Goodness of fit.

January 25, 2019

The purpose of this short note is to develop the concept of goodness of fit and specially, how to compare the relative value of two models with different number of parameters. These concepts of statistics are based on the concept of probabilities and a short introduction to the necessary tools of probability is given. It is however assumed that the reader is familiar with the basic concepts of probabilities.

1 Introduction.

On of the basic tasks of any researcher is to measure some data and compare it to a given model. This is called *fitting*, and many questions might arise :

1. Are my data and model compatible ?
2. Having a choice between two alternating models, which one better describes my data ?

The last question seems to many people the hardest one when different models have different number of parameters. Obviously, the model with more parameters would better fit the data, but how many more parameters can be accepted as reasonable[1] ?

Many people have been traumatized by their courses in statistics and find these questions difficult. However, they are not harder than the question "I have rolled a dice 100 times and the average I get is 4.23; should I be suspicious of this dice ?" In the following, we'll try to answer this kind of question, but let us note that the answer will depend on the two key parameters : "100 times" and the value 4.23.

2 The difference between probability and statistics.

The main assumption behind probabilities is the following : you have a random variable (say a dice, or you're partner attitude before dinner) and you make an infinite number[2] of measurements (realizations). Based on these measurements, you establish the probability P_i, or the probability density $p(x)$ that your random variable would produce the value i or fall between x and $x + dx$. Of course, no finite being can make an infinite number of measurements[3], but mathematicians don't really care. An other way of finding the probability law of a random

[1] As the famous saying goes : "give me five parameters, I'll fit an elephant ; give me 10, I'll make it dance" .

[2] I *mean* it : not a large number, an infinite one.

[3] We can get close, as any macroscopic thermodynamic system at equilibrium will confirm that all microscopic states follow the Gibbs law.

process is, instead of making measurements, making assumptions. Assuming that the dice is unbiased, I deduce that the probability of getting 5 is 1/6. Assuming that the number of incoming calls now is independent of incoming calls in the past will give you a Poisson process, and so on and so forth.

Now comes the hard work of the statistician : by making a finite number of measurements (realizations), he has to make an assumption about the validity of the hypothesis : outright reject it or not reject it (it will always be hard to make him accept something).

So here the fundamental rift between these peoples : tell me how many measurement you do and I'll tell you who you are.

3 A little probabilities.

3.1 Means.

Let us suppose that we have a random variables X, described by a probability density $p(x)$. What is its expectation $\langle X \rangle$? To answer this question exactly, we should measure this variable an infinite number of time, sum up them and divide them by the number of measurements[4]. As we had done the infinite measurement in order to establish $p(x)$, we can skip the measurements and use directly our knowledge of $p(x)$:

$$\langle X \rangle = \int_I x p(x) dx$$

What we are doing mathematically is to collect all measurements which fall between x and $x + dx$ (this step was already done when establishing $p(x)$) , multiply them by x and sum them up. A french teacher measuring the average grades of his students does the same thing : collect all term papers which have grade 0, 1, 2, ... 20 ; multiply each grade by the proportion of term papers and sum them up.

Of course, the usual mean is nothing exceptional, we could have measured other kind of means, like $\langle f(X) \rangle$: again, make infinite measurements, add a column to your table and for each measurement x in your table, compute $f(x)$. Then sum up this last column and divide by the number of measurements :

$$\langle f(X) \rangle = \int_I f(x) p(x) dx$$

There are two widely other kind of means used by mathematicians : the Variance $\left\langle (X - \langle X \rangle)^2 \right\rangle = \langle X^2 \rangle - \langle X \rangle^2$ and the characteristic function

$$\phi(s) = \left\langle e^{isX} \right\rangle$$

or some other variants such as the probability generating function $\phi(z) = \left\langle z^X \right\rangle$. Basically, the characteristic function is another name for the Fourier (or Laplace) transform of probability densities.

[4] dividing infinities is done through a limit process $\lim_{N \to \infty} \sum_{i=1}^{N} x_i / N$

Examples.

- For a binary, equiprobable process with values ± 1,

$$\phi(s) = \frac{1}{2}e^{is} + \frac{1}{2}e^{-is} = \cos(s)$$

- For a discrete Poisson process $p(n;\lambda) = e^{-\lambda}\lambda^n/n!$,

$$\phi(z) = e^{\lambda(z-1)}$$

- For a Normal distribution

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  The characteristic function is

$$\phi(s) = e^{-i\mu s}e^{-s^2\sigma^2/2}$$

  This is one reason for the love affair between probabilists and the Normal distribution : the characteristic function is also a Gaussian, which plays the role of a *stable* function for the FT.

