(i) to compare the distributions with expectations from theories, and (ii) to extract from them physical quantities of interest like masses, widths, couplings, spins and so on. We call **fit** the method to do both these important things.

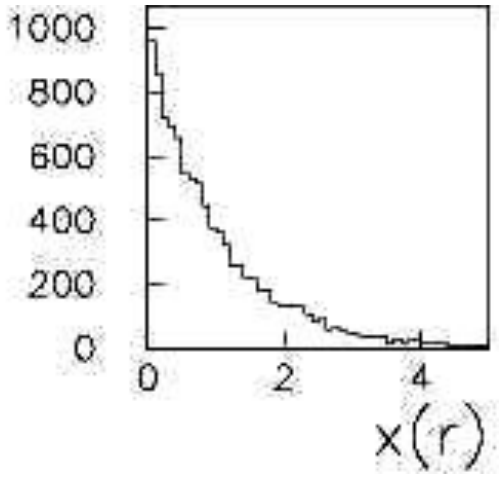
Analysis of event distributions: the fit

(1) First of all we have to define the hypothesis. It can be the theoretical function $y(x/\underline{\theta})$, x being the variable or the set of variables, and $\underline{\theta}$ a set of K **parameters** . K could be even 0, in this case the theory makes an "absolute prediction" and there is no need to adjust parameters to compare it to theory.

(2) Then we have to define a **test statistics** t, that is a variable depending on the data that, if the hypothesis is correct, has a known distribution function (in the following we use **pdf** to indicate probability distribution functions). The meaning of this pdf is the following: if we repeat the experiment many times and if every time we evaluate t, if the hypothesis is correct the histogram of the sample statistics will follow the pdf within the statistical errors of the sample.

(3) Finally we do the experiment. In case the theory depends on few parameters, we adjust the parameters in such a way to get the best possible agreement between data and theory. From this we obtain the **estimates** of the parameters with their uncertainties. We evaluate then the actual value of t, let's call it t^* from the data after parameter adjustment, and see if in the t pdf this value corresponds to a region of high or low probability. In case it is in a region of high probability, it's likely that the theory is correct, so that we conclude that the experiment **corroborates** the theory. In case it corresponds to a region of low probability it's unlikely that the theory is correct, so that we say that the experiment **falsifies** the theory, or, in other words, that we have not found any parameter region that allows an acceptable agreement.

Choice of test statistics: binned data



Histogram: $\sum_{i=1}^{M} n_i = N$

Theory: $y=y(x/\underline{\theta})$ θ_i, i=1….K

Prediction of the theory in bin i:

1) Value of the function at the center $\overline{x}_i$ of the bin multiplied by the bin width δx (note: [y]=[dN/dx]) $y_i = y(\overline{x}_i/\underline{\theta})\delta x$

2) or more exactly integrating y over the bin i $y_i = \int_{\overline{x}_i-\delta x/2}^{\overline{x}_i+\delta x/2} y(x/\underline{\theta})dx$

The predicted total number of events is: $\sum_{i=1}^{M} y_i = N_0$

The two definitions are equivalent in the limit of small bin size wrt to the typical scale of variations in the distribution

Choice of test statistics: binned data

Which statistics for the n_i data in the histogram?
two possibilities:

- We repeat the experiment holding the total number of events N fixed. In this case n_i has a multinomial distribution. The joint distribution of the n_i, with i=1,...,M is

$$p(n_1, ..n_M) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!}$$

where p_i is the probability associated to the bin i. Notice that the joint distribution cannot be factorized in a product of single bin probability distributions, since the fixed value of events N determines a correlation between the bin contents.

$$\begin{aligned} E[n_i] &= Np_i \\ Var[n_i] &= Np_i(1-p_i) \\ cov[n_i, n_j] &= -Np_ip_j \end{aligned}$$

Correlation negligible for events distributed over a large number of bins

Choice of test statistics: binned data

Which statistics for the n_i data in the histogram?
two possibilities:

