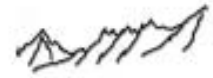
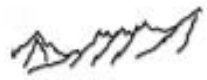


Hypothesis testing and how to deal with a fine structure

- Historical example : the tea taster
- Generalization and definitions
- Practical example : counting events
- The confidence levels
- Neyman-Pearson theorem
- Simple and composite hypotheses testing
- Goodness-of-fit testing
- example : Observation of a fine structure

The tea taster (1)

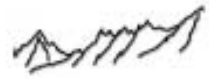


A person claims that he (or she) can recognize, simply tasting a cup of tea, if milk has been poured first or not. 8 cups are given to him (her) and he (she) is asked to designate the 4 cups where milk has been poured first. The person recognizes correctly 3 cups (out of 4). Yes or no, does he own the claimed talent ?

It's clearly a decision process, which should be built in order to give a false answer with a low probability.

- The candidate selects 4 cups out of 8. He has $\left(\begin{array}{c} 8 \\ 4 \end{array} \right) = 70$ ways to make his choice. Among these 70, only one corresponds to the 4 "good" cups.
- An ordinary (without any particular talent) man would have :
  - 1 chance out of 70 for giving the good choice
  - 16 chances out of 70 (4×4) to give 3 good cups.
  - 36 chances out of 70 (6×6) to give 2, etc.

The tea taster (2)

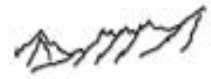


- One has to choose the hypothesis we want to test (H_0). One should be able to compute all probabilities under this hypothesis. There is no real choice here. H_0 = no special talent !
- An alternative hypothesis (H_1) is needed also. It's not necessary to be able to compute probabilities under that hypothesis (dissymmetry). Here, H_1 is simply the negation of H_0.
- All possible outcomes of the experiment (tea tasting) are separated into 2 categories, C and $\not{C}$: we put in C the observed result and all (more extreme) which suggest that H_1 may be true.

$$\mathcal{P}(C|H_1) > \mathcal{P}(C|H_0) = p_c = 17/70 = 0.24$$

- Finally, we apply a decision rule :
  - if p_c is judged too low, one rejects H_0
  - if p_c is sufficiently large, one cannot exclude H_0

Generalisation



Check and decide if ONE hypothesis better explains the data than another hypothesis (or any other hypothesis).
The two hypotheses are traditionally called :

$$\mathrm{H}_0 : \text{the null hypothesis, and}$$
$$\mathrm{H}_1 : \text{the alternative hypothesis.}$$

and we want to test H_0 against H_1 (dissymmetry).
As always, we know $\mathcal{P}(\text{data}|\mathrm{H}_0)$, not necessarily $\mathcal{P}(\text{data}|\mathrm{H}_1)$.

If W is the space of all possible data, we must find a critical region w inside W in which we reject H_0, thus including all possible data that suggest that H_0 may be wrong.
In practice, the determination of a multidimensional critical region may be difficult, so one often chooses a single test statistic $t(X)$ instead.

Errors of first and second kinds

Usually one adjusts the size of the critical region so as to obtain a desired level of significance α, defined as the probability of X falling in w when H_0 is true

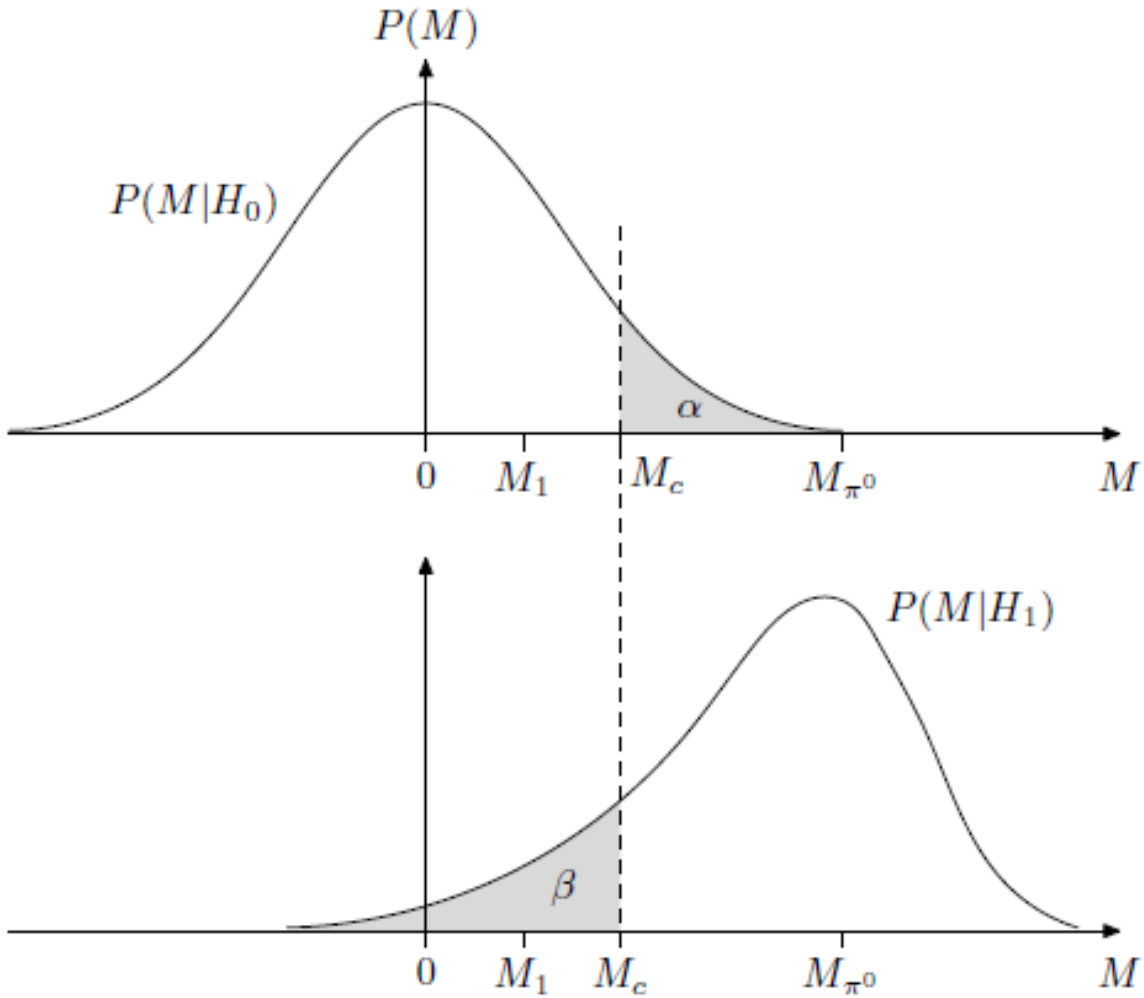
$$\mathcal{P}(X \in w | H_0) = \alpha$$

Note that β can be calculated only if $\mathcal{P}(\text{data}|H_1)$ is known.
The usefulness of a test depends on its ability to discriminate against the alternative hypothesis H_1. The measure of this usefulness is the power of the test, defined as the probability $1 - \beta$ of X falling into the critical region when H_1 is true.
$\mathcal{P}(X \in w | H_1) = 1 - \beta$ or equivalently :