3.2 Sum and product of random variables.

Having two *independent*[5] RV X and Y, we can make new RVs such as $Z = X+Y$ and $Z' = XY$. Each time we make a measure of X and Y, we sum up (or multiply) these values and call it a measurement for Z (or Z'). It comes to adding a new column to our measurements table. Its fairly easy to show that

$$\begin{aligned} \langle X+Y\rangle &= \langle X\rangle + \langle Y\rangle \\ \langle XY\rangle &= \langle X\rangle \langle Y\rangle \end{aligned}$$

We can fairly easily compute the probability density for Z or Z'. For example,

$$p_Z(z) = \int_I p_X(x)p_Y(z-x)dx$$

which is just a convolution product. What is nice about the convolution product is that in the Fourier space, it transforms into a normal product.

Here we had in mind two *different* RVs. What about the sum $Z = X + X$? Again, from a measurement point of view, this consists of making a measurement for X, then another measurement for X and then summing them. Obviously, this is very different from measuring X and then multiplying it by 2. We can generalize that to multiple addition:

$$Z_N = \sum_{i=1}^{N} X_i$$

where X_i are the same RVs. Let us set $\mu = \langle X\rangle$ and $\sigma^2 = \langle X^2\rangle - \langle X\rangle^2$. It is straightforward to show that

$$\langle Z_N\rangle = N\langle X\rangle$$

[5]The independence is not necessary for the sum.

On the other hand,

$$\begin{aligned}\left\langle Z_N^2 \right\rangle &= \sum_{i=1}^{N} \langle X_i \rangle^2 + \sum_{i,j=1, i\neq j}^{N} \langle X_i \rangle \langle X_j \rangle \\ &= N\left(\sigma^2 + \mu^2\right) + N(N-1)\mu^2 \\ &= N\sigma^2 + N^2\mu^2\end{aligned}$$

So for the variance we have

$$\mathrm{Var}(Z_N) = N\mathrm{Var}(X)$$

The very important result is that when grouping the results of a RV by packets of N elements, both the mean and the variance are multiplied by N : they grow linearly as the packet size !

We need one more elementary concept : the RVs $X + a$ and aX, where a is a constant.

$$\begin{aligned}\langle X + a \rangle &= \langle X \rangle + a \\ \langle aX \rangle &= a \langle X \rangle\end{aligned}$$

So let us define $Z = aX$. Obviously, we have

$$\left\langle Z^2 \right\rangle = a^2 \left\langle X^2 \right\rangle$$

Now having all that in hand, we can look at the average of a packet of N identical independent random variables :

$$\bar{Z}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

from what we said, we easily get the fact that

$$\langle \bar{Z}_N \rangle = \langle X \rangle \quad (1)$$

$$\mathrm{Var}(\bar{Z}_N) = \frac{1}{N}\mathrm{Var}(X) \quad (2)$$

We see the crucial fact here : the variance of the packet average is *reduced* by the size of the packet. This is why when we have imperfect measure apparatus, we make N measurements and average them.

The Normal distribution. Usually, computing the real probability distribution of a sum of random variables is cumbersome. An exception is the Normal distribution. We now that the convolution product of two Gaussian is again a Gaussian, so if $X = \mathcal{N}(\mu, \sigma^2)$, based on what we said above

$$\bar{Z}_N = \frac{1}{N} \sum_{i=1}^{N} X_i = \mathcal{N}(\mu, \sigma^2/N)$$

As we will often sum RVs, this is another reason for our love of this function. As the Gaussian is the *fixed point* of addition, it is not hard to show that whatever a reasonable RV X, $\bar{Z}_N$ becomes a Normal distribution when N is large. This most celebrated result is called the central limit theorem.