- We repeat the experiment holding fixed the integrated luminosity or the observation time of the experiment. In this case N is not fixed and fluctuates in general between an experiment and another. The n_i are independent and have poissonian distributions:

$$p(n_1, ..n_M) = \prod_{i=1}^{M} \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!}$$

where λ_i is the expected counting in each bin.

$$\begin{aligned} E[n_i] &= \lambda_i \\ Var[n_i] &= \lambda_i \\ cov[n_i, n_j] &= 0 \end{aligned}$$

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

Definition of the test statistics t :

Neiman χ^2

$$\chi^2_N = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{n_i}$$

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

Definition of the test statistics t :

Neiman χ^2

$$\chi^2_N = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{n_i}$$

Pearson χ^2

$$\chi^2_P = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{y_i}$$

Gaussian pdf and the Central Limit Theorem

The Gaussian pdf is so useful because almost any random variable that is a sum of a large number of small contributions follows it. This follows from the Central Limit Theorem:

For n independent r.v.s x_i with finite variances σ_i^2, otherwise arbitrary pdfs, consider the sum

$$y = \sum_{i=1}^{n} x_i$$

In the limit $n \to \infty$, y is a Gaussian r.v. with

$$E[y] = \sum_{i=1}^{n} \mu_i \qquad V[y] = \sum_{i=1}^{n} \sigma_i^2$$

Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian r.v.s.

Central Limit Theorem (2)

The CLT can be proved using characteristic functions (Fourier transforms), see, e.g., SDA Chapter 10.

For finite n, the theorem is approximately valid to the extent that the fluctuation of the sum is not dominated by one (or few) terms.

Beware of measurement errors with non-Gaussian tails.

Good example: velocity component v_x of air molecules.

OK example: total deflection due to multiple Coulomb scattering. (Rare large angle deflections give non-Gaussian tail.)

Bad example: energy loss of charged particle traversing thin gas layer. (Rare collisions make up large fraction of energy loss, cf. Landau pdf.)

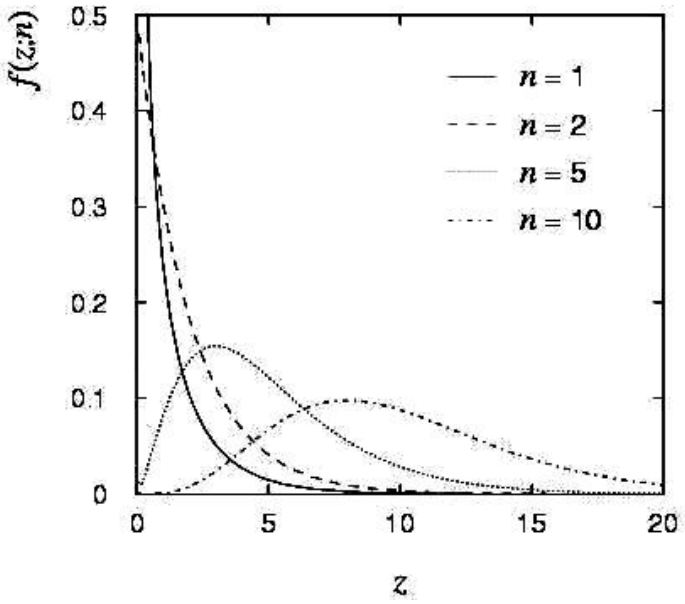
Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

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

$n = 1, 2, \ldots =$ number of 'degrees of freedom' (dof)

$$E[z] = n\,, \qquad V[z] = 2n\,.$$



For independent Gaussian x_i, $i = 1, \ldots, n$, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$ follows χ^2 pdf with n dof.

Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

Definition of the test statistics t :

Pearson χ^2

$$\chi^2_P = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{y_i}$$

In case of n_i being poissonian variables in the gaussian limit, the Pearson χ^2 is a statistics following a χ^2 distribution with a number of degrees of freedom equal to $M - K$. Infact we know that a χ^2 variable is the sum of the squares of standard gaussian variables, so that if $y_i=E[n_i]$ holds, this is the case for χ^2_P. However we know that the gaussian limit is reached for n_i at least above $10 \div 20$ counts. If we have histograms with few counts, and we are far from the gaussian limit, the pdf of χ^2_P is not exactly a χ^2 so that care is needed in the result interpretation.