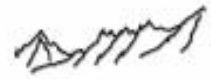
$$\mathcal{P}(X \in W - w | H_1) = \beta$$

	H_0 TRUE	H_1 TRUE
$X \notin w$ ACCEPT H_0	Acceptance good Prob $= 1 - \alpha$	Contamination Error of the second kind Prob $= \beta$
$X \in w$ REJECT H_0	Loss Error of the first kind Prob $= \alpha$	Rejection good Prob $= 1 - \beta$

Separation of two classes



Simple and Composite Hypotheses



When the hypotheses H_0 and H_1 are completely specified, i.e. with no free parameters, they are called simple hypotheses.

The theory of hypothesis testing for these simple hypotheses is well known and holds for any size of samples. It is the domain of applicability of the Neyman-Pearson theorem.

When a hypothesis contains one or more free parameters, it is a composite hypothesis, for which there is only an asymptotic theory. Unfortunately, most of the interesting problems involve composite hypotheses.

Example : signal search with background

We have a counting experiment, and we observe N events, while the expectations are in mean b background events and s signal events (if the expected signal is there). The possible observations (random variable n) follow a Poisson with parameters b or $s+b$ depending on the hypotheses.
The Poisson's law is :

$$\mathcal{L}(n|\lambda) = \lambda^n \exp(-\lambda)/n!$$

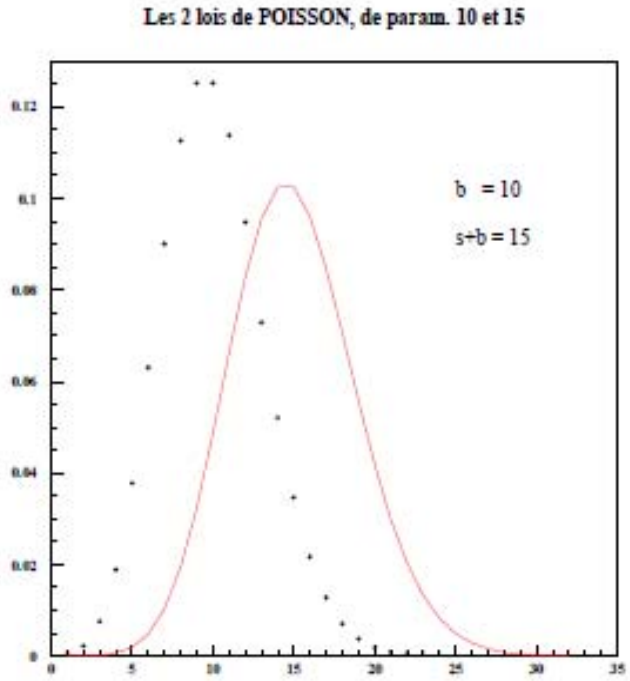
Thus :

$$\mathcal{L}(n|b) = b^n \exp(-b)/n!$$

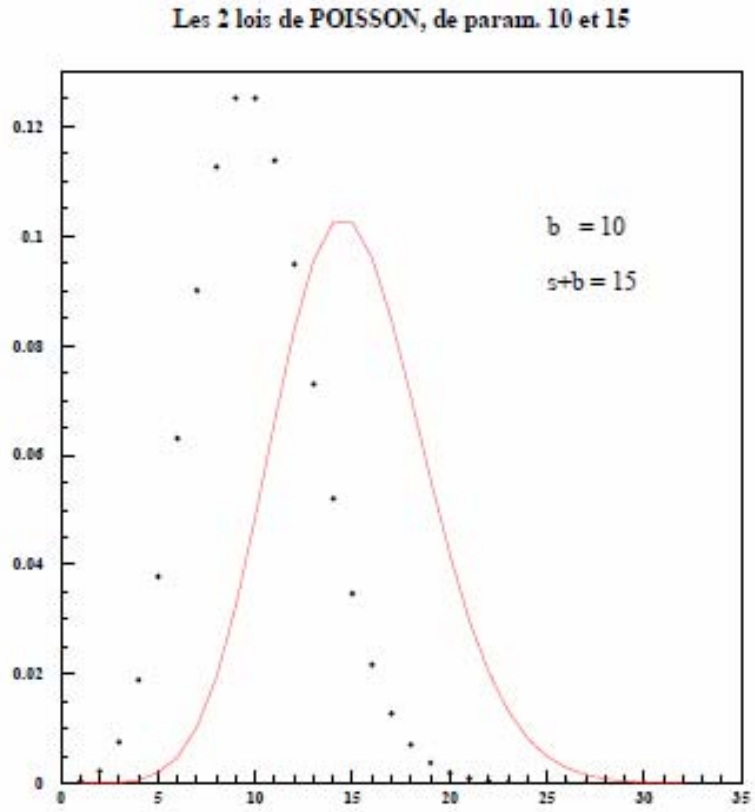
under the hypothesis "background only", and

$$\mathcal{L}(n|s+b) = (s+b)^n \exp(-(s+b))/n!$$

under the hypothesis "signal + background"



We observe $N = 16$



test $H_0 \equiv$ background

$p_c = \mathcal{P}(n \geq N|b) = 4.9\%$

Definition :

$p_c \equiv 1\text{-CL}_b$

Convention (physics) :

