\hat{x^2}] - (E[\hat{x}])^2 = E[(\frac{1}{n}\sum x_i)(\frac{1}{n}\sum x_j)] - \mu^2 \quad (2.147)$$

$$= \frac{1}{n^2}\sum_{i,j} E[x_i x_j] - \mu^2 \quad (2.148)$$

$$= \frac{1}{n^2}[(n^2-n)\mu^2 + n(\mu^2+\sigma^2)] - \mu^2 = \frac{\sigma^2}{n} \quad (2.149)$$

where in the last step we have employed $E[x_i x_j] = \mu^2$ for $i \neq j$ and $E[x_i^2] = \mu^2 + \sigma^2$ (and that there are $n^2 - n$ combinations $i \neq j$ and n of $i = j$). The same result can be readily obtained by the law of error propagation, which in this case is exact since the mean is a linear combination of random variables. This is a very important result: the standard deviation of the mean is a factor $1/\sqrt{n}$ smaller wrt the standard deviation of a single measure. So if I perform another measurement, I expect it to deviate by $\sim \sigma$ wrt the mean; but if I take another mean of a set of n similar measurements, then I expect the new mean to deviate from the old one only by $\sim \sigma/\sqrt{n}$.

Exercise.

Show that (see eq. 2.142)

$$E[\frac{1}{n-1}\sum_i (x_i - \hat{x})^2] = \sigma^2 \quad (2.150)$$

One has

$$E[\frac{1}{n-1}\sum_i (x_i - \hat{x})^2] = \frac{1}{n-1}\sum_i \int (x_i - \frac{\sum x_j}{n})^2 f(x_1, x_2, ..) d^n x \quad (2.151)$$

$$= \frac{1}{n-1}\sum_i \int (x_i^2 + \frac{1}{n^2}\sum_{jk} x_k x_j - \frac{2}{n} x_i \sum_j x_j) f(x_1, x_2, ..) d^n x \quad (2.152)$$

$$= \frac{1}{n-1}\sum_i [\sigma^2 + \mu^2 + \frac{n^2-n}{n^2}\mu^2 + n\frac{(\sigma^2+\mu^2)}{n^2} - \frac{2}{n}(\sigma^2 + \mu^2 + (n-1)\mu^2)] \quad (2.153)$$

where we employed the relation

$$\int \sum_{jk} x_j x_k f(x) d^n x = (n^2-n)\int x_j x_k f(x_1, x_2, ..) d^n x + n \int x^2 f(x_1, x_2, ..) d^n x = (n^2-n)\mu^2 + n(\sigma^2+\mu^2) \quad (2.154)$$

Finally we obtain

$$E[\frac{1}{n-1}\sum_i (x_i - \hat{x})^2] = \frac{1}{n-1}\sum_i [(\sigma^2+\mu^2)(\frac{n-1}{n}) + \mu^2(\frac{1-n}{n}))] \quad (2.155)$$

$$= \sigma^2 \quad (2.156)$$

QED

Chapter 3

The likelihood function and the Fisher matrix

The general problem of parameter estimation is solved using Bayes' theorem. This allows to obtain the estimators for any parameter and their region of confidence. In this Chapter we introduce several Bayesian tools, from the Fisher matrix to model comparison. Some application to real cases are also presented.

3.1 From prior to posterior

Let us suppose we know, or have good reasons to suspect, that a random variable x, e.g., the apparent magnitude of a supernova, has a probability distribution function (PDF) $f(x;\theta)$ that depends on an *unknown* parameter θ, e.g., the absolute magnitude. The ";" is meant to distinguish the random variables x from the parameter θ. Such a probability is called a *conditional probability* of having the data x given the theoretical parameter θ. We may for instance suppose that the apparent magnitude m is distributed as a Gaussian variable with a given variance σ^2 (the observational error on m), but we do not know one of the cosmological parameters that enter the expected value $m_{\rm th} = 5\log_{10} d_L(z;\Omega_m^{(0)},\Omega_\Lambda^{(0)}) + \text{constant}$, where d_L is the luminosity distance.

If we repeat the measure and we obtain $x_1, x_2, x_3...$, then the law of joint probability tells us that the probability of obtaining x_1 in the interval $\mathrm{d}x_1$ around x_1, x_2 in the interval $\mathrm{d}x_2$ around x_2 and so forth is

$$f(x_i;\theta)\mathrm{d}^n x_i \equiv \prod_i f_i(x_i;\theta)\mathrm{d}x_i = f_1(x_1;\theta)f_2(x_2;\theta)f_3(x_3;\theta)...\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3...\,, \tag{3.1}$$

if the measures are independent of each other. Clearly, for every θ this multivariate PDF will assume a different value. It is logical to *define* the best θ as the value for which $\prod_i f(x_i;\theta)$ is maximal. Indeed, if we generate random variables distributed as $f(x;\theta)$, the most likely outcome for x is that value maximizing $f(x;\theta)$. Conversely, if we have a particular outcome x, then our best bet is to assume that θ is such as to maximize the occurrence of that x. We used as an example independent data and a single parameter but this is by no means necessary. We define the best θ_α as those parameters that maximizes the joint function $f(x_1, x_2, ..x_n; \theta_1, \theta_2, ...\theta_m)$. Since in general we have many parameters to estimate, we write the function simply $f(x_i;\theta_\alpha)$, meaning all the x_i's and all the θ_α's.

The maximum likelihood method of parameter estimation consists therefore in finding the parameters that maximize the *likelihood function* $f(x_i;\theta_\alpha)$ by solving the system

$$\frac{\partial f(x_i;\theta_\alpha)}{\partial\theta_\alpha} = 0\,, \qquad \alpha = 1,..,m\,. \tag{3.2}$$

Let us denote the solutions of these equations as $\hat{\theta}_\alpha$. They are functions of the data x_i and therefore are random variables, just as the data are. So the classical *frequentist* approach would try to determine the distribution of the $\hat{\theta}_\alpha$s knowing the distribution of the x_is; if this is possible, one can assign probabilities to $\hat{\theta}_\alpha$'s ranges, for instance determine the interval of $\hat{\theta}_\alpha$ that contains 95% probability that a particular set of data has been drawn from the theoretical distribution (we will se this in later chapters). One problem with this approach is that it is often

too difficult to derive the $\hat{\theta}_j$'s distribution analytically and very demanding to derive them numerically through simulated datasets. But the main problem is that this approach does not take into account what we already know concerning the theoretical parameters, for instance the result of previous experiments. To handle this information properly we need to switch to the *Bayesian* approach. Instead of looking for the probability $f(x_i;\theta_\alpha)$ of having the data given the model, we estimate the probability $L(\theta_\alpha;x_i)$ of having the model given the data (see Fig. 3.2).

This problem is solved by the fundamental theorem of conditional probabilities, the so-called Bayes' theorem[1]:

$$P(T;D) = \frac{P(D;T)P(T)}{P(D)}\,, \tag{3.3}$$

where we denote the known data x_i with D and the unknown theory (that is, the theoretical parameters θ_α) with T. On the r.h.s., $P(D;T)$ is the conditional probability of having the data given the theory; $P(T)$ and $P(D)$ are the probability of having the theory and the data, respectively; finally, on the l.h.s., $P(T;D)$ is the conditional probability of having the theory given the data. Bayes' theorem is a consequence of the definition of conditional probability $P(A;B) \equiv P(A,B)/P(B)$ and of the symmetry of the joint probability $P(A,B)$ (the probability of having both A and B) under the exchange of A, B.

It follows that

$$L(\theta_\alpha;x_i) = \frac{f(x_i;\theta_\alpha)p(\theta_\alpha)}{g(x_i)}\,, \tag{3.4}$$

where $p(\theta_\alpha)$ is called the *prior* probability for the parameters θ_α, while $g(x_i)$ is the PDF of the data x_i. The final function $L(\theta_\alpha;x_i)$ (or simply $L(\theta_\alpha)$ for shortness) is called *posterior* although sometimes it is also loosely called likelihood just as $f(x_i;\theta_\alpha)$ and generally denoted as L. The posterior contains the information we are looking for: the probability distribution of the parameters given that we observed the data x_i and that we have some prior knowledge about the parameters themselves. In fact the whole method in the Bayesian context should be called "the posterior method" rather than the "likelihood" method. Of course, the frequentist and the Bayesian estimator coincide if the prior is uniform; however, the PDF of the parameter would still be different (see an example later on).

Once we use Bayes' theorem to convert the parameters into random variables, the data themselves are no longer regarded as random variables, but just as given data, fixed by the experiment once and for all. There is no need to think in terms of frequencies of occurrence anymore, and therefore no need to define probability in terms of an infinite sequence of experiments. Of course, we still have to assume that the data are distributed according to some class of functions, e.g. that they are Gaussian distributed, in order to form a likelihood. In principle, however, we could define a probability distribution with many free parameters that represents practically all possible distributions, or even a distribution that is parametrized by its value in many small intervals. Whatever our choice is, it is an assumption, not the result of an hypothetical infinite number of previous experiments. The Bayesian methods are then said to express our degree of belief in our assumptions concerning the data and the priors.

Since $L(\theta_\alpha;x_i)$ is a probability distribution function for θ_α, it has to be normalized to unity:

$$\int L(\theta_\alpha;x_i)\mathrm{d}^n\theta_\alpha = 1 = \frac{\int f(x_i;\theta_\alpha)p(\theta_\alpha)\mathrm{d}^n\theta_\alpha}{g(x_i)}\,, \tag{3.5}$$

and therefore

$$\int f(x_i;\theta_\alpha)p(\theta_\alpha)\mathrm{d}^n\theta_\alpha = g(x_i)\,. \tag{3.6}$$

As we will see in the next section the integral on the l.h.s. is called *evidence* and the same name is sometimes given also to $g(x_i)$. The function $g(x_i)$ does not depend on the parameters θ_α and therefore it is of no help in estimating the parameters. From the point of view of $L(\theta_\alpha)$ it is just a normalization factor. The prior $p(\theta_\alpha)$ is also often unknown. Normally we do not know the probability distribution of theories, that is, whether the ΛCDM model is more probable, from an absolute point of view, than a modified gravity model or whether $\Omega_\Lambda^{(0)} = 0$ is more probable than $\Omega_\Lambda^{(0)} = 0.7$. However, we often *do know* something which, while not quite absolute in any sense, still represents some form of information independent of the data at hand. Namely, we know the results of previous experiments. If an experiment convincingly excluded, say, $\Omega_m^{(0)} < 0.1$, then we could use this information, putting

[1] Reverend Thomas Bayes (1702–1761) studied what in modern terminology is the binomial distribution and introduced the concept of conditional probability. His work was published posthumously in 1763.

$p(\Omega_m^{(0)} < 0.1) = 0$. If instead we believe that $h = 0.72 \pm 0.08$, then we could use as $p(h)$ a Gaussian with mean 0.78 and standard deviation 0.08. These are typical *prior distributions*.

Priors can be of many kinds. Beside including other experiments, we could simply exclude unphysical values, e.g., $\Omega_m^{(0)} < 0$ or weigh down some regions of parameter space that we, perhaps subjectively, consider less likely. What matters is not so much what we decide to include as prior but rather that we make this decision explicit to the reader and to the potential user of our results. Every posterior, sooner or later, will become a prior for us or for somebody else, and it is our responsibility to make it explicit which prior information we adopted, no less to avoid that a future user includes twice the same information. The easiness of including prior information of all kinds is one of the major advantage of the Bayesian approach.

There are two important facts to note about priors. First, priors matter. Clearly the final result depends on the prior, just as our bet on the result of a football match depends on what we know about the teams based on previous games (and on our personal interpretation of those results). One can say that priors quantify our physical intuition. Second, priors are unavoidable. Even if we are not consciously choosing a prior, the way we manage the statistical problem at hand *always* implies some form of prior. No prior on a parameter means in fact $p(\theta) = 1$ in the domain where θ is defined and $p(\theta) = 0$ outside. Even when θ is defined in the whole real range we are still choosing a "flat" prior, $p(\theta) = 1$, over other possible choices. One could push this argument as far as saying that our choice of theory and its parameters θ already constitute a strong prior. So, again, the important issue is to specify exactly what prior is employed. An improper prior, i.e. one which is not normalized to unity, can also be employed. For instance, one can assume a uniform prior in the entire range from $-\infty$ to $+\infty$.

Once we have $L(\theta_\alpha)$ we need to search the maximum to obtain the maximum likelihood estimators (MLE) $\hat{\theta}_i$. Because of the priors, this will differ in general from the maximum of $f(x_i; \theta_\alpha)$. Equation (3.2) is then replaced by

$$\frac{\partial L(\theta_\alpha)}{\partial \theta_\alpha} = 0\,, \qquad \alpha = 1, .., n\,. \tag{3.7}$$

If, as usually the case, we discard the denominator $g(x_i)$ in Eq. (3.4), the posterior L is not normalized and its normalization has to be recalculated. The overall normalization N is the integral over the parameter space:

$$N = \int L(\theta_\alpha)\, \mathrm{d}^n\theta_\alpha\,, \tag{3.8}$$

where the integral extends to the whole parameter domain. From the normalized likelihood [i.e. $L(\theta_\alpha)/N$ which we keep calling $L(\theta_\alpha)$], we can derive the regions of confidence (or *belief*) for the parameters. These are defined as regions $R(\alpha)$ of constant $L(\theta_\alpha)$ for which

$$\int_{R(\alpha)} L(\theta_\alpha)\, \mathrm{d}^n\theta = \alpha\,. \tag{3.9}$$

The region $R(\alpha)$ is the region for which the integral above evaluates to $0 < \alpha < 1$ (remember that now L is normalized and therefore its integral over the whole domain is 1). To find R one evaluates

$$\int \hat{L}(L_i)\, \mathrm{d}^n\theta = \alpha_i\,, \tag{3.10}$$

where $\hat{L}(L_i) = L$ if $L > L_i$ and 0 elsewhere (i.e. the volume lying within the curve of "height" L_i, smaller than the peak of L). By trial and error (or by interpolating over a grid of L_i) one finds the preferred α_i. The typical choices are $\alpha = 0.683, 0.954, 0.997$ (also denoted as $1, 2, 3\sigma$, respectively, but sometimes other reference values are also employed. The value L_i that corresponds to α_i is the level at which we have to cut L to find the region $R(\alpha_i)$.

3.2 Marginalization

Often, we are interested in some subset of parameters and consider the others as "nuisance" of which we would gladly get rid of. For instance, if we are analyzing a set of supernovae apparent magnitudes m_i and comparing them to the theoretical predictions $m_{\rm th} = 5 \log_{10} d_L(z; \Omega_m^{(0)}, \Omega_\Lambda^{(0)}) + C$, we may be interested in $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$ but not in the constant C. This constant depends on the K correction and on the standard absolute magnitude M, to which

we can add also the constant $\log_{10} H_0^{-1}$. Our general likelihood is therefore a function of $C, \Omega_m^{(0)}, \Omega_\Lambda^{(0)}$ but we can transform it into a function of $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$ alone by integrating out C:

$$L(\Omega_m^{(0)}, \Omega_\Lambda^{(0)}) \equiv \int L(C, \Omega_m^{(0)}, \Omega_\Lambda^{(0)})\, \mathrm{d}C\,, \tag{3.11}$$

where the integration extends over the domain of definition of C, which in absence of better information could as well be from $-\infty$ to $+\infty$ [there should be no confusion by denoting both the "old" and the "new" likelihood by the same symbol in Eq. (3.11)]. This very common procedure is called *marginalization*.

Often one wants to marginalize a multidimensional problem down to a more manageable and plottable 2-dimensional likelihood. Also, one could quote final confidence regions by marginalizing in turn to single parameters, e.g.,

$$L(\Omega_\Lambda^{(0)}) = \int_0^\infty L(\Omega_m^{(0)}, \Omega_\Lambda^{(0)})\, \mathrm{d}\Omega_m^{(0)}\,. \tag{3.12}$$

For instance, if the maximum likelihood estimator of $\Omega_m^{(0)}$ is 0.3 and

$$\int_R L(\Omega_m^{(0)}) \mathrm{d}\Omega_m^{(0)} = 0.683\,, \tag{3.13}$$

when R is the interval $\Omega_m^{(0)} = [0.1, 0.4]$, we will write as our final result $\Omega_m^{(0)} = 0.3^{+0.1}_{-0.2}$ at 68.3 % confidence level (or, less precisely, at 1σ: notice that this will absolutely not imply that at 2σ one should expect -0.1 as lower limit of $\Omega_m^{(0)}$!).

In the common case in which we want to marginalize over a constant offset or over a multiplicative factor one can often obtain an analytical result. Here we work out the first case. Taking again the example of supernovae, suppose that we have N standard candle sources at redshifts z_i with apparent magnitudes m_i and that our preferred cosmological model predicts magnitudes $m_{\mathrm{th},i} = M + 5\log_{10} d_L(z_i;\theta_\alpha) + 25$, where $d_L(z_i;\theta_\alpha)$ is the luminosity distance measured in Megaparsecs. The luminosity distance is proportional to $1/H_0$. We can therefore take this factor out of the logarithm and write $m_{\mathrm{th},i} = \alpha + \mu_i$, where $\mu_i = 5\log_{10} \hat{d}_L(z_i;\theta_\alpha)$ and $\alpha = M + 25 - 5\log_{10} H_0$ and $\hat{d}_L$ is $d_L H_0$. We have very little information on α, so we decide to marginalize it over:

$$L(\theta_\alpha) = N \int \mathrm{d}\alpha\, \exp\left[-\frac{1}{2}\sum_i \frac{(m_i - \mu_i - \alpha)^2}{\sigma_i^2}\right]\,, \tag{3.14}$$

where N is an unimportant normalization factor. Then we have

$$\begin{aligned} L(\theta_\alpha) &= N \int \mathrm{d}\alpha \exp\left[-\frac{1}{2}\sum_i \frac{(m_i-\mu_i)^2 + \alpha^2 - 2\alpha(m_i-\mu_i)}{\sigma_i^2}\right] \\ &= N\exp(-S_2/2)\int \mathrm{d}\alpha \exp(\alpha S_1 - \alpha^2 S_0/2) \\ &= N\exp\left[-\frac{1}{2}\left(S_2 - \frac{S_1^2}{S_0}\right)\right]\int \mathrm{d}\alpha\, \exp\left[-\frac{1}{2}\left(\alpha - \frac{S_1}{S_0}\right)^2 S_0\right]\,, \end{aligned} \tag{3.15}$$

where $S_0 = \sum(1/\sigma_i^2)$, $S_1 = \sum y_i/\sigma_i^2$, $S_2 = \sum y_i^2/\sigma_i^2$, and $y_i = m_i - \mu_i$. The integration in the region $(-\infty, +\infty)$ gives a constant independent of μ_i and therefore independent of the theoretical parameters that we absorb in N:

$$L(\theta_\alpha) = N\exp\left[-\frac{1}{2}\left(S_2 - \frac{S_1^2}{S_0}\right)\right]\,. \tag{3.16}$$

This is then the new likelihood marginalized over the nuisance additive parameter α. Notice that the parameters θ_α ended up inside μ_i which are inside S_1, S_2. A similar analytic integration can get rid of multiplicative parameters. If the analytical integration is impossible, then one has to marginalize numerically.

Sometimes one prefers to fix a parameter, rather than marginalizing over it, perhaps because one wants to see what happens for particularly interesting values of that parameter. So for instance one may fix $\Omega_\Lambda^{(0)}$ to be $\Omega_\Lambda^{(0)} = 0$ and evaluate $L(\Omega_m^{(0)}, \Omega_\Lambda^{(0)} = 0)$. Then the result will obviously depend on the fixed value. When that value is the maximum likelihood estimator, the likelihood is said to be *maximized* (as opposed to *marginalized*) with respect to that parameter.

3.3 Some examples

If this is your first encounter with maximum likelihood methods, warm up by proving that if we have the Gaussian likelihood $f(x_i; \mu, \sigma^2)$

$$f(x_i; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_i^n \frac{(x_i-\mu)^2}{\sigma^2}\right], \tag{3.17}$$

then the MLE of μ is given by

$$\hat{\mu} = \frac{1}{n}\sum_i^n x_i \,. \tag{3.18}$$

Analogously, you can prove that the variance MLE is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i^n (x_i - \hat{\mu})^2 \,. \tag{3.19}$$

You may notice that this falls short of the standard result according to which the estimate of the sample variance has $(n-1)$ instead of n at the denominator. In this case in fact the maximum likelihood estimator is biased, which means that its expectation value does not equal the "true" or "population" value. Indeed, maximum likelihood estimators are not necessarily unbiased although under some general conditions they are asymptotically (i.e. for $n \to \infty$) unbiased.

The MLE for correlated Gaussian variables can also be easily obtained. Assuming a common mean μ for all variables, the likelihood is

$$f(x_i) = (2\pi)^{-n/2}|C|^{-1/2} \exp\left[-\frac{1}{2}(x_i - \mu u_i)C_{ij}^{-1}(x_j - \mu u_j)\right], \tag{3.20}$$

where $u_i = (1, 1, 1...)$. Then we obtain

$$\hat{\mu} = \frac{u_i C_{ij}^{-1} x_j}{u_i C_{ij}^{-1} u_j} \tag{3.21}$$

If the correlation matrix is diagonal this becomes

$$\hat{\mu} = \frac{\sum x_j \sigma_j^{-2}}{\sum \sigma_j^{-2}} \tag{3.22}$$

ie. a weighted sum, where each variable is weighted by the inverse of its variance. This estimator is the minimum variance estimator. Using the error propagation formula, which is exact for Gaussian independent variables, we see that the variance of the mean estimator $\hat{\mu}$ is then

$$\sigma_{\hat{\mu}}^2 = \frac{1}{\sum \sigma_j^{-2}} \tag{3.23}$$

So both in $\hat{\mu}$ and in $\sigma_{\hat{\mu}}^2$, data points with large error weigh less than those with small errors.

Notice that if we follow the frequentist approach (and therefore don't use Bayes' theorem) we could still derive a PDF for the variance (here we need to distinguish between the particular ML value of the estimator, σ_{ML}^2, and its generic value as random variable, which I denote with $\hat{\sigma}^2$). Let us take the simplest case of constant variance (which now, according to frequentist practice, is to be replaced by the estimator value σ_{ML}^2), and let us normalize $\hat{\sigma}^2$ by defining the new variable z:

$$z \equiv \frac{n\hat{\sigma}^2}{\sigma_{ML}^2} = \frac{n^2}{S_2}\hat{\sigma}^2 \tag{3.24}$$

where we defined $S_2 = \sum(x_i - \mu)^2$ and $\sigma_{ML}^2 = S_2/n$. Since $\hat{\sigma}^2$ is a quadratic function of Gaussian variables, the normalized variable z follows a χ^2 distribution with n degrees of freedom (we assume here that μ is known in advance, i.e. is not estimated from the data)

$$P(z = \hat{\sigma}^2 n^2/S_2) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \tag{3.25}$$

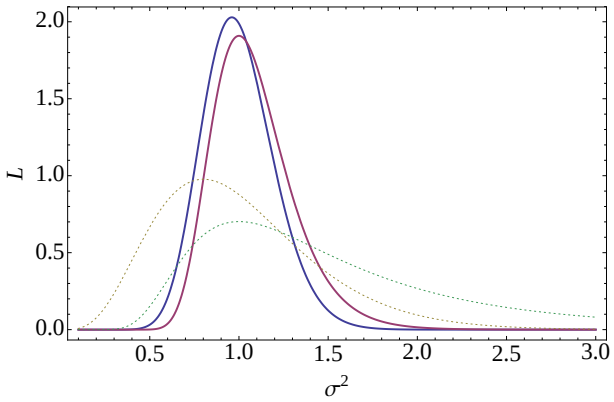

Figure 3.1: Frequentist distribution of $\hat{\sigma}^2$ (in blue) compared with the Bayesian distribution (in red) for $n = 50$ (thick lines) and $n = 10$ (dotted lines), for $S_2 = n$.

Since $dz/d\hat{\sigma}^2 = n^2/S_2$, the distribution of $\hat{\sigma}^2$is finally

$$P(\hat{\sigma}^2) = P(z = \frac{\hat{\sigma}^2 n^2}{S_2})\frac{n^2}{S_2} = \frac{n^2}{S_2 2^{n/2}\Gamma(n/2)}(\frac{\hat{\sigma}^2 n^2}{S_2})^{n/2-1}e^{-\frac{\hat{\sigma}^2 n^2}{2S_2}} \tag{3.26}$$

However using the Bayesian approach, (with uniform prior, so that the maximum likelihood estimator is the same for frequentists and Bayesians) we would obtain that $\hat{\sigma}^2$ is distributed as

$$P(\hat{\sigma}^2) = \frac{N}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_i^n \frac{(x_i-\mu)^2}{\hat{\sigma}^2}\right] \tag{3.27}$$

where the normalization N (normalized in $\hat{\sigma}^2$, not in x_i!) is

$$N^{-1} = \int_0^\infty d\hat{\sigma}^2 (2\pi\hat{\sigma}^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_i^n \frac{(x_i-\mu)^2}{\hat{\sigma}^2}\right] = \frac{1}{2}\pi^{-n/2}S_2^{1-n/2}\Gamma(\frac{n}{2}-1) \tag{3.28}$$

By construction, the Bayesian PDFs has the peak (or *mode*) at the ML estimator, while, again by construction, the frequentist distribution has the *mean* at the ML estimator (since $\langle\hat{\sigma}^2\rangle = \sigma^2$ for fixed μ) and the mode at $(n-2)\sigma^2_{ML}/n$. The two distributions are clearly different (see Fig. 3.1). It is only in the $n \to \infty$ limit that they become identical Gaussian distributions. On the other hand, the frequentist and Bayesian distributions for the mean parameter $\hat{\mu}$ of Gaussian variables are identical (if the prior is uniform).

Let us conclude on a philosophical tone. One could say that the use of priors constitutes the whole difference between the Bayesian approach and the so-called *frequentist* one. The frequentist approach prefers not to deal with priors at all and therefore refuses to use Bayes' theorem to convert theoretical parameters into random variables. Once a frequentist finds a maximum likelihood estimator (which as any other estimator is a function of data and therefore is a random variable), he or she tries to determine its distribution as a function of the assumed distribution of the data. In most cases, this is done by generating numerically many mock datasets and calculating for each dataset the estimator, deriving then its approximate distribution. This Monte Carlo approach is the hallmark of the frequentist approach. It is powerful, objective and general but by rejecting priors fails to take into account previous knowledge. It is therefore suitable only when one can afford not to fully consider previous knowledge. This applies for instance when new experiments are much better than previous ones so that priors do not really matter and when each experiment measures only one parameter, say the mass of a particle, so that the outcome does not depend on other poorly measured parameters. Both features characterize most particle physics experiments and this explains why most particle physicists are frequentist. Astrophysics and cosmology live in another experimental world: data are hard to come by, observations cannot be twisted and repeated as easily as in a laboratory, models are characterized by many correlated parameters and every drop of previous information, even loosely related to a given parameter, has to be taken into account. Most of the evidence for dark energy comes from *combining* CMB and supernovae priors, each of them measuring many correlated parameters at once. It is no surprise that Bayesian methods are so popular in astrophysics and cosmology.