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

Definition of the test statistics t :

Neiman χ^2

$$\chi^2_N = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{n_i}$$

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

Definition of the test statistics t :

Neiman χ^2

$$\chi^2_N = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{n_i}$$

The Neyman χ^2 is less well defined. In fact a χ^2 variable requires the gaussian σ in each denominator. By putting n_i we make an approximation[18]. However in case of large values of n_i to a good approximation the Neyman χ^2 has also a χ^2 distribution. A specific problem of the Neyman χ^2 is present when $n_i = 0$. But again, for low statistics histogram a different approach should be considered.

[18]The Neyman χ^2 was widely used in the past, since it makes simpler the calculation, the parameters being only in the numerator of the formula. With the present computing facilities there are no strong motivations to use it.

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

More general test statistics t : Likelihood method

N fixed (multinomial case) $(y_i = N_0 p_i)$
(negligible bin correlation assumed)

$$L_m(\underline{n}/\underline{y}) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!} = N! \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i! N_0^{n_i}}$$

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

More general test statistics t : Likelihood method

N not fixed (Poisson case) $y_i = \lambda_i$

$$L_p(\underline{n}/\underline{y}) = \prod_{i=1}^{M} \frac{e^{-y_i} y_i^{n_i}}{n_i!}$$

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

More general test statistics t : Likelihood method

N not fixed (poisson case) $y_i = \lambda_i$

$$L_p(\underline{n}/\underline{y}) = \prod_{i=1}^{M} \frac{e^{-y_i} y_i^{n_i}}{n_i!}$$

$$L_m(\underline{n}/\underline{y}) = N! \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i! N_0^{n_i}} = \frac{N!}{N_0^N} \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i!}$$

$$L_p(\underline{n}/\underline{y}) = e^{-N_0} \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i!} = \frac{e^{-N_0} N_0^N}{N!} L_m(\underline{n}/\underline{y})$$

L_p is essentially L_m multiplied by the poissonian fluctuation of N with mean N_0

Choice of test statistics: binned data

Fit: we impose the condition $y_i = E[n_i]$

More general test statistics t : Likelihood method

Which test statistics for the Likelihood function?

The pdf of a likelihood function in general depends on the specific problem, and can be evaluated by means of a MonteCarlo simulation of the situation we are considering (TOY MC), i.e. simulations done for different values of the parameters θ_i

Choice of test statistics: binned data

WILKS THEOREM

expectation values $\nu_i = E[n_i]$ of the contents of each bin

$$\chi^2_\lambda = -2 \ln \frac{L(\underline{n}/\underline{y})}{L(\underline{n}/\underline{\nu})}$$

has a χ^2 pdf with $M - K$ degrees of freedom in the asymptotic limit

(ν_i gaussians)

$\Rightarrow$ We can use Likelihood ratios as test statistics with known pdf, more general than Pearson $\chi 2$, it holds in asymp. limit but whatever is the stat. model.

Connection with the Neyman-Pearson Lemma

$$P(type - I errors) = 1 - \epsilon = \alpha$$

$$P(type - II errors) = \frac{1}{R} = \beta$$

Given the two hypotheses H_s and H_b and given a set of K discriminating variables x_1, x_2,...x_K, we can define the two "likelihoods"

$$L(x_1, ..., x_K/H_s) = P(x_1, ...x_K/H_s) \tag{66}$$

$$L(x_1, ..., x_K/H_b) = P(x_1, ...x_K/H_b) \tag{67}$$

equal to the probabilities to have a given set of values x_i given the two hypotheses, and the **likelihood ratio** defined as

$$\lambda(x_1, ...x_K) = \frac{L(x_1, ..., x_K/H_s)}{L(x_1, ..., x_K/H_b)} \tag{68}$$

Neyman-Pearson Lemma:
For fixed α value, a selection based on the discriminant variable λ has the lowest β value.

=> The "likelihood ratio" is the most powerful quantity to discriminate between hypotheses.