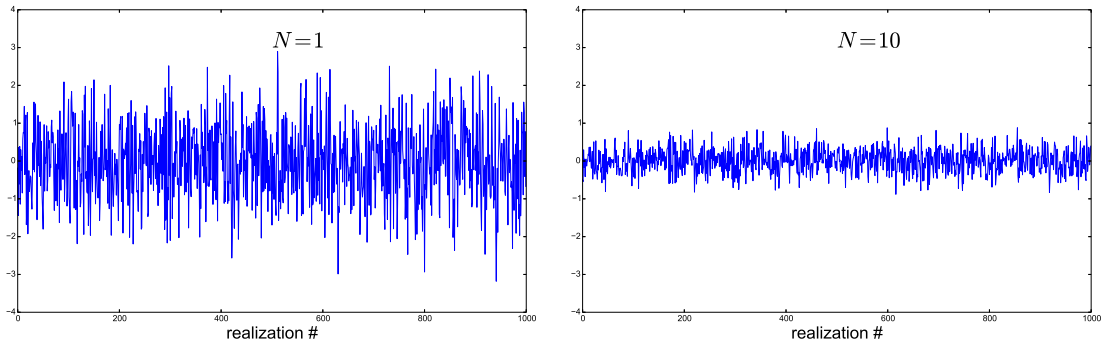


Figure 1: The effect of packet size. In the left panel, 1000 measurements of a random variable have been made and their value duly plotted. On the right panel, 10000 measures have been made ; these data have been grouped in packet of size $N = 10$ (1000 packets), and the average of each packet is reported on the plot. Note the decrease in the spread.

3.3 Change of variable.

Let us again consider a RV X. Each time we make a measurement, we report the value x_i of X for this measure, and then in a new column, the value $y_i = f(x_i)$. We have made a new random variable $Y = f(X)$. knowing the probability density $p_X(x)$ of the original variable, what is the probability density $p_Y(y)$ of the new variable ? The answer is straightforward:

$$\begin{aligned} p_X(x)dx &= P(x < X < x + dx) \\ &= P(f(x) < Y < f(x + dx)) \\ &= P(f(x) < Y < f(x) + f'(x)dx) \\ &= p_Y(f(x))f'(x)dx \end{aligned}$$

Keeping in mind that $y = f(x)$, we have

$$p_X(x)dx = p_Y(y)dy \tag{3}$$

Or in a slightly more complicated notation

$$p_Y(y) = p_X\left(f^{-1}(y)\right)\frac{dx}{dy} \tag{4}$$

Of course, some care should be taken when

- $f'(x) < 0$. Looking back at our derivation, we see that we have to use $|f'(x)|$
- The function $f^{-1}(x)$ is multivalued. As usual, we have to cut the space into parts where $f^{-1}(.)$ is univaluate and add the results.

Example 1. Let $Y = 2X$. Here, $y = f(x) = 2x$ and $x = f^{-1}(y) = (1/2)y$. So

$$p_Y(y) = \frac{1}{2}p_X(y/2)$$

If for example $X = \mathcal{N}(\mu, \sigma^2)$, then $Y = \mathcal{N}(2\mu, (2\sigma)^2)$.

Example 2. Consider the continuous Poisson process T, which for example gives the probability density for the time $t \in [0, \infty[$ between two successive incoming call :

$$p_T(t) = \mu e^{-\mu t}$$

Consider now the change of variable $Y = e^T$ or equivalently $T = \log(Y)$. Note that $Y \in [1, \infty[$. Then

$$\begin{aligned} p_Y(y) &= p_T(\log y)/y \\ &= \frac{\mu}{y^{\mu+1}} \end{aligned}$$

We see here something dangerous. As $\mu > 0$, this is indeed a probability density because

$$\int_1^\infty p_Y(y)dy = 1$$

But this is a *long tail* kind of probability which decrease very slowly. In particular, moments $\langle Y^n \rangle$ don't exist if $n > \mu$. We see that how we measure a RV can induces biases and philosophical problems.

Example 3 : χ^2. Consider a Normal RV $X = \mathcal{N}(0, 1)$ and the variable $Y = X^2$ or $X = \pm\sqrt{Y}$. We have here a multivalued inverse function so we have to consider $X > 0$ and $X < 0$ separately. Let us first look at positive X :

$$\begin{aligned} p_Y(y) &= p_X(\sqrt{y})\frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}} \end{aligned}$$

doing the same thing for the other half of the space and adding these we have

$$p_Y(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-y/2}$$

This is a most important RV for statistics and it is called the χ^2 distribution with one degree of freedom.