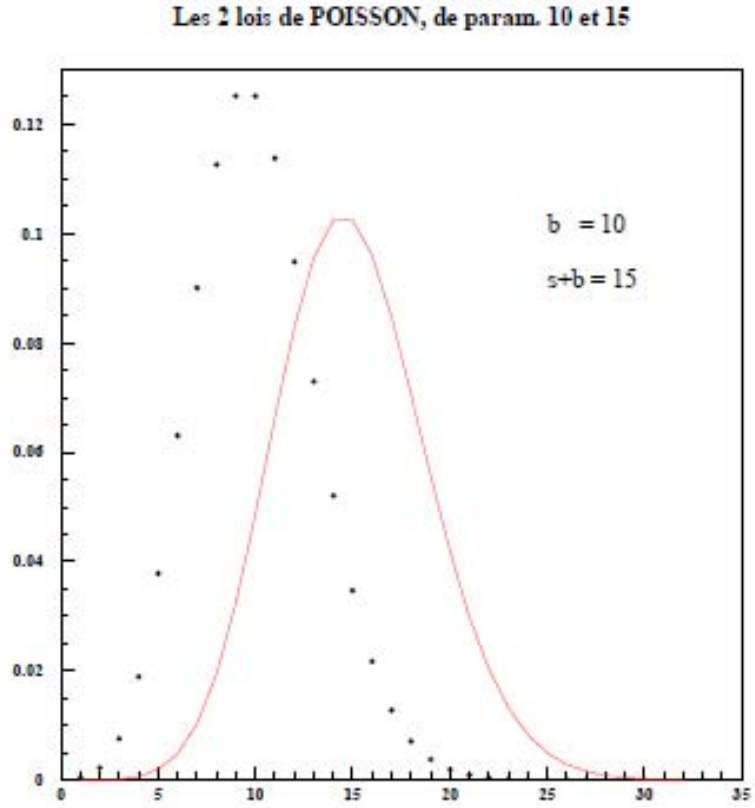
discovery $\equiv$ rejection H_0

if $p_c < 5.7 \times 10^{-7}$

($5\ \sigma$ discovery)

H_0 **cannot** be rejected here.

We observe $N = 16$



test $\mathrm{H}_0 \equiv s + b$

$p_c = \mathcal{P}(n \leq N | s + b) = 66.3\%$

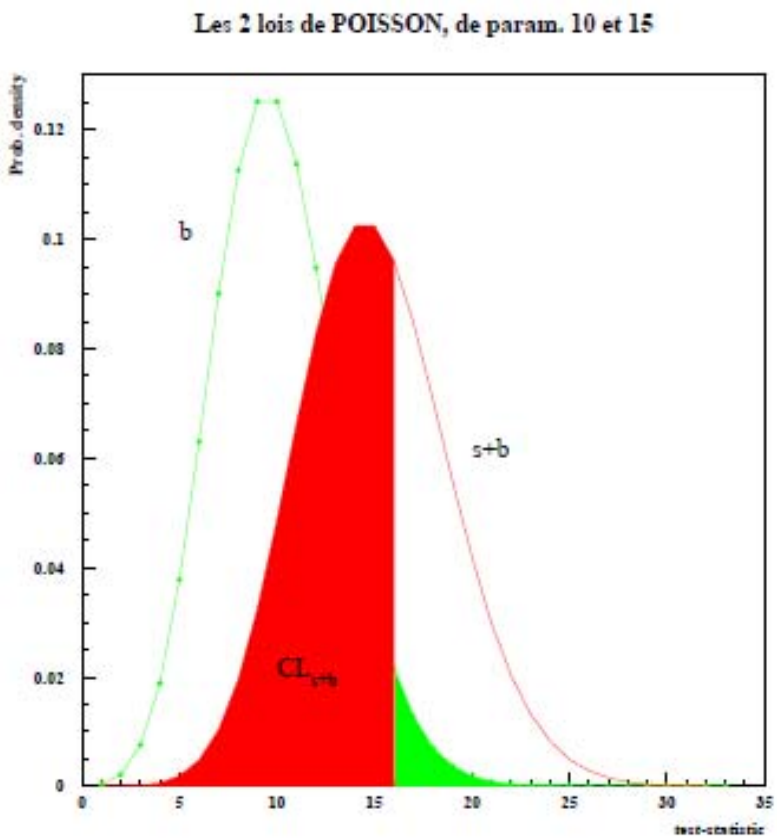
Definition :

$p_c \equiv \mathrm{CL}_{s+b}$

Convention (statistics) :
no signal $\equiv \mathrm{H}_0$ rejection
if $p_c < 5. \times 10^{-2}$
(95% confidence level limit)

H_0 **cannot** be rejected here.

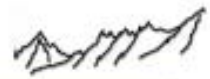
The confidence levels



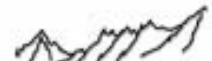
- 1-CL_b measures the non-compatibility with the "b" hypothesis.
- CL_{s+b} measures the non-compatibility with the "s+b" hypothesis
- The notion of CL_s is not standard statistics

1 - CL_b	# σ
0.1587	1 σ
0.0228	2 σ
0.00135	3 σ
$3.15\ 10^{-5}$	4 σ
$2.85\ 10^{-7}$	5 σ

(unilateral)

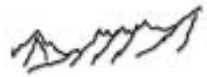


- A test thus consists in a definition of two hypotheses, a choice of a test-statistics (as discriminant as possible), and the fixing of a threshold α (falses rejections).
- A large window (from N = 8 to N = 25) of possible observations lead to a situation such that neither the b hypothesis, nor the $s + b$ hypothesis can be rejected ! (consequence of the dissymmetry) This is due to the fact that the standard deviation of the Poisson' law is $\sqrt{\lambda}$.
- Within a given test-statistic and identical analyses, an improvement can arise only from an increase in luminosity, ie b AND $s + b$.



- When the hypotheses are chosen, and α fixed, we have still to find the "best" test-statistic. This is based on the discrimination power : we have to choose the test-statistic which minimizes β or (equivalently) maximizes the power $1 - \beta$.
- Neyman-Pearson theorem In the case of simple hypothesis against simple hypothesis, with the same probability model (eg Poisson in both H, there is a test which is Uniformly (whatever the value of α) Most Powerful. This test-statistic is the ratio of the likelihood functions under the two hypotheses.

Likelihood ratio



- Take again our example : recall

$$\mathcal{L}(n|b) = b^n \exp(-b)/n!$$

$$\mathcal{L}(n|s+b) = (s+b)^n \exp(-(s+b))/n!$$

Thus :

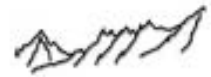
$$Q = \mathcal{L}(n|s+b)/\mathcal{L}(n|b) = (1+s/b)^n \times e^{-s}$$

$$-2\ln Q = 2s - 2n\ln(1+s/b)$$

Note that Q or $-2\ln Q$ have the same monotony behaviour than n. They have the same properties as statistics are concerned. n. is thus the optimal test-statistic in this case.