WILKS THEOREM

In the following we evaluate χ^2_λ for the poissonian histogram.

$$\chi^2_\lambda = -2\ln\prod_{i=1}^{M}\frac{e^{-y_i}y_i^{n_i}}{n_i!} + 2\ln\prod_{i=1}^{M}\frac{e^{-\nu_i}\nu_i^{n_i}}{n_i!} \tag{110}$$

Notice that the first term includes the theory (through the y_i), while the second requires the knowledge of the expectation values of the data. If we make the identification $\nu_i = n_i$, we get:

$$\chi^2_\lambda = -2\sum_{i=1}^{M}\left(n_i\ln\frac{y_i}{n_i} - (y_i - n_i)\right) = -2\sum_{i=1}^{M}\left(n_i\ln\frac{y_i}{n_i}\right) + 2(N_0 - N) \tag{111}$$

By imposing $\nu_i = n_i$ eq. χ^2_λ is the ratio of the likelihood of the theory to the likelihood of the data. The lower is χ^2_λ the better is the agreement between data and theory. For $y_i = n_i$ (perfect agreement) $\chi^2_\lambda = 0$.

If we make the same calculation for the multinomial likelihood we obtain the same expression but without the $N_0 - N$ term that corresponds to the fluctuation of the total number of events. This term is only present when we allow the total number of events to fluctuate, as in the poissonian case.

WILKS THEOREM

5.2.2. *Study of a functional dependence.* A likelihood function can also be easily defined in another context widely used in experimental physics. We consider the case of M measurements z_i all characterized by gaussian fluctuations with uncertainties σ_i done for different values of an independent variable x. If the theory predicts a functional dependence between z and x given by the function $z = f(x/\underline{\theta})$ possibly depending on a set of parameters $\underline{\theta}$, in case of no correlation between the measurements z_i, and completely neglecting possible uncertainties on x, we can build a gaussian likelihood:

$$L_g(\underline{z}/\underline{\theta}) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\left(z_i - f(x_i/\underline{\theta})\right)^2}{2\sigma_i^2}}$$

Choice of test statistics: binned data

WILKS THEOREM

Let's now apply the Wilks theorem to this case. For the gaussian measurements we make the identification $\nu_i = E[z_i] = z_i$ and we get:

(113)

$$\chi^2_\lambda = -2\ln\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(z_i-f(x_i/\underline{\theta}))^2}{2\sigma_i^2}} + 2\ln\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(z_i-z_i)^2}{2\sigma_i^2}} = \sum_{i=1}^{M}\frac{(z_i - f(x_i/\underline{\theta}))^2}{\sigma_i^2}$$

The test statistics obtained here is a χ^2, typically used in the context of the so called **least squares method**. So we have proved essentially that the least square method can be derived through the Wilks theorem by a gaussian likelihood ratio model.

Choice of test statistics: unbinned data

5.2.3. *Unbinned data.* In case we have a limited number N of events so that any binning will bring us to small values of bin contents, a different approach can be used, equally relying on the likelihood method: we can fit the **unbinned** data. In other words we build our likelihood function directly considering the probability of each single event. If we call H our hypothesis (eventually depending on a set of K parameters $\underline{\theta}$), x_i with i=1, ...N the values of the variable x for the N events and $f(x/\underline{\theta})$ the pdf of x given the hypothesis H, the likelihood can be written as:

$$L(\underline{x}/H) = \prod_{i=1}^{N} f(x_i/\underline{\theta}) \tag{114}$$

valid in case the events are not correlated. Notice that in this case the product runs on the events, not on the bins as in the previous case. If N is not fixed but fluctuates we can include "by hand" in the likelihood, the poissonian fluctuation of N around an expectation value that we call N_0 (eventually an additional parameter to be fit)[20]:

$$L(\underline{x}/H) = \frac{e^{-N_0} N_0^N}{N!} \prod_{i=1}^{N} f(x_i/\underline{\theta}) \tag{115}$$

This is called **extended likelihood**.