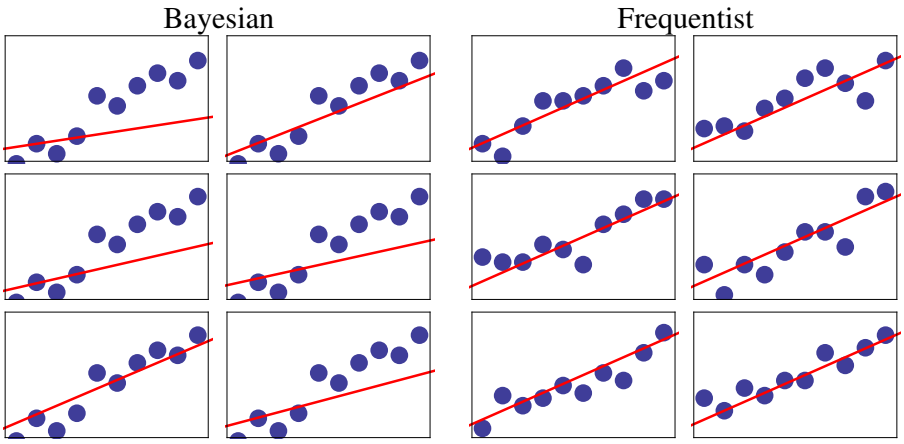

Figure 3.2: In the Bayesian view (left), the data are held fixed while the theoretical prediction, the red line, is distributed randomly according to the posterior; in the frequentist view (right), the data are random variables distributed according to the likelihood, while the theoretical line is held fixed.

3.4 Sampling the posterior

The posterior L is in general a surface in a N-dimensional parameter space. In order to evaluate its maximum and its general shape, one needs to estimate its value at many points in a high-dimensional space: this is called *sampling*. If N is large, its computation can be very lenghty. That is why maximum likelihood estimators have become popular only after the availablilty of fast computers in the last decades.

In general, there are two methods to evaluate the posterior: grids or random sampling. In a grid approach, one samples L in a regular or irregular grid of points, and then interpolates between the points. The grid can be equally spaced in each dimension, can have different spacings in different dimensions, can be irregular if one has reason to suspect that certain regions of the parameter space are more interesting than others, or can be adaptive, that is, automatically adjust so as to better sample the interesting regions near the peaks or in fast-varying zones. The problem with the grid approach is that if one needs e.g. 10 values per dimension, with say $N = 10$ dimensions, one has 10^{10} computations to perform. The complexity grows exponentially with the number of dimensions. On the other hand, the implementation is very simple, the geometry is under control, and the algorithm is very well parallelizable.

The random-based methods are generally called Monte Carlo methods. The idea here is that a grid wastes lots of time sampling uninteresting regions of the posterior, e.g. the tails far from the peak. A method that automatically samples better the peak regions and more sporadically the tail regions, might improve the speed by orders of magnitude. Most such models are based on Markov chains, which simply means that the i-th sampling point depends only on the $i-1$-th point. So we speak of Monte Carlo Markov chains, i.e. MCMC methods.

A typical and widely employed sampling scheme is called *Metropolis-Hastings*. The basic idea is very simple. Let us choose a random point $\boldsymbol{x}_0$ in the parameter space, i.e. a vector. The next point $\boldsymbol{x}_1 = \{x_{1i}\}$ will be chosen by drawing (typically) from a multidimensional Gaussian distribution

$$\exp -\frac{1}{2}\sum_i \frac{(x_{1i} - x_{0i})^2}{\sigma^2} \tag{3.29}$$

with some σ^2 that has to be chosen according to some criteria, as we discuss later. This distribution is called *proposal distribution*. In practice, a generator of random Gaussian numbers will give me the next point $\mathbf{x}_1$, given $\mathbf{x}_0$ and σ^2. Now we evaluate the posterior ratio

$$a = \frac{P(\boldsymbol{x}_1)}{P(\boldsymbol{x}_0)} \tag{3.30}$$

A very good property of this ratio is that is independent of the normalization of L, which is a computationally demanding operation. If this ratio is larger than 1, it means $\mathbf{x}_1$ is higher up than $\mathbf{x}_0$, so we are moving towards the

peak. This is good, and we move to $\mathbf{x}_1$. If $a < 1$, we move to $\mathbf{x}_1$ only $a\%$ of the times, and stay in $\mathbf{x}_0$ the remaining cases. That is, we generate a uniform random variable r between 0 and unity and move to $\mathbf{x}_1$ only if $r < a$, which of course happens $a\%$ of the time. If we stay in $\mathbf{x}_0$ we start over again by generating a new candidate point x_1, until we move to $\mathbf{x}_1$ for good. This procedure tends to move towards the peak but does not discard completely the tail regions. The expectation is that the peak regions are well sampled, while the tail regions are only sporadically sampled, which is what we want.

Now, the number of sample points in any given region x is proportional to the average $P(x)$ in that region. That is, the density of points is proportional to P. So, isodensity regions correspond to isoprobable values of the posterior, and the fraction of point inside these regions give the probability of finding the parameters in that contour (confidence regions or credible regions). The absolute normalization of the posterior never enters the algorithm. The marginalized probabilities are obtained simply by discarding from every sampling point the dimensions to be marginalized over; that is, if I have the collection of points $(x_1, y_1, z_1), (x_2, y_2, z_2)...$, the chain marginalized over z is simply $(x_1, y_1), (x_2, y_2)...$

All this is straightforward. However, there are several caveats. First, the initial point $\mathbf{x}_0$ can be really far from the peak, so it might take quite some time before the sampling starts climbing up, and there will be many points near $x_0$even if this point is far into the tail. To avoid this, the first n sampling points are discarded (*burn-in*). This value of n is arbitrary and typically of the order of thousands. Second, the value of σ in the proposal distribution is also arbitrary. One should aim at a value such that $30-50\%$ of the candidate x_{i+1} are accepted. This can be tested in a first phase of the chain and then tuned. Third, the chain should be stopped when finally the distribution does not change any longer (stationary distribution): criteria for this might vary and depend a bit on visual inspection. Fourth, sometimes the chain gets trapped near secondary peaks; the chain must then be restarted from a different point or σ^2 increased. Also, in some codes not every sampling point is accepted but only, say, every 10th or 100th, in order to reduce the correlation among points. All in all, there is a considerable amount of black magic in every MCMC code.

A variant of the Metropolis-Hastings is the Gibbs sampling. The idea behind the Gibbs sampling is that it is sometime easier to move (that means, faster to evaluate) with the Markov chain along fixed dimensions than in the full N-dimensional space. This also allows to choose a different σ^2 in the proposal distribution for each direction. The algorithm works as follows. When proposing a candidate $\mathbf{x}_{i+1}$ from a previous $\mathbf{x}_i$, we should come up with the N components of the $\mathbf{x}_{i+1}$ vector. Gibbs sampling chooses the first component of $\mathbf{x}_{i+1}$ the same way as in the Metropolis-Hastings method, as if the problem were unidimensional. Then we choose the second component, again with the Metropolis-Hastings method, but possibly with a different σ^2, keeping the first that we already obtained and all the subsequent ones fixed. That is, the components that come before the k-th component are updated, while those that come after and therefore still to be generated, remain the previous ones. Then we repeat until we have a full new vector. In other words, at every step, we choose the j-th component of the $i+1$ vector as a Metropolis-Hastings algorithm with the conditional likelihood

$$P(x_{i+1,j}|x_{i+1,1}, x_{i+1,2}, ..., x_{i,j+1}, x_{i,j+2}, ...) \tag{3.31}$$

All this is equivalent to performing the rejection-acceptance test in many unidimensional steps rather than all at once with the new point. Beside the advantage of tuning the σ^2 according to the dimension (the tuning might be achieved separately in each dimension adjusting σ^2 after a number of steps), the Gibbs sampling is particularly useful if along some parameter dimension the likelihood is particularly simple, for instance analytical.

Many other algorithms have been proposed. The goal is always the same: sample the posterior fast and accurately. The choice of the method depends a lot on what we know about, or reasonably expect from, the true posterior.

3.5 Fisher matrix

As straightforward and versatile as the likelihood method is, it is still often too complicated or computing-expensive to implement, especially when there are more than a few parameters involved. In fact there are some cases in which several tens or hundreds of parameters are present.

One could think that a model with more than 3 or 4 free parameters does not deserve the name of model and even less that one of "theory". However every theory begins by representing a vast dataset with a smaller set of numbers. And since cosmological experiments may easily collect terabytes of data, reducing them to 10, 100, or 1000 numbers should be seen already as a great progress towards a unified description (if there is one!).

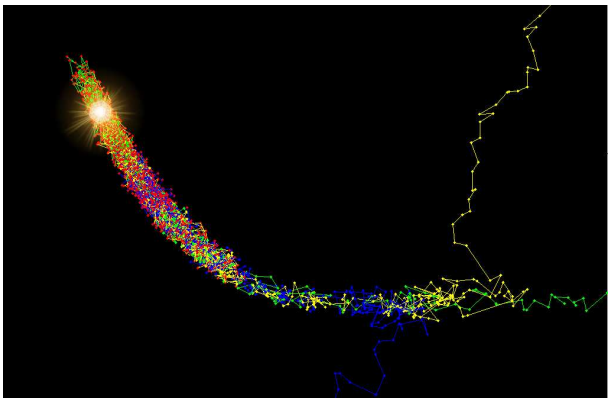

Figure 3.3: Distribution of points in a Metropolis-Hastings Markov chain. The initial burn-in phase is clearly visible on the right. The peak of the posterior i illuminated (from Wikipedia Commons, by Vasileios Zografos CC BY-SA 3.0.)

Anyway, the problem with the likelihood is that we need to evaluate $L(\theta_\alpha)$ for every θ_α, or at least for *many* θ_α, e.g., for a grid of, say, ten values for each dimension in parameters space. If there are 10 parameters, this means 10^{10} different evaluations. If each evaluation takes a second (say, a run of a CMB code), we are in for a waiting time of a 300 years...

One way out is to use a Monte Carlo approach, as we have already seen. Instead of building a full grid, one explores the landscape with random jumps. The size of the jumps in turn may be related to the steepness of the function (smaller jumps over rough terrain, larger ones over flatlands). This technique will grow with the number D of dimensions (parameters) as D, instead of exponentially as in full grid method. But this might still be a lot: a typical Markov chain exploration can take hundred of thousands of computations.

It is time to think of something faster: the Fisher matrix. The idea is straightforward: to approximate the full likelihood with a (multivariate) Gaussian distribution,

$$L \approx N \exp\left[-\frac{1}{2}(\theta_\alpha - \hat{\theta}_\alpha)F_{\alpha\beta}(\theta_\beta - \hat{\theta}_\beta)\right], \tag{3.32}$$

where the values $\hat{\theta}_\beta$, the maximum likelihood estimators, are function of the data, and $F_{\alpha\beta}$, the Fisher (or information) matrix, is the inverse of the correlation matrix among the parameters evaluated at $\hat{\theta}$. It is crucial to pay attention to the fact that the likelihood is now supposed to be a Gaussian function of the *parameters*, not (or not necessarily) of the data. We often assumed in the previous sections the data to be Gaussian but now we require the same for the parameters. The form (3.32) is of course a crude approximation. One could hope however that it is a reasonable approximation at least near the peak of the distribution, given that around a local maximum every smooth function (in this case $\ln L$) can be approximated as a quadratic function. Therefore we expect this approximation to work better for θ_α close to their estimators $\hat{\theta}_\alpha$.

If both the estimators *and* the data are Gaussian, then the estimators must be linear combinations of the data (as e.g. the mean is) and the frequentist distribution of the estimators is exactly Eq. (3.32). In this case, the estimators $\hat{\theta}$ have the same frequentist distribution as the Bayesian distribution of the theoretical parameters θ (if the prior is uniform) and the distinction between the two approaches becomes blurred.

Expanding the exponent of a generic likelihood up to second order near its peak (i.e. near the maximum likelihood (ML) value $\hat{\theta}_\alpha$ of the parameters) as

$$\ln L(\boldsymbol{x};\theta_\alpha) \approx \ln L(\hat{\theta}_\alpha) + \frac{1}{2}\frac{\partial^2 \ln L(\theta_\alpha)}{\partial\theta_\alpha \partial\theta_\beta}\bigg|_{\rm ML} (\theta_\alpha - \hat{\theta_\alpha})(\theta_\beta - \hat{\theta}_\beta)\,, \tag{3.33}$$

(naturally the first derivatives are absent because they vanish at the peak) we find, comparing with Eq. (3.32), that

the normalization $N = L(\hat{\theta}_i)$ depends only on the data and that the *Fisher matrix* (FM) is defined as

$$F_{\alpha\beta} \equiv -\frac{\partial^2 \ln L(\boldsymbol{x}_0;\boldsymbol{\theta})}{\partial\theta_\alpha \partial\theta_\beta}\bigg|_{\rm ML} . \tag{3.34}$$

where $\boldsymbol{x}_0$ are the particular set of observed data. If the prior is also Gaussian, the Fisher matrix for the posterior will be obtained by simply adding to $F_{\alpha\beta}$ the Fisher matrix for the prior, see below.

For Gaussian data we can write down the Fisher-approximated likelihood more explicitly. In this case in fact the peak of the likelihood $L(\hat{\theta}_\alpha)$ coincides with the smallest χ^2, so we can write

$$L(\boldsymbol{x};\theta_\alpha) \approx \frac{1}{(2\pi)^{N/2}|C|^{1/2}} e^{-\frac{1}{2}\chi^2_{min}} e^{-\frac{1}{2}(\theta_\alpha - \hat{\theta}_\alpha)F_{\alpha\beta}(\theta_\beta - \hat{\theta}_\beta)} \tag{3.35}$$

This expression will be repeatedly used in the following.

You may say now that in order to find the ML estimator we still have to build the full likelihood: does this again require the 10^{10} evaluations of $L(\theta_\alpha)$ that we mentioned above? Well, we could answer that there are fast numerical methods to search for maxima in a multi-dimensional function without spanning the whole parameter space. For instance, in one dimension, if we can guess that the parameter is near $\theta^{(0)}$ then we can expand the derivative of the log-likelihood $\mathcal{L} = -\ln L$ as follows

$$\mathcal{L}_{,\theta}(\theta) \approx \mathcal{L}_{,\theta}(\theta^{(0)}) + \mathcal{L}_{,\theta\theta}(\theta - \theta^{(0)}) , \tag{3.36}$$

and estimate the minimum of $\mathcal{L}$ (i.e. the maximum of L) by putting $\mathcal{L}_{,\theta}(\theta) = 0$. Then we find the approximation

$$\theta^{(1)} = \theta^{(0)} - \frac{\mathcal{L}_{,\theta}}{\mathcal{L}_{,\theta\theta}}\bigg|_{\theta^{(0)}} , \tag{3.37}$$

which could be iterated by assuming as new guess $\theta^{(1)}$ instead of $\theta^{(0)}$. This method, called Newton-Raphson or Levenbury-Marquardt, is extremely fast for well-behaved likelihood functions and can be directly generalized to the multi-dimensional case. However perhaps the most useful application of the Fisher formalism is to the cases in which we do not need to search for the likelihood peak because we already know from the start the ML estimator: when we are *simulating* an experiment.

In this case we want to produce an estimate of the covariance matrix of the parameters averaging over several possible future datasets. We need then the expected value of the FM over the data, i.e.

$$F_{\alpha\beta} \equiv -\left\langle \frac{\partial^2 \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\alpha\partial\theta_\beta} \right\rangle_{ML} = \left[-\int \frac{\partial^2 \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\alpha\partial\theta_\beta} L(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} \right]_{ML} . \tag{3.38}$$

Of course if the FM is constant, i.e. if the parameters appear linearly in the likelihood exponent, the result of the average is the same FM. We can also write

$$\begin{aligned} F_{\alpha\beta} &\equiv -\left\langle \frac{\partial^2 \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\alpha\partial\theta_\beta} \right\rangle_{ML} = -\left\langle \frac{\partial^2 L(\boldsymbol{x};\boldsymbol{\theta})}{L(\boldsymbol{x};\boldsymbol{\theta})\partial\theta_\alpha\partial\theta_\beta} \right\rangle_{ML} + \left\langle \frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\alpha} \frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\beta} \right\rangle_{ML} \\ &= \left\langle \frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\alpha} \frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\beta} \right\rangle_{ML} \end{aligned} \tag{3.39}$$

since $\langle \frac{L_{,\alpha\beta}}{L} \rangle = \int L_{,\alpha\beta} d^n x = (\int L d^n x)_{,\alpha\beta} = 0$.

Suppose we want to forecast how well a future supernovae experiment, which is supposed to collect $n = 10,000$ supernovae light curves and to derive their peak magnitude m_i with errors σ_i, is capable of constraining the cosmological parameters $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$. Let us start by assuming that the n random variables $m_i(z_i)$ follow a PDF with known variance σ_i and mean $m_{\rm th}(z_i;\Omega_m^{(0)},\Omega_\Lambda^{(0)}) = 5\log_{10} d_L(z_i;\Omega_m^{(0)},\Omega_\Lambda^{(0)}) + C$. Here we take the PDF to be Gaussian but we could also assume any other PDF that we have any reason to describe the data. Since the data PDF is assumed to be Gaussian we can immediately form the likelihood (neglecting the normalization constant):

$$L_m \approx \exp\left[-\frac{1}{2} \sum_i \frac{(m_i - m_{\rm th}(z_i))^2}{\sigma_i^2} \right] = \exp\left(-\frac{1}{2} \mu_i C_{ij}^{-1} \mu_j \right) . \tag{3.40}$$

Here we have expressed the argument of the exponential in a slightly more general way: we have introduced the vector $\mu_i \equiv m_i - m_{\rm th}(z_i)$ and the correlation matrix C_{ij}, that in this particular case is rather trivial

$$\boldsymbol{C} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2 ...)\,. \tag{3.41}$$

When we discuss dark energy, we are interested in the parameters such as $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$. So we wish to produce a likelihood function of $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$, something in the form of Eq. (3.32) like

$$L(\Omega_m^{(0)}, \Omega_\Lambda^{(0)}) = \exp\left[-\frac{1}{2}(\Omega_\alpha^{(0)} - \hat{\Omega}_\alpha^{(0)}) F_{\alpha\beta} (\Omega_\beta^{(0)} - \hat{\Omega}_\beta^{(0)})\right]\,, \tag{3.42}$$

where $F_{\alpha\beta}$ is of course our Fisher matrix and α, β run over the subscripts m, Λ. Since real data are not yet present, we do not have the ML estimators $\hat{\Omega}_\alpha^{(0)}$.However we are simulating the future experiment, so we may take for estimators the values $m_{\rm th}(z_i; \Omega_m^{(0)F}, \Omega_\Lambda^{(0)F})$ obtained using some fiducial cosmology $\Omega_m^{(0)F}, \Omega_\Lambda^{(0)F}$, for instance $\Omega_m^{(0)F} = 0.3, \Omega_\Lambda^{(0)F} = 0.7$. This means that we will find the confidence regions only around this particular parameter set. If we decide to change fiducial values, we have to redo our calculations and all our results will change in some way.

The Fisher matrix of the likelihood (3.40) is then, using (3.39):

$$F_{\alpha\beta} = -\left.\frac{\partial^2 \ln L_m}{\partial\Omega_\alpha^{(0)} \partial\Omega_\beta^{(0)}}\right|_F = \sum_i \frac{1}{\sigma_i^2} \left.\frac{\partial m_{\rm th}(z_n; \Omega_m^{(0)}, \Omega_\Lambda^{(0)})}{\partial\Omega_\alpha^{(0)}} \frac{\partial m_{\rm th}(z_n; \Omega_m^{(0)}, \Omega_\Lambda^{(0)})}{\partial\Omega_\beta^{(0)}}\right|_F\,. \tag{3.43}$$

Notice that $F_{\alpha\beta}$ is not diagonal even if the original correlation matrix C_{ij} was. Since the same $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$ appear in all $m_{\rm th}(z_n)$, we vary the likelihood of obtaining *all* m_i by varying $\Omega_{m,\Lambda}^{(0)}$. We can now use Eq. (3.42) to derive the confidence errors for $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$. In practice, what we have developed so far is a formalism to propagate the errors from the observational errors σ_i to the cosmological parameters. The errors σ_i, in turn, must be based on the expected performance of the experiment and often their derivation is the most complicated step, involving many fine details of the observations. Calculating numerically the second order partial derivatives in the Fisher matrix requires only a few estimations of the likelihood for each of the parameters; if we have 10 parameters this makes few tens of calculations instead of the 10^{10} we mentioned at the beginning of this section.

Once we have reduced our likelihood into a Gaussian form, the Fisher matrix is all we need to derive all the properties. The next section is concerned with various ways to manipulate the Fisher matrix to achieve several results.

3.6 Manipulating the Fisher matrix

Suppose we decide to switch from a set of parameters p_β to another one $q_\alpha(p_\beta)$, for instance from $\Omega_m^{(0)}, \Omega_\Lambda^{(0)}$ to the spatial curvature $\Omega_K^{(0)} = 1 - \Omega_m^{(0)} - \Omega_\Lambda^{(0)}$ and their ratio $R_{m\Lambda} = \Omega_m^{(0)}/\Omega_\Lambda^{(0)}$. If we know the Fisher matrix for p_i, the approximated likelihood is

$$L = \exp\left(-\frac{1}{2}\tilde{p}_\alpha F_{\alpha\beta}^{(x)} \tilde{p}_\beta\right)\,, \tag{3.44}$$

where $\tilde{p}_\alpha = x_\alpha - x_\alpha^{\rm ML}$. Approximating q_α near $q_\alpha^{\rm ML}$ as

$$q_\alpha \approx q_\alpha^{\rm ML} + \left.\frac{\partial q_\alpha}{\partial p_\beta}\right|_{\rm ML} (p_\beta - p_\beta^{\rm ML})\,, \tag{3.45}$$

where $q_\alpha^{\rm ML} \equiv q_\alpha(p^{\rm ML})$, we can write

$$\tilde{q}_\alpha \equiv q_\alpha - q_\alpha^{\rm ML} = J_{\alpha\beta}^{-1} \tilde{p}_\beta\,. \tag{3.46}$$

Here $J_{\alpha\beta} \equiv (\partial p_\alpha/\partial q_\beta)_{\rm ML}$ is the transformation Jacobian evaluated on the ML estimators. Then we have

$$\tilde{p}_\alpha = J_{\alpha\beta} \tilde{q}_\beta\,, \tag{3.47}$$

and we can find the new Fisher matrix by substituting into Eq. (3.44) simply as

$$F^{(y)}_{\alpha\beta} = J_{\alpha\gamma} F^{(x)}_{\gamma\sigma} J_{\sigma\beta}\,, \tag{3.48}$$

which is summed over indices. We can say that the Fisher matrix transforms as a tensor. Notice that the Jacobian matrix does not need to be a square matrix. The old parameters p_β can be projected in fact onto a smaller number of new parameters q_α.

One may wonder why the Jacobian does not enter also in the transformation from the volume element $\mathrm{d}p_i \mathrm{d}p_2 ...$ to the new element $\mathrm{d}q_i \mathrm{d}q_2 ...$, so that $L(q_\alpha) = |J| L[p_\beta(q_\alpha)]$. This would imply an additional logarithmic term $\ln |J|$ in the transformed probability function, spoiling the Gaussian approximation altogether. However near the ML values we can approximate $|J|$ with $|J_{\rm ML}|$ and include this constant factor in the overall normalization. That is, forget about it.

What if we want to *maximize* the likelihood with respect to some parameter? This means, if you remember, to fix one of the parameters to its maximum likelihood estimator. With the Fisher matrix this is really trivial, since fixing a parameter to its maximum likelihood estimator means putting the difference $\theta_\alpha - \hat{\theta}_i = 0$ and therefore to discard all entries in the Fisher matrix related to the i-th parameter. In practice, this means that one removes from the Fisher matrix the rows and columns of the maximized parameters.

What about *marginalization* then? Take a general 2-dimensional Gaussian PDF

$$G(x_1, x_2) = N \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} - 2\frac{\rho x_1 x_2}{\sigma_1 \sigma_2} \right) \right]\,, \tag{3.49}$$

where ρ is the correlation factor. This PDF can be written as

$$G(X_i) = N \exp\left[-\frac{1}{2} (X_i C^{-1}_{ij} X_j) \right]\,, \tag{3.50}$$

where $X_i \equiv x_i - \mu_i$ (generalizing to non-zero μ's), and

$$\boldsymbol{C} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\,. \tag{3.51}$$

Let us now evaluate the integral $\int G(x_1, x_2)\, \mathrm{d}x_2$ over the whole real domain. The result is given by

$$G(x_1) = \tilde{N} \exp[-x_1^2/(2\sigma_1^2)]\,, \tag{3.52}$$

where $\tilde{N}$ is a new normalization constant. The new correlation "matrix" is now simply $C_{11} = \sigma_1^2$.

In terms of the Fisher matrix $\boldsymbol{F} = \boldsymbol{C}^{-1}$ we see that the outcome of the marginalization has been the removal from $\boldsymbol{F}^{-1} = \boldsymbol{C}$ of the rows and columns related to the second parameter. This trick remains true for any number of dimensions: to marginalize over the j-th parameter, one simply needs to remove from the *inverse* of the Fisher matrix $\boldsymbol{F}^{-1}$ the j-th row and column; to marginalize at once over several parameters, one removes all the rows and columns related to those parameters. As a consequence, the diagonal of the inverse Fisher matrix contains the *fully-marginalized* 1σ errors of the corresponding parameters (i.e. the errors one gets on the α-th parameter after marginalizing over all the others)

$$\sigma_\alpha^2 = (\boldsymbol{F}^{-1})_{\alpha\alpha}\,. \tag{3.53}$$

This latter property is probably the most useful and time-saving feature of the whole Fisher method. Be warned however that the procedure of inverting and striking out rows and columns is in general numerically unstable if the matrix contains small eigenvalues. There are more stable algorithms that perform this operation.