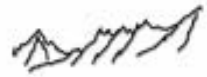
- Asymptotic property of the likelihood ratio : under some conditions concerning the regularity of the hypotheses pdfs, $-2\ln Q$ behaves asymptotically as a χ^2 with 1 degree of freedom under the numerator hypothesis.

A few remarks (physics)



- Up to now, we have treated b and s as exactly known. In fact, they are nuisance parameters, since they are estimated through our simulations.
- It is relatively easy (at least conceptually) to take into account the errors on b and s by convoluting these errors with the Poisson' laws.
- If the estimators of b and s are pretty precise, only a small degradation of the results will be observed.
- If the errors are rather big (and/or if some badly controlled systematics play a role) that's no more true ! It is very important to estimate b and s as precisely as possible.

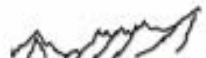
Towards the test used for Higgs search at LEP



- First step : a counting experiment (we just have seen that !)
- A simple counting is not optimal per se since all kept events have somehow the same weight, which means that the "last cut", the one which decides is an event is a "candidate" or not, has potentially an enormous impact on the final decision.

- Indeed, we have much more information about the events than just a "yes or no" counting. In most of the cases, we can use a bi-dim. info, eg
  - a reconstructed mass
  - a global variable summarizing the info for that event (ANN output or similar)

Towards the test used for Higgs search at LEP (2)



- We can thus estimate the background b_i and the expected signal s_i in every point of the plane, and then write again the likelihoods and their ratio :.

$$-2\ln Q = 2S_{tot} - 2\sum_{i=1}^{N} \ln(1 + s_i/b_i)$$

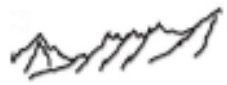
where N is the number of observed events.
Or we can bin the 2-dim plane :

$$-2\ln Q = 2S_{tot} - 2\sum_{i=1}^{N_{bins}} N_i \ln(1 + s_i/b_i)$$

where N_i is the number of observed events in the bin i.

- A precise estimation of b_i and s_i becomes mandatory.
- Each event comes thus with a weight $\ln(1 + s_i/b_i)$.

Likelihood Ratio Test



This is the extension of the Neyman-Pearson Test to composite hypotheses. Unfortunately, its properties are known only asymptotically.

Let the observations $\mathbf{X}$ have a distribution $f(\mathbf{X}|\theta)$, depending on parameters, $\theta = (\theta_1, \theta_2, \ldots)$. Then the likelihood function is

$$L(\mathbf{X}|\theta) = \prod_{i=1}^{N} f(X_i|\theta)\,.$$

In general, let the total θ-space be denoted θ, and let ν be some subspace of θ, then any test of parametric hypotheses (of the same family) can be stated as

$$H_0 : \theta \in \nu$$

$$H_1 : \theta \in \theta - \nu$$

Likelihood Ratio Test (cont.)

We can then define the maximum likelihood ratio, a test statistic for H_0:

$$\lambda = \frac{\max\limits_{\theta \in \nu} L(\mathbf{X}|\theta)}{\max\limits_{\theta \in \theta} L(\mathbf{X}|\theta)}.$$

If H_0 and H_1 were simple hypotheses, λ would reduce to the Neyman-Pearson test statistic, giving the UMP test. For composite hypotheses, we can say only that λ is always a function of the sufficient statistic for the problem, and produces workable tests with good properties, at least for large sets of observations.

Likelihood Ratio Test (cont.)

The importance of the maximum likelihood ratio comes from the fact that asymptotically:

if H_0 imposes r constraints on the $s + r$ parameters in H_0 and H_1, then

$$-2 \ln \lambda \text{ is distributed as } \chi^2(r) \text{ under } H_0$$

This means we can read off the confidence level α from a table of χ^2.

However, the bad news is that this is only true asymptotically,
and there is no good way to know how good the approximation is
except to do a Monte Carlo calculation.

Likelihood Ratio Test - Example

Problem: Find the ratio X of two complex decay amplitudes:

$$X = \frac{A(\text{reaction 1})}{A(\text{reaction 2})}.$$

In the general case, X may be any complex number, but there exist three different theories which predict the following for X:

- A: If Theory A is valid, $X = 0$.
- B: If Theory B is valid, X is real and $Im(X) = 0$.
- C: If Theory C is valid, X is purely imaginary and non-zero.

We decide that the value of X is interesting only in so far as it could distinguish between the hypotheses A, B, C or the general case.
Therefore, we are doing hypothesis testing, not parameter estimation.

Hypothesis A is simple,
Hypothesis B is composite, including hypothesis A as a special case.
Hypothesis C is also composite, and separate from A and B.
The alternative to all these is that Re (X) and Im (X) are both non-zero.