The - logarithm of the likelihood is used in most cases[21]:

$$-\ln L(\underline{x}/H) = -\sum_{i=1}^{N} \ln f(x_i/\underline{\theta}) \tag{116}$$

Choice of test statistics: correlations

5.2.4. *Fit of correlated data.* By using the product of the probability functions to write down the likelihood, we are assuming no correlation between bins (in case of histograms) or between events (in case of unbinned fits). In general it is possible to take into account properly the correlation between measurements in the definition of a likelihood function. We see how this happens in a simple case. Assume that our gaussian measurements of z_i (see above) are not independent. In this case the likelihood cannot be decomposed in the product of single likelihoods, but a "joint likelihood" $L(\underline{z}, /\underline{\theta})$ is defined, including the covariance matrix V_{ij} between the measurements. The covariance matrix has the parameters variances in the diagonal elements and the covariances in the off-diagonal elements. Starting from the joint likelihood of the measurement, we build the likelihood ratio and in the end we are left with the final χ^2:

$$\chi^2 = \sum_{j,k=1}^{M} (z_j - f(x_j/\underline{\theta}))\, V_{jk}^{-1}\, (z_k - f(x_k/\underline{\theta})) \tag{117}$$

that is still a χ^2 variable with $M - K$ degrees of freedom.

Goodness-of-fit test : P-value

Test of hypothesis H_0 (**null hypothesis**)

Fit done (best estimate of θ_i) => t* obtained for the test statistics

Suppose pdf of test statistics t known => $f(t|H_0)$

P-value $$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\,dt$$

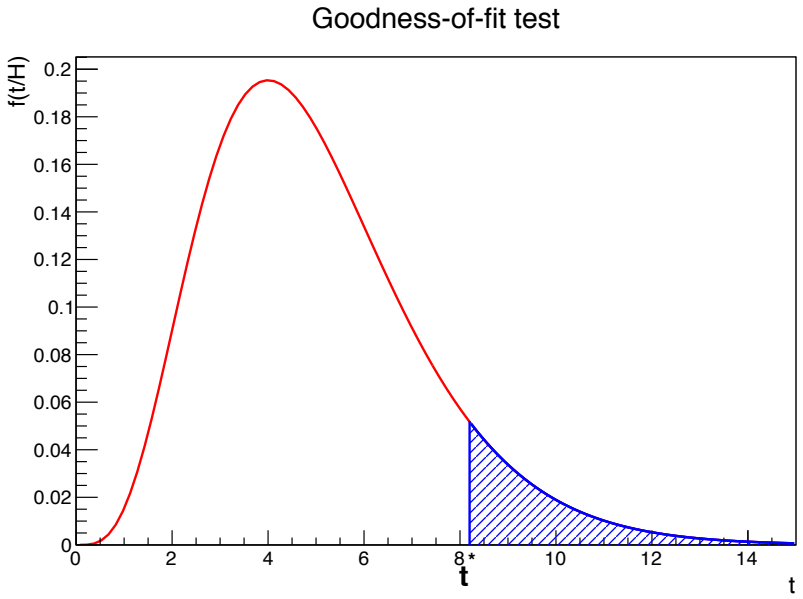


FIGURE 9. χ^2 distribution for 5 degrees of freedom. The case of $t^* = 8.2$ is illustrated. The blue hatched area correspond to the p_0 value.

Goodness-of-fit test : P-value

Meaning of P-value

$$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\,dt$$

Probability that - if H_0 is true - the result t of the experiment will fluctuate more than t*.

Repeating the experiment N times, p_0 is the fraction in which we get t>t*

$t > t^*$. If this number is low, either the hypothesis is wrong or there was an anomalous large fluctuation. In other words we are on the right tail of the distribution. So we can put a limit on the acceptable values of p_0: if p_0 is less than, say 5% or 1% we will reject the null hypothesis, if it is larger than the same limit we will say on the contrary that the null hypothesis is corroborated. The choice of the limit (5, 1 or 0.1%) depends on the nature of the problem, and on the degree we decide to be severe with the results we are considering.