Often we want to reduce the Fisher matrix to a 2×2 matrix $\boldsymbol{F}_2$ for two parameters, say θ_1, θ_2, because then it is easy to plot the resulting 2-dimensional confidence regions, defined as the regions of constant likelihood that contains a predetermined fraction of the total likelihood volume. Since the problem has been reduced from the start to gaussianity, we will necessarily have ellipsoidal confidence regions on the plane θ_1, θ_2. Looking at the form of the 2-dimensional Gaussian PDF (2.109), you will realize that the semiaxes of the ellipses are oriented along the eigenvectors of $\boldsymbol{F}_2^{-1}$, that is, they form an angle

$$\tan 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\,, \tag{3.54}$$

with the coordinate axes. Moreover, the semiaxes are proportional to the square root of the eigenvalues of $\mathbf{F}^{-1}$. The length of the semiaxes depends clearly on the level of confidence. If we take the semiaxes length along the i-th eigenvector equal to $\sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue, we are finding the 1σ region, but because we are in two dimensions, this level does not contain 68.3% of the probability but rather less than 40%. Instead, we find by integrating a 2-dimensional Gaussian that the *one-dimensional* "1σ" region corresponding to 68.3% of probability content is found for semiaxes which are roughly 1.51 times the square root of the eigenvalues. Regions at 95.4% and 99.7% correspond to semiaxes 2.49 and 3.44 times the the square root of the eigenvalues, respectively. The area of the 68.3% ellipses is πab, if a and b are the semiaxes length, that is 1.51 times the the square root of the eigenvalues. The area is therefore equal to $(1.51)^2\pi(\det \boldsymbol{F}_2)^{-1/2}$. Since an experiment is more constraining when the confidence region is smaller, one can define a simple but useful figure of merit (FOM) as

$$\mathrm{FOM} = \sqrt{\det \boldsymbol{F}_2}\,. \tag{3.55}$$

Notice however that the FOM is often defined to be the area at 95%, or some other similar but not equivalent choice.

The FOM is particularly relevant to dark energy parameters such as w_0, w_1 . The FOM naturally depends on how many parameters have been marginalized. Every parameter marginalization increases (or more exactly, does not reduce) the amount of uncertainty with respect to a maximized likelihood and therefore decreases the available information and the FOM of the final set of parameters.

All these simple rules are really good news for practical work. The bad news comes when they do not work. The major problem, in practice, is when the Fisher matrix itself is singular. Then there is no inverse and no marginalization. But the Fisher matrix can be singular only when rows or columns are not linearly independent. It is easy to see when this happens. If $L(\theta_1, \theta_2)$ depends on the two parameters through a constant combination, e.g., $a\theta_1 + b\theta_2$, then the Fisher matrix will be singular.

Let us turn this bug into a feature. If the Fisher matrix is singular, then it means that there is a linear combination of two or more parameters hidden somewhere in the likelihood. Therefore, we can substitute a new parameter $\hat{\theta}$ in place of that combination, e.g., $\hat{\theta} = a\theta_1 + b\theta_2$ and remove the singularity by restricting ourselves to $\hat{\theta}$ instead of the original pair. Actually we should have done this from the start, since if the physics depends only on the combination $a\theta_1 + b\theta_2$ there is no way we can distinguish between θ_1, θ_2. It is only this combination that matters and we should replace it by $\hat{\theta}$. We say in this case that there is a *degeneracy* between θ_1 and θ_2. Sometimes, however, it is not obvious at all that this was the case and the singularity of the Fisher matrix is a warning for us to look better or to find a prior (e.g., other experiments) that give separate information on one of the quasi-degenerate parameters and break the degeneracy.

This brings us to another advantage of the Fisher matrix approach. How do we add priors to a Fisher matrix F_{ij}? If the prior is the outcome of another experiment and we have the Fisher matrix $F^{(p)}_{\alpha\beta}$ of that experiment, then the problem reduces to multiplying a Gaussian likelihood by another Gaussian likelihood, obtaining a new Gaussian likelihood. If the experiments have the same ML estimators or the same fiducial model, as in the case in which we simulate them, the new Fisher matrix is given by

$$F^{(\mathrm{tot})}_{\alpha\beta} = F_{\alpha\beta} + F^{(p)}_{\alpha\beta}\,. \tag{3.56}$$

As simple as this: combining the information from two forecasts (with the same fiducial model) means summing their Fisher matrices. In so doing one has to ensure that the parameters and their order is exactly the same for both matrices: trivial, but a most likely source of practical confusion. If one of the experiments constrains only a subset of the total parameters (for instance, supernovae experiments do not constrain the primordial perturbation slope n_s), it means that it contains no information on that subset, and therefore the corresponding rows and columns are to be put to zero. This means that the two Fisher matrices are rendered of the same rank by filling the one with less parameters (say $\boldsymbol{F}^{(p)}$) with zeros in the correct position. For instance if we only want to add the information that the single m-th parameter comes with an error σ_m then we add the Fisher matrix (no sum on m)

$$F^{(p)}_{\alpha\beta} = \frac{\delta^m_\alpha \delta^m_\beta}{\sigma^2_m}\,. \tag{3.57}$$

So you see that in this case $\boldsymbol{F}^{(p)}$ would be utterly singular but the total $\boldsymbol{F}^{(\mathrm{tot})}$ is not (unless of course $\boldsymbol{F}$ was singular as well for the same parameter, bad luck really).

Let us mention the final point about the Fisher matrix. A statistical theorem known as Cramer-Rao inequality states that the variance of an unbiased estimator cannot be less than $(\boldsymbol{F}^{-1})_{\alpha\alpha}$ (which means first to take the inverse and *then* take the α-th term on the diagonal, i.e. our fully marginalized variances). In this sense the Fisher matrix gives the minimal error one can hope to achieve. If you are very optimist then the Fisher matrix is your tool. Notice, however, that the maximum likelihood estimators need not be unbiased estimators at all, although they are unbiased for large samples (asymptotically unbiased) otherwise they would be of little utility. So we could end up in producing the best possible error estimate for some unbiased estimators which we do not know how to determine!

Once we accept the Gaussian approximation, the Fisher matrix embodies all the information we have on the problem. The manipulation of the Fisher matrix therefore is all we need. To recapitulate, there are five golden rules of *fisherology*:

1. To *transform* variables, multiply the Fisher matrix on the right and on the left by the transformation Jacobian.
2. To *maximize* over some parameters, remove from the matrix the rows and the columns related to those parameters.
3. To *marginalize* over some parameters, remove from the *inverse* matrix the rows and the columns related to those parameters (being careful about the numerical instability pointed out above).
4. To *combine* Fisher matrices from independent experiments with the same fiducial, sum the corresponding Fisher matrices, ensuring the same order of parameters, and, if necessary, inserting rows and columns of zeros for unconstrained parameters.
5. The *ellipsoidal confidence regions* have semiaxes lengths equal to the square root of the eigenvalues of the *inverse* Fisher matrix, while the semiaxes are oriented along the corresponding eigenvectors. The *area* of the ellipse (or volume of ellipsoid) is proportional to the square root of the determinant of the inverse Fisher matrix. The determinant of the Fisher matrix is an indicator of performance or a figure of merit.

If one wishes, one could define a new set of parameters by diagonalizing the Fisher matrix, obtaining circular (or spherical) confidence regions. In some cases this is useful because it reveals hidden properties (see Sec. 4.4). There are other cases in which the new parameters are so remote from any physical direct meaning that the exercise is futile. Notice that the confidence region volume (and therefore the FOM) does not change under the diagonalization.

3.7 An application to cosmological data

Let us apply the transformation technique to an interesting problem. In cosmology one uses extensively the parametrization around $a_0 = 1$ of the equation of state $w_{\rm DE}(a) = w_0 + w_1(1-a)$. We could however have expanded $w_{\rm DE}(a)$ around any other point a_p and write instead $w_{\rm DE}(a) = w_p + w_1(a_p - a)$, where

$$w_p = w_0 + w_1(1-a_p)\,. \tag{3.58}$$

We can now ask the question whether the constraint we obtain on w_p (i.e. $\sigma^2_{w_p}$) is tighter than the one on w_0, that is whether we can better rule out say $w_{\rm DE} = -1$ at a_p than at a_0. The problem consists therefore in finding the value a_p (called *pivot point*) that minimizes the variance of $w_{\rm DE}(a)$. Denoting the maximum likelihood estimators (or fiducial values) with $\hat{w}_0, \hat{w}_1$, this occurs for the value of a which is the solution of the following equation,

$$\begin{aligned}\frac{\rm d}{{\rm d}a}\left[\langle[(w_0-\hat{w}_0)+(1-a)(w_1-\hat{w}_1)]^2\rangle\right] &= \frac{\rm d}{{\rm d}a}\left[\sigma^2_{w_0}+(1-a)^2\sigma^2_{w_1}+2(1-a)\rho\sigma_{w_0}\sigma_{w_1}\right] \\ &= -2(1-a)\sigma^2_{w_1}-2\rho\sigma_{w_0}\sigma_{w_1}=0\,. \end{aligned}\tag{3.59}$$

Here $\sigma^2_{w_i} \equiv \langle(w_i-\hat{w}_i)^2\rangle$ for $i=0,1$ and $\rho \equiv \langle(w_0-\hat{w}_0)(w_1-\hat{w}_1)\rangle/(\sigma_{w_0}\sigma_{w_1})$ is the correlation coefficient. Then we obtain

$$a_p = 1 + \frac{\rho\sigma_{w_0}}{\sigma_{w_1}}\,. \tag{3.60}$$

In terms of the two-dimensional Fisher matrix F_{ij} for w_0, w_1, we can write

$$\sigma^2_{w_0} = (\boldsymbol{F}^{-1})_{11}\,, \qquad \sigma^2_{w_1} = (\boldsymbol{F}^{-1})_{22}\,, \qquad \rho\sigma_{w_0}\sigma_{w_1} = (\boldsymbol{F}^{-1})_{12}\,. \tag{3.61}$$

The transformation from $\boldsymbol{p} = (w_0, w_1)$ to $\boldsymbol{q} = (w_p, w_1)$ is achieved by using Eq. (3.48) with the transformation matrix

$$\boldsymbol{J} = \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}} = \begin{pmatrix} 1 & a_p - 1 \\ 0 & 1 \end{pmatrix} . \tag{3.62}$$

It is straightforward to verify that with this transformation the new matrix $\boldsymbol{F}_p = \boldsymbol{J}^t \boldsymbol{F} \boldsymbol{J}$ is diagonal (the superscript t denotes transpose) and its inverse is:

$$\boldsymbol{F}_p^{-1} = \begin{pmatrix} \sigma_{w_0}^2 (1-\rho^2) & 0 \\ 0 & \sigma_{w_1}^2 \end{pmatrix} . \tag{3.63}$$

The parameters w_p, w_1 are therefore uncorrelated and their confidence regions are circular. Moreover, as expected, the error on w_p, $\sigma_{w_p}^2 \equiv \sigma_{w_0}^2 (1-\rho^2)$, is always smaller than $\sigma_{w_0}^2$.

3.8 The Fisher matrix for the power spectrum

Now we have all the tools to derive a very useful result, the Fisher matrix for an experiment that measures the galaxy power spectrum.

Suppose a future experiment will provide us with the Fourier coefficients $\delta_{\boldsymbol{k}}$ of a galaxy distribution and their power spectrum calculated for a set of m wavenumbers $\boldsymbol{k}_i$ in some redshift bin $z, z + \Delta z$. Our theory predicts the spectrum $P(k, z; p_\alpha)$ as function of, say, $p_\alpha \equiv \Omega_m^{(0)}, \Omega_b^{(0)}, h, n_s$ etc. In any real survey with a galaxy density $n(z)$, however, the power spectrum will include the Poisson noise part :

$$\Delta_{\boldsymbol{k}}^2 \equiv \langle \delta_{\boldsymbol{k}} \delta_{\boldsymbol{k}}^* \rangle = \langle \delta_{\boldsymbol{k}} \delta_{-\boldsymbol{k}} \rangle = P(\boldsymbol{k}, z) + \frac{1}{n} . \tag{3.64}$$

Since the average galaxy density is estimated from the survey itself we have by construction $\langle \delta(x) \rangle = 0$ and therefore $\langle \delta_{\boldsymbol{k}_i} \rangle = 0$ for any $\boldsymbol{k}_i$. The coefficients $\delta_{\boldsymbol{k}_i}$ are complex variables in which the real and imaginary parts obey the same Gaussian statistics. So now we calculate the Fisher matrix for only, say, the real parts of $\delta_{\boldsymbol{k}_i}$ and the Fisher matrix for the whole $\delta_{\boldsymbol{k}_i}$ is simply the sum of two identical Fisher matrices, i.e. twice the result for the real parts. However when we count the total number of independent modes we have to remember that only half of them are statistically independent since $\delta_{\boldsymbol{k}}^* = \delta_{-\boldsymbol{k}}$ so in fact we should finally divide by two the final result. That is, we can forget both factors.

If we assume the galaxy distribution to be well approximated by a Gaussian we can write the likelihood:

$$L = \frac{1}{(2\pi)^{m/2} \Pi_i \Delta_i} \exp \left[-\frac{1}{2} \sum_i^m \frac{\delta_i^2}{\Delta_i^2} \right] , \tag{3.65}$$

(where to simplify notation we write $\Delta_i = \Delta_{\boldsymbol{k}_i}$, $\delta_i = \mathrm{Re}\, \delta_{\boldsymbol{k}_i}$) assuming that the measures at every $\boldsymbol{k}_i$ are statistically independent. When we simulate a future experiment, $P(k, z)$ is taken to be the theoretical spectrum of our fiducial model described by the parameters $p_\alpha^{(F)}$. Then we have

$$\mathcal{L} = -\ln L = \frac{m}{2} \ln(2\pi) + \sum_i \ln \Delta_i + \sum_i \frac{\delta_i^2}{2\Delta_i^2} . \tag{3.66}$$

We further simplify the notation by suppressing the index i running over the k bins from Δ_i, δ_i and denote the differentiation with respect to the α-th parameter as $\Delta_{,\alpha}$. Now from Eq. (3.34) the Fisher matrix for a particular z bin is

$$\begin{aligned} F_{\alpha\beta} &= \left\langle \frac{\partial^2 \mathcal{L}}{\partial p_\alpha \partial p_\beta} \right\rangle = \sum_i \left[\frac{\Delta_{,\alpha\beta}}{\Delta} - \frac{\Delta_{,\alpha} \Delta_{,\beta}}{\Delta^2} - \langle \delta^2 \rangle \left(\frac{\Delta_{,\alpha\beta}}{\Delta^3} - 3 \frac{\Delta_{,\alpha} \Delta_{,\beta}}{\Delta^4} \right) \right] \\ &= \frac{1}{2} \sum_i \frac{\partial \ln P_i}{\partial p_\alpha} \frac{\partial \ln P_i}{\partial p_\beta} \left(\frac{n P_i}{1 + n P_i} \right)^2 , \end{aligned} \tag{3.67}$$

[where we used $\langle \delta^2 \rangle = \Delta^2$ from Eq. (3.64)] calculated on the fiducial model.

For a more compact expression we can now approximate the sum with an integral over k. To do this we need to count how many modes lie in the bin defined by the modulus interval $k, k + \mathrm{d}k$ and cosine interval $\mathrm{d}\mu$,, i.e. in the Fourier volume $2\pi k^2 \mathrm{d}k\mathrm{d}\mu$. The number of modes we can really use is limited by two factors: the size of the volume and the shot noise. Modes larger than the survey volume cannot be measured. Short modes sampled by only a few galaxies cannot be reliably measured either.

To take into account these limitations we discretize the Fourier space into cells of volume $V_{\rm cell} = (2\pi)^3/V_{\rm survey}$, so that we have $2\pi k^2 \mathrm{d}k\mathrm{d}\mu/V_{\rm cell} = (2\pi)^{-2} V_{\rm survey} k^2 \mathrm{d}k\mathrm{d}\mu$ modes in the survey volume. The integral form of the Fisher matrix is therefore given by

$$F_{\alpha\beta} = \frac{1}{8\pi^2} \int_{-1}^{+1} \mathrm{d}\mu \int_{k_{\rm min}}^{k_{\rm max}} k^2 \mathrm{d}k \frac{\partial \ln P(k,\mu)}{\partial p_\alpha} \frac{\partial \ln P(k,\mu)}{\partial p_\beta} \left[\frac{nP(k,\mu)}{nP(k,\mu)+1}\right]^2 V_{\rm survey}\,. \tag{3.68}$$

The factor

$$V_{\rm eff} = \left[\frac{nP(k,\mu)}{nP(k,\mu)+1}\right]^2 V_{\rm survey}\,, \tag{3.69}$$

can be seen as an effective survey volume. When $nP \gg 1$ the sampling is good enough to derive all the cosmological information that can be extracted from the survey and there is no need of more sources. For $nP \ll 1$ the effective volume is severely reduced. If we subdivide the data into several z independent bins, we can simply sum the Fisher matrices for every bin.

3.9 The Fisher matrix for general Gaussian data

It is straightforward to extend the Fisher matrix calculation to a more general Gaussian likelihood with full correlation. Consider a set of n Gaussian data $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$ distributed according to the likelihood

$$L = \frac{1}{(2\pi)^{n/2}\sqrt{\det \boldsymbol{C}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]\,, \tag{3.70}$$

where t denotes the transpose. We define the data matrix $\boldsymbol{D} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^t$. Then the covariance matrix is defined in all generality as the expected value of $\boldsymbol{D}$:

$$\langle \boldsymbol{D} \rangle = \boldsymbol{C}\,. \tag{3.71}$$

We can write, up to a constant

$$\mathcal{L} = -\ln L = \frac{1}{2}[\ln\det \boldsymbol{C} + \mathrm{Tr}\, \boldsymbol{C}^{-1}\boldsymbol{D}] = \frac{1}{2}\mathrm{Tr}\,[\ln \boldsymbol{C} + \boldsymbol{C}^{-1}\boldsymbol{D}]\,, \tag{3.72}$$

where we used the matrix identity: $\ln\det \boldsymbol{C} = \mathrm{Tr}\ln \boldsymbol{C}$. We suppose now that the theoretical parameters $\boldsymbol{\theta}$ are *both* in $\boldsymbol{\mu}$ and in $\boldsymbol{C}$. The Fisher matrix is then the expected value

$$F_{\alpha\beta} = \left\langle \frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle \equiv \langle \mathcal{L}_{,\alpha\beta} \rangle\,, \tag{3.73}$$

To calculate $\langle \mathcal{L}_{,\alpha\beta}\rangle$ we use the fact that for Gaussian data $\langle \boldsymbol{x}\rangle = \boldsymbol{\mu}$, and consequently

$$\langle \boldsymbol{D}_{,\alpha}\rangle = 0\,, \quad \langle \boldsymbol{D}_{,\alpha\beta}\rangle = \boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}^t_{,\beta} + \boldsymbol{\mu}_{,\beta}\boldsymbol{\mu}^t_{,\alpha} = 2\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}^t_{,\beta}\,. \tag{3.74}$$

Notice that $\langle \boldsymbol{D}_{,\alpha}\rangle \neq \langle \boldsymbol{D}\rangle_{,\alpha}$. Then we have (since $(C^{-1})_{,\alpha} = -C^{-1}C_{,\alpha}C^{-1}$)

$$2\mathcal{L}_{,\alpha} = \mathrm{Tr}\,[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}(\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}) + \boldsymbol{C}^{-1}\boldsymbol{D}_{,\alpha}]\,, \tag{3.75}$$

($\boldsymbol{I}$ is the identity matrix) which averages to zero,

$$\langle \mathcal{L}_{,\alpha}\rangle = 0\,. \tag{3.76}$$

This result is actually true for any distribution, not just Gaussian, since it corresponds to the derivative with respect to the parameters of the norm of the distribution. Notice that the average only acts on $\boldsymbol{D}$ since the random

variables, the data, are only there, while of course derivatives act only on $\boldsymbol{C}$ and $\boldsymbol{\mu}$ since parameters are only there. To evaluate $\langle \mathcal{L}_{,\alpha\beta} \rangle$ we notice that all first derivatives $\langle \boldsymbol{D}_{,\alpha} \rangle$ vanish and that $\langle \boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D} \rangle = 0$. Then we are finally left with

$$F_{\alpha\beta} \equiv \langle \mathcal{L}_{,\alpha\beta} \rangle = \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta} + \boldsymbol{C}^{-1}\langle \boldsymbol{D}_{,\alpha\beta} \rangle\right] = \frac{1}{2}C^{-1}_{\ell m}\frac{\partial C_{mn}}{\partial\theta_\alpha}C^{-1}_{np}\frac{\partial C_{p\ell}}{\partial\theta_\beta} + C^{-1}_{\ell m}\frac{\partial\mu_\ell}{\partial\theta_\alpha}\frac{\partial\mu_m}{\partial\theta_\beta}\,, \tag{3.77}$$

(sum over repeated indices) where in the last equality we have written down the full index expression to be more explicit. Equation (3.67) is recovered when $\boldsymbol{\mu} = \boldsymbol{0}$ and $C_{\ell m} = \Delta^2_m\delta_{\ell m}$. This expression is extremely useful and allows for an impressive range of applications.

3.10 Model selection

So far we have been working within a given model. When we choose a model to test, we also select some free functions that define the model and that we parametrize in some convenient way. If we decide to change a model, e.g., from the uncoupled dark energy model with $w_{\rm DE} =$ constant to a specific $f(R)$ model, we have to start a new process so that the likelihood will give us a new set of best fit parameters. But how do we decide whether the $f(R)$ model is better than the dark energy model with $w_{\rm DE} =$ constant?

This is a problem of *model selection*, rather than model optimization. One possibility (the *frequentist* approach) is to simply evaluate the "goodness of fit": once we have the best fit parameters for models A and B, we calculate the χ^2 statistics of the model prediction with respect to data and choose the one with better χ^2 statistics (which is not necessarily the one with lowest χ^2 because the χ^2 statistics depends also on the number of degrees of freedom, namely on the number of independent data minus the number of free parameters). Beside the intrinsic problem of any frequentist approach (e.g., lack of priors), this is often too rough a guide to selection, mostly because if the model B includes a parameter that is poorly constrained by the data it would not help in the fit but it would still be counted as an extra degree of freedom and this would unfairly penalize it. Imagine for instance two very similar dark energy models, A and B, with two parameters each. Suppose that the model B predicts some peculiar feature at the redshift $z = 3$, e.g., cluster abundance, and that feature depends on a third parameter. The model B is interesting also because of this unique prediction but it would be unfairly penalized by current constraints, since we have very limited knowledge of high-redshift clusters so far. A χ^2 test would presumably conclude that the model A fits existing data as well as the model B but with one parameter less and therefore it would win.

To overcome this problem we can instead use another model selection procedure, called *evidence* or marginal likelihood. Let us consider again Bayes theorem for the parameters (θ) and data (D), and let us add now the specification that the probabilities are taken *given a model* M

$$P(\theta; D, M) = \frac{L(D;\theta, M)p(\theta; M)}{E(D; M)} \tag{3.78}$$

This form can be obtained by combining the probabilities of obtaining the data and the parameters given a model

$$P(D, \theta; M) = P(D;\theta, M)P(\theta; M) \tag{3.79}$$
$$P(\theta, D; M) = P(\theta; D, M)P(D; M), \tag{3.80}$$

and equaling the two since "data and parameters" equals "parameters and data".

By integration of eq. (3.78) over the parameters we obtain the normalization unity on the rhs, from which the *evidence*

$$E(\boldsymbol{x}; M) = \int f(\boldsymbol{x}; \theta^M_\alpha)p(\theta^M_\alpha)\,\mathrm{d}^n\theta^M_\alpha\,, \tag{3.81}$$

where as before $\boldsymbol{x} = (x_1, x_2, ...)$ are random data, θ^M_α are n theoretical parameters that describe the model M, f is the likelihood function, and p is the prior probability of the parameter θ^M_α. Note that we have added a superscript M to remember that the parameters refer to some model M. One can see that the evidence is then the likelihood averaged over the entire parameter space.

Now if we have any reason to weigh the models in some way, we can assign a model prior $p(M_j)$ and use Bayes' theorem again to write

$$L(M; \boldsymbol{x}) = E(\boldsymbol{x}; M)\frac{p(M)}{p(\boldsymbol{x})}\,, \tag{3.82}$$

i.e. the probability of having model M given the data. We can finally use this probability to compare quantitatively two models taking the ratio of probabilities (so that $p(\boldsymbol{x})$ cancels out):

$$\frac{L(M_1;\boldsymbol{x})}{L(M_2;\boldsymbol{x})} = B_{12}\frac{p(M_1)}{p(M_2)}\,. \tag{3.83}$$

where we introduced the Bayes ratio or odds

$$B_{12} = \frac{\int f(\boldsymbol{x};\theta_\alpha^{M_1})p(\theta_\alpha^{M_1})\mathrm{d}^n\theta_\alpha^{M_1}}{\int f(\boldsymbol{x};\theta_\alpha^{M_2})p(\theta_\alpha^{M_2})\mathrm{d}^n\theta_\alpha^{M_2}}\,. \tag{3.84}$$

Often, however, one assumes that $p(M_1) = p(M_2)$. A Bayes ratio $B_{12} > 1$ (< 1) says that current data favors the model M_1 (M_2).

Suppose now that a certain parameter θ_n is very poorly constrained by the data x_i. This implies that the likelihood $f(x_i;\theta_\alpha)$ is practically independent of θ_n, that is, f remains almost constant when varying θ_n. Then if the prior is factorizable (which is often the case) so that $p(\theta_\alpha) = \Pi_\alpha p_\alpha(\theta_\alpha)$, we see that the integral over θ_n decouples. Since the priors are just standard probability distribution functions we have $\int p_n(\theta_n)\mathrm{d}\theta_n = 1$, so that as expected θ_n does not enter the evidence integral. The evidence therefore correctly discards poorly constrained parameters and does not penalize models for introducing them. The blame is where it belongs: poor data.