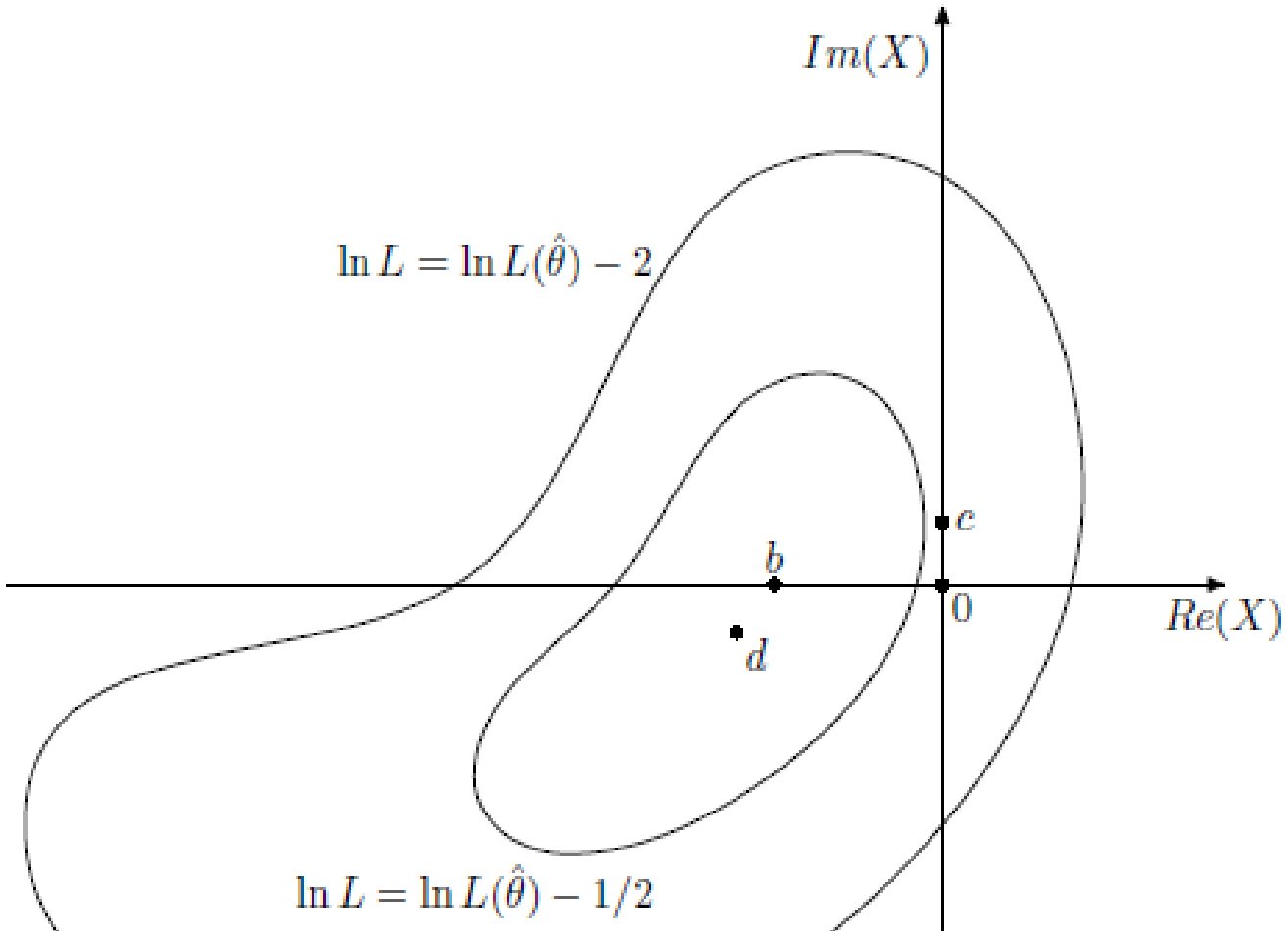
Likelihood Ratio Test - Example

The contours of the log-likelihood function $\ln L(X)$ near its maximum.

$X = d$ is the point where $\ln L$ is maximal.

$X = b$ is the maximum of $\ln L$ when $Im(X) = 0$.

$X = c$ is the maximum of $\ln L$ when $Re(X) = 0$.



Likelihood Ratio Test - Example

The maximum likelihood ratio for hypothesis A versus the general case is

$$\lambda_a = \frac{L(0)}{L(d)}\,.$$

If hypothesis A is true, $-2\ln\lambda_a$
is distributed asymptotically as a $\chi^2(2)$, and this give the usual test for Theory A.
To test Theory B, the m.l. ratio for hypothesis B versus the general case is

$$\lambda_b = \frac{L(b)}{L(d)}\,.$$

If B is true, $-2\ln\lambda_b$ is distributed asymptotically as a $\chi^2(1)$. Finally, Theory C can be tested in the same way, using $L(c)$ in place of $L(b)$.

Goodness-of-Fit Testing (GOF)

As in hypothesis testing, we are again concerned with the test of a null hypothesis H_0 with a test statistic T, in a critical region w_α, at a significance level α.

Unlike the previous situations, however, the alternative hypothesis, H_1 is now the set of all possible alternatives to H_0.
Thus H_1 cannot be formulated, the risk of second kind, β, is unknown, and the power of the test is undefined.

Since it is in general impossible to know whether one test is more powerful than another, the theoretical basis for goodness-of-fit (GOF) testing is much less satisfactory than the basis for classical hypothesis testing.

Nevertheless, GOF testing is quantitatively the most successful area of statistics. In particular, Pearson's venerable Chi-square test is the most heavily used method in all of statistics.

GOF Testing: From the test statistic to the P-value.

Goodness-of-fit tests compare the experimental data with their p.d.f. under the null hypothesis H_0, leading to the statement:

> *If H_0 were true and the experiment were repeated many times, one would obtain data as far away (or further) from H_0 as the observed data with probability P.*

The quantity P is then called the P-value of the test for this data set and hypothesis. A small value of P is taken as evidence against H_0, which the physicist calls a bad fit.

From the test statistic to the P-value.

It is clear from the above that in order to construct a GOF test we need:

1. A test statistic, that is a function of the data and of H_0, which is a measure of the "distance" between the data and the hypothesis, and
2. A way to calculate the probability of exceeding the observed value of the test statistic for H_0. That is, a function to map the value of the test statistic into a P-value.

If the data X are discrete and our test statistic is $t = t(X)$ which takes on the value $t_0 = t(X_0)$ for the data X_0, the P-value would be given by:

$$P_X = \sum_{X:t \geq t_0} P(X|H_0)\,,$$

where the sum is taken over all values of X for which $t(X) \geq t_0$.

Example: Test of Poisson counting rate

Example of discrete counting data:
We have recorded 12 counts in one hour, and we wish to know if this is compatible with the theory which predicts $\mu = 17.3$ counts per hour.

The obvious test statistic is the absolute difference $|N - \mu|$, and assuming that the probability of n decays is given by the Poisson distribution, we can calculate the P-value by taking the sum in the previous slide.

$$P_{12} = \sum_{n:|n-\mu|\geq 5.3} \frac{e^{-\mu}\mu^n}{n!} = \sum_{n=1}^{12} \frac{e^{-17.3}17.3^n}{n!} + \sum_{n=23}^{\infty} \frac{e^{-17.3}17.3^n}{n!}$$

Evaluating the above P-value, we get $P_{12} = 0.229$.

The interpretation is that the observation is not significantly different from the expected value, since one should observe a number of counts at least as far from the expected value about 23% of the time.

Distribution-free Tests

When the data are continuous, the sum becomes an integral:

$$P_X = \int_{X:t>t_0} P(X|H_0)\,, \tag{1}$$

and this now becomes so complicated to compute that one tries to avoid using this form. Instead, one looks for a test statistic such that the distribution of t is known independently of H_0.