3.4 The χ^2 and Student's distribution.

The χ^2 distribution plays such an important role in statistics that we must spend a little time developing it. Consider a Normal RV $X = \mathcal{N}(0, 1)$ and the new RV

$$Y_N = \sum_{i=1}^{N} X_i^2$$

Which is called the χ^2 with N degree of freedom. As we have shown above, the probability density of Y_1 is given by

$$p_1(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-y/2} \tag{5}$$

from which we can compute[6] the mean and variance of Y_1:

$$\langle Y_1 \rangle = 1 \;\; ; \;\; \mathrm{Var}(Y_1) = 2$$

A small amount of computation[7] shows that its characteristic function is

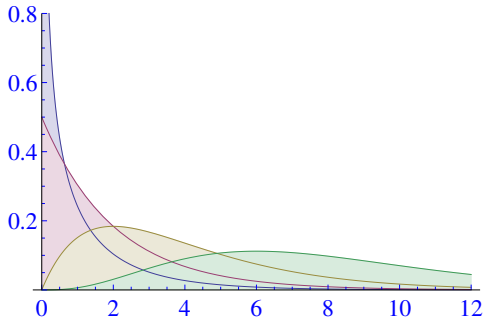
$$\phi_1(s) = \left\langle e^{isy} \right\rangle = (1 - 2is)^{-1/2}$$

The characteristic function of Y_N is therefore

$$\phi_N(s) = (1 - 2is)^{-N/2}$$

and taking the inverse Fourier transform thus gives us

$$p_N(y) = A_N y^{N/2-1} e^{-y/2} \tag{6}$$



A_N is a normalization constant which is[8]

$$A_N = \frac{1}{2^{N/2}\Gamma(N/2)}$$

The figure shows the distribution for $N = 1, 2, 4, 8$.

Finally, from what we know from the addition of random variables we get

$$\langle Y_N \rangle = N \;\; ; \;\; \mathrm{Var}(Y_N) = 2N$$

We will see the χ^2 distribution popping up at at many places. If N is large, we can even drop the exact expression (6) and use a Normal distribution $\mathcal{N}(N, 2N)$, as we know that the central limit theorem

One more distribution we will also encounter is the Student's distribution. Let's $X = \mathcal{N}(0, 1)$ and Y_N a χ^2 RV with N degree of freedom. Then the Student's RV Z is defined by

$$Z = \frac{X}{\sqrt{Y/N}}$$

and its probability density given by

$$p_N(z) = B_N \left(1 + \frac{z^2}{N}\right)^{-\frac{N+1}{2}} \tag{7}$$

where B_N is a normalization constant

$$B_N = \frac{1}{\sqrt{\pi N}} \frac{\Gamma\left((N+1)/2\right)}{\Gamma\left(N/2\right)}$$

[6] It is not hard to show that

$$\int_0^\infty y^n p_1(y) dy = \frac{1}{\sqrt{\pi}} 2^n \Gamma(n + 1/2)$$

by the very definition of the function $\Gamma(x)$.

[7] This involves only integration of Gaussians, when a correct change of variable has been made.

[8] The continuous function $\Gamma(x)$ generalizes the factorial : $\Gamma(n+1) = n\Gamma(n) = n!$. The two particular values of interest are $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

4 A little basic statistics.

4.1 The main question.

As we said at the beginning, the statistics, in contrast to probability, is concerned with a finite number of measures. From example, I suppose that a RV follows a given law, I make *one* measurements x and then ask "what is the probability of obtaining x" ? The answer to this question, if the RV is continuous, is ... zero of course. The probability of getting 0.1265398753 from a normal distribution is zero.

A better formulated question is "supposing that I know the RV, what is the probability of getting a value *as large* as the one I observe ?". If the probability density is given by $p(u)$ and I measure a value x, then the probability of observing a value that large (or larger) is

$$P(>x) = \int_x^\infty p(u)du$$

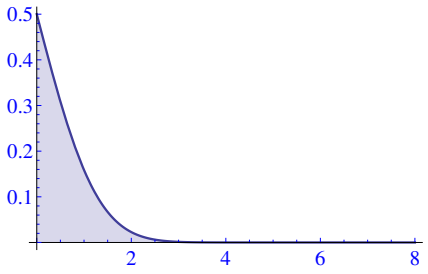
Suppose that we suppose a $\mathcal{N}(\mu, \sigma^2)$ random variable, and then measure a value $x > \mu$. Then the probability for a value that large is

$$\begin{aligned} P(>x) &= \int_x^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-(u-\mu)^2/2\sigma^2} du \\ &= \int_{(x-\mu)/\sigma}^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \end{aligned} \tag{8}$$

We see that the answer to this question is pretty simple and depends only on how relatively far we are from μ. For a normal distribution, being 1σ larger[9] than the average is about 0.15 (0.30 if we disregard the sign), 2σ is 0.02 (about 5% without the sign) and 10^{-3} for larger than 3σ. If we have measured something at 5σ, then we can have a serious doubt about the validity of our hypothesis. 5σ is the gold standard in particle physics and this is how many particles such as Z^0 and W were discovered, by discarding the hypothesis that the observed trajectories could be explained by the known particle of this time.