If the likelihood and the prior can both be approximated by Gaussian distributions *in the parameters*, we can evaluate the evidence analytically. Let us assume then an uncorrelated Gaussian likelihood with best fit parameters $\theta_\alpha^{(B)}$ and variances $\sigma_{B,i}$ and an uncorrelated Gaussian prior with means $\theta_\alpha^{(P)}$ and variances $\sigma_{P,i}$. The posterior can be written as

$$\begin{aligned}
L(\theta_\alpha) &= \prod_\alpha f(\boldsymbol{x};\theta_\alpha)p(\theta_\alpha) \\
&= L_{max}\prod_\alpha (2\pi\sigma_{P,\alpha}^2)^{-1/2}\exp\left[-\frac{(\theta_\alpha-\theta_\alpha^{(B)})^2}{2\sigma_{B,\alpha}^2} - \frac{(\theta_\alpha-\theta_\alpha^{(P)})^2}{2\sigma_{P,\alpha}^2}\right] \\
&= L_{max}\prod_\alpha (2\pi\sigma_{P,\alpha}^2)^{-1/2}\exp\left[-\frac{1}{2}\frac{(\theta_\alpha-\theta_\alpha^*)^2}{\sigma_{\alpha*}^2}\right]\exp\left[-\frac{1}{2}\frac{(\theta_\alpha^{(B)}-\theta_\alpha^{(P)})^2}{\sigma_{B,\alpha}^2+\sigma_{P,\alpha}^2}\right]\,,
\end{aligned} \tag{3.85}$$

where $L_{\max}$ is the likelihood maximum and where the posterior mean and variance for each i are

$$\theta_\alpha^* = \frac{\sigma_{B,\alpha}^2\theta_\alpha^{(P)} + \sigma_{P,\alpha}^2\theta_\alpha^{(B)}}{\sigma_{B,\alpha}^2+\sigma_{P,\alpha}^2}\,, \tag{3.86}$$

$$\sigma_{\alpha*}^2 = \frac{\sigma_{P,\alpha}^2\sigma_{B,\alpha}^2}{\sigma_{B,\alpha}^2+\sigma_{P,\alpha}^2}\,. \tag{3.87}$$

The evidence is therefore

$$\begin{aligned}
E &= \int\prod_\alpha f(\boldsymbol{x};\theta_\alpha)p(\theta_\alpha)\mathrm{d}\theta_\alpha \\
&= L_{max}\prod_\alpha\frac{\sigma_{\alpha*}}{\sigma_{P,\alpha}}\exp\left\{-\frac{1}{2}\left[\left(\frac{\theta_\alpha^{(B)}}{\sigma_{B,\alpha}}\right)^2 + \left(\frac{\theta_\alpha^{(P)}}{\sigma_{P,\alpha}}\right)^2 - \left(\frac{\theta_\alpha^*}{\sigma_{\alpha*}}\right)^2\right]\right\}\,.
\end{aligned} \tag{3.88}$$

If the data x_i are Gaussian with mean $\hat{x}_i$ and correlation matrix C_{ij} then (see Eq. 3.35)

$$L_{max} = N_L e^{-\frac{\chi^2_{min}}{2}} \tag{3.89}$$

where N_L is the likelihood normalization (independent of the model) and χ^2_{min} is the minimum of $(x_i-\hat{x}_i)C_{ij}^{-1}(x_j-\hat{x}_j)$, i.e. the usual χ^2 minimum.

We see that the evidence is determined by three factors. They embody three requirements on what a good model should do:

1. it should give a good fit to the data,

2. it should do so with a small number of parameters with respect to the data to be fitted,

3. it should not be too discordant with the prior.

Indeed, the first factor in Eq. (3.88), $L_{\max}$, is the likelihood maximum and expresses how well the model fits the data. In forming the Bayes ratio one would get the likelihood ratio, which is the basis of the frequentist approach to model selection (i.e. for Gaussian distributed variables $-2\log(L_{max}/N_L) = \chi^2$). The second factor is a ratio of parameter volumes: if we take the variance as a measure of the available parameter space for the i-th parameter, this factor expresses how the parameter volume changes from the prior to the posterior. Every factor $\sigma_{\alpha *}/\sigma_{P,\alpha} = \sigma_{B,\alpha}/(\sigma_{B,\alpha} + \sigma_{P,\alpha})^{1/2}$ is smaller than unity, so adding more parameters penalizes the evidence, quantifying Occam's razor argument. If however the data do not constrain the i-th parameter, i.e. if $\sigma_{B,\alpha} \gg \sigma_{P,\alpha}$, then the α-th factor $\sigma_{\alpha *}/\sigma_{P,\alpha}$ is close to unity and there is no penalization. Finally the third factor (the exponential) penalizes the evidence if the best-fit α-th parameter or the prior mean differs appreciably from the posterior mean θ^*_α: although the new data might justify that parameter, the overall agreement including the prior does not seem to require it. The model is then penalized because of its inconsistency with the prior. Here again, if data constraints are very weak (large $\sigma_{B,\alpha}$) then there is no penalization.

It is a matter of straightforward algebra to extend the expression to correlated Gaussian parameters. If the evidence integral is

$$\begin{aligned} E &= \int f(\boldsymbol{x};\theta_\alpha)p(\theta_\alpha)\mathrm{d}\theta_\alpha \\ &\approx N_L e^{-\frac{\chi^2_{min}}{2}} \int \exp\left[-\frac{1}{2}(\theta_\alpha - \theta^{(B)}_\alpha)L_{\alpha\beta}(\theta_\beta - \theta^{(B)}_\beta) - \frac{1}{2}(\theta_\alpha - \theta^{(P)}_\alpha)P_{\alpha\beta}(\theta_\beta - \theta^{(P)}_\beta)\right]\mathrm{d}\theta_\alpha\,, \end{aligned} \tag{3.90}$$

where $\theta^{(B)}_\alpha$ are the best fit estimates, $\theta^{(P)}_\alpha$ are the prior means, $L_{\alpha\beta}$ in the exponential factor is the inverse of the covariance matrix of the likelihood (i.e. the Fisher matrix) and $P_{\alpha\beta}$ is the inverse of the covariance matrix of the prior, we obtain

$$E = N_L e^{-\frac{\chi^2_{min}}{2}} \frac{|\boldsymbol{P}|^{1/2}}{|\boldsymbol{F}|^{1/2}} \exp\left[-\frac{1}{2}(\theta^{(B)}_\alpha L_{\alpha\beta}\theta^{(B)}_\beta + \theta^{(P)}_\alpha P_{\alpha\beta}\theta^{(P)}_\beta - \tilde{\theta}_\alpha F_{\alpha\beta}\tilde{\theta}_\beta)\right]\,, \tag{3.91}$$

where $\boldsymbol{F} = \boldsymbol{P} + \boldsymbol{L}$ and $\tilde{\theta}_\alpha = (\boldsymbol{F}^{-1})_{\alpha\beta}[L_{\beta\gamma}\theta^{(B)}_\gamma + P_{\beta\gamma}\theta^{(P)}_\gamma]$. If the prior is very weak the final exponential term reduces to unity and for Gaussian data we can write the Bayes ratio as

$$B_{AB} = e^{-\frac{1}{2}(\chi^2_A - \chi^2_B)} \frac{|P_A F_B|^{1/2}}{|F_A P_B|^{1/2}} \tag{3.92}$$

where $\chi^2_{A,B}$ are the minimum χ^2.

Example 1. We want to find the Bayes ratio for two models: model A predicting that a quantity $\theta = 0$ with no free parameters, and model B which assigns θ a Gaussian prior distribution with zero mean and variance Σ^2. Therefore the prior of model A is a Dirac δ_D function centered on $\theta = 0$, whereas the prior of model B is $e^{-\theta^2/2\Sigma^2}/\sqrt{2\pi\Sigma^2}$. Notice that here the models differ in their prior, rather than in the parameters. We assume that we performed a measurement of θ described by a normal likelihood of standard deviation σ, and with the maximum likelihood value lying λ standard deviations away from 0, i.e. $|\theta_{\rm ML}/\sigma| = \lambda$. The data are described by a Gaussian $e^{-(\theta-\theta_{\rm ML})^2/2\sigma^2}$ with $\theta_{\rm ML} = \lambda\sigma$. We calculate Bayes' ratio as

$$\begin{aligned} B_{AB} &= \frac{\int f(\boldsymbol{x};\theta^{M_1}_\alpha)p(\theta^{M_1}_\alpha)\mathrm{d}\theta^{M_1}_\alpha}{\int f(\boldsymbol{x};\theta^{M_2}_\alpha)p(\theta^{M_2}_\alpha)\mathrm{d}\theta^{M_2}_\alpha} \\ &= \frac{\int e^{-(\theta-\theta_{\max})^2/2\sigma^2}\delta(\theta)\mathrm{d}\theta}{(2\pi\Sigma^2)^{1/2}\int e^{-(\theta-\theta_{\max})^2/2\sigma^2}e^{-\theta^2/2\Sigma^2}\mathrm{d}\theta} \\ &= \sqrt{1+r^{-2}}\, e^{-\frac{\lambda^2}{2(1+r^2)}}\,, \end{aligned} \tag{3.93}$$

where $r = \sigma/\Sigma$. We can identify the limiting cases:

- If the best-fit parameter $\theta_{\max}$ is many σ away from the predicted $\theta = 0$ (i.e. $\lambda \gg 1$), then it follows that $B_{AB} \ll 1$, favoring model B that allows for the extra freedom Σ.
- If λ is not too large and $r \ll 1$, i.e. the data is much more peaked than the B prior and close to the predicted value, then we have $B_{AB} \approx 1/r \gg 1$ so that the extra parameter introduced by model B is not needed and A is favored. This is in touch with Occam's razor argument.
- If $r \gg 1$, then $B_{AB} \approx 1$ and hence there is not enough information to decide between A and B. Although B has more parameters, the fact that the data have a large error and are too poor to constrain θ implies that no preference must be given to either A or B.

The evidence is often not easy to evaluate because it requires a multidimensional integration over the whole parameter space. Several approximation or alternative model selection techniques have been proposed (see for instance the excellent review by Trotta [1]). They are however only justified in specific cases and may give conflicting results, sometimes leading to controversies. Whenever possible, the evidence integral should be used instead.

Let us now come back to the Bayes factors, i.e. the ratio of the evidences. Once we have calculated this ratio we are still to decide how to gauge it in favor of the model A or B. There is no absolute way to achieve this: large or small factors should incline us towards one of the two models over the other one, but there is no absolute "statistics" to associate to any specific level. The scale most used in literature is called Jeffreys' scale. If $|\ln B_{12}| < 1$ there is no evidence in favor of any of the models ("inconclusive evidence"); if $|\ln B_{12}| > 1$ there is a "weak evidence"; $|\ln B_{12}| > 2.5$ means "moderate evidence"; $|\ln B_{12}| > 5$ means "strong evidence". Of course this terminology is purely suggestive and not to be taken literally. We can consider it as a practical bookkeeping device. When the data promote a model from weakly to moderately to strongly "evident", it is time to take it seriously and challenge aggressively.

Example 2.

A set of n data is distributed as Gaussian variables with zero mean and unknown variance σ^2, with uniform prior up to some arbitrary large value of σ^2 (model A). Estimating the variance with the maximum likelihood method, one obtains

$$\hat{\sigma}^2 = \frac{S_2}{n} \tag{3.94}$$

where $S_2 = \sum x_i^2$. The standard frequentist measure of goodness of fit (see next Chapter) is then $\chi^2/n = S_2/\hat{\sigma}^2 n = 1$, and therefore the test fails completely, in the sense that one obtains χ^2/n always unity by construction. Compare now a second theoretical models in which there are two variances, σ_1^2 for the first half of the data and σ_2^2 for the second half, both of them taken as free parameters, and same priors as before (model B). Clearly one has $S_2 = S_{2(1)} + S_{2(2)}$, where $S_{2(i)}$ are the respective partial sums, and $\hat{\sigma}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$, where $\hat{\sigma}_i^2 = S_{2(i)} n/2$. The frequentist χ^2/n is still unity in both cases, so we cannot say which model is a better fit. Since the priors are uniform up to a very large $\sigma = \Delta$, we assume the integral over σ^2 can be extended up to infinity, and the prior for every parameter is just an overall constant Δ^{-2}. The Bayesian evidences are then

$$E_A = \frac{1}{\Delta^2} \int_0^\infty (2\pi\sigma^2)^{-n/2} \exp(-\frac{1}{2}\frac{S_2}{\sigma^2}) d\sigma^2 = \frac{1}{2\pi\Delta^2} (\pi n \hat{\sigma}^2)^{1-n/2} \Gamma(\frac{n}{2} - 1) \tag{3.95}$$

and

$$E_B = \frac{1}{\Delta^4} \int (2\pi\sigma_1^2)^{-n/4} \exp(-\frac{1}{2}\frac{S_{2(1)}}{\sigma_1^2}) d\sigma_1^2 \int (2\pi\sigma_2^2)^{-n/4} \exp(-\frac{1}{2}\frac{S_{2(2)}}{\sigma_2^2}) d\sigma_2^2 = \frac{1}{(2\pi)^2\Delta^4} (\pi^2 \frac{n^2}{4} \hat{\sigma}_1^2 \hat{\sigma}_2^2)^{1-n/4} \Gamma(\frac{n}{4} - 1)^2 \tag{3.96}$$

and the ratio is

$$B_{BA} = \frac{E_B}{E_A} = \frac{(\hat{\sigma}_1^2 \hat{\sigma}_2^2)^{1-n/4}}{\Delta^2 (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^{1-n/2}} \frac{n\Gamma(-1 + \frac{n}{4})^2}{4\Gamma(-1 + \frac{n}{2})} \tag{3.97}$$

For large n this becomes

$$\log B_{BA} \to -\log \Delta^2 + (1 - \frac{n}{4}) \log(\hat{\sigma}_1^2 \hat{\sigma}_2^2) - (1 - \frac{n}{2}) \log(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) + \log \frac{n}{4} + \frac{n}{2} \log \frac{1}{2} \tag{3.98}$$

$$\approx \log(\frac{\hat{\sigma}_1^2 \hat{\sigma}_2^2}{\Delta^2 (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}) + \frac{n}{2} \log(\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2\hat{\sigma}_1 \hat{\sigma}_2}) \tag{3.99}$$

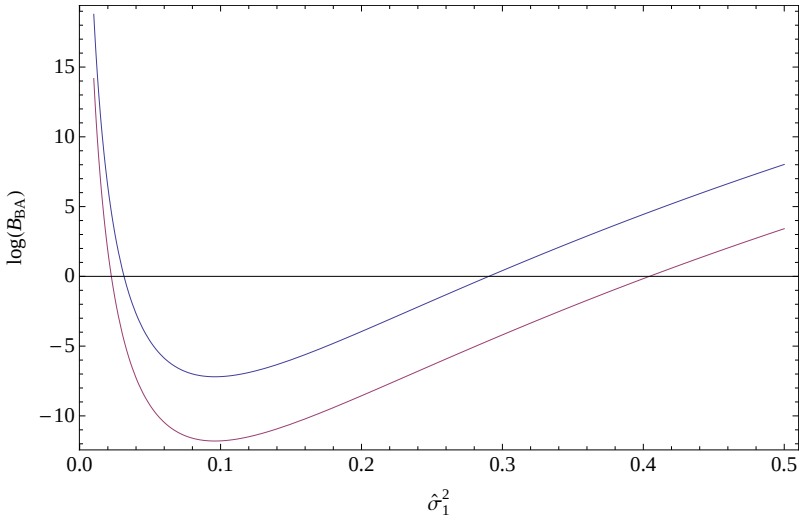

Figure 3.4: Bayesian ratio for Example 2, plotted versus $\hat{\sigma}_1^2$, for two values of the prior range, $\Delta = 10$ (upper blue line) and $\Delta = 100$ (lower red line), fixing $n = 100, \hat{\sigma}_2^2 = 0.1$. Near the minimum at $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$ model A is favoured, otherwise model B is favoured.

The Bayesian ratio has a minimum for large n at $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$ for which

$$B_{BA}^{(min)} \approx 4\sqrt{\frac{\pi}{n}}\frac{\hat{\sigma}_1^2}{\Delta^2} \tag{3.100}$$

which is only weakly dependent of n and smaller than unity (so A is favoured) since we assumed $\Delta \gg \hat{\sigma}_1$ and large n. As expected, near the minimum, where the best fit variances are similar, B_{BA} is minimal, and therefore A is preferred; if the variances are very different, then B is preferred. However the exact region where B_{BA} is actually larger or smaller than unity depends on the value of the prior range Δ, as is the absolute value of B_{BA}, as shown in Fig. (3.4). One can also see that increasing the number of data, and fixing all the rest, the region where A is favoured narrows down.

3.11 A simplified measure of evidence

Evaluating the full multidimensional evidence integral can be cumbersome, so many proposals have been advanced to approximate the evidence with simpler expressions (see full discussion in [1]). No one of these is general enough to be applied indiscriminately but often they give a quick useful estimate. The simplest one is the so-called Bayesian Information Criterion (BIC), proposed by Schwarz in 1978. If the unnormalized (i.e., without the denominator $p(D)$) posterior $P(T) = L(D;T)p(T)$ can be approximated with the Fisher matrix, then

$$P(\theta) = P_{\max} \exp -\frac{1}{2}(\theta - \hat{\theta})_\alpha F_{\alpha\beta}(\theta - \hat{\theta})_\beta \tag{3.101}$$

and the evidence can be written as

$$E = \int P(\theta) d^k\theta = P_{\max} p (2\pi)^{k/2} |\mathbf{F}|^{-1/2} \tag{3.102}$$

where k is the number of parameters. Then we have

$$\log E = \log P_{\max} + \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{F}| \tag{3.103}$$

For a large number $N \to \infty$ of data points, the maximum θ_{max} of the posterior tends to the maximum of the likelihood, since in this limit the prior is not important. Then $P_{max} \approx L_{max} p(\theta_{max})$ and

$$\log E = \log L_{\max} + \log p(\theta_{\max}) + \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{F}| \tag{3.104}$$

Let us further assume that the data have been binned in n_b bins, each containing N/n_b points, for a total of N data. The variance in each bin scales then as σ^2/N. The first term is then

$$\log L_{\max} = -\frac{n_b}{2}\log(2\pi) - n_b \log\sigma + \frac{n_b}{2}\log N - \frac{1}{2}\chi^2_{\min} \tag{3.105}$$

and therefore depends on N, since $\langle\chi^2\rangle = N$. Let us now take the expression (3.77) for the Fisher matrix, and assume the parameters are only in μ

$$F_{\alpha\beta} = C^{-1}_{\ell m}\frac{\partial\mu_\ell}{\partial\theta_\alpha}\frac{\partial\mu_m}{\partial\theta_\beta} \tag{3.106}$$

Then we have $|F| = |C|^{-1}|\frac{\partial\mu_\ell}{\partial\theta_\alpha}\frac{\partial\mu_m}{\partial\theta_\beta}|$. The entries on the diagonal of $C_{\ell m}$ are the variances, and according to our assumptions they scale with N so that $C_{ii} \sim \sigma^2/N$. On the other hand, the matrix F is a $k \times k$ matrix, so if F is approximately diagonal, its determinant will be composed of k factors, each of which scaling as $\sim N$, so $|\boldsymbol{F}| \sim N^k$. This allows us to approximate

$$\log|F| \approx \text{const} + k\log N \tag{3.107}$$

So we have

$$\log E = \log L_{\max} - \frac{k}{2}\log N + \text{const} \tag{3.108}$$

where the first two terms increase with N, while the last one collects all the terms that are independent of N. The definition of the BIC is finally $-2\log E$ i.e.

$$\text{BIC} = -2\log L_{\max} + k\log N \tag{3.109}$$

and minimizing the BIC is equivalent to maximizing the evidence.

Other approximate criteria have been introduced in the literature. For some of them and for a general discussion and cautionary remarks, see Trotta [1].

3.12 Robustness

The entire likelihood approach to parameter fitting and confidence regions have a clear problem: they are insensitive to the possibility that the data are inconsistent with each other. If we have two datasets d_1 and d_2 that comes from two different experiments, we combine them into a likelihood without asking ourself whether they agree with each other. It could be that one of the two datasets is actually biased by some unknown systematic effect, for instance if the supernovae Ia employed to measure the dark energy equation of state are heavily contaminated by some other class of sources. Let's us discuss here a possible way to address this problem within the context of the Fisher matrix approximation [2].

Suppose we have two datasets D_i with $i = 1, 2$, both described by Gaussian likelihoods that can be approximated by a Gaussian function of the theoretical parameters

$$L_i = L^i_o e^{-\frac{1}{2}(\mu_i-\theta)^t L_i(\mu_i-\theta)} \tag{3.110}$$

where L_i is the Fisher matrix and θ the vector of theoretical parameters centered around the ML estimators μ_i. Notice that in this section latin indices run over the number of datasets, not over the individual data points. If we have also a Gaussian prior centered on the origin and with correlation inverse P then the posterior for each dataset is another Gaussian with Fisher matrix

$$F_i = L_i + P \tag{3.111}$$

and mean

$$\bar{\mu}_i = F_i^{-1} L_i \mu_i \tag{3.112}$$

Finally, the combined posterior will be a Gaussian with Fisher matrix and mean

$$F = L_1 + L_2 + P \tag{3.113}$$

$$\mu = F^{-1}\sum_i L_i\mu_i \tag{3.114}$$

(notice that when we combine the two datasets there is only one prior, not two). We see that F does not depend on μ, i.e. the Fisher matrix (and consequently the FOM) does not depend on the distance between the two experiments confidence regions (although it depends on their orientation): the datasets could be very much different, and therefore very likely incompatible, but the FOM would be exactly the same.

The evidence for such a combined posterior is (see eq. 3.91)

$$E_{comb} = L_{\max}\frac{|\boldsymbol{P}|^{1/2}}{|\boldsymbol{F}|^{1/2}}\exp\left[-\frac{1}{2}(\sum_i \bar{\mu}_i^t L \bar{\mu}_i - \mu^t F \mu)\right], \tag{3.115}$$

where $L = L_1 + L_2$ (notice that the prior is supposed to be centered on the origin).

However, if the data are actually coming from completely different distributions, and therefore independent, their evidence should be written as the product of the separate evidences

$$E_{ind} = E_{d_1} E_{d_2} \tag{3.116}$$

In other words, we are assuming here that one of the datasets, say d_2 actually depends on a totally different sets of parameters (eg the properties of the SN Ia progenitors or galaxy environment) and therefore we are testing two models, one with the cosmological parameters only, the other with added systematics. However we erroneously interpret the new hidden parameters as the real ones and therefore we employ the same parameter names and the same prior.

We can then form the Bayes ratio

$$R = \frac{E_{comb}}{E_{ind}} = \frac{p(D)}{p(d_1)p(d_2)} = \frac{p(d_1,d_2)}{p(d_1)p(d_2)} \tag{3.117}$$

$$= \frac{p(d_2;d_1)p(d_1)}{p(d_1)p(d_2)} = \frac{p(d_2;d_1)}{p(d_2)} \tag{3.118}$$

We can write

$$p(d_2;d_1) = \int p(d_2;\theta)p(\theta;d_1)d\theta \tag{3.119}$$

where the first term in the integral is the likelihood of the second probe and the second term is the posterior of the first probe. Their ratio can be evaluated analytically and we obtain

$$R = \left(\frac{|F_1 F_2|}{|FP|}\right)^{1/2}\exp\left[-\frac{1}{2}(\bar{\mu}_1^t F_1 \bar{\mu_1} + \bar{\mu}_2^t F_2 \bar{\mu_2} - \mu^t F \mu)\right] \tag{3.120}$$

The first factor, formed out of the determinants, express the Occam's razor factor of parameter volumes, while the second penalizes R if the two probes are very different from each other (so the hypothesis that they come from different models, or equivalently that systematics are important, is favored). We denote R as *robustness*. From its definition, we expect the robustness to be a measure of how much the probes overlap: the more they do, the more consistent the two datasets are. This is exactly orthogonal to the statistical FOM, which is maximized when the probes have little overlap!

To make progress we have now to assume some systematic bias among the two probes. For simplicity we assume now that experiment 1 represents our current knowledge and is unbiased wrt the prior, i.e. we set $\mu_1 = 0$. We wish to design experiment 2 so that it gives a high Bayes factor R when some hypothetical bias is present:

$$\mu_2 = b \tag{3.121}$$

The bias itself is typically the result of some physical effect, either in the detectors or in the sources, which can be parametrized by some systematic parameters s_k. The resulting bias vector $\delta\mu_2$ on the parameter space of interest (i.e. the cosmological parameter space) is given in explicit component form by the general projection formula (see Appendix)

$$\delta\mu_{2\alpha} = -F^{-1}_{\alpha\beta}F^*_{\beta\gamma}\delta s_\gamma \tag{3.122}$$

where

$$F^*_{\beta\gamma} = \frac{\partial^2 \log L}{\partial\mu_\beta \partial s_\gamma} \tag{3.123}$$

is a sort of systematic Fisher matrix (which in general is not a square matrix).

Once we have the projected bias μ_2 we define

$$\bar{\mu}_2 = F_2^{-1} L_2 b \tag{3.124}$$
$$\mu = F^{-1} L_2 b \tag{3.125}$$

The last relation quantifies how much the best estimates μ of the combined posterior moves when there is a systematic bias b in one of the probes. This gives

$$\ln R = -\frac{1}{2} b^t F^* b - \frac{1}{2} \ln \frac{|FP|}{|F_1 F_2|} \tag{3.126}$$

where

$$F^* = L_2 (F_2^{-1} - F^{-1}) L_2 \tag{3.127}$$

For simplicity, we now redefine the robustness in such a way that $R = 1$ is achieved when experiment 2 is identical to 1 and unbiassed, i.e. $b = 0$:

$$R_N = \frac{R}{R^*} \tag{3.128}$$

where

$$R^* = \frac{|F_1|}{(|2L_1 + P||P|)^{1/2}} \tag{3.129}$$

This can be seen as the reference experiment: a trivial repetition of the experiment 1. With this arbitrary normalization, R_N can be used to design models that maximize or minimize the robustness. If it is maximized, it means that the datasets are more likely to come from the same distribution and that therefore there is no need of suspect a systematic bias. If it is minimized, on the other hand, we have good reason to investigate systematic biases in one of the experiments. In other words, given an expected bias vector, we could design an experiment to reduce the impact of bias on the parameter estimation (high robustness) or to increase the sensitivity to the bias in order to detect it. Finally, we have

$$\ln \frac{R}{R^*} = -\frac{1}{2} b^t F^* b - \frac{1}{2} \ln \frac{|F_1||L_1 + L_2 + P|}{|F_2||2L_1 + P|} \tag{3.130}$$

which tends to

$$\ln \frac{R}{R^*} = -\frac{1}{2} b^t F^* b - \frac{1}{2} \ln \frac{|L_1 + L_2|}{2^{N_p} |L_2|} \tag{3.131}$$

where N_p is the number of parameters. In this limit there is no dependence on the prior. If we consider two experiments with no bias, $b = 0$, and with Fisher matrices orthogonal to each other and diagonal (i.e. with semiaxes aligned with the axes) with two entries $L_1 = \mathrm{diag}(s, \ell)$ and $L_2 = \mathrm{diag}(\ell, s)$ with $s \ll \ell$, this gives $\ln(R/R^*) \approx -0.5 \ln(\ell/s) \ll 1$. Statistically, two orthogonal probes maximise the constraints, but are not robust to systematics.

The problem can be further simplified by assuming that the axes in the 2D parameter space are rotated so that they lie along the direction of the bias vector. Then the product $b^t F^* b$ becomes simply $|b|^2 D_{11}$, where D_{11} is the $x-$component of the matrix F^* in the basis that diagonalizes it. Moreover, we can evaluate an average R_N along the bias direction with a Gaussian weight function $W(x) = (\sqrt{2\pi}|b|)^{-1} \exp -\frac{1}{2} \frac{x^2}{|b|^2}$,

$$\langle R_N \rangle = \int W(x) R_N dx = \frac{(|F_2||2L_1 + P|)^{1/2}}{|F F_1|^{1/2}} (|b|^2 D_{11} + 1)^{-1/2} \tag{3.132}$$

A very manageable expression is obtained if the two probes are aligned along the minor or major axis and if they are identical up to a roto-translation (so $|F_1| = |F_2|$). This is not an uncommon situation since experiments with widely different statistical power do not need to be combined. Then we have

$$\langle R_N \rangle \approx \frac{2\sigma_{2,x}\sigma_{2,y}}{(\sigma_{1,y}^2 + \sigma_{2,y}^2)^{1/2} (b^2 + \sigma_{1,x}^2 + \sigma_{2,x}^2)^{1/2}} \tag{3.133}$$

where $\sigma_{1,x}$ means the error on the x-axis of the probe 1, and similarly for the other cases. This expression contains many possible cases. Assume now for a further simplification the ellipses to be relatively orthogonal or parallel. In the first case, $\sigma_{2,y} = \sigma_{1,x}$ and $\sigma_{2,x} = \sigma_{1,y}$ and we have

$$R_\perp = \frac{2r}{1+r^2}(1 + \frac{b^2 r^2}{\sigma_{2,x}^2(1+r^2)})^{-1/2} \tag{3.134}$$

where $r = \sigma_{2,x}/\sigma_{2,y}$ while in the second case

$$R_\parallel = (1 + \frac{b^2}{2\sigma_{2,x}^2})^{-1/2} \tag{3.135}$$

In both cases, the maximal value is 1, ie. the reference unbiased experiment identical to 1. If $B = b/\sigma_{2,x}$ is small, parallel probes are more robust than the orthogonal ones; ortho probes can be more robust than parallel ones only if

$$\sqrt{\frac{1}{1+B^2}} < r < 1 \tag{3.136}$$

i.e. only if $\sigma_{2,y} > \sigma_{2,x}$, that is when probe 2 is elongated perpendicularly to the bias vector. This is indeed what one expects from a measure of overlapness.

Appendix

Eq. (3.122) can be proved as follows.

The maximum likelihood estimator $\bar{\theta}_\alpha$, given a likelihood function $L(\theta_\alpha)$ for the parameters θ_α, is obtained by solving the system of equations

$$\mathcal{L}_{,\alpha} = 0 \tag{3.137}$$

where $\mathcal{L} = -\log L$. In the Fisher matrix approximation one has $\mathcal{L} = \frac{1}{2} D_i D_j P_{ij} - \frac{1}{2} P$ where $P = C^{-1}$ is the precision matrix (the inverse of the data correlation matrix) and $D_i = d_i - t_i$ is the vector of data minus theory points.

Suppose now L depends also on a parameter s (that we refer to as a systematic parameter) that has been assigned a particular value, and we want to estimate how $\bar{\theta}_\alpha$ changes when s is shifted to another value $s + \delta s$, where δs is assumed very small. Then we have

$$\mathcal{L}(s + \delta s) = \mathcal{L}(s) + \mathcal{L}_{,s}\delta s \tag{3.138}$$

and the equations that give the new maximum likelihood estimator vector $\hat{\theta} = \bar{\theta} + \delta\theta$ become to first order in δs and $\delta\theta$

$$\mathcal{L}(s+\delta s)_{,\alpha} = \mathcal{L}_{,\alpha} \mid_{\hat{\theta}} + (\mathcal{L}_{,\alpha s} \mid_{\bar{\theta}})\delta s \tag{3.139}$$

$$= \mathcal{L}_{,\alpha} \mid_{\bar{\theta}} + (\mathcal{L}_{,\alpha\beta} \mid_{\bar{\theta}})\delta\theta_\beta + (\mathcal{L}_{,\alpha s} \mid_{\bar{\theta}})\delta s \tag{3.140}$$

$$= (\mathcal{L}_{,\alpha\beta} \mid_{\bar{\theta}})\delta\theta_\beta + (\mathcal{L}_{,\alpha s} \mid_{\bar{\theta}})\delta s = 0 \tag{3.141}$$

where we employed Eq. (3.137). Finally we obtain

$$\delta\theta_\alpha = -\mathcal{L}^{-1}_{,\alpha\beta}\mathcal{L}_{,\beta s}\delta s \tag{3.142}$$

(sum over β). If there are several systematic parameters, then a sum over the s parameters is understood. Now, once we average over many data realization, $\langle \mathcal{L}_{,\alpha\beta} \rangle = F_{\alpha\beta}$ is the Fisher matrix, and $\langle \mathcal{L}_{,\beta s} \rangle = F_{\beta s}$ is a sort of systematic Fisher matrix. In practice, this means that one includes the systematic parameters s_i in a general Fisher matrix that contains also the parameters θ_α, and then selects the $\alpha\beta$ and the βs_i submatrices and produces the sum over β and s_i, namely

$$\delta\theta_\alpha = -F^{-1}_{\alpha\beta} F_{\beta s_i}\delta s_i \tag{3.143}$$

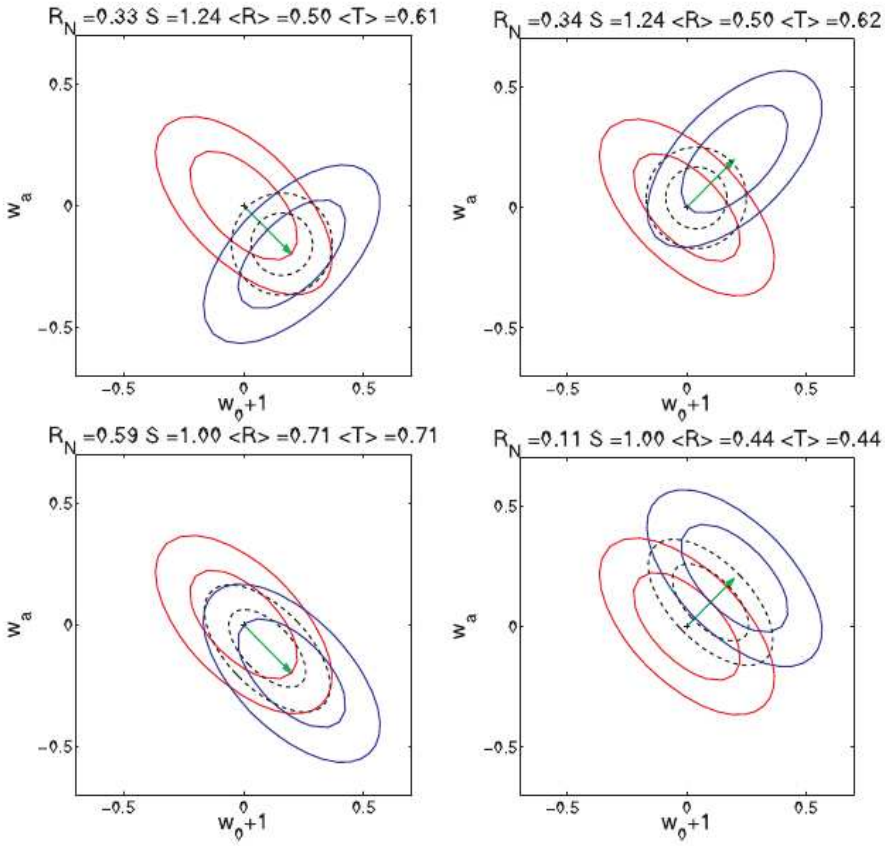

Figure 1. Illustration of statistical and robustness FoM for a future probe (blue ellipses, 68 and 95 per cent C.L.) which is systematically biased with respect to the present-day constraints (red ellipses) in the direction given by the green bias vector. The black dotted ellipses represent the combined constraints. Notice that the statistical FoM S does not change in the presence of a systematic bias.

Figure 3.5: From MNRAS 415, 143 (2011).

Exercise. Suppose x_i are Gaussian variables with mean x_0 and variances σ_i^2. Due to some systematic, the variances might be affected by a constant offset, not well accounted for, $\sigma_i^2 \to \sigma_i^2 + \delta$. How would the estimate of the mean x_0 be affected by an offset $\delta \ll \sigma_i^2$ (neglect the presence of δ in the normalizing determinant)?

In absence of systematic error δ, the variables are distributed as

$$L \sim \exp -\frac{1}{2} \sum_i \frac{(x_i - x_0)^2}{\sigma_i^2} = \exp -\frac{1}{2}(S_{22} - 2x_0 S_{12} + x_0^2 S_{02}) \tag{3.144}$$

where

$$S_{nm} = \sum_i \frac{x_i^n}{\sigma_i^m} \tag{3.145}$$

The maximum likelihood estimator of x_0 is then

$$x_0 = \frac{S_{22}}{S_{02}} \tag{3.146}$$

When we switch on δ, we have the new likelihood

$$L \sim \exp -\frac{1}{2} \sum_i \frac{(x_i - x_0)^2}{\sigma_i^2 + \delta} \approx \exp -\frac{1}{2} \sum_i \frac{(x_i - x_0)^2}{\sigma_i^2} (1 - \frac{\delta}{\sigma_i^2}) \tag{3.147}$$

and the new estimator equation $d(\log L)/dx_0 = 0$, from which, at first order in δ,

$$\delta x_0 \equiv x_0^{(new)} - x_0 = -\frac{\delta}{S_{02}}(S_{14} - \frac{S_{12} S_{04}}{S_{02}}) \tag{3.148}$$

This can be seen to coincide with Eq. (3.122). If $\sigma_i = \sigma$, constant for all i, the bias vanishes: the estimation of the mean is then independent of the variance.

Chapter 4

Fitting with linear models

An important class of cases in which maximum likelihood estimators and confidence regions take a particularly simple form is the case in which the data are Gaussian and the model parameters appear linearly in the mean. Let us assume we have N data d_i , one for each value of the *independent* variable x_i (which are *not* random variables) and that

$$d_i = f_i + e_i \tag{4.1}$$

where e_i are errors (random variables) which are assumed to be distributed as Gaussian variables. Here f_i are theoretical functions that depend linearly on a number of parameters A_α

$$f_i = \sum_\alpha A_\alpha g_{i\alpha} \tag{4.2}$$

where $g_{i\alpha}(x_i)$ are functions of the variable x_i. Eg the data could be galaxy magnitudes ($d_i = m_i$) as a function of redshifts ($x_i = z_i$), temperatures ($d_i = T_i$) at different times ($x_i = t_i$), etc. The expression for f could be then

$$f(x) = A_0 + A_1 x + A_2 x^2 + A_3 x^3 ... \tag{4.3}$$

but could also include any other function, e.g. $f(x) = A_0 + A_1 e^x + A_2 \sin x$ etc. For instance, for a quadratic fit $f(x) = A_0 + A_1 x + A_2 x^2$ with four data points, the matrix $g_{i\alpha}$ will have the form

$$g = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{pmatrix} \tag{4.4}$$

Since f is not a random variable, if e_i are gaussian variables (not necessarily independent) then also d_i are. Goal of this section is to find the maximum likelihood estimators for A_α and their confidence regions.

In Fig. (4.1) you can see an example of a simple linear fit with a straight line (notice that "linear fit" does not necessarily mean "straight line"!). In Fig. (4.2) a far more interesting example of a non-linear fit. In the case of parameters that are non-linear, e.g. $f(x) = \exp(ax)$, the likelihood and the best fit is normally obtained numerically with the methods seen in the previous Section.

4.1 The simplest case: Fitting with a straight line.

Let us start with the simplest example, fitting the N data d_i with a straight line

$$f(x) = ax + b \tag{4.5}$$

where x will take the values x_i corresponding to the N data, i.e. we assume that each data point d_i is distributed as an indepedent Gaussian variables with error σ_i (assumed known) and mean given by our model $f_i = ax_i + b$. We assume the variables x_i have negligible error (for instance, they correspond to the epoch at which the measurements

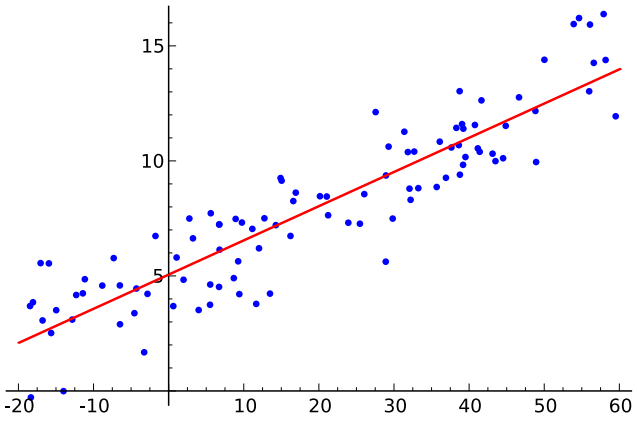

Figure 4.1: Example of a linear fit with a straight line (By Sewaqu - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11967659)

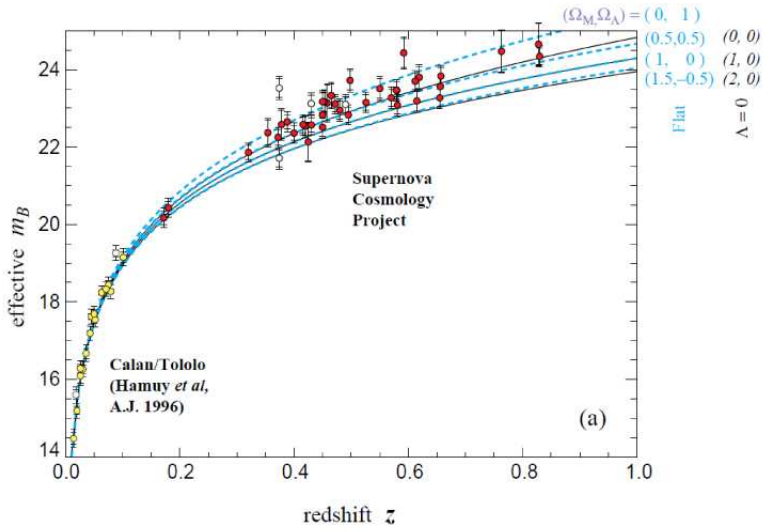

Figure 4.2: A Nobel-prize example of non-linear fit. The fit of the supernovae data of the Supernovae Cosmology Project (Perlmutter et al 1998) depends in a complicate, non-linear way on the cosmological parameters Ω_m, Ω_Λ. The best fit (upper continuous black curve) revealed the cosmic acceleration.

are done). The problem is to find the best estimator of a, b and their covariance matrix. So we need to maximize the likelihood

$$P(D; a, b) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp -\frac{1}{2} \sum_i \frac{(d_i - (ax_i + b))^2}{\sigma_i^2} \tag{4.6}$$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp -\frac{1}{2} \sum_i \frac{d_i^2 + (ax_i + b)^2 - 2d_i(ax_i + b)}{\sigma_i^2} \tag{4.7}$$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp -\frac{1}{2} \frac{\sum_i d_i^2 + \sum_i (ax_i + b)^2 - 2\sum_i d_i(ax_i + b)}{\sigma_i^2} \tag{4.8}$$

with respect to a, b. We can rewrite $-2 \log P$ (up to irrelevant constants that do not depend on a, b) as

$$-2 \log P = const + \sum_i \frac{(a^2 x_i^2 + b^2 + 2abx_i - 2ad_i x_i - 2d_i b)}{\sigma_i^2} \tag{4.9}$$

$$= const + a^2 S_2 + b^2 S_0 + 2abS_1 - 2aS_{dx} - 2bS_d \tag{4.10}$$

where

$$S_0 = \sum_i \frac{1}{\sigma_i^2} \tag{4.11}$$

$$S_1 = \sum_i \frac{x_i}{\sigma_i^2} \tag{4.12}$$

$$S_2 = \sum_i \frac{x_i^2}{\sigma_i^2} \tag{4.13}$$

$$S_{dx} = \sum_i \frac{x_i d_i}{\sigma_i^2} \tag{4.14}$$

$$S_d = \sum_i \frac{d_i}{\sigma_i^2} \tag{4.15}$$

To maximize P, or equivalently to minimize $-2 \log P$, we have to solve the system of equations $\partial P/\partial a = 0$ and $\partial P/\partial b = 0$, that is

$$bS_0 + aS_1 = S_d \tag{4.16}$$

$$bS_1 + aS_2 = S_{dx} \tag{4.17}$$

which is solved by

$$\bar{\mathbf{A}} = \mathbf{G}^{-1}\mathbf{D} \tag{4.18}$$

if we denote with $\bar{\mathbf{A}} = (\bar{b}, \bar{a})^T$ the vector of the maximum likelihood estimators of b, a, with $\mathbf{D} = (S_d, S_{dx})^T$ the vector of the data points and with

$$\mathbf{G} = \begin{pmatrix} S_0 & S_1 \\ S_1 & S_2 \end{pmatrix} \tag{4.19}$$

the *design* matrix, that depends entirely on our choice of the fitting function and on the errors σ_i. This solves the problem of the maximum likelihood estimators. For the variances, one has to note that a, b are linear functions of the data contained in D, and since the data d_i are Gaussian variables, so are a, b. Instead of performing this step, now we generalize the whole procedure in the next Section.

4.2 Normal equations for linear fitting

If the data are independent and they all have variance σ then the joint likelihood is

$$P(D; A_\alpha, I) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp -\frac{Q}{2\sigma^2} \tag{4.20}$$

where

$$Q = \sum_i^N (d_i - f_i)^2 \tag{4.21}$$

$$= d_i d_i + \sum_{\alpha,\beta} A_\alpha A_\beta g_{\alpha i} g_{\beta i} - 2 \sum_\alpha A_\alpha d_i g_{\alpha i} \tag{4.22}$$

sum over Latin indexes implied. The priors for the M parameters A_α can be taken to be uniform

$$P(A_\alpha; I) = \prod_\alpha \frac{1}{\Delta A_\alpha} \tag{4.23}$$

where ΔA_α is the theoretically expected range of variation of the parameter A_α. From Bayes' theorem we obtain

$$P(A_\alpha; D, I) = C e^{-Q/2\sigma^2} \tag{4.24}$$

where C does not depend on A's:

$$C = \left(\int d^M A_\alpha e^{-Q/2\sigma^2} \right)^{-1} \tag{4.25}$$

The best estimate for A's is obtained then when the posterior $P(A_\alpha)$ is maximized:

$$\frac{\partial Q}{\partial A_\alpha} = 2 \sum_\beta A_\beta g_{\beta i} g_{\alpha i} - 2 d_i g_{\alpha i} = 0 \tag{4.26}$$

These equations are called *normal equations*. If we define the matrix and vector

$$G_{\alpha\beta} \equiv g_{\beta i} g_{\alpha i} \tag{4.27}$$

$$D_\alpha \equiv d_i g_{\alpha i} \tag{4.28}$$

(remember that we always sum over repeated Latin indexes) then we can write the normal equation in matrix form as

$$\mathbf{G}\mathbf{A} = \mathbf{D} \tag{4.29}$$

(where $\boldsymbol{A} = \{A_0, A_1, ...\}$) and solve as

$$\bar{\mathbf{A}} = \mathbf{G}^{-1}\mathbf{D} \tag{4.30}$$

For instance, if we fit with a straight line we have $g_1 = 1, g_2 = x$ and for instance $G_{11} = \sum_i g_{1i} g_{1i} = \sum_i 1 = N$ etc:

$$\mathbf{G} = \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \tag{4.31}$$

$$\mathbf{D} = \begin{pmatrix} \sum_i d_i \\ \sum_i d_i x_i \end{pmatrix} \tag{4.32}$$

and finally we obtain the solution

$$\bar{\mathbf{A}} = \begin{pmatrix} G_{22} D_1 - G_{12} D_2 \\ G_{11} D_2 - G_{12} D_1 \end{pmatrix} \frac{1}{G_{11} G_{22} - G_{12}^2} \tag{4.33}$$

This was obtained for uncorrelated data with constant variance σ^2. The generalization to correlated data with arbitrary variances is then straightforward. In this case we have in fact

$$Q = \sum_{i,j}^N (d_i - f_i) C_{ij}^{-1} (d_j - f_j) \tag{4.34}$$

$$= d_i C_{ij}^{-1} d_j + \sum_{\alpha,\beta} A_\alpha A_\beta g_{\alpha i} C_{ij}^{-1} g_{\beta j} - 2 \sum_\alpha A_\alpha d_i C_{ij}^{-1} g_{\alpha j} \tag{4.35}$$

and we simply have to redefine

$$G_{\alpha\beta} \equiv g_{\beta i} C_{ij}^{-1} g_{\alpha j} \tag{4.36}$$

$$D_\alpha \equiv d_i C_{ij}^{-1} g_{\alpha j} \tag{4.37}$$

to obtain again formally the same equations (4.29) and the same solution $\bar{\mathbf{A}} = \mathbf{G}^{-1}\mathbf{D}$. Notice again that $\bar{\mathbf{A}}$ is linear in the data d_i; therefore the maximum likelihood estimator $\hat{\mathbf{A}}$ for a linear problem is a linear functions of the random variables and is distributed as a multivariate Gaussian.

4.3 Confidence regions

If the prior is uniform in an infinite range (improper prior), the parameters in the linear problem have a Gaussian posterior with mean $\bar{\mathbf{A}}$ and correlation matrix given by the inverse of the Fisher matrix that we already calculated in Sec. 3.9. Since in the linear model the correlation matrix does not depend on the parameters, we have

$$F_{\alpha\beta} \equiv C_{ij}^{-1} \frac{\partial \mu_i}{\partial \theta_\alpha} \frac{\partial \mu_j}{\partial \theta_\beta} \,, \tag{4.38}$$

In the present context the means are $\mu_i = \sum_\alpha A_\alpha g_{i\alpha}$ and $\theta_\alpha = A_\alpha$, so we obtain

$$F_{\alpha\beta} \equiv C_{ij}^{-1} g_{\alpha i} g_{\beta j} = G_{\alpha\beta} \,, \tag{4.39}$$

This is a very nice result: the design matrix is also the Fisher matrix. The marginalized errors on the parameters are therefore

$$\sigma_\alpha^2 = (\mathbf{G}^{-1})_{\alpha\alpha} \tag{4.40}$$

This reduces to the case of the straight line and of uncorrelated points with variances σ_i^2 (the same case already seen in Sec. 4.1) as follows. First, we have $C_{ij} = \sigma_i^2 \delta_{ij}$ (no sum over i here) and therefore $C_{ij}^{-1} = \sigma_i^{-2}\delta_{ij}$ and we find $G_{11} = \sum_i \sigma_i^{-2} g_{1i} g_{1i} = \sum_i \sigma_i^{-2}$, $G_{12} = G_{21} = \sum_i \sigma_i^{-2} g_{1i} g_{2i} = \sum_i \sigma_i^{-2} x_i$, and $G_{22} = \sum_i \sigma_i^{-2} x_i^2$. Then, If we define

$$S_n = \sum \frac{x_i^n}{\sigma_i^2} \tag{4.41}$$

we obtain

$$\mathbf{F} = \mathbf{G} = \begin{pmatrix} S_0 & S_1 \\ S_1 & S_2 \end{pmatrix} \tag{4.42}$$

and $\det \mathbf{F} = S_0 S_2 - S_1^2$ so that

$$\mathbf{F}^{-1} = \frac{1}{\det \mathbf{F}} \begin{pmatrix} S_2 & -S_1 \\ -S_1 & S_0 \end{pmatrix} \tag{4.43}$$

Finally, if the σ_i^2 are all equal to σ^2 we obtain $\det \mathbf{F} = S_0 S_2 - S_1^2 = \sigma^{-4}(N \sum x_i^2 - (\sum x_i)^2) = \sigma^{-4} N^2 s_x^2$ where

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N} \tag{4.44}$$

is the x sample variance and $\bar{x} = \sum x_i / N$ the sample mean. Then we have

$$\mathbf{F}^{-1} = \frac{\sigma^2}{N^2 s_x^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix} = \frac{\sigma^2}{N s_x^2} \begin{pmatrix} \bar{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \tag{4.45}$$

(where we used the notation $\bar{f(x)} = \frac{\sum_i f(x_i)}{N}$). This shows that for a straight line fit $A_0 + A_1 x$ the variances are

$$(F^{-1})_{11} = \mathrm{Var}[A_0] = \frac{\sigma^2}{N s_x^2} \bar{x^2} \tag{4.46}$$

$$(F^{-1})_{22} = \mathrm{Var}[A_1] = \frac{\sigma^2}{N s_x^2} \tag{4.47}$$

$$(F^{-1})_{12} = \mathrm{Cov}[A_0 A_1] = -\frac{\sigma^2}{N s_x^2} \bar{x} \tag{4.48}$$

As is intuitive, the errors are smaller when the data have small variance σ^2 and large dispersion along the x-axis, s_x^2 or large number of points N. Notice again that the parameters are generally correlated even if the data points were not.

In this special case (linear model, Gaussian data, uniform prior), the distribution of the estimators and their posterior are the same. The frequentist and the Bayesian approach therefore coincide.

4.4 Principal component analysis

So far we have assumed very specific models, for instance a linear model, and have proceeded to get constraints on the parameters. The likelihood method will certainly find some constraints, no matter how wrong is our modeling. For instance, take the expansion $f(z) = a_0 + a_1 z + a_2 z^2 + ...$ and suppose that we stop at a_1. Given a dataset, we could end up with very good constraints on a_0 but very loose on a_1. We may content ourselves with that and blame the experimenters for their poor data. However, how can we be sure that the data do not contain good constraints on, say, a_2 or some other higher-order parameters? If the data do not extend very far we do not expect this, but still it would be nice to quantify which parameters (and how many) we can reasonably constrain for a given dataset. In other words we would like to *find* the best parametrization, rather than to *assume* one.