Such a test is called a distribution-free test. We consider only distribution-free tests, such that the P-value does not depend on the details of the hypothesis H_0, but only on the value of t, and possibly one or two integers such as the number of events, the number of bins in a histogram, or the number of constraints in a fit.

Then the mapping from t to P-value can be calculated once for all and published in tables, of which the well-known χ^2 tables are an example.

Pearson's Chi-square Test

The obvious way to measure the distance between the data and the hypothesis H_0 is to

1. Determine the expectation of the data under the hypothesis H_0 .
2. Find the metric in the space of the data to measure the distance of the data from its expectation under H_0 .

When the data consists of measurements $\mathbf{Y} = Y_1, Y_2, \ldots, Y_k$ of quantities which, under H_0 are equal to $\mathbf{f} = f_1, f_2, \ldots, f_k$ with covariance matrix $\underset{\sim}{V}$, the distance between the data and H_0 is clearly:

$$T = (\mathbf{Y} - \mathbf{f})^T \, \underset{\sim}{V}^{-1} \, (\mathbf{Y} - \mathbf{f})$$

This is just the Pearson test statistic usually called chi-square, because it is distributed as $\chi^2(k)$ under H_0 if the measurements $\mathbf{Y}$ are Gaussian-distributed. That means the P-value may be found from a table of $\chi^2(k)$, or by calling PROB(T,k).

Pearson's Chi-square test for histograms

Karl Pearson made use of the asymptotic Normality of a multinomial p.d.f. in order to find the (asymptotic) distribution of:

$$T = (\mathbf{n} - N\mathbf{p})^T \, \underset{\sim}{V}^{-1} \, (\mathbf{n} - N\mathbf{p})$$

where $\underset{\sim}{V}$ is the covariance matrix of the observations (bin contents) $\mathbf{n}$ and N is the total number of events in the histogram.

In the usual case where the bins are independent, we have

$$T = \frac{1}{N} \sum_{i=1}^{k} \frac{(n_i - N p_i)^2}{p_i} = \frac{1}{N} \sum_{i=1}^{k} \frac{n_i^2}{p_i} - N \, .$$

This is the usual χ^2 goodness-of-fit test for histograms. The distribution of T is generally accepted as close enough to $\chi^2(k-1)$ when all the expected numbers of events per bin (Np_i) are greater than 5. Cochran relaxes this restriction, claiming the approximation to be good if not more than 20% of the bins have expectations between 1 and 5.

Tests on Unbinned Data

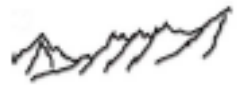
By combining events into histogram bins (called data classes in the statistical literature), some information is lost: the position of each event inside the bin. The loss of information may be negligible if the bin width is small compared with the experimental resolution, but in general one must expect tests on binned data to be inferior to tests on individual events.

Unfortunately, the requirement of distribution-free tests restricts the choice of tests for unbinned data, and we will consider only those based on the order statistics (or empirical distribution function). Moreover, this class is limited to data depending on only one random variable, and to hypotheses H_0 which do not depend on parameters θ to be estimated from the data.

When data are combined into histograms, more tests are available, but they may be valid only for a large number of events per bin.

Such considerations seriously limit the use of goodness-of-fit tests in many dimensions.

Order statistics

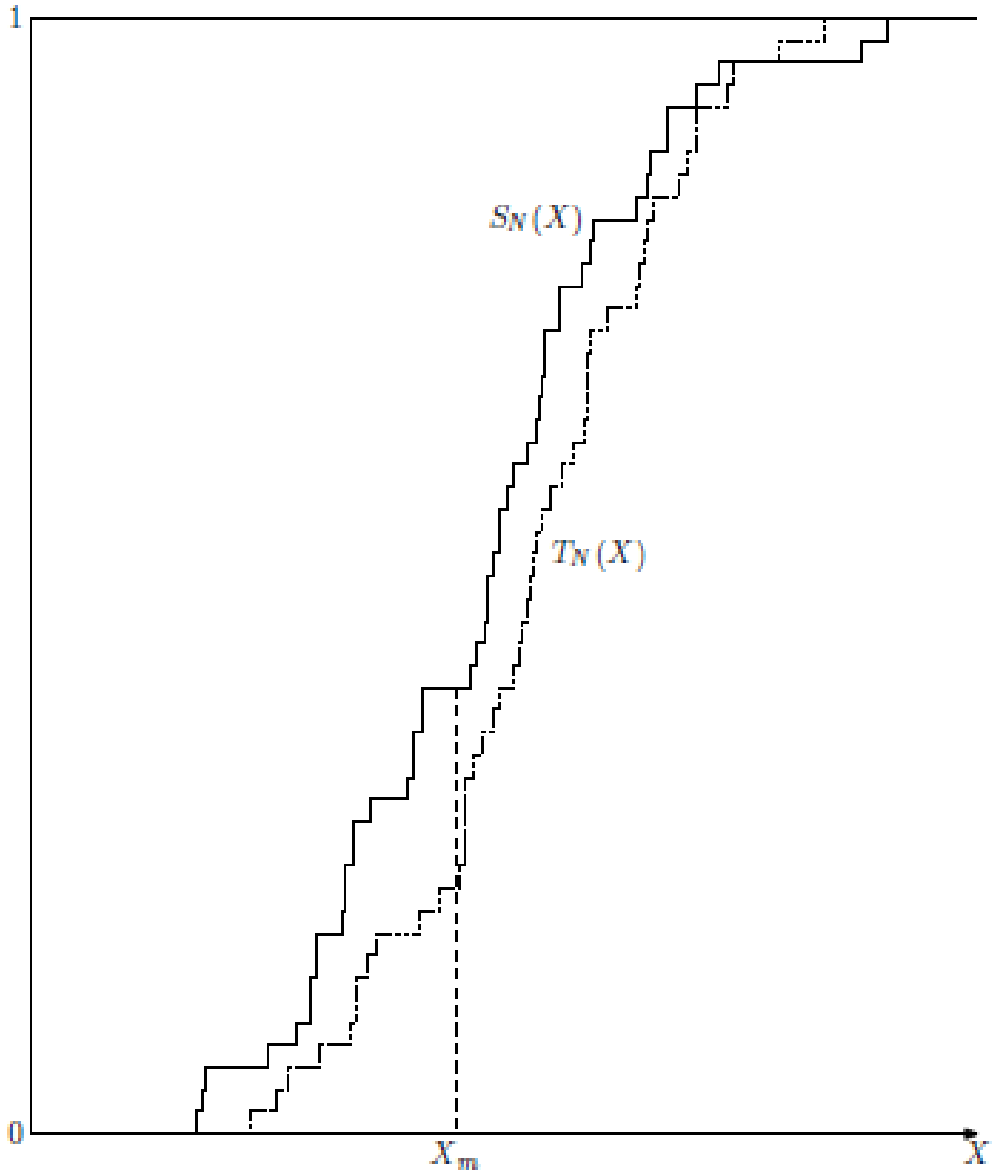


Given N independent observations $X_1, \ldots, X_N$ of the random variable X, let us reorder the observations in ascending order, so that $X_{(i)} \leq X_{(2)} \leq, \ldots, \leq X_{(N)}$ (this is always permissible since the observations are independent). The ordered observations $X_{(i)}$ are called the order statistics. Their cumulative distribution is called the empirical distribution function or EDF.