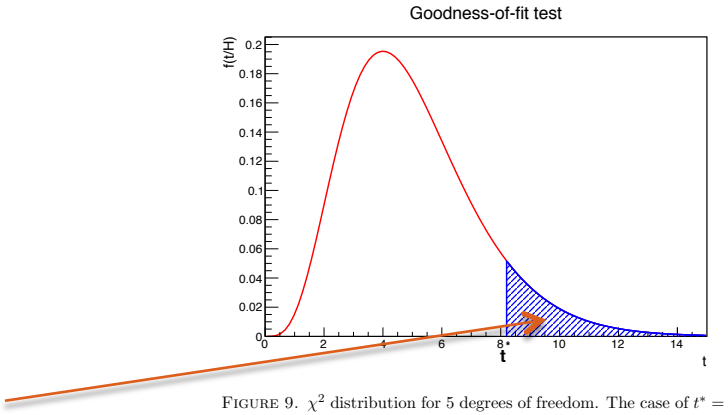


FIGURE 9. χ^2 distribution for 5 degrees of freedom. The case of $t^* = 8.2$ is illustrated. The blue hatched area correspond to the p_0 value.

$p_0 \approx 0$ => rejection of null H_0 hypothesis,
i.e. scarce agreement data-theory

Goodness-of-fit test : P-value

Meaning of P-value

$$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\,dt$$

f(t) pdf of t
g(F) pdf of F primitive of f

The P-value is a random variable itself uniformly distributed between 0 and 1:

$$g(F)dF = f(t)dt$$

$$\text{by definition } dF/dt = f(t) \qquad g(F) = \frac{f(t)}{dF/dt} = \frac{f(t)}{f(t)} = 1$$

All p-values are equally probable! e.g. $p_0 \approx 0$ or $p_0 \approx 1$

Goodness-of-fit test : P-value

Meaning of P-value

$$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\,dt$$

f(t) pdf of t
g(F) pdf of F primitive of f

The P-value is a random variable itself uniformly distributed between 0 and 1:

$$g(F)dF = f(t)dt$$

$$\text{by definition } dF/dt = f(t) \qquad g(F) = \frac{f(t)}{dF/dt} = \frac{f(t)}{f(t)} = 1$$

All p-values are equally probable! e.g. $p_0 \approx 0$ or $p_0 \approx 1$
If H_0 is true, if H_0 is false usually $p_0 \approx 0$.

Goodness-of-fit test : P-value

Meaning of P-value

$$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\,dt$$

The P-value is a random variable itself uniformly distributed between 0 and 1:

All p-values are equally probable! e.g. $p_0 \approx 0$ or $p_0 \approx 1$
If H_0 is true, if H_0 is false usually $p_0 \approx 0$.

What if $p_0 \approx 1$?

$p_0 \approx 1$ => underfluctuations of experimental points or overestimate of the uncertainties , i.e. scarce self-consistency of data