The integral in (8) is so often used that it has received a name and its values have been tabulated:

$$\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \frac{1}{2}\text{erfc}(z/\sqrt{2})$$



Many questions we will ask in the following would be of this kind. The crucial thing about statistics is to formulate the question correctly, and then reject the hypothesis if the measurements is far out of our acceptable range for these hypothesis.

4.2 Measuring.

A typical question for the experimentalist is : "I have measured this value for my physical variables ; is it compatible with my hypothesis" ?

[9] For RVs having positive and negative values, the question to be evaluated is more often "what is the probability of getting a value outside the region $](\mu - \delta)/\sigma, (\mu + \delta)/\sigma[$.

Let us make this question more precise. Let us suppose that each measurements gives me back a value blurred with the precision of my measurements. If the "true" value is x, the value I measure is

$$y = x + \eta$$

where η is called the noise (or uncertainties) of the measure. We suppose first that the random variable η is centered, or else there is something very wrong with the measurement apparatus. Next we will suppose that (Hypothesis H1) η is a Normal variable, *i.e.* $\eta = \mathcal{N}(0, \sigma^2)$. Because we know the apparatus, we suppose that we know the value of σ. The random variable Y is obviously $\mathcal{N}(x, \sigma^2)$.

Supposing that the measurement noise follows a Normal law is widespread. There are many reasons behind it. Very often, an apparatus is a sum of different parts, each having its own noise, and we know that the sum of many independent variables converges to a Normal distribution. The other reason is the fact that we know how to handle normal law, so we stick to it. This hypothesis can be revised without too much hassle if we know very well the noise of our apparatus.

Let us first calibrate our instruments by measuring a known quantity x. We make N measurements y_1, ...,y_N. If our hypotheses are correct, we know that the random variable

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

should behave as $\mathcal{N}(x, \sigma^2/N)$.

Therefore, we average our N measures

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

and evaluate how probable it is to obtain a value that large, given the fact that $\bar{y}$ is drawn from a distribution $\mathcal{N}(x, \sigma^2/N)$.

The quantity $\sigma/\sqrt{N}$ is called the standard error (σ itself is called as you know the standard deviation). Therefore, if our $\bar{y}$ is at 5 standard error, we can be pretty sure that something is fishy in our hypothesis : (i) the real quantity is not x as we had supposed ; (ii) the standard deviation of our apparatus in not σ, but something bigger ; (iii) the noise of the apparatus is not normally distributed. Statistics cannot answer these questions.

The standard error. The standard error is one of the most crucial parameters when doing statistics, and one can often bypass more serious computations such as the χ^2 or the Student's $t-$or $f-$ test by having a fast evaluation of how plausible a hypothesis is. Of course, when you are writing a paper, you have to show that you have done these more serious tests and have a nice $p-$values. However, most often than not, people don't understand these $p-$values (we'll come to that later) and are confident that the black box program which produced it can be trusted.

As a practical example, consider that you have two classes, each composed of N students. The average grades in your first class is $\bar{y}_1$ (say 11/20) and in your second class is $\bar{y}_2$ (9/20). Is something very different between these classes (quality of the teacher or students, severity of the exams, some cheating, ...) or is this just random ? A fast estimation would be the following

: We make the hypothesis that (H1) these classes are similar, (H2) a student's grade follow a normal distribution.

We then mix all the data to get an estimation of the mean and the variance

$$\mu = \frac{1}{2N}\sum_{i=1}^{2N} y_i \ ; \ \sigma^2 = \frac{1}{2N-1}\sum_{i=1}^{2N}(y_i-\mu)^2$$

We now compute the average of each classes

$$\bar{y}_1 = \frac{1}{N}\sum_{i=1}^{N} y_i \ ; \ \bar{y}_2 = \frac{1}{N}\sum_{i=N+1}^{2N} y_i$$

If our hypotheses are correct, $\bar{y}_1$ and $\bar{y}_2$ are drawn from $\mathcal{N}(\mu, \sigma^2/N)$, so their difference $\bar{y}_1 - \bar{y}_2$ is drawn from $\mathcal{N}(0, 2\sigma^2/N)$.