$$S_N(X) = \begin{cases} 0 & X < X_{(1)} \\ i/n \quad \text{for} & X_{(i)} \leq X < X_{(i+1)}, \quad i = 1, \ldots, \quad N-1. \\ 1 & X_{(N)} \leq X \end{cases}$$

Note that $S_N(X)$ always increases in steps of equal height, N^{-1}.

Order statistics



Example of two cumulative distributions, $S_N(X)$ and $T_N(X)$

For these two data sets, the maximum distance $S_N - T_N$ occurs at $X = X_m$.

We shall consider different norms on the difference $S_N(X) - F(X)$ as test statistics.

Smirnov - Cramér - von Mises test

Consider the statistic

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - F(X)]^2 f(X) dX \,,$$

where $f(X)$ is the p.d.f. corresponding to the hypothesis H_0, $F(X)$ is the cumulative distribution, and $S_N(X)$ is defined as above, which gives

$$\begin{aligned} W^2 &= \int_{-\infty}^{X_1} F^2(X) dF(X) + \sum_{i=1}^{N-1} \int_{X_i}^{X_{i+1}} \left[\frac{i}{N} - F(X)\right]^2 dF(X) \\ &\quad + \int_{X_N}^{\infty} [1 - F(X)]^2 dF(X) \\ &= \frac{1}{N}\left\{\frac{1}{12N} + \sum_{i=1}^{N}\left[F(X_i) - \frac{2i-1}{2N}\right]^2\right\} , \end{aligned}$$

using the properties $F(-\infty) \equiv 0$, $F(+\infty) \equiv 1$.

Smirnov - Cramér - von Mises test

The Smirnov-Cramér-von Mises test statistic W^2 has mean and variance

$$E(W^2) = \frac{1}{N}\int_0^1 F(1-F)dF = \frac{1}{6N}$$
$$V(W^2) = E(W^4) - [E(W^2)]^2 = \frac{4N-3}{180N^3}.$$

Smirnov has calculated the critical values of NW^2

Test size α	Critical value of NW^2
0.10	0.347
0.05	0.461
0.01	0.743
0.001	1.168

It has been shown that, to the accuracy of this table, the asymptotic limit is reached when $N \geq 3$.

Smirnov - Cramér - von Mises test

When H_0 is composite, W^2 is not in general distribution-free. When X is many-dimensional, the test also fails, unless the components are independent. However, one can form a test to compare two distributions, $F(X)$ and $G(X)$. Let the number of observations be N and M, respectively, and let the hypothesis be H_0: $F(X) = G(X)$. Then the test statistic is

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - S_M(X)]^2 d\left[\frac{NF(X) + MG(X)}{N + M}\right].$$

Then the quantity

$$\frac{MN}{M + N} W^2$$

has the critical values shown in the table above.

Kolmogorov test

The test statistic is now the maximum deviation of the observed distribution $S_N(X)$ from the distribution $F(X)$ expected under H_0. This is defined either as

$$D_N = \max \; |S_N(X) - F(X)| \qquad \text{for all } X$$

or as

$$D_N^{\pm} = \max \; \{\pm[S_N(X) - F(X)]\} \quad \text{for all } X\,,$$

when one is considering only one-sided tests. It can be shown that the limiting distribution of $\sqrt{N}D_N$ is

$$\lim_{N\to\infty} P(\sqrt{N}D_N > z) = 2\sum_{r=1}^{\infty}(-1)^{r-1}\exp(-2r^2z^2)$$

and that of $\sqrt{N}D_N^{\pm}$ is

$$\lim_{N\to\infty} P(\sqrt{N}D_N^{\pm} > z) = \exp(-2z^2)\,.$$

Kolmogorov Test

Alternatively, the probability statement above can be restated as

$$\lim_{N\to\infty} P[2N(D_N^{\pm})^2 \leq 2z] = 1 - e^{-2z^2}.$$

Thus $4N(D_N^{\pm})^2$ have a $\chi^2(2)$ distribution.
The limiting distributions are considered valid for $N \approx 80$.

We give some critical values of $\sqrt{N}D_N$.

Test size α	Critical value of $\sqrt{N}D_N$
0.01	1.63
0.05	1.36
0.10	1.22
0.20	1.07

Two-Sample Kolmogorov Test

The equivalent statistic for comparing two distributions $S_N(X)$ and $S_M(X)$ is

$$D_{MN} = \max |S_N(X) - S_M(X)| \qquad \text{for all } X$$

or, for one-sided tests

$$D^{\pm}_{MN} = \max \{\pm[S_N(X) - S_M(X)]\} \qquad \text{for all } X\,.$$

Then $\sqrt{MN/(M+N)}D_{MN}$ has the limiting distribution of $\sqrt{N}D_N$ and $\sqrt{MN/(M+N)}D^{\pm}_{MN}$ have the limiting distribution of $\sqrt{N}D^{\pm}_N$.

Kolmogorov Test

Finally, one may invert the probability statement about D_N to set up a confidence belt for $F(X)$. The statement

$$P\{D_N = \max|S_N(X) - F(X)| > d_\alpha\} = \alpha$$

defines d_α as the α-point of D_N. If follows that

$$P\{S_N(X) - d_\alpha \leq F(X) \leq S_N(X) + d_\alpha\} = 1 - \alpha\,.$$

Therefore, setting up a belt $\pm d_\alpha$ about $(S_N(X)$, the probability that $F(X)$ is entirely within the belt is $1 - \alpha$ (similarly d_α can be used to set up one-sided bounds). One can thus compute the number of observations necessary to obtain $F(X)$ to any accuracy. Suppose for example that one wants $F(X)$ to precision 0.05 with probability 0.99, then one needs $N = (1.628/0.05)^2 \sim 1000$ observations.