2-tails test vs 1-tail test

e.g. Accept H_0 if $5\% < p_0 < 95\ \%$

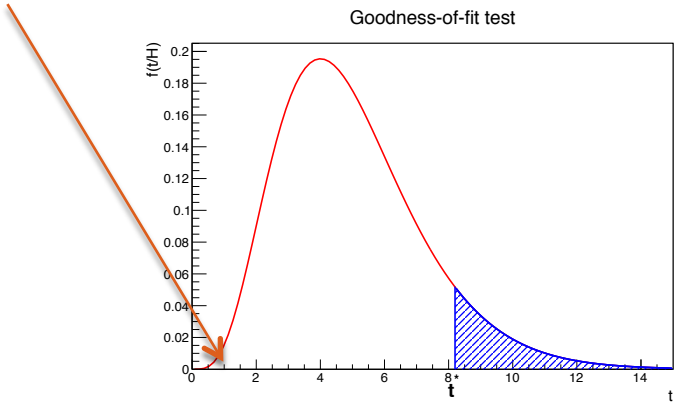


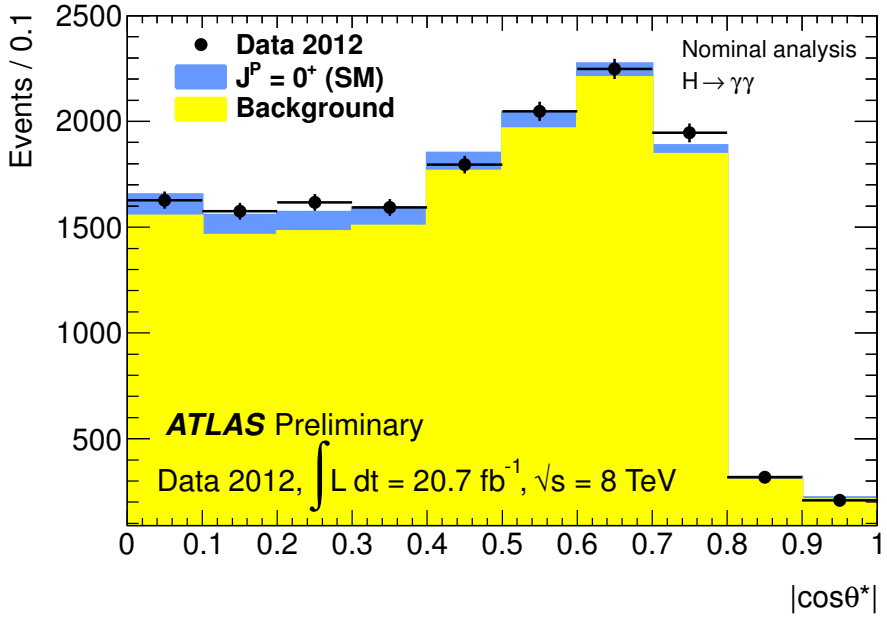
FIGURE 9. χ^2 distribution for 5 degrees of freedom. The case of $t^* = 8.2$ is illustrated. The blue hatched area correspond to the p_0 value.

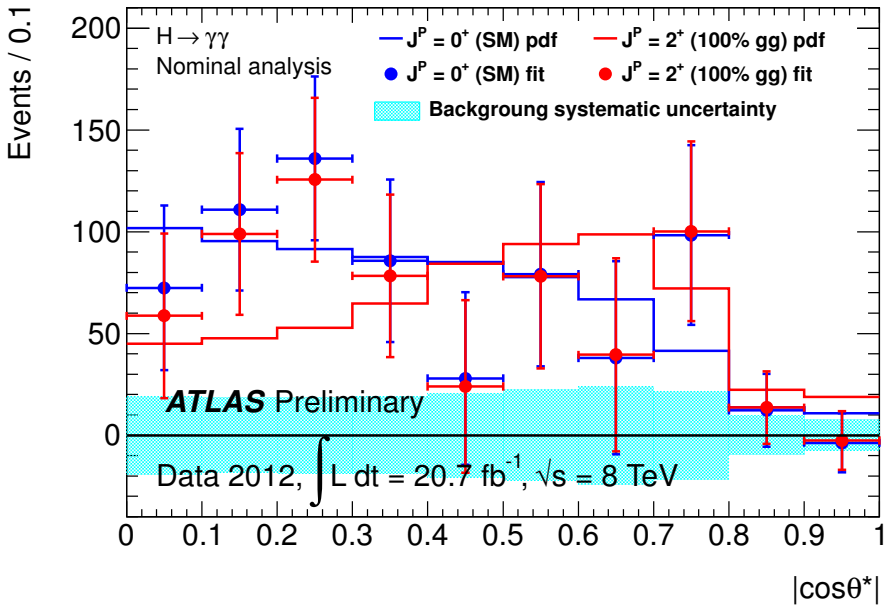
Example of two alternate hypotheses H_0 and H_1

In the two-body decay $H \rightarrow \gamma\gamma$, the spin information is extracted from the distribution of the polar angle θ^* of the photons with respect to the z-axis of the Collins-Soper frame.

$$\cos\theta^* = \frac{\sinh(\eta_{\gamma_1} - \eta_{\gamma_2})}{\sqrt{1 + \left(p_{\mathrm{T}}^{\gamma\gamma}/m_{\gamma\gamma}\right)^2}} \cdot \frac{2 p_{\mathrm{T}}^{\gamma_1} p_{\mathrm{T}}^{\gamma_2}}{m_{\gamma\gamma}^2}$$

With this choice, the impact of initial state radiation is expected to be minimized and a better discrimination power compared to other choices of axis, such as the beam axis or the boost axis of the particle, is achieved. A spin-0 particle decays isotropically in its rest frame; before any acceptance cuts, the distribution $dN/d\cos\theta^*$ is thus uniform. The corresponding distribution for a spin-2 particle follows a combination of Wigner functions for the production and decay whose probabilities are specified in particular models.