Suppose that we have measured $\sigma = 5$ and the two classes are composed of 10 students. The standard error for the difference is $\sqrt{2}5/\sqrt{10} = 5/\sqrt{5} \approx 2.2$; on the other hand, the difference between the average of the two classes $11 - 9 = 2$, which is around one standard error. We cannot reject the hypothesis that this difference is just random.

On the other hand, if the two classes were composed of 100 students, the standard error of the difference would be $\sqrt{2}5/\sqrt{100}\approx 0.7$ and this time the difference is at 3 standard errors, which should grab our attention. We can then do much finer analysis of the situation.

4.3 Measuring II.

Let us suppose that we have calibrated our preceding apparatus in the lab and we are now making real measurements. We have measured N data y_i and wonder if these measurements are sound and can be trusted, based on what we know from our apparatus. This time, we don't know the real value of the quantity we are measuring, so we make an estimate for it

$$x = \frac{1}{N}\sum_{i=1}^{N} y_i$$

We can know for example make an estimate of the variance of the data and compare it to the value σ of our apparatus. This first estimation however is not very good as it will crucially depend on the number of measurements we have made. Let us go one step further and compute

$$V = \sum_{i=1}^{N} \frac{(y_i - x)^2}{\sigma^2}$$

If our hypotheses are correct, V is drawn from a χ^2 distribution with $N-1$ degree of freedom. The reason behind $N-1$ instead of N is that we have already used the data once to estimate the mean[10]. Now we can ask this more precise question : "what is the probability that we could have obtained a value that large (or larger) :

$$P(\geq V) = \int_V^\infty p_{N-1}(u)du$$

where $p_{N-1}(u)$ is the p.d. of the χ^2 distribution with $N-1$ degree of freedom given by relation (6).

[10] We can make this argument much more sound mathematically.

Example. Consider that our apparatus in $\mathcal{N}(0,1)$ and we have measured five values $5, 8, 12, 9, 9$. Here we have $x = 8.6$ and $V = 25.2$. Looking up a table for the cumulative distribution of χ^2 distribution or performing an integration by a computer, we find that

$$P(\geq 25.2) = 4 \times 10^{-5}$$

which is a pretty low value. So we can reject the hypothesis that these data our compatible with our apparatus measurements. Something has happened, for example there is environmental noise added to the noise of apparatus. We note that if $\sigma = 2$, then

$$P(\geq 25.2) = 0.18$$

which cannot be used to reject the hypothesis: we can have some trust in our measures. Note that the standard deviation of our data is 2.5, which is much higher than 1 (in the first case), but not much different from 2 (in the second case). Computing the $p-$values allows us to estimate how much this difference is relevant.

5 Fitting : goodness of fit.

Let us now consider that the signal x we measure by our apparatus depends on some parameter t : $x = x(t)$. We make different measurements for different value of our parameter $y_i = x(t_i) + \eta$. For the moment, let us suppose that our apparatus has the same noise σ for all value of t.

We have a nice theory which gives us the exact form of the function $x_1(t; a, b)$ where a, b are some parameters of the model. We want first estimate a, b and second estimate if the theory has any credibility. As before, we compute

$$V(a,b) = \sum_{i=1}^{N} \frac{(y_i - x_1(t_i; a, b))^2}{\sigma^2}$$

and find a, b by minimizing $V(a, b)$,*i.e.*

$$\frac{\partial V}{\partial a} = 0 \quad ; \quad \frac{\partial V}{\partial b} = 0$$

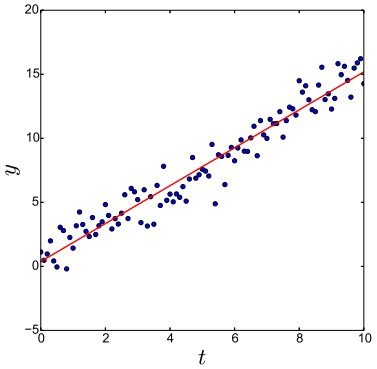
OK, well done, this is what we do all the time. But how good is our fit ? How much can we trust it ? Is it enough for the fit to be graphically sound ?

Well, we know the answer : V is drawn from a χ^2 distribution with $N - 2$ degree of freedom, so it is straightforward to estimate the goodness of fit and give it a $p-$value.