The likelihood function is not a good test statistic

Unfortunately, the value of the likelihood does not make a good GOF test statistic. This can be seen in different ways, but the first clue should come when we judge whether the likelihood is a measure of the "distance" between the data and the hypothesis.

At first glance, we might expect it to be a good measure, since we know the maximum of the likelihood gives the best fit to the data.

But in m.l. estimation, we are using the likelihood for fixed data as a function of the parameters in the hypothesis, whereas in GOF testing we use the likelihood for a fixed hypothesis as a function of the data, which is very different.

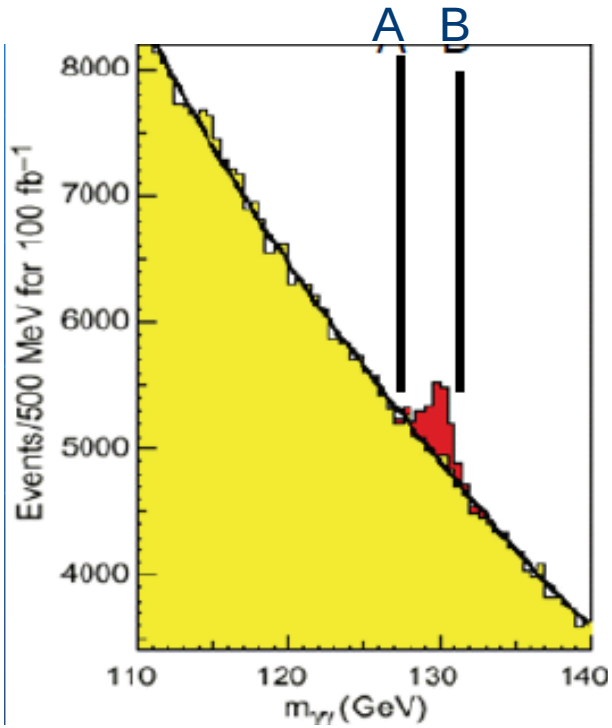
Observation of a fine structure

The situation often arises that a new phenomenon manifests itself as a relatively narrow signal superimposed on a smooth background.

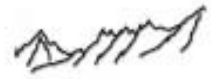
- The first question is then: do the observations suggest fine structure in the region AB ?
- If yes, the next problem will be to estimate parameters (size and position) of the signal.

H_0 = no structure, background only
$H_1 = s + b$ if we can estimate s,
otherwise, not H_0 (GOF test)

In both cases, we have to estimate b



Observation of a fine structure (2)



Let us describe the background by a function $b(X, \theta)$ of the observations X and unknown parameters θ. The estimates $\hat{\theta}$ and their covariance matrix V can be obtained by the methods of point estimation, excluding the region AB, to give

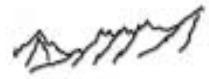
$$\hat{b}_{AB} = \int_A^B b(X, \hat{\theta}) \mathrm{d}X$$

and since $\hat{b}_{AB}$ is a function of $\hat{\theta}$, its variance is obtained by the usual methods of change of variable.
Let the number of observations in AB be n_{AB}. The natural test statistic for determining whether n_{AB} is significantly different from $\hat{b}_{AB}$ is

$$T = (n_{AB} - \hat{b}_{AB})^2 / V(n_{AB} - \hat{b}_{AB})$$

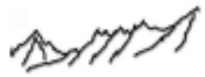
Under the H_0 hypothesis, $E(n_{AB}) = V(n_{AB}) = b_{AB}$, estimated by $\hat{b}_{AB}$. Thus, $V(n_{AB} - \hat{b}_{AB}) \approx \hat{b}_{AB} + \hat{\sigma}^2_{AB}$ since the covariance term is null (we excluded AB when estimating θ), and $T \approx (n_{AB} - \hat{b}_{AB})^2 / (\hat{b}_{AB} + \hat{\sigma}^2_{AB})$, which behaves asymptotically as a χ^2 with 1 dof under H_0. One often expresses T in terms of *standard deviations* $d = \sqrt{T}$.

Observation of a fine structure (3)



Until now, we have implicitly assumed the region AB to be *selected independently of the observations*. However, if the choice of region AB is based on the data, all the computation is no longer appropriate, since we did not account for the probability of the occurence of a signal in any arbitrary place of the full region. To illustrate this, consider signals which are only one bin wide. Let p be the probability of exceeding d standard deviations in a given bin. When no bin is specified in advance, the probability of exceeding d standard deviations in at least one bin out of k bins is obviously $q = 1 - (1-p)^k \approx kp$. For instance, in a histogram of 40 bins, a 3 standard deviation effect in a given bin has the same significance as a 4 standard deviation effect in any one (unspecified) bin.

The statistical significance is different whether the "signal" is expected at a given place or not.



Suppose now that we have rejected the H_0 hypothesis (no structure) and accepted the idea that a signal is there. One can still estimate the size of the signal by $s = n_{AB} - \hat{b}_{AB}$, but the variance of the estimate is no longer the same, since we do not test the same hypothesis; our test is now : true physical effect of size s and background b_{AB} in region AB. The variance of s is thus

$$V(n_{AB} - \hat{b}_{AB}) \approx n_{AB} + \hat{\sigma}^2_{AB}$$

Signal significance

Which is the significance of an observation $x = 178$ events in a region "signal-like", when the expected background is $b = 100$ with a 10% error.

$s/b^{0.5}$	7.8 σ
$s/(b+\sigma_b)^{0.5}$	7.4 σ
Variance H_0	5.5 σ
Variance H_1	4.7 σ
TDR ATLAS	5.5 σ
Cousins	5.0 σ
Profile Like.	5.0 σ

(too) Many formulae !

ATLAS and CMS should negotiate and use the same method (CMS uses « Cousins ») otherwise ATLAS would need less luminosity than CMS to claim a discovery !

CONCLUSION

Should physicists be Do-it-yourself Statisticians?

Physicists should not be inventing new statistical methods.

However, if you can't find what you want, do the following:

1. Follow the procedures and rules of statistics (example: don't integrate under likelihood functions.)
2. Verify the properties of your method (like coverage).
3. When you have a good idea of what you want to do, search the statistics literature or ask the advice of a professional. If the method is a good one, you will probably find that it already exists.