Example of two alternate hypotheses H_0 and H_1

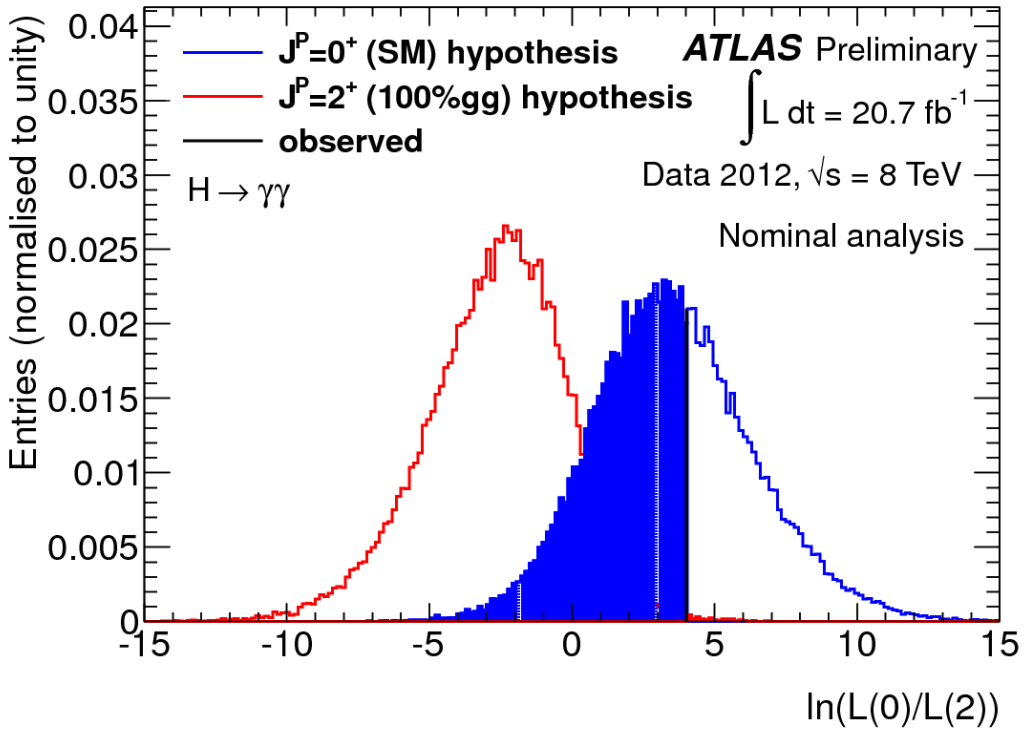


FIGURE 10. One of the results of the ATLAS experiment for the study of the spin of the Higgs boson. The pdf's of the test statistics q (defined as the logarithm of the likelihood ratio) are shown for two alternative hypotheses: spin 0 and spin 2. The black vertical line corresponds to the experimental value of the test statistics. The blue hatched area is the 1-p-value. (taken from ATLAS Collaboration, ATLAS-CONF-2013-029).

Two alternate hypotheses H_0 and H_1

Define t_{cut}
If $t^* < t_{cut}$ => accept the null hypothesis

If $t^* > t_{cut}$ => accept the alternate hypothesis

By applying a cut we accept type-I and type-II errors (similarly to single events…)

$$\alpha = \int_{t_{cut}}^{\infty} f(t/H_0)dt$$
$$\beta = \int_{-\infty}^{t_{cut}} f(t/H_1)dt$$

Apply Neyman-Pearson lemma, i.e. construct a Likelihood ratio variable
as best test statistics