Statistical Tests and Limits

Lecture 1: general formalism

IN2P3 School of Statistics

Autrans, France

17—21 May, 2010

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

→ Lecture 1: General formalism

Definition and properties of a statistical test

Significance tests (and goodness-of-fit) , *p*-values

Lecture 2: Setting limits

Confidence intervals

Bayesian Credible intervals

Lecture 3: Further topics for tests and limits

More on systematics / nuisance parameters

Look-elsewhere effect

CLs

Bayesian model selection

Hypotheses

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim f(x|H)$.

Possible values of x form the sample space S (or “data space”).

Simple (or “point”) hypothesis: $f(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Definition of a test

Consider e.g. a simple hypothesis H_0 and alternative H_1.

A test of H_0 is defined by specifying a critical region W of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W \mid H_0) \leq \alpha$$

If x is observed in the critical region, reject H_0.

α is called the size or significance level of the test.

Critical region also called "rejection" region; complement is acceptance region.

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P $(0.117 < \alpha_s < 0.121)$,

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)\,dH}$$

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximimum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in \mathrm{S} - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Statistical test in a particle physics context

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons,

x_2 = mean p_t of jets,

x_3 = missing energy, ...

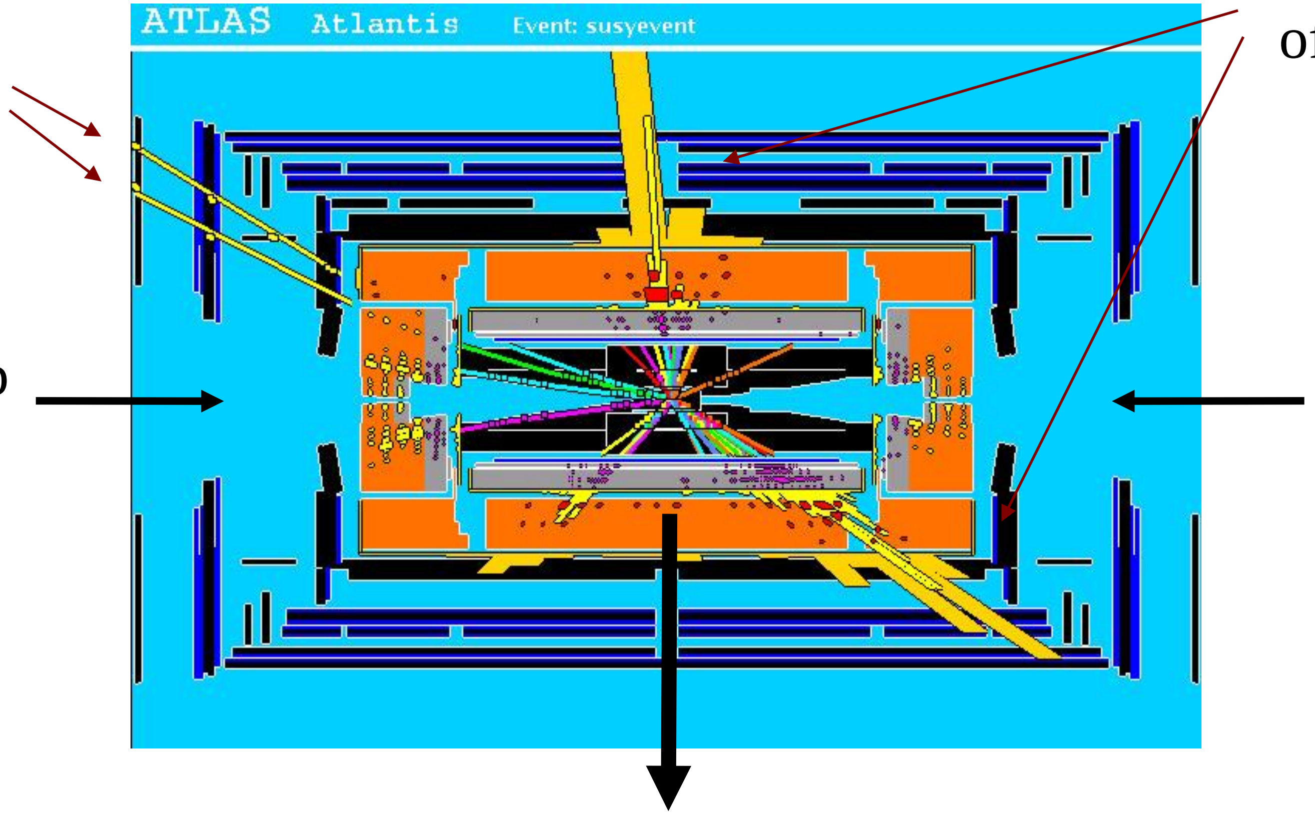
$\vec{x}$ follows some *n*-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$\mathrm{pp} \rightarrow t\bar{t}\ , \qquad \mathrm{pp} \rightarrow \tilde{g}\tilde{g}\ , \ldots$$

For each reaction we consider we will have a hypothesis for the pdf of $\vec{x}$, e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

Often call H_0 the background hypothesis (e.g. SM events); H_1, H_2, ... are possible signal hypotheses.