Example. Consider the plot of our 101 measurements, when the apparatus noise amplitude is $\sigma = 1$. We can suspect that the data can be described by the model

$$x_1(t) = a + bt$$

Following the preceding steps, we find that $a = 0.38$ and $b = 1.48$. Pushing the evaluation, we find that $V = 123.2$. Comparing it



to a χ^2 with 99 degree of freedom, its $p-$value
is

$$P(\geq V) = 0.05$$

This not exceptionally good or bad, we cannot reject it out of hand. For a Normal distribution, this is the probability of being at 2σ.

Varying amplitude of the noise. We have supposed that the amplitude of the noise is constant for all values of the parameter t. This may not be the case and we could have $\sigma = \sigma(t)$. This is not a serious problem, as the quantity

$$V(a,b) = \sum_{i=1}^{N} \frac{(y_i - x_1(t_i;a,b))^2}{\sigma(t_i)^2}$$

generalizes our discussion.

6 Fitting II : best model.

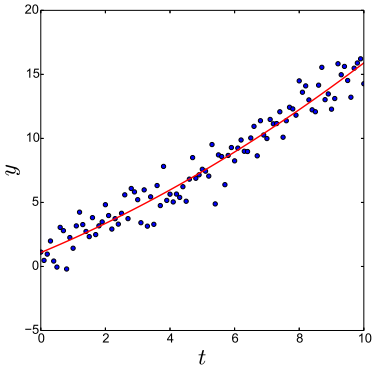
More often than not, we are faced with the question of deciding between two alternative models $x_1(t)$ and $x_2(t)$ to describe our data. If the data have the same parameters, we can estimate which one significantly enhance the $p-$value. But very often, the choice is between models with different number of parameters. For example, we want to compare $x_1(t;a,b)$ with $x_2(t;a,b,c)$. Obviously, the model with 3 parameters fit better, but it does that at the cost of adding one more parameter. Is the cost worth it ?

We have all the tools now to answer this question : Evaluate the $p-$values of the first model with $N-2$ degree of freedom and the second model with $N-3$ degree of freedom and compare these $p-$values now.

Let us come back to our preceding example and use now a quadratic model :

$$x_2(t) = a + bt + ct^2$$

where we get this time $a = 1.1$, $b = 1.1$ and $c = 0.04$. Graphically, it seems to be slightly nicer. Pushing the computations, we get $V = 112.4$. So the value is obviously reduced. Is it worth it ? Comparing that to a χ^2 with 98 degree of freedom, its $p-$value is



$$P(\geq V) = 0.16$$

Indeed, we have enhanced the goodness of fit.

Consider now a third model

$$x_3(t) = a + bt + ct^2 + dt^3$$

Doing all the optimization, we find that the value V doesn't change and stays at 112 and we decrease our $p-$values with this model to 0.14. So, there is no need to use 4 parameters and three is enough.

Now full disclosure : the data was generated by the formula $y = 1 + t + 0.05t^2 + \mathcal{N}(0, 1)$.

Did we however really had to go to all these computations ? For large degree of freedom N, V is a Normal random variable with mean N and standard deviation $\sqrt{2N}$. In the first case, with $N = 99$, this gives a standard deviation of 14, while the value we obtained for V was 123, so we were at $24/14\sigma = 1.7\sigma$. In the second case, $N = 98$ and $\sigma = 14$, $V = 112$ so we were precisely at 1σ.

7 Effect of noise amplitude.

Very often, the amplitude of the experimental noise σ is not known by the experimentalist. All the discussion we had before crucially depended on knowing this parameter, so what can be salvaged ?

Let us come back to our preceding section where we had supposed $\sigma = 1$. What if, given the same data, we had supposed $\sigma = a$? Obviously, the χ^2 realization V we had measured become $V' = V/a^2$. If $a \ll 1$, both models have to be rejected and if $a \gg 1$, both model become acceptable.

To fix the Idea, let us first suppose $a = 2$. For the first model with 99 degree of freedom, we would have $V_1' = 30.8$ and for the second one with 98 degree of freedom, $V_2' = 28$. For both these values, the $p-$value becomes practically 1 and they are as plausible.

On the other hand, if we had $a = 1/2$, both $p-$values become nearly zero and both model have to be rejected.

So what to do ? There is no way of getting around this problem, except estimating somehow the noise before fitting. We can for example make many measures at each value of t and use these spread to estimated the noise.