One way of doing this is to approximate the function $f(z)$ in the range z_a, z_b with many stepwise constant values:

$$f(z) = \sum_{i=1}^{N} \theta_i(z) w_i \,, \tag{4.49}$$

where $\theta_i = 1$ for z inside the bin $(z_i, z_i + \Delta z)$ and 0 outside. Here we make an exception to our rule of using Greek indices for parameters in order to stress that in this case we can have as many parameters as binned data. So now we have $N(\gg 1)$ parameters w_i instead of two or three. Technically, this is just a bigger Fisher matrix problem and we could proceed as before. In this case, however, it would be really nice to have uncorrelated errors on the parameters, since they all measure the same quantity, $f(z)$, and it will be difficult to compare different experiments if the errors are correlated (and compactifying to a single FOM would discard too much information). What we would like is in fact an expansion

$$f(z) = \sum_{i=1}^{N} \alpha_i e_i(z) \,, \tag{4.50}$$

where the coefficients α_i are uncorrelated. Since uncorrelated parameters mean a diagonal Fisher matrix, the problem is solved by diagonalizing the Fisher matrix for the original N parameters w_i, thus obtaining a diagonal F_{ij}^D. This is always possible since F_{ij} is a real symmetric non-degenerate matrix. The orthogonal basis functions $e_i(z)$ will be then the eigenvectors, with N eigenvalues λ_i (which are all positive since F_{ij} is positive definite). The new parameters α_i will have the variance $\sigma_i^2 = 1/\lambda_i = [(\boldsymbol{F}^D)^{-1}]_{ii}$ (i.e. the elements on the diagonal of the inverse Fisher matrix).

Now, a parameter with a large error is a badly measured parameter. It means that the data are not able to measure that parameter very well. On the contrary, a parameter with small error is well measured. Therefore we can rank the parameters α_i according to their errors, that is, according to the magnitude of the eigenvalues of F_{ij}. The highest eigenvalues (smallest errors) are called "principal components" and the whole method is called *principal component analysis* (PCA). This method is based on the fact that every well-behaved function can be expanded in piecewise constant fragments and that every non-singular Fisher matrix can be diagonalized. That is, the PCA can always be used when we need to reconstruct an unknown function.

So we have now a few well-measured components plus many others with large errors. The eigenvectors $e_i(z)$ associated with the principal components are functions of z, built up by linear combinations of the $\theta_\alpha(z_i)$. They tell us the range of z which is *best measured* by the data. We can plot them and have at once a view of the range of z most sensitive to that particular dataset. This is perhaps the best feature of the PCA since it allows us to optimize an experiment towards any range we are interested in.

The coefficient α_i themselves are rarely interesting. They can be evaluated by employing the property that the eigenvectors are orthogonal. Let us also normalize them by

$$\int e_i^2(z)\, \mathrm{d}z = 1 \,, \tag{4.51}$$

where the integration is taken in the whole z_a, z_b region. Multiplying Eq. (4.50) by $e_i(z)$ and then integrating, we obtain

$$\alpha_i = \int f(z) e_i(z) \mathrm{d}z \,. \tag{4.52}$$

In comparing different experiments the PCA might help, but care has to be taken when interpreting the results. In general the distribution of eigenvalues can be very different among different experiments and it is not obvious whether it is preferable to have few well-measured components in a small range or many not-so-well measured components in a large range. Reducing everything to a single FOM would kill the whole spirit of the PCA method and at the end of the day a sensible theoretical expectation is *the* principal component of any analysis.

Chapter 5

Frequentist approach: parameter estimation, confidence regions and hypothesis testing

In this section we leave for a moment the Bayesian approach and explore a few results in the frequentist context. There are two reasons for doing this. One is that many researchers and papers adopt a frequentist methodology and one should know what techniques they use in order to understand their results. The second is that in some case one might be completely unable to choose a prior or to agree on one with collaborators. In these cases a frequentist approach could be useful because it does not rely on subjective choices.

Any function of data only (and not of unknown parameters) is called a *statistics*. The sample mean, sample variance etc, are all statistics. In this section we will find the PDF of some statistics (the mean, the variance, and some of their combination, i.e. the normalized variable and the variance ratio) always assuming that the data are *independent Gaussian variates*. More general cases will be discussed in the next Section. These PDFs will in general depend on one or more unknown parameters. Once we have the PDF of a statistics, we can answer two questions:

1. Which is the confidence region of the unknown parameters?
2. How likely is to find the particular value of the statistics we got?

5.1 Distribution of the sample mean

If we have N data x_i assumed to be independent Gaussian variates $G(\mu, \sigma)$, any linear combination of x_i is a Gaussian variable. Therefore the sample mean statistics

$$\hat{x} = \frac{1}{N} \sum_i x_i \tag{5.1}$$

is a Gaussian variable $G(\mu, \sigma/\sqrt{N})$.

5.2 Distribution of the sample variance

We have already seen that the combination

$$Y = \sum_i \frac{(x_i - \mu)^2}{\sigma^2} = \sum_i Z_i^2 \tag{5.2}$$

where $x_i \sim G(\mu, \sigma)$ is a χ^2 variable with N dof. Therefore the random variable combination

$$(N-1)\frac{S^2}{\sigma^2} \tag{5.3}$$

where S^2 is the sample variance, defined as,

$$S^2 = \sum_i \frac{(x_i - \hat{x})^2}{N-1}\,, \tag{5.4}$$

is a χ^2_{N-1} variable. In fact we have

$$\begin{aligned}(N-1)S^2 &= \sum[(x_i-\mu)-(\hat{x}-\mu)]^2 &(5.5)\\ &= \sum(x_i-\mu)^2 - N(\hat{x}-\mu)^2 &(5.6)\end{aligned}$$

from which

$$(N-1)\frac{S^2}{\sigma^2} = \sum \frac{(x_i-\mu)^2}{\sigma^2} - \frac{(\hat{x}-\mu)^2}{(\sigma^2/N)} \tag{5.7}$$

Now the first term on the rhs is χ^2_N while the last term is χ^2_1; a general theorem says that the sum/difference of two χ^2 variables with dof ν_1, ν_2 is a χ^2 variable with $\nu_{tot} = \nu_1 \pm \nu_2$ (if $\nu_{tot} > 0$). Therefore

$$(N-1)\frac{S^2}{\sigma^2} \sim \chi^2_{N-1} \tag{5.8}$$

It follows

$$\langle (N-1)\frac{S^2}{\sigma^2}\rangle = N-1 \tag{5.9}$$

i.e. $\langle S^2\rangle = \sigma^2$. This last statement is actually true for x_i belonging to any distribution.

Notice that if we knew μ, so that $S^2 = \sum_i \frac{(x_i-\mu)^2}{N-1}$, then $NS^2/\sigma^2 \sim \chi^2_N$. Notice also that S^2 is a statistics (depends only on measured data) while $(N-1)S^2/\sigma^2$ is a statistics only if σ^2 is known.

5.3 Distribution of normalized variable (t-Student distribution).

If $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$ then one can show that the variable

$$T = \frac{Z}{\sqrt{X/\nu}} \tag{5.10}$$

is distributed as the $t-$Student distribution (see Fig. 5.1)

$$f(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\left[1+\left(\frac{t^2}{\nu}\right)\right]^{-\frac{\nu+1}{2}} \tag{5.11}$$

with $-\infty < t < \infty$ and $\nu > 0$. One has

$$\begin{aligned}\langle T\rangle &= 0 &(5.12)\\ \mathrm{Var}(T) &= \frac{\nu}{\nu-2} &(5.13)\end{aligned}$$

if $\nu > 2$.

Now a statistics T can be constructed if we have N data $x_i \sim N(\mu,\sigma)$ and we form the variables

$$Z = \frac{\hat{x}-\mu}{\sigma/\sqrt{N}} \tag{5.14}$$

which is $N(0,1)$ and

$$X = (N-1)\frac{S^2}{\sigma^2} \tag{5.15}$$

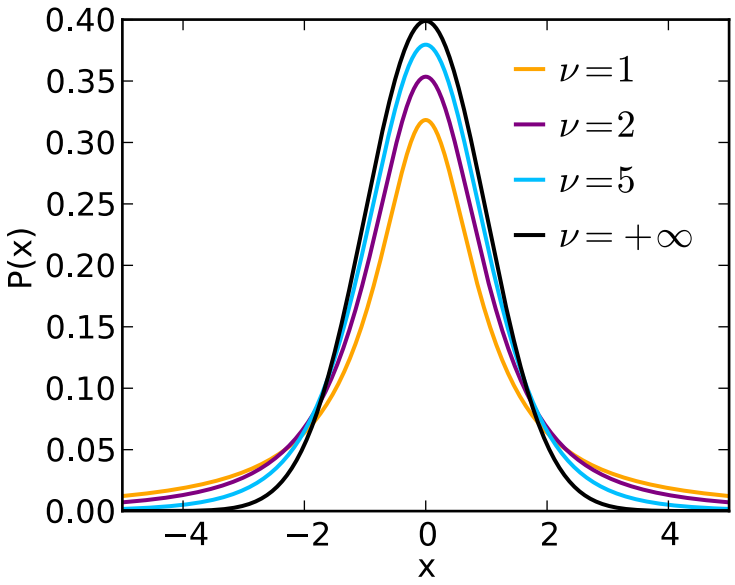

Figure 5.1: PDF of the t-Student variable (By Skbkekas - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9546828)

which is χ^2_{N-1}. Then

$$T = \frac{Z}{\sqrt{X/(N-1)}} = \frac{\hat{x} - \mu}{S/\sqrt{N}} \tag{5.16}$$

is a t-Student variable with $\nu = N - 1$ (and is a statistics if μ is known).

If we have two datasets, we can form a new variable that is *approximately* a t-Student variable:

$$T = \frac{\hat{x}_1 - \hat{x}_2 - (\mu_1 - \mu_2)}{S_D} \tag{5.17}$$

where

$$S_D = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{5.18}$$

and the distribution has a number of d.o.f. equal to

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{5.19}$$

If $\mu_1 - \mu_2$ is known (eg, the two populations are supposed to have the same mean), then T is a statistics.

5.4 Distribution of the ratio of two variances (F-distribution).

Let us state the following theorem: If X, Y are two independent χ^2 variables with ν_1, ν_2 dof, then

$$F = \frac{X/\nu_1}{Y/\nu_2} \tag{5.20}$$

is distributed as

$$P(F; \nu_1, \nu_2) = \frac{\Gamma[(\nu_1 + \nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{F^{\frac{\nu_1 - 2}{2}}}{(1 + F\nu_1/\nu_2)^{(\nu_1+\nu_2)/2}} \tag{5.21}$$

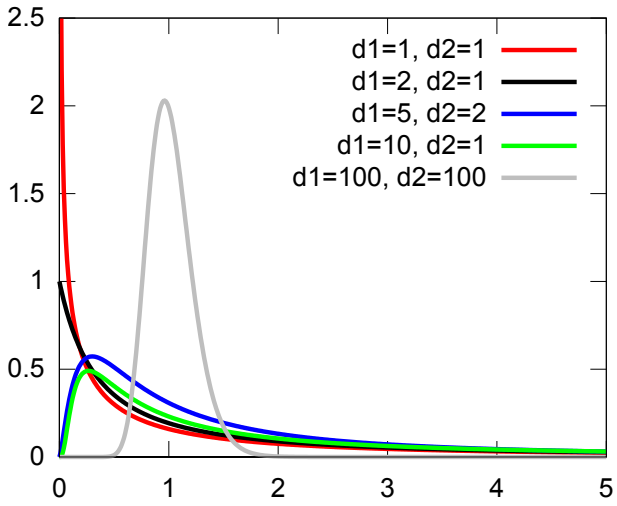

Figure 5.2: PDF of the F variable (By IkamusumeFan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=34777108)

Statistics (sample quantities)	PDF
mean	$G(\mu/\sigma/\sqrt{N})$
variance	χ^2_{N-1}
mean/variance$^{1/2}$	t-Student
variance1/variance2	F-distribution

Figure 5.3: Schematic list of PDF statistics

if $F > 0$ and 0 elsewhere (see Fig. 5.2). One has

$$\langle F \rangle = \frac{\nu_2}{\nu_2 - 2} \tag{5.22}$$

$$\mathrm{Var}(F) = \frac{\nu_2^2(2\nu_1 + 2\nu_2 - 4)}{\nu_1(\nu_2 - 1)^2(\nu_2 - 4)} \tag{5.23}$$

for $\nu_2 > 2$ and> 4, respectively.

Now, if we have two sample variances

$$X \equiv (n_1 - 1)\frac{S_1^2}{\sigma_1^2} \tag{5.24}$$

$$Y \equiv (n_2 - 1)\frac{S_2^2}{\sigma_2^2} \tag{5.25}$$

then the ratio

$$F_{12} = \frac{X/\nu_1}{Y/\nu_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{5.26}$$

is a $F(\nu_1, \nu_2)$ variable (and is a statistics if σ_1/σ_2 is known, eg. the variances of the two populations are the same).

In Table 5.3 we list schematically the four statistics we have introduced, with their PDF.

5.5 Confidence regions

The PDF of the statistics θ will in general depend on some unknown parameter: for instance $\hat{x}$ is distributed as $G(\mu, \sigma/\sqrt{N})$, and in general we do not know μ, σ. Once we know the PDF of a statistics θ, we can answer two questions: 1) how likely is to find the unknown parameters (e.g. μ, σ) in a given region (inference); and 2) how

likely is to obtain the particular value $\bar{\theta}$ that we find in a given experiment, assuming that the unknown paramater is fixed to some hypothetical value (hypothesis testing).

The second question is a well-posed one. The first one however is problematic: how can we determine if $\bar{\theta}$ is likely or not if we don't know the full PDF? The frequentist trick is to replace the unknown parameter, e.g. μ, by its estimate $\hat{x}$. We see how this works in this section, leaving hypothesis testing to the next section.

Let us evaluate $\theta_{1-\alpha/2}$ and $\theta_{\alpha/2}$ defined in this way:

$$\int_{\theta_{1-\alpha/2}}^{\infty} f(\theta)d\theta = \frac{\alpha}{2} \tag{5.27}$$

$$\int_{-\infty}^{\theta_{\alpha/2}} f(\theta)d\theta = \frac{\alpha}{2} \tag{5.28}$$

i.e. as the value of θ that delimits an area equal to $\alpha/2$ or $1-\alpha/2$ when the PDF is integrated from $-\infty$ (or from the lowest value of the domain) to θ. One has therefore

$$P(\theta_{\alpha/2} < \theta < \theta_{1-\alpha/2}) = 1-\alpha \tag{5.29}$$

The region within $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ is the confidence region for θ at level of confidence $1-\alpha$ (see Fig. 5.4).

For instance, suppose we have measured N data and obtained the particular value sample mean $\bar{x}$. If the data are $G(\mu,\sigma)$ the sample mean $\hat{x}$ is $G(\mu,\sigma/\sqrt{n})$. We suppose we know σ and need to find the confidence region for μ. We have then

$$P(\mu - \frac{\sigma}{\sqrt{n}} < \hat{x} < \mu + \frac{\sigma}{\sqrt{n}}) = 0.68 \tag{5.30}$$

and therefore

$$P(\hat{x} - \frac{\sigma}{\sqrt{n}} < \mu < \hat{x} + \frac{\sigma}{\sqrt{n}}) = 0.68 \tag{5.31}$$

However, $\hat{x}$ is a random variable, while we only know the particular value $\bar{x}$ and we cannot in principle trade one for the other. This is a fundamental problem with the frequentist approach, and of the main reason to use the Bayesian approach, which gives directly the distribution of the theoretical parameters. Nevertheless, we *define* the confidence region for μ to be

$$\bar{x} - \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} \tag{5.32}$$

and this will be the frequentist answer one obtains about the confidence region of μ.

Another example. If we have measured the variance $\bar{S}^2$ we can find the confidence region for σ^2 by exploiting the fact that $(n-1)S^2/\sigma^2$ is distributed as a χ^2 variable. Therefore we obtain

$$\chi^2_{\alpha/2} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{1-\alpha/2} \tag{5.33}$$

from which we obtain the confidence region

$$(n-1)\frac{\bar{S}^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < (n-1)\frac{\bar{S}^2}{\chi^2_{\alpha/2}} \tag{5.34}$$

where again we replaced S^2 with $\bar{S}^2$.

5.6 Hypothesis testing

The basic idea of hypothesis testing is to employ the same techniques of this Chapter to answer a related question. Instead of finding the confidence region for a parameter, we try to answer the question of whether the particular value of a given statistics is likely or not when we assume a specific hypothesis on the value of a parameter. The workflow is like this:

- enunciate an hypothesis H_0 concerning one or more parameters of the distribution of the random variables: eg, the mean is zero, the variance is smaller than something etc. We want to test H_0, that is, to see whether it is consistent with the data

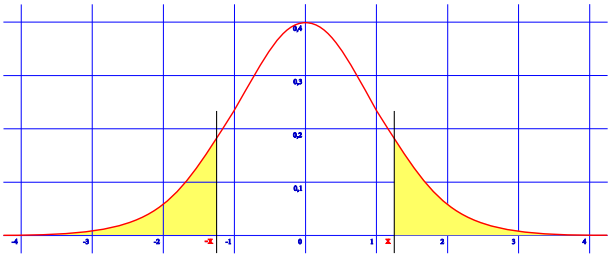

Figure 5.4: Region of confidence: the two yellow regions contain a probability fraction $\alpha/2$ each (By User:HiTe - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=2084849)

- enunciate the alternative hypothesis H_1; normally H_1 is simply: H_0 is false
- define a statistics and measure its particular value in an experiment
- assuming the hypothesis to be true, evaluate the p-value (see below) of the statistics, i.e. the probability that H_0 is true
- if p is smaller than a pre-defined threshold α (or $\alpha/2$ for two-tail tests), the hypothesis is rejected (to $1-p$-level of confidence); if not, we cannot rule it out (which is not the same as saying "H_0 is true"!)

Let us start with an example (see Gregory 7.2.1). Suppose we have the flux from a radiogalaxy as a function of time. We make the hypothesis

- H_0 : the galaxy flux is constant in time
- H_1 : the galaxy flux is not constant in time

We suppose the random variable (the flux) is distributed as a Gaussian with known variance σ (perhaps because we have some extra information, previous experiments etc). We evaluate for our dataset the statistics

$$\bar{\chi}^2 = \sum_i^N \frac{(x_i - \bar{x})^2}{\sigma^2} = 26.76 \tag{5.35}$$

where $N = 18$. A high χ^2 supports H_1, a low one supports H_0.

In order to performt the test of H_0 we need to choose between 1-tail (measuring p using only one side of the distribution, the upper one or the lower one) and 2-tail tests (using both sides). The decision depends entirely on the nature of H_1. If H_1 is supported by high (low) values of the statistics, then we need to employ a upper (lower) 1-tail test to reject H_0. If it is supported by both high and low values we need to adopt a 2-tail test.

Since in our present case H_1 is supported by the high values of χ^2, we employ a upper 1-tail test to estimate the p-value. We know that if H_0 is true, and therefore $\langle x \rangle = \hat{x}$, this statistics is distributed as χ^2_{N-1} and we find the 1-tail p-value, ie. the probability that $\chi^2 > 26.76$ as

$$P(\chi^2_{N-1} > 26.76) = 0.02 \tag{5.36}$$

Therefore we reject the hypothesis H_0 that the flux is constant to 98% confidence level. This means that if we repeat the experiment 50 times and H_0 is true, we will obtain only once a χ^2 larger than this; therefore, we risk making the error of rejecting H_0, while in fact it is true, only 2% of the times. This error is called of Type I.

If we want to minimize this error, we should select a lower threshold to decide if we want to accept H_0 or not, say $\hat{p} = 0.01$: in this case, we would not have rejected H_0. In so doing, however, we risk committing an error of Type II, i.e. failing to reject H_0 while in fact the hypothesis is false. The p-value expresses then the possibility of making errors of two kinds, as we see in Fig. 5.5. Normally one considers an error Type I more serious that Type II, so the value of p should be rather small, e.g. 5% or less.

On the other hand, if the hypothesis to test was H_0 : the flux is variable, then we should have used a lower 1-tail test and obtain as p-value $P(\chi^2_{N-1} < 26.76) = 0.98$ and conclude that we cannot rule out H_0.

Decision	Reality	error type
Reject H_0	actually H_0 is true	Type I (*conviction*)
Reject H_0	indeed $H_0$is false	-
Fail to reject H_0	indeed H_0 is true	-
Fail to reject H_0	actually H_0 is false	Type II (*acquittal*)

Figure 5.5: Table of Errors (adapted from P. Gregory)

In other cases we could need a 2-tailed test. For instance if we want to compare the mean μ_1 of a data sample with the mean μ_2 of another data sample then we can build a $t-$Student test by forming the combination given in Eq. (5.17)

$$T = \frac{\hat{x}_1 - \hat{x}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/N_1 + S_2^2/N_2}} \tag{5.37}$$

Suppose now the hypothesis to test is H_0 : the means are equal, i.e. $\mu_1 = \mu_2$. Then T becomes a statistics (i.e depends on data alone) and we can estimate its value from the datasets. Hypothesis H_0 is clearly supported by a T close to $\langle T \rangle = 0$, while a deviation on either side supports H_1. We need to use then a 2-tail test. This means that if the assigned threshold for rejection of H_0 is α, the null hypothesis is rejected if the value of the statistics lies either in the $\alpha/2$ upper tail (i.e. in the region $T < T_{\alpha/2}$ such that the integral of the PDF from $T_{p/2}$ to infinity is $\alpha/2$) or in $\alpha/2$ lower tail. In practice, the only difference between a 1-tail and a 2-tail test, is that in the second case the threshold should be set at $\alpha/2$ in order to claim a $(1-\alpha)$-confidence level for rejection.

Finally, let us note that the important value to provide out of your data analysis is the p-value. The decision on rejection or otherwise depends on a pre-set level α that normally is purely conventional and depends on the consensus in a particular discipline.

5.7 Testing a linear fit

One of the most useful example of hypothesis testing occurs when we obtain the best fit of a linear model (see Sect. 4.2). We could ask in fact how likely is the hypothesis that the data come from a population described by the best fit parameters.

Suppose that analyzing N Gaussian data Y_i we obtained M best fit parameters $\bar{A}_\alpha$. Defining $T_i = \bar{A}_\alpha g_{\alpha i}$ (sum over α) the best fit function evaluated at the location of the N data points, we can form the statistics

$$Z = (Y_i - T_i) C_{ij}^{-1} (Y_j - T_j) \tag{5.38}$$

As we have seen in Sec. (4.2) the best fit parameters are linear functions of the data Y_i and therefore Z is a quadratic function of the data. By diagonalizing the matrix C_{ij}^{-1} one can rewrite Z as a sum of $N-M$ uncorrelated squared normalized Gaussian data, which shows that Z is indeed a χ^2 variable with $N-M$ degrees of freedom. Since we know the distribution of Z we can now proceed to hypothesis testing. The hypothesis is H_0: the data Y_i are extracted from a multivariate Gaussian distribution with covariance matrix C_{ij} and mean T_i. If the hypothesis is true then Z should be close to its expected value $\langle Z \rangle = N - M$, otherwise we should reject H_0. Notice that a perfect fit, $Z = 0$ or very small, is in contrast with our hypothesis because if the data have some variance then they are not expected to be perfectly aligned on the best fit curve. So we should adopt a two tail test and if $\bar{Z}$ is the particular value we obtain in an experiment, evaluate

$$P(\chi^2 > \bar{Z}) = \int_{\bar{Z}}^{\infty} P_{N-M}(\chi^2) d\chi^2 = p \tag{5.39}$$

when $\bar{Z} > \langle Z \rangle$ and the same integral but in the region $(0, \bar{Z})$ in the opposite case. If the p-value we obtain is smaller than the threshold $\alpha/2$ we have chosen, we should reject H_0. Since a χ^2_ν distribution approximates a Gaussian with mean ν and variance 2ν, a quick indication that the data are not likely to have been extracted from the underlying distribution is obtained if $Z/(N-M)$ deviates from unity in either directions by more than a few times $1/\sqrt{N-M}$.

The function Z is employed also in a slightly different manner, i.e. as a goodness-of-fit indicator. If Z is very small the fit is clearly good. Then we can use its value as a quick estimate of how good a fit is, regardless

of any hypothesis testing. As we have seen in Sect. 3.10, this is just one of the three factors that enter the Bayesian evidence, since indeed $Z = \chi^2_{min}$. It is often the dominant one, which explains why it can be used as an approximation to the full evidence.

5.8 Analysis of variance

A particular case of hypothesis testing is called analysis of variance or ANOVA. Suppose we have p samples from Gaussian distributions with the same variance σ^2 but supposedly different unknown means μ_α, $\alpha = 1, .., p$. Each sample has a different size n_α, so we have $N = \sum_\alpha n_\alpha$ data, that we denote as

- $x_{\alpha i}$, i-th data from α-th sample

We would like to test H_0 : the means are equal, against $H_1 : H_0$ is false. Let us define the sample means

$$\bar{x}_\alpha = \frac{1}{n_\alpha} \sum_i x_{\alpha i} \tag{5.40}$$

and the sample variances

$$\bar{\sigma}_\alpha = \frac{1}{n_\alpha} \sum_i (x_{\alpha i} - \bar{x}_\alpha)^2 \tag{5.41}$$

(notice we use the maximum likelihood estimator here). Then we define the overall mean and variance

$$\bar{x}_{tot} = \frac{1}{N} \sum_{\alpha,i} x_{\alpha i} \tag{5.42}$$

$$\bar{\sigma}^2_{tot} = \frac{Q^2_{tot}}{N} \tag{5.43}$$

where

$$Q^2_{tot} = \sum_{\alpha,i} (x_{\alpha i} - \bar{x}_{tot})^2$$

Now after some algebra we find that we can partition the sum of squares in this way

$$Q^2_{tot} = \sum_{\alpha,i} (x_{\alpha i} - \bar{x}_\alpha)^2 + \sum_\alpha n_\alpha (\bar{x}_\alpha - \bar{x}_{tot})^2 \equiv Q^2_{res} + Q^2_{set} \tag{5.44}$$

i.e. as the sum of the squared residuals inside each set Q^2_{res} plus the sum of squared residuals from set to set, Q^2_{set}. Now Q^2_{res}/σ^2 is a χ^2_{N-p} variable. If H_0 is true, moreover, it can be shown that Q^2_{set}/σ^2 is a χ^2_{p-1} variable. We have then that the statistics

$$F = \frac{Q^2_{set}/(p-1)}{Q^2_{res}/(N-p)} \tag{5.45}$$

is a $F(p-1, N-p)$ variable. Then the particular value F that we obtain in a given experiment can be employed to test hypothesis H_0. In fact, if the means $\bar{x}_\alpha$ are equal, $\bar{x}_\alpha = \bar{x}_{tot}$ so that $Q_{set} \to 0$ and $F \to 0$. A low F supports therefore the hypothesis that the samples come from the same distribution. Since H_1 is supported by high values of F we need to employ a upper 1-tail test.

5.9 Numerical methods

The distributions we have seen in this Chapter are all analytical, due to the simple assumption of Gaussian data. In general, the PDF of a function of the data can be impossible to derive analytically. In this case, one can obtain the numerical PDF of a function of the data $\hat{\theta} = f(d_i)$ by generating many random sets of data d_i extracted from their PDF, evaluating each time $\hat{\theta}$. The histogram of the values of $\hat{\theta}$ gives the approximate numerical PDF of the estimator. Of course one needs to know the data PDF. If one does not have this information then one can still use the non-parametric methods of the next Chapter.

Chapter 6

Frequentist approach: Non-parametric tests

In this section we continue with the frequentist approach to parameter estimation but here we do not assume Gaussian data. We still however assume that the data are *independent and identically distributed* (IID). This case generally applies to data sampling, i.e. a series of independent measurement of some random variable that obeys some unknown distribution. Once we have the statistics distribution, the same method of confidence regions and hypothesis testing we have seen in the previous section applies.

6.1 Pearson χ^2 test for binned data

We start by asking whether the data come from a given distribution. Take a sample of N data d_i, divided into k mutually exclusive bins, with N_i data each. Suppose we know from previous tests or from some theory that a fraction p_k of data should go in bin k. Let us then take as hypothesis

- H_0: for every i, the probability that d_i falls in bin k is p_k.

Under this hypothesis, the number of data expected in bin i is $n_i = Np_i$ and of course $N = \sum n_i$. Let us start with $k = 2$ for simplicity. Then we have a single parameter since the probability that x is in bin 1 is p_1 and in bin 2 is $1 - p_1$. The probability of finding a number x of data in bin 1 given that we expect p_1 is given by the binomial

$$P(x; N, p) = \frac{N!}{(N-x)!x!} p_1^x (1-p_1)^{n-x} \tag{6.1}$$

From this we know that the variance of x is $\sigma^2 = Np_1(1-p_1)$ and that $\langle x \rangle = Np_1$. Then we can form the standardized variable

$$Y = \frac{n_1 - Np_1}{\sqrt{Np_1(1-p_1)}} \tag{6.2}$$

and claim that Y approximates a Gaussian variable with $N(0,1)$. Therefore Y^2 approximates a χ^2_1. After some algebra we find that one can write

$$Y^2 = \sum_{i=1}^{2} \frac{(n_i - Np_i)^2}{Np_i} \tag{6.3}$$

This helps to generalize the procedure to $k > 2$. In fact we find simply that for any k

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{6.4}$$

is a χ^2_{k-1} variable. The alternative H_1 is supported by high values of χ^2 so we employ a upper 1-tail test. Therefore, the hypothesis H_0 that the data come from the distribution p_k is rejected to $1-\alpha$ confidence level if χ^2_{k-1} is found to lie in the upper $\alpha\%$ of the distribution.

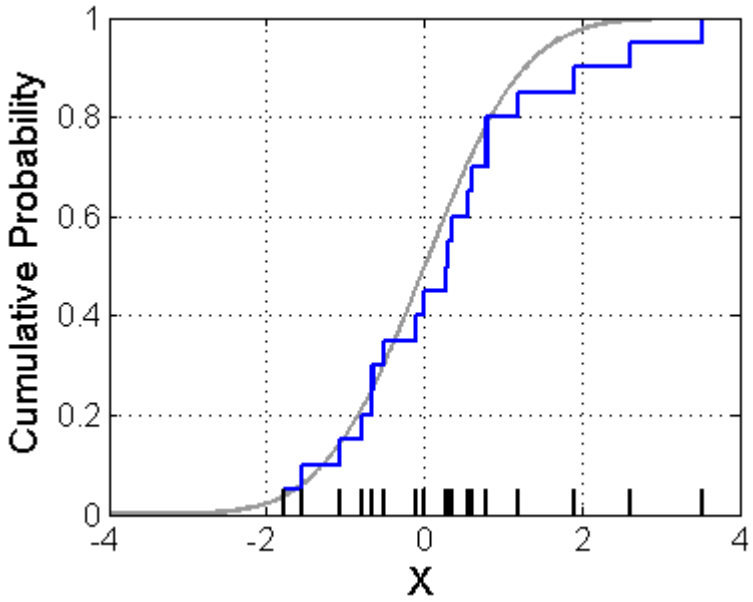

Figure 6.1: Sample cumulative function (by Bscan, commons.wikimedia.org/wiki/File:Empirical_CDF.png).

6.2 Kolmogorov-Smirnov test

In this section, like in the previous one, we do not assume any given distribution for the random variables (the data). Contrary to the previous case, we do not require any binning. Suppose we have n samples x_i from a given unknown distribution whose cumulative function is $F(x)$. Let us define the *sample (or empirical) cumulative function* as

$$F_n(x) = k/n \tag{6.5}$$

where k is the number of data points $x_i \leq x$. For instance if $x_i = 0.1, 0.3, 0.35, 0.7, 0.8, 0.9$, then $F_n(0.4) = 1/2$ since there are 3 values less than 0.4 out of 6 in total. Of course $F_n(x)$ is a highly discontinuous function, i.e. a step-wise constant function (see e.g. Fig. 6.1). For any x less than the lowest observed value, $F_n(x) = 0$ and for any x higher than any observed value $F_n(x) = 1$. If we had the entire set of possible observations, $F_n(x)$ would coincide with the cumulative distribution function. Let us now define the largest deviation of F_n from the true F as

$$D_n = \sup |F_n(x) - F(x)| \tag{6.6}$$

in the entire allowed range of x. An important theorem due to Kolmogorov and Smirnov says that

$$\lim_{n\to\infty} P(n^{1/2} D_n \leq t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 t^2} \equiv H(t) \tag{6.7}$$

regardless of $F(x)$. Under a specific hypothesis on $F(x)$, the quantity D_n is a statistics, i.e. can be evaluated using only the observed data. Clearly the hypothesis H_0 : the distribution is $F(x)$, is supported by a small D_n, while the alternative H_1 by a high D_n, so we use a upper 1-tail test. For instance, if we find that $n^{1/2} D_n = 1.36$ for a given dataset (and a given hypothesis on $F(x)$), we can estimate that $H(1.36) = 0.95$, i.e. the hypothesis H_0 would be rejected at 95% confidence level for any value of $n^{1/2} D_n$ larger than 1.36.

6.3 Kolmogorov-Smirnov test for two samples

One can use the same Kolmogorov-Smirnov statistics also to test if two samples come from identical distributions, without specifying which. In this case the hypotheses will be

- H_0: $F(x) = G(x)$
- H_1: the hypothesis H_0 is not true.

The sample distribution we reconstruct from two given samples of, respectively, m, n data are $F_m(x)$ and $G_n(x)$. It can be shown that the statistics

$$(\frac{nm}{n+m})^{1/2} D_{nm} \tag{6.8}$$

where

$$D_{mn} = \sup |F_m(x) - G_n(x)| \tag{6.9}$$

has a cumulative function $H(t)$ defined as in Eq. (6.7), i.e.

$$\lim_{n\to\infty} P((\frac{nm}{n+m})^{1/2} D_{nm} \leq t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2t^2} = H(t) \tag{6.10}$$

A test using D_{mn} is called a Kolmogorov-Smirnov two-sample test. Here again, one should use a upper 1-tail test.

6.4 Wilcoxon test

Another way to test that two samples come from the same distribution, i.e. to test the same hypothesis H_0 of the previous section, is the Wilcoxon test. Suppose we have m data from a sample X_i and n from a second sample Y_j (as usual all the data are assumed independent). Let's order all of them in a single list from the smallest to the largest one. To each observations we assign a rank r_i from 1 to $m+n$ according to the order in the list. The average of the ranks will of course be $(m + n + 1)/2$. If the hypothesis H_0 is true, then the ranks of the m observations X_i should occur randomly and not being concentrated anywhere in particular, and the same for the n values Y_j. More exactly, r_i should be a uniform random variable in the range $1, m + n$. This implies that the sum S of the ranks of X_i should be close to $E[S] = m(m + n + 1)/2$, i.e. m times the average. Similarly, employing the property of uniform variates Eq. (2.12), one can show that under H_0

$$\mathrm{Var}[S] = \frac{mn(m + n + 1)}{12} \tag{6.11}$$

Now for the central limit theorem, the distribution of S should tend, for large m, n , to a Gaussian with mean $E[S]$ and variance $\mathrm{Var}[S]$, more or less regardless of the true distribution of the data. So H_0 is rejected if

$$Z \equiv \frac{|S - m(m + n + 1)/2|}{[\frac{mn(m+n+1)}{12}]^{1/2}} \geq c \tag{6.12}$$

where if the level of significance is chosen to be α, the constant c is chosen so that

$$c = \Phi^{-1}(1 - \alpha/2) \tag{6.13}$$

where $\Phi(x)$ is the cumulative function of the Normal. Notice that since H_1 is supported by both a large and a small value of Z, we need to use a 2-tail test.

6.5 Bootstrap

We have seen so far several distributions for parameter estimators. Now we consider an estimator of the distribution itself, that was actually already introduced in the Kolmogorov-Smirnov test. Suppose we have n datapoints x_i from a given unknown distribution whose cumulative function is $F(x)$. Let us define the *sample (or empirical) cumulative function* as

$$F_n(x) = k/n \tag{6.14}$$

where k is the number of data points $x_i \leq x$. The sample PDF corresponding to this cumulative function is

$$f_n(x_i) = \frac{1}{n\Delta x_i} \tag{6.15}$$

for $x_i \in (x_i, x_{i+1})$ and where $\Delta x_i = x_{i+1} - x_i$. In fact, $\int_{x_1}^{x_k} f_n(x)dx = \sum_{i=1}^{k} f_n(x_i)\Delta x_i = k/n$ i.e. the cumulative function. In the limit of large n we can expect the sample PDF to approximate the true PDF. Then, we can estimate the distribution of *any* statistics $\hat{\theta}$ function of a sample of size m by simply drawing many m-sized samples from the sample PDF f_n. That is, if we have a statistics that is a function of $m \leq n$ data (normally howevere $m = n$), for instance the sample mean

$$\hat{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{6.16}$$

we generate numerically a large number N of samples of the same size m drawing from the distribution f_n (obtained from the real measurements) and obtain therefore a number N of values of $\hat{x}$ (let us denote them as $\bar{x}_\alpha$, $\alpha = 1, N$). The distribution of the $\bar{x}_\alpha$ approximates the distribution of $\hat{x}$. This can be done for any function of the data, i.e. any statistics.

In fact, the method can be further simplified. The samples to be generated only have to obey the sample PDF, with no restriction on how they are generated. In fact, any value of x within the interval (x_i, x_{i+1}) is as good as any other. So we can restrict ourselves to sampling only among the original n data points. This is called *resampling with replacement*. That is, we can take the n data points, put them in a box, and extract randomly a new set of n values from the box, always "replacing" in the box the values that are extracted. This means that the new set of n points will contain some duplicates of the original dataset and will correspondingly lack some of the original points. We can generate thousands of resampled set in this way, evaluate the statistics for each of them and obtain the PDF of the statistics.

The advantage of this method, called *bootstrap* (from the idiomatic expression "pull yourself up by your bootstraps"), are manifold. First, it applies to any random data without any hypothesis on their distribution (but we still need to assume that the data are independent and identically distributed). Second, it applies to any statistics, i.e. any function of the data. Third, is very easy, requiring no more than a modest computational capacity. The main disadvantage however is that it is only asymptotically unbiased (in the sense that it requires both the sample size m and the number of resamplings N to be very large) and there are no general criteria for deciding how close to the asymptotic distribution we are.

6.6 Sufficient statistics

A statistics $T(d_i)$ such that the joint data probability distribution, characterized by a parameter θ, can be written as

$$P(d_i; \theta) = h(d_i)g(\theta, T(d_j)) \tag{6.17}$$

is called a sufficient statistics for θ for that particular data PDF. Intuitively, this means that the parameter is fully characterized by $T(d_j)$; including more data in the statistics (among the same dataset d_i!) will not improve or change the estimation of θ. Notice, incidentally, that in the Bayesian context this is obvious since $h(d_i)$ does not depend on the parameter and therefore can be absorbed into the posterior normalization. For instance, the maximum likelihood estimator will depend only on maximizing g and therefore only on $T(d_j)$. The sample mean and variance are sufficient statistics for Gaussian data. If (weirdly enough!) we use only $N - 1$ data to evaluate the mean instead of the full set of N data, the sample mean will not be sufficient.

Chapter 7

Random fields: Correlation function and power spectrum

In this Chapter we will present various methods to quantify the degree of correlation of random fields, i.e. random data distributed over a manifold. Typical random fields are the spatial distribution of particles or galaxies in three dimensions or telecommunication signals distributed over a time axis.

7.1 Definition of the correlation functions

Common statistical descriptors for spatially distributed data are the $n-$point correlation functions. Take N particles in a volume V. Let $\rho_0 dV$ be the average number of particles in an infinitesimal volume dV, being $\rho_0 = N/V$ the average number density. Let n_a be the number of particles in a small volume dV_a. Then by definition $\langle n_a \rangle = \rho_0 dV_a$. If $dN_{ab} = \langle n_a n_b \rangle$ is the average number of *pairs* in the volumes dV_a and dV_b(i.e., the product of the number of particles in one volume with the number in the other volume), separated by r_{ab}, then the 2-point correlation function $\xi(r_{ab})$ is defined as

$$dN_{ab} = \langle n_a n_b \rangle = \rho_0^2 dV_a dV_b (1 + \xi(r_{ab})) \tag{7.1}$$

If the points are uncorrelated, then the average number of pairs is exactly equal to the product of the average number of particles in the two volumes, and the correlation ξ vanishes; if there is correlation among the volumes, on the other hand, then the correlation is different from zero. The correlation function is also defined, equivalently, as the spatial average of the product of the density contrast $\delta(r_a) = n_a/(\rho_0 dV) - 1$ at two different points

$$\xi(r_{ab}) = \frac{dN_{ab}}{\rho_0^2 dV_a dV_b} - 1 = \langle \delta(r_a)\delta(r_b) \rangle \tag{7.2}$$

In practice it is easier to derive the correlation function as the average density of particles at a distance r from another particle. This is a *conditional* density, that is the density of particles at distance r given that there is a particle at $r = 0$. The number of pairs is then the number of particles in both volumes divided by the number of particles $n_a = \rho_0 dV_a$ in the volume dV_a at $r = 0$:

$$dn_b|_{n_a} = dN_{ab}/n_a = \rho_0^2 dV_a dV_b (1 + \xi(r_{ab}))/n_a = \rho_0 dV_b (1 + \xi(r_b)) \tag{7.3}$$

Identifying $dN_b|_{n_a}$ with the conditional number dN_c, the correlation function can then be defined as

$$\xi(r) = \frac{dN_c(r)}{\rho_0 dV} - 1 = \frac{\langle \rho_c \rangle}{\rho_0} - 1 \tag{7.4}$$

i.e. as the average number of particles at distance r from any given particle (or number of neighbors), divided by the expected number of particles at the same distance in a uniform distribution, minus 1, or *conditional density contrast.* If the correlation is positive, there are then more particles than in a uniform distribution: the distribution is then said to be positively clustered. This definition is purely radial, and does not distinguish between isotropic and anisotropic distributions. One could generalize this definition by introducing the anisotropic correlation function

as the number of pairs in volumes at distance r and a given longitude and latitude. This is useful whenever there is some reason to suspect that the distribution is indeed anisotropic, as when there is a significant distortion along the line-of-sight due to the redshift.

If the average density of particles is estimated from the sample itself, i.e. $\rho_0 = N/V$, it is clear that the integral of $dN_c(r)$ must converge to the number of particles in the sample :

$$\int_0^R dN_c(r) = \int \rho(r) dV = N \tag{7.5}$$

In this case the correlation function is a sample quantity, and it is subject to the integral constraint (Peebles 1980)

$$\int_0^R \xi_s(r) dV = N/\rho_0 - V = 0 \tag{7.6}$$

Assuming spatial isotropy this is

$$4\pi \int_0^R \xi_s(r) r^2 dr = 0 \tag{7.7}$$

If the sample density is different from the true density of the whole distribution, we must expect that the $\xi_s(r)$ estimated in the sample differs from the true correlation function. From Eq. (7.4), we see that $g(r) = 1 + \xi(r)$ scales as ρ_0^{-1}. Only if we can identify the sample density ρ_0 with the true density the estimate of $\xi(r)$ is correct. In general, the density is estimated in a survey centered on ourselves, so that what we obtain is in reality a conditional density.

The conditional density at distance r from a particle, averaged over the particles in the survey, is often denoted in the statistical literature as $\Gamma(r)$; we have therefore from Eq. (7.4)

$$\Gamma(r) \equiv \langle \rho_c \rangle = \rho_0 (1 + \xi) \tag{7.8}$$

The average in spherical cells of radius R and volume V of this quantity is denoted as

$$\Gamma^*(R) \equiv \langle \rho_c \rangle_{sph} = \rho_0 (1 + \widehat{\xi}) \tag{7.9}$$

where

$$\widehat{\xi} = V^{-1} \int \xi dV \tag{7.10}$$

To evaluate $\Gamma^*(R)$ one finds the average of the number of neighbors inside a distance R from any particle contained in the sample.

7.2 Measuring the correlation function in real catalog

Consider now the estimator (7.4). It requires the estimation of the density ρ_c inside a shell of thickness dr at distance r from every particle. In other words, it requires the estimation of the volume of every shell. In practice, a direct estimation of the shell volume is difficult because of the complicate boundary that a real survey often has. Moreover, if we are working on a magnitude-limited sample, the expected density ρ_0 must take into account the selection function. The simplest way to measure ξ is to compare the real catalog to a MonteCarlo (i.e. random, or more exactly Poissonian) catalog with exactly the same number of particles, the same boundaries and the same selection function. Then, the estimator can be written as

$$\xi = \frac{DD}{DR} - 1 \tag{7.11}$$

where DD means we center on a real galaxy (data D), count the number DD of galaxies at distance r, and divide by the number of galaxies DR at the same distance but in the MonteCarlo catalog (label R). In other words, instead of calculating the volume of the shell, which is a difficult task in realistic cases, we estimate it by counting the galaxies in the Poissonian MonteCarlo realization. In this way, all possible boundaries and selection function can be easily mimicked in the Poisson catalog, and will affect DD and DR in the same way (statistically). To reduce the effect of the Poisson noise in the MC catalog, we can in fact use a higher number of artificial particles, say α times the real data, and then multiply DD/DR by α.

7.3 Correlation function of a planar distribution

Let us estimate now the CF of a planar distribution. Consider two large spherical volumes of radius R_s. Let us distribute in one N particles uniformly on a plane passing through the center, and in the other the same N particles but now uniformly in the whole volume. The latter is our MonteCarlo artificial catalog. We have to estimate DD in a spherical shell at distance r from the planar distribution, and DR in the artificial one. In the planar world, the spherical shell cuts a circular ring of radius $r \ll R_s$ and thickness dr, so we have, on average

$$DD = \text{superf. density} \times 2\pi r dr = \frac{N}{\pi R_s^2} 2\pi r dr \tag{7.12}$$

In the uniform world we have

$$DR = \text{density} \times 4\pi r^2 dr = \frac{3N}{4\pi R_s^3} 4\pi r^2 dr \tag{7.13}$$

Then we get

$$\xi = \frac{2R_s}{3} r^{-1} - 1 \tag{7.14}$$

This is the CF of a planar distribution. As we can see, $1 + \xi$ goes as r^{-1}, and its amplitude depends on the size of the "universe" R_s. It is clear that, in this case, the amplitude of the correlation function is not a measure of the amount of inhomogeneity of the content, but rather a measure of the geometry of the container.

Notice that the constraint 7.7 is satisfied:

$$\int_0^{R_s} \xi r^2 dr = \frac{2R_s}{3} \int^{R_s} r dr - \int^{R_s} r^2 dr = \frac{R_s^3}{3} - \frac{R_s^3}{3} = 0$$

7.4 Correlation function of random clusters

Consider now the following model: there are m clumps of N particles each, uniformly distributed inside cubes of side R_c, and the cubes themselves are distributed uniformly in the universe. The total number of particles is mN . The total volume is mD^3, if $D \gg R_c$ is the mean intercluster distance. For $r \ll R_c$, each particle sees around it a uniform distribution with density $\rho_c = N/R_c^3$, while the global mean density is $\rho = mN/(mD^3) = N/D^3$. It follows

$$\xi(r \ll R_c) = \frac{N}{R_c^3} \frac{D^3}{N} - 1 = \left(\frac{D}{R_c} \right)^3 - 1 \tag{7.15}$$

On the other hand, for $r \gg D$, the distribution of particles is essentially random, and

$$\xi(r \gg D) = 0 \tag{7.16}$$

There are therefore three regimes: at very small scales, the CF is constant and positive; at large scales, the model is homogeneous, and at intermediate scales it decreases from one plateau to the other. Notice however that, in order to verify the integral constraint, the CF must become negative at some intermediate scale. This corresponds to the fact that outside the clusters there are less particles that in a uniform distribution. Notice also that now the CF amplitude does not depend on the universe volume, but only on the fixed parameters D and R_c.

7.5 The angular correlation function

Sometimes we need to project the 3D correlation function onto a 2D plane. For instance, because the angular position of the galaxies is so much easier to determine than their distance, the angular correlation function has been often employed in astronomy. Here we write down the relation between the two correlations, that is the Limber equation, in order to show some properties.

Let $\Phi(r; m_{\rm lim})$ denote the radial selection function,

$$\Phi(r; m_{\rm lim}) = \int_{-\infty}^{M(r, m_{\rm lim})} \phi(M) dM \tag{7.17}$$

that is the density of galaxies at distance r in a magnitude-limited, such that $\int \Phi dV = N$ is the total number of sources selected. Since the density at distance r is $\Phi(r)$, instead of the constant ρ_0, the number of pairs in volumes dV_1, dV_2 at positions $\mathbf{r}_1, \mathbf{r}_2$ is now modified as follows

$$dN_{12} = dV_1 dV_2 (1 + \xi(r_{12}))\Phi(r_1)\Phi(r_2) \tag{7.18}$$

where

$$r_{12} = |\mathbf{r}_1 - \mathbf{r}_2| = \left(r_1^2 + r_2^2 - 2r_1 r_2 \cos\theta\right)^{1/2}$$

Now, the number of pairs dN_θ which appear to be separated by an angle θ in the sky is clearly the integral of dN_{12} over all positions $\mathbf{r}_1, \mathbf{r}_2$ provided that their angular separation θ is constant. Then we have

$$dN_\theta = \int dN_{12} = \int dV_1 dV_2 (1 + \xi(r_{12}))\Phi(r_1)\Phi(r_2) \tag{7.19}$$

The angular correlation function is defined, in analogy to the spatial correlation

$$w(\theta) = \frac{dN_\theta}{\rho_s^2 dA_1 dA_2} - 1 \tag{7.20}$$

where ρ_s is the surface density, and $\rho_s dA_1 = \left(\int_{V_1} \Phi dV\right)$ is the expected number of particles in the area dA that subtends the volume V_1 (e.g., dA_1 is a circular patch of angular radius α and V_1 is the line-of-sight cone of beam size α). Then we obtain the relation between spatial and angular correlation functions:

$$w(\theta) = \frac{\int dV_1 dV_2 \xi(r_{12})\Phi(r_1)\Phi(r_2)}{\left(\int \Phi dV\right)^2} \tag{7.21}$$

In the limit of small separations, this equation can be simplified. If $\xi(r_{12})$ declines rapidly for large separations, we might assume that the integral is important only if $r_1 \simeq r_2 \simeq r$; if we also take a small θ we have

$$r_{12}^2 = (r_1 - r_2)^2 + r^2\theta^2 = u^2 + r^2\theta^2 \tag{7.22}$$

where $u = r_1 - r_2$. Passing from r_1, r_2 to u, r in the integral, and integrating out the angular variables, we get the Limber equation

$$w(\theta) = \frac{\int_0^\infty r^4 \Phi(r)^2 dr \int_{-\infty}^\infty du \xi(x)}{\left(\int r^2 \Phi dr\right)^2} \tag{7.23}$$

where $x^2 = u^2 + r^2\theta^2$. A simple use of this equation is when a power law approximation holds, $\xi = Ar^{-\gamma}$. Then we can define a variable z such that $u = \theta r z$, and we obtain

$$\begin{aligned} w(\theta) &= \frac{\int_0^\infty r^4 \Phi(r)^2 dr \int_{-\infty}^\infty \theta r dz \left[(\theta r z)^2 + (r\theta)^2\right]^{-\gamma/2}}{\left(\int r^2 \Phi dr\right)^2} \\ &= \frac{\int_0^\infty r^4 \Phi(r)^2 dr \int_{-\infty}^\infty \theta^{1-\gamma} r^{1-\gamma} dz \left[z^2 + 1\right]^{-\gamma/2}}{\left(\int r^2 \Phi dr\right)^2} = B\theta^{1-\gamma} \end{aligned} \tag{7.24}$$

where the coefficient B is just some number that depends on $m_{\rm lim}$ and γ

$$B = \frac{\int_0^\infty r^{5-\gamma} \Phi(r)^2 dr \int_{-\infty}^\infty dz \left[z^2 + 1\right]^{-\gamma/2}}{\left(\int r^2 \Phi dr\right)^2} \tag{7.25}$$

Eq. (7.24) reveals that, in the limit of small angular scales and negligible correlations at large distances, the angular power law is $1 - \gamma$. This is in fact roughly confirmed in several angular catalogues, although in a limited range of angular scales.

7.6 The n-point correlation function and the scaling hierarchy

The correlation function can be generalized to more than two points. The 3-point function for instance is defined as

$$\varsigma(r_a, r_b, r_c) = \langle \delta(r_a)\delta(r_b)\delta(r_c) \rangle \tag{7.26}$$

In terms of the counts in infinitesimal cells we can write

$$\begin{aligned} \varsigma(r_a, r_b, r_c) &= \langle \left(\frac{n_a}{\rho_0 dV_a} - 1 \right) \left(\frac{n_b}{\rho_0 dV_b} - 1 \right) \left(\frac{n_c}{\rho_0 dV_c} - 1 \right) \rangle \\ &= \frac{\langle n_a n_b n_c \rangle}{\rho_0^3 dV_a dV_b dV_c} - \xi_{ab} - \xi_{bc} - \xi_{ac} - 1 \end{aligned} \tag{7.27}$$

so that we obtain the useful relation

$$\langle n_a n_b n_c \rangle = \rho_0^3 dV_a dV_b dV_c (1 + \xi_{ab} + \xi_{bc} + \xi_{ac} + \varsigma_{abc}) \tag{7.28}$$

7.7 The power spectrum

One of the most employed statistical estimator for density fields is the power spectrum. In recent years it has been used to quantify the clustering properties in many galaxy surveys. The main reason is that almost all theories of structure formation predict a specific shape of the spectrum, because the plane waves evolve independently in the linear approximation of the gravitational equations.

Unless otherwise specified, the conventions for the 3D Fourier transforms is

$$\begin{aligned} f(x) &= \frac{V}{(2\pi)^3} \int f_k e^{ikx} d^3k \\ f_k &= \frac{1}{V} \int f(x) e^{ikx} d^3x \end{aligned} \tag{7.29}$$

and it is always understood that $ikx = i\mathbf{k} \cdot \mathbf{x}$. With this conventions, $f(x)$ and f_k have the same dimensions. However, the Dirac delta will be defined as customary as

$$\begin{aligned} \delta_D(x) &= \frac{1}{(2\pi)^3} \int e^{ikx} d^3k \\ \delta_D(k) &= \frac{1}{(2\pi)^3} \int e^{ikx} d^3x \\ \int \delta_D(k) d^3k &= \int \delta_D(x) d^3x = 1 \end{aligned} \tag{7.30}$$

Let $\delta(x)$ be the density contrast of a density field and

$$\delta_k = \frac{1}{V} \int \delta(x) e^{ikx} dV \tag{7.31}$$

its Fourier transform.

Notice that δ_k is a complex quantity but that $\delta_{\boldsymbol{k}}^* = \delta_{-\boldsymbol{k}}$. If $\delta(x)$ is sampled in cells of size V_c that form a grid of size V (e.g. the output of a N-body simulation), then the k grid will be defined as cells of size $(2\pi)^3/V$ that form a grid of size $(2\pi)^3/V_c$. The total number of random variables δ_k is the same as the total number of random variables $\delta(x)$: although δ_k is a complex quantity, the fact that $\delta_{\boldsymbol{k}}^* = \delta_{-\boldsymbol{k}}$ means half of the coefficients are redundant.

The sample power spectrum is defined as

$$P(k) = V \delta_k \delta_k^* \tag{7.32}$$

Notice that the power spectrum has the dimension of a volume. It follows

$$P(k) = \frac{1}{V} \int \delta(x)\delta(y) e^{ik(x-y)} dV_x dV_y \tag{7.33}$$

Now, we put $r = x - y$, and we identify the average $\langle ... \rangle$ with a spatial average, i.e. we assume that the sample is so large that distant regions are completely uncorrelated and act therefore as indepedent realizations. Then

$$\xi(r) = \langle \delta(y+r)\delta(y) \rangle = \frac{1}{V} \int \delta(y+r)\delta(y) dV_y \tag{7.34}$$

then,

$$P(k) = \int \xi(r) e^{ikr} dV \tag{7.35}$$

Therefore, the power spectrum is the Fourier transform of the correlation function (Wiener-Khintchin theorem). The inverse property is

$$\xi(r) = (2\pi)^{-3} \int P(k) e^{ikr} d^3k \tag{7.36}$$

(notice that here, following most literature, the Fourier volume factor is not included). Finally, assuming spatial isotropy, i.e. that the correlation function depends only on the modulus $|r|$, we obtain

$$P(k) = 4\pi \int \xi(r) \frac{\sin kr}{kr} r^2 dr \tag{7.37}$$

A more general definition of power spectrum can also be given, but this time we have to think in terms of ensemble averages, rather than volume averages. Consider in fact the ensemble average of $V\delta_k \delta^*_{k'}$:

$$V\langle \delta_k \delta^*_{k'} \rangle = \frac{1}{V} \int \langle \delta(y)\delta(y+r) \rangle e^{i(k-k')y + ikr} dV_r dV_y \tag{7.38}$$

Performing ensemble averages, one has to think of fixing a positions and making the average over the ensemble of realizations. Then the average can enter the integration, and average only over the random variables δ. Then we obtain

$$V\langle \delta_k \delta^*_{k'} \rangle = \frac{1}{V} \int \xi(r) e^{i(k-k')y + ikr} dV_r dV_y = \frac{(2\pi)^3}{V} P(k) \delta_D(k - k') \tag{7.39}$$

The definition (7.39) states simply that modes at different wavelengths are uncorrelated if the field is statistically homogeneous (that is, if ξ does not depend on the position in which is calculated but only on the distance r). This will often be useful later.

These definitions refer to infinite samples and to a continuous field. In reality,we always have a finite sample and a discrete realization of the field, i.e.. a finite number of particles. Therefore, we have to take into account the effects of both finiteness and discreteness.

Consider now a discrete distribution of N particles, each at a position x_i; then we can use Dirac's delta and write

$$\rho(x) = \sum_i \delta_D(x - x_i) \tag{7.40}$$

so that $\int \rho(x) dV = N$, as it should. The Fourier transform of the density contrast $\delta = \rho(x)/\rho_0 - 1$ is now

$$\delta_k = \frac{1}{V} \int \frac{V}{N} \sum_i \delta_D(x - x_i) e^{ikx} dV - (2\pi)^3 \delta_D(k) = \frac{1}{N} \sum_i e^{ikx_i} \tag{7.41}$$

whete the last equality is only valid for $k \neq 0$. Now, the expected value of the power spectrum is

$$P(k) = V\langle \delta_k \delta^*_k \rangle \tag{7.42}$$

that is

$$P(k) = \frac{V}{N^2} \sum_{ij} e^{ik(x_i - x_j)} \tag{7.43}$$

Finally, if the positions x_i and x_j are uncorrelated, we can pick up only the terms with $i = j$, so that we obtain the pure noise spectrum

$$P_n(k) = \frac{V}{N^2} \sum_i 1 = \frac{V}{V} \tag{7.44}$$

We now redo the same calculation including the effect of the *window function*$W(x)$, a function that expresses the way in which the particles are selected. A typical selection procedure is to take all particles within a given region, and no particles elsewhere. In this case, the function will be a constant inside the survey, and zero outside. We will always consider such a kind of window function in the following, and normalize it so that

$$\int W(x)dV = 1 \tag{7.45}$$

With this normalization, $W(x) = 1/V$ inside the survey. The density contrast field we have in a specific sample is therefore the universal field times the window function (times the sample volume V because of the way we normalized W)

$$\delta_s = \delta(x)VW(x) \tag{7.46}$$

Let us now again express the field as a sum of Dirac functions

$$\delta(x) = \left(\frac{\rho(x)}{\rho_0} - 1\right) VW(x) = \frac{V}{N}\sum_i w_i \delta_D(x - x_i) - VW(x) \tag{7.47}$$

where $w_i = VW(x_i)$. The Fourier transform is

$$\delta_k = \frac{1}{V}\int \left(\frac{V}{N}\sum_i w_i \delta_D(x - x_i) - VW(x)\right) e^{ikx} dV = \frac{1}{N}\sum_i w_i e^{ikx_i} - W_k \tag{7.48}$$

where we introduced the $k-$space window function

$$W_k = \int W(x) e^{ikx} dV \tag{7.49}$$

normalized so that $W_0 = 1$. The most commonly used window function is the so-called top-hat function, which is the FT of the simple selection rule

$$\begin{aligned} W(x) &= 1/V \quad \text{inside a spherical volume } V \text{ of radius } R \\ W(x) &= 0 \quad \text{outside} \end{aligned} \tag{7.50}$$

We have then

$$\begin{aligned} W_k &= \int W(x) e^{ikx} dV = V^{-1} \int e^{ikx} dV \\ &= \frac{3}{4\pi} R^{-3} \int^R r^2 dr \int_{-\pi}^{\pi} e^{ikr\cos\theta} d\cos\theta d\phi \\ &= \frac{3}{2} R^{-3} \int^R \left(e^{ikr} - e^{-ikr}\right) \frac{r^2}{ikr} dr \\ &= 3R^{-3} \int^R \frac{r \sin kr}{k} dr = 3\frac{\sin kR - kR\cos kR}{(kR)^3} \end{aligned}$$

Notice that $W_0 = 1$, and that the WF declines rapidly as $k \to \pi/R$ (see Fig. 7.1). Now, the expected value of the power spectrum is

$$P(k) = V\langle \delta_k \delta_k^* \rangle \tag{7.51}$$

that is

$$P(k) = \frac{V}{N^2}\sum_{ij} w_i w_j e^{ik(x_i - x_j)} - VW_k^2 \tag{7.52}$$

We used the relation

$$\langle \frac{1}{N}\sum_i w_i e^{ikx_i} \rangle = \frac{1}{N}\sum_i \int W(x) e^{ikx} dV = W_k \tag{7.53}$$

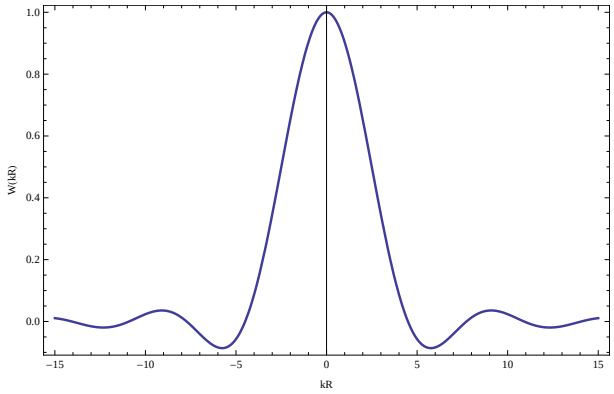

Figure 7.1: Top-hat spherical window function.

Finally, if the positions x_i and x_j are uncorrelated, we can pick up only the terms with $i = j$, so that, neglecting the window function, which is important only for $k \to 0$, we obtain the pure noise spectrum

$$P_n(k) = \frac{V}{N^2} \sum_i w_i^2 = V/N \tag{7.54}$$

where the last equality holds only if $w_i = 1$ for all particles. The noise spectrum is negligible only for large densities, $\rho_0 = N/V \to \infty$. In general, the noise is not always negligible and has to be subtracted from the estimate. For the power spectrum applies the same consideration expressed for the moments: the power spectrum does not characterize completely a distribution, unless we know the distribution has some specific property, e.g. is Gaussian, or Poisson, etc.

7.8 From the power spectrum to the moments

The power spectrum is often the basic outcome of the structure formation theories, and it is convenient to express all the other quantities in terms of it. Here we find the relation between the power spectrum and the moments of the counts in random cells.

Consider a finite cell. Divide it into infinitesimal cells with counts n_i either zero or unity. We have by definition of ξ

$$\langle n_i n_j \rangle = \rho_0^2 dV_i dV_j [1 + \xi_{ij}] \tag{7.55}$$

The count in the cell is $N = \sum n_i$. The variance is then $M_2 = (\langle N^2 \rangle - N_0^2)/N_0^2$ where

$$\begin{aligned} \langle N^2 \rangle &= \langle \sum n_i \sum n_j \rangle = \sum \langle n_i^2 \rangle + \sum \langle n_i n_j \rangle = \\ & N_0 + N_0^2 \int dV_i dV_j W_i W_j [1 + \xi_{ij}] \end{aligned} \tag{7.56}$$

where $N_0 = \rho_0 V$ is the count average, and $\xi_{ij} \equiv \xi(|\mathbf{r}_i - \mathbf{r}_j|)$. Let us simplify the notation by putting

$$W_i dV_i = dV_i^*$$

We define the integral (by definition $\int W dV = \int dV^* = 1$ for any window function)

$$\sigma^2 = \int dV_1^* dV_2^* \xi_{12} \tag{7.57}$$

Inserting the power spectrum we have

$$\sigma^2 = (2\pi)^{-3} \int P(k) e^{i\mathbf{k}(\mathbf{r}_1 - \mathbf{r}_2)} W_1 W_2 d^3k d^3r_1 d^3r_2 \tag{7.58}$$

This becomes, for *spherical cells*,

$$\sigma^2 = (2\pi^2)^{-1} \int P(k) W^2(k) k^2 dk \tag{7.59}$$

Finally we obtain the relation between the power spectrum (or the correlation function) and the second-order moment of the counts:

$$M_2 = N_0^{-2} \langle (\Delta N)^2 \rangle = N_0^{-2} (\langle N_i^2 \rangle - N_0^2) = N_0^{-1} + \sigma^2 \tag{7.60}$$

where $\Delta N = N - N_0$. The first term is the noise, the second term is the count variance in the continuous limit.

For the third order moment we proceed in a similar fashion:

$$\langle N^3 \rangle = \langle \sum n_i \sum n_j \sum n_k \rangle = \sum \langle n_i^2 \rangle + 3 \sum \langle n_i^2 \rangle \sum n_i + \sum \langle n_i n_j n_k \rangle = \tag{7.61}$$

$$N_0 + 3N_0^2 + N_0^3 \int dV_i^* dV_j^* dV_k^* [1 + \xi_{ij} + \xi_{ik} + \xi_{jk} + \varsigma_{ijk}] \tag{7.62}$$

where in the last equality we used the definition of the three point correlation given in Eq. (7.28)

$$\langle n_i n_j n_k \rangle = \rho_0^3 dV_i dV_j dV_k [1 + \xi_{ij} + \xi_{ik} + \xi_{jk} + \varsigma_{ijk}] \tag{7.63}$$

The third order moment is then

$$M_3 = N_0^{-3} \langle (\Delta N)^3 \rangle = N_0^{-2} + \int dV_i^* dV_j^* dV_k^* \varsigma_{ijk} \tag{7.64}$$

If we can assume the scaling relation $\varsigma_{ijk} = Q[\xi_{ij}\xi_{jk} + \xi_{ij}\xi_{ik} + \xi_{ik}\xi_{jk}]$ then we can express M_3 in terms of $P(k)$ and of the new parameter Q. In the limit of large N_0, a Gaussian field ($M_3 = 0$) has $Q = 0$.

7.9 Bias in a Gaussian field

Consider a Gaussian density field with correlation $\xi(r)$. By definition, we have that the fluctuation density contrast field $\delta = \delta\rho/\rho$ obey the rules

$$\langle \delta_1 \delta_2 \rangle = \xi(r) \tag{7.65}$$

$$\langle \delta_1^2 \rangle = \xi(0) = \sigma^2 \tag{7.66}$$

The density δ at each point is distributed then as

$$P(\delta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{\delta^2}{2\sigma^2}\right]$$

where by definition $\sigma^2 = \int \delta_1^2 P(\delta) d\delta$. The probability that the fluctuation field is above a certain threshold $\nu\sigma$, where σ is the field variance, is

$$P_1 = \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{\nu\sigma} \exp\left[-\frac{\delta^2}{2\sigma^2}\right] d\delta \tag{7.67}$$

Now, the joint probability that the density at one point is δ_1 and the density at another is δ_2,

$$P(\delta_1, \delta_2) = \left[(2\pi)^2 \det M\right]^{-1/2} \exp\left[-\frac{1}{2}\delta^i \delta^j M_{ij}\right] \tag{7.68}$$

where $\delta^i = \{\delta_1, \delta_2\}$ and where r is the distance between the two points. The covariance matrix is

$$M^{-1} = \begin{pmatrix} \sigma^2 & \xi(r) \\ \xi(r) & \sigma^2 \end{pmatrix} \tag{7.69}$$

We can write then the probability that the x field is above the threshold ν at both location as

$$\begin{aligned} P_2 &= \left[(2\pi)^2 \det M\right]^{-1/2} \int_{\nu\sigma}\int_{\nu\sigma} \exp\left[-\frac{1}{2}\delta^i\delta^j M_{ij}\right] d\delta_i d\delta_2 \\ &= \left[(2\pi)^2 \left(\sigma^4-\xi^2\right)\right]^{-1/2} \int_{\nu\sigma}\int_{\nu\sigma} \exp\left[-\frac{\sigma^2\delta_1^2+\sigma^2\delta_2^2-2\xi\delta_1\delta_2}{2\left[\sigma^4-\xi^2\right]}\right] d\delta_1 d\delta_2 \end{aligned} \tag{7.70}$$

Now, suppose there are N particles in the field; the number of particles in regions above threshold is $N_1 = P_1 N$, while the number of pairs in regions above threshold is $N_2 = P_2 N^2$. The correlation function of the regions above threshold is, by definition of correlation function

$$1+\xi_\nu = \frac{N_2}{N_1^2} = \frac{P_2}{P_1^2} \tag{7.71}$$

The integral can be done easily numerically, but an interesting approximation is to take the limit for $\xi(r) \ll 1$ and $\nu \gg 1$, i.e. at large scales and for high peaks. Using the large ν approximation (Abramovitz-Stegun 7.1.23)

$$P_1 = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \int_{\nu\sigma} \exp\left[-\frac{\delta^2}{2\sigma^2}\right] d\delta \simeq \frac{1}{\left(2\pi\nu^2\right)^{1/2}} e^{-\nu^2/2} \tag{7.72}$$

and expanding

$$\begin{aligned} \exp\left[-\frac{\sigma^2\delta_1^2+\sigma^2\delta_2^2-2\xi\delta_1\delta_2}{2\left[\sigma^4-\xi^2\right]}\right] &\simeq \exp\left[-\frac{\delta_1^2+\delta_2^2}{2\sigma^2}\right] \exp\left[\frac{\xi\delta_1\delta_2}{\sigma^4}\right] \\ &\simeq \exp\left[-\frac{\delta_1^2+\delta_2^2}{2\sigma^2}\right]\left(1+\frac{\xi\delta_1\delta_2}{\sigma^4}\right) \end{aligned} \tag{7.73}$$

we get

$$\frac{P_2}{P_1^2} \simeq 1 + \frac{\nu^2}{\sigma^2}\xi e^{-\nu^2} \int_{\nu\sigma} e^{-\frac{\delta_1^2+\delta_2^2}{2\sigma^2}} \delta_1\delta_2 \frac{d\delta_1 d\delta_2}{\sigma^4}$$

and finally (Kaiser 1984)

$$\xi_\nu \simeq \frac{\nu^2}{\sigma^2}\xi \tag{7.74}$$

This shows that peaks are more correlated than the background density field. This is the same effects one observes on mountain ranges: near a peak there is very likely another peak. Eq. (7.74) gives some credibility to the approximation usually made that the galaxy density field is a scale-independent-biased version of the mass density field, but it should be noticed that this is expected only for large ν and at large scales, and that the whole mechanism relies on the assumption that there is only one object per threshold region.

Eq. (7.74) can be applied to galaxy clusters. Suppose first we smooth the field on scales of, say, 5 Mpc/h, so that the variance σ on such scales is of order unity. It is found observationally that the cluster correlation function is roughly ten times larger than the galaxy correlation. This would imply a $\nu \simeq 3$, which is not unreasonable. Notice that some level of biasing is necessary: collapsed object form only where $\delta > 1$.

7.10 Poissonian noise

Among infinite possible ways to characterize a distributions of particles, the n-point correlation functions, or their integral average, the moments, are often selected because of their straightforward definition, and because of their easy numerical computation. It is often necessary to think of a distribution of particles as a *finite* and *discrete* sample drawn from an underlying field. We need then to distinguish between the properties of the underlying field, that we sometimes refer to as the "universe", and the properties of the sample under investigation: the sample gives only an estimate of the universe. If we want to infer the properties of the universe from those of the sample, we need to take into account both the finiteness and the discreteness. In particular, we need to assume a model for the behavior of the field *beyond* the sample, and "*beneath* " the particles that sample the field. Two typical assumptions are that the field outside the sample looks like the field inside (fair sample hypothesis), and that the particles are a

Poisson sampling of the continuous field (Poisson sampling hypothesis). Both assumptions can be tested only when is possible to obtain a larger and denser sample. Lacking this possibility, which is often the case in cosmology, the assumptions need to be treated with great caution.

Let us begin with the moments. Suppose we partition a distribution of particles into m cells, for instance spherical cells of radius r, *randomly located*, and count the number of particles inside each cell. This gives the counts N_i, with $i = 1, .., m$. Then we can form the number density contrasts

$$\delta_i = (N_i - N_0)/N_0 \tag{7.75}$$

where N_0 is the average count

$$N_0 = \sum N_i/m \tag{7.76}$$

and we can form the p-th order central moment

$$M_p = \langle {\delta_i}^p \rangle = m^{-1} \sum_i {\delta_i}^p \tag{7.77}$$

By definition, $M_0 = 1, M_1 = 0$. Suppose now that the probability to have a density contrast between δ and $\delta + d\delta$ is $P(\delta)d\delta$, where $P(\delta)$ is the probability density function (PDF) of the counts. The moments M_p of the particle distribution are an *estimate* of the moments of the PDF: in fact, in the limit in which we sample the full distribution, the moments M_p coincide with the moment of the PDF. For instance, the second order moment, M_2 is an estimate of the variance of the number density contrasts. The third order moment is called skewness, while the combination

$$K \equiv M_4 - 3M_2^2 \tag{7.78}$$

is the kurtosis. If the PDF is a gaussian

$$P(\delta) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \tag{7.79}$$

then all its moments depend on σ and both the skewness and the kurtosis vanish. These moments are therefore the simplest estimator of deviation from gaussianity.

In practice we estimate the moments from a finite distribution of particles, i.e. a discrete random sampling of a continuous underlying field, for instance the dark matter field. The number of particles at any given point is then a function of the continuous field. In cosmology, this function is established by the physical processes that led to the formation of the discrete particles, the galaxies, and can be in general extremely complicated. As already mentioned, the simplest assumption we can make is that the galaxies are a Poisson sampling of the mass field, that is, the galaxies trace the mass. In this case, the average density of galaxies in any given region is proportional to the average density of the underlying field. A slightly more complicated assumption can be that galaxies are a Poisson sampling not everywhere, but only when the underlying field is above a certain threshold. This is what is often referred to as biased formation . It is clear that the true physical process can be much more complicated than this, for instance the threshold may vary in space, or the sampling function can be non-local, etc.. In most of what we will say here, the simplest Poisson assumption is always understood.

Assuming Poisson sampling, we immediately encounter the problem of Poisson noise. That is, the number of particles N at a point in which the density field is ν, is a random variable distributed as a Poisson variable with mean proportional to ν, say equal to $\eta = \beta\nu$, that is

$$P(N;\eta) = \frac{e^{-\eta}\eta^N}{N!} \tag{7.80}$$

If η is distributed as $f(\eta)$, then the PDF of N is

$$P(N) = \int f(\eta) P(N;\eta) d\eta$$

The moments of N are then a function of the moments of $f(\eta)$ and of $P(N;\eta)$. If we are interested in the properties of the underlying field, we need to estimate the moments of η, and of the density contrast $\delta\eta/\eta_0$, from the moments

of N. This can be done easily exploiting the properties of the generating functions, defined above. We have in fact that

$$\int \frac{\eta^N}{N!} dN = e^{\eta} \tag{7.81}$$

so that the GF of $P(N)$ is

$$\begin{aligned} G(\phi) &= \int e^{\phi N} f(\eta) \frac{e^{-\eta}\eta^N}{N!} d\eta dN = \int e^{-\eta} f(\eta) d\eta \int \frac{\left(e^{\phi}\eta\right)^N}{N!} dN \\ &= \int e^{-\eta} f(\eta) d\eta e^{e^{\phi}\eta} = \int e^{\psi\eta} f(\eta) d\eta \end{aligned} \tag{7.82}$$

where

$$\psi = e^{\phi} - 1 \tag{7.83}$$

Then, we have obtained the useful result that the GF of a PDF convolved with the Poisson distribution is the GF of the PDF with the change of variable $\phi \to e^{\phi} - 1$. Then we get

$$M_2(N) = \frac{d^2 G(\psi(\phi))}{d\phi^2}|_{\phi=0} = \frac{d\psi}{d\phi}\frac{d}{d\psi}\left(\frac{d\psi}{d\phi}\frac{dG(\psi)}{d\psi}\right)|_{\phi=0} = M_2^* + M_1^* \tag{7.84}$$

where the moments refer to the PDF of N, and the starred moments refer to the underlying field. . Since the first moment of the PDF of N is the mean N_0, we get finally that the variance including the Poisson sampling is

$$M_2(N) = M_2^*(N) + N_0 \tag{7.85}$$

and that, consequently, the variance of the underlying field is obtained as

$$M_2^*(N) = M_2(N) - N_0 \tag{7.86}$$

The moments of the density contrast of the continuous field (labelled with an asterisk) in terms of the moments of the discrete realization of N_0 particles can be obtained further dividing by N_0^p:

$$M_2^* = M_2 - N_0^{-1}, \tag{7.87}$$

$$M_3^* = M_3 - 3M_2 N_0^{-1} + 2N_0^{-2}, \tag{7.88}$$

$$M_4^* = M_4 - 3M_2^2 - 6M_3 N_0^{-1} + 11M_2 N_0^{-2} - 6N_0^{-3}, \tag{7.89}$$

where the terms in N_0 are the Poisson terms.

In conclusion, a simple way to describe a distribution of particles is to estimate the lowest moments in cells of varying radius, that is to evaluate $M_2(r), M_3(r), K(r)$ and so on. However, no finite amount of moments do characterize completely the distribution, unless of course we know already that the distribution depends on a finite amount of parameters, e.g. is Gaussian or Poisson, etc.

Bibliography

[1] Trotta R., *Bayes in the Sky,* Contemporary Physics, 49, 71

[2] March M., R. Trotta, L. Amendola, D. Huterer, *Robustness to systematics*, MNRAS 415, 143 (2011)

[3] L. Amendola & S. Tsujikawa, *Dark Energy*, CUP 2010

[4] P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press

[5] M. DeGroot and M. Schervish, *Probability and Statistics*, Addison Wesley