A simulated SUSY event in ATLAS



high p_T muons

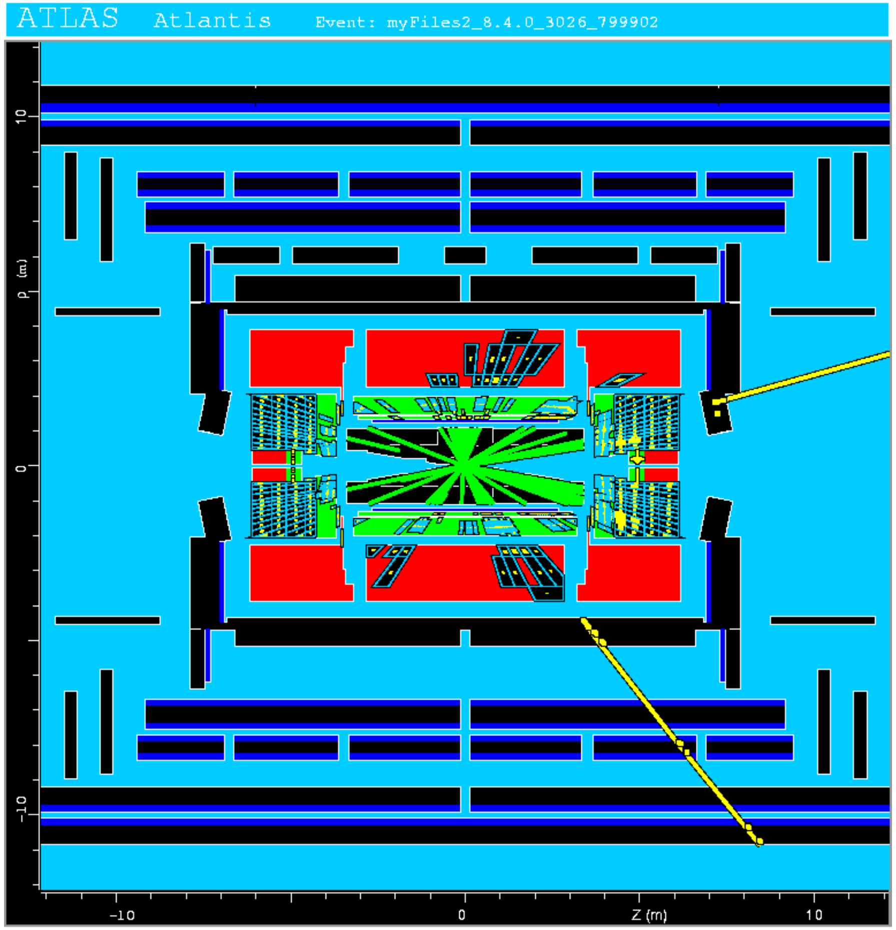
high p_T jets of hadrons

p

p

missing transverse energy

Background events



This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

Selecting events

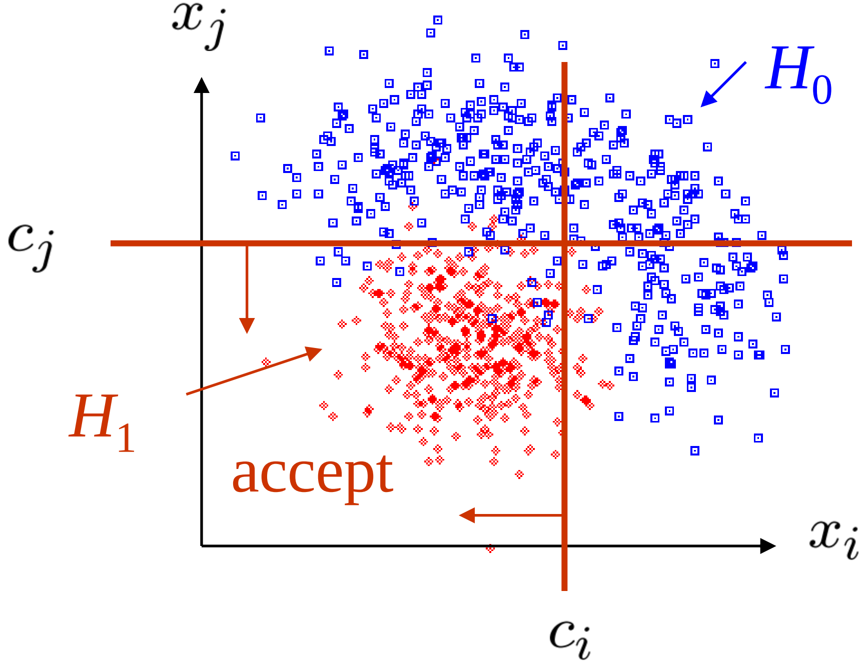
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_0.

Each event is a point in $\vec{x}$ space. What 'decision boundary' should we use to accept/reject events as belonging to event type H_0?

Perhaps select events with 'cuts':

$$x_i < c_i$$

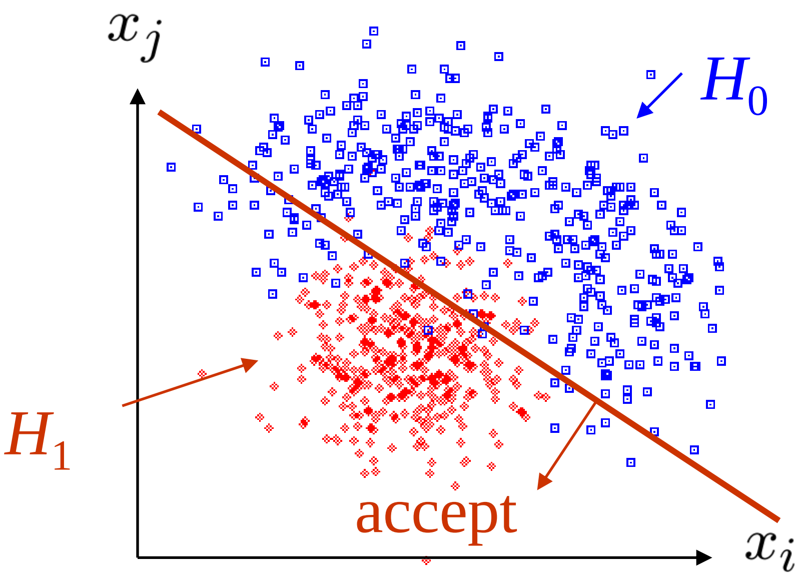
$$x_j < c_j$$



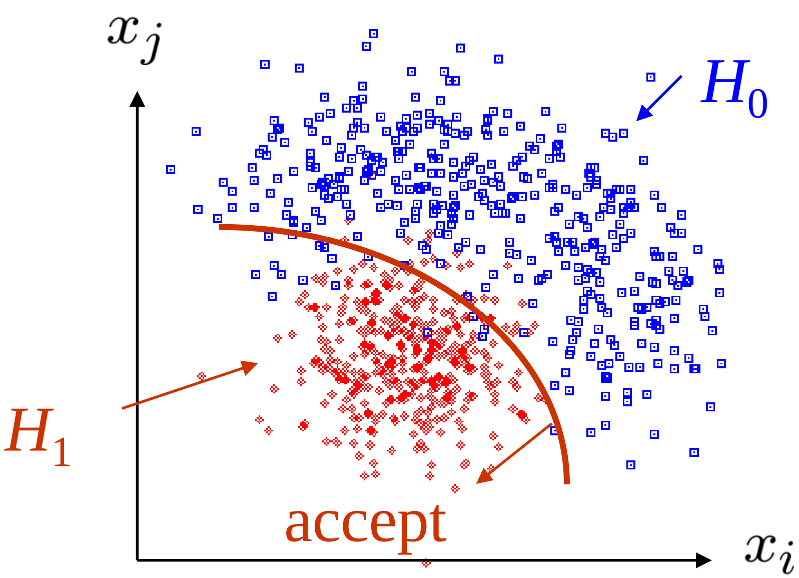
Other ways to select events

Or maybe use some other sort of decision boundary:

linear



or nonlinear



How can we do this in an ‘optimal’ way?

Test statistics

Construct a 'test statistic' of lower dimension (e.g. scalar)

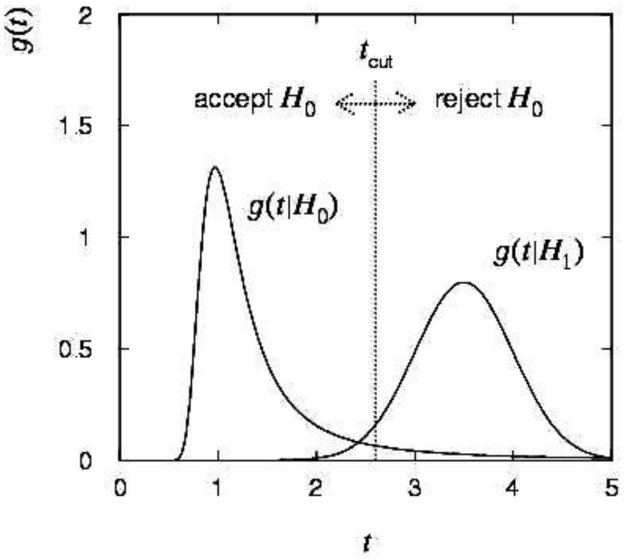
$$t(x_1, \ldots, x_n)$$

Goal is to compactify data without losing ability to discriminate between hypotheses.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on *t*.

This effectively divides the sample space into two regions, where we accept or reject H_0 and thus defines a statistical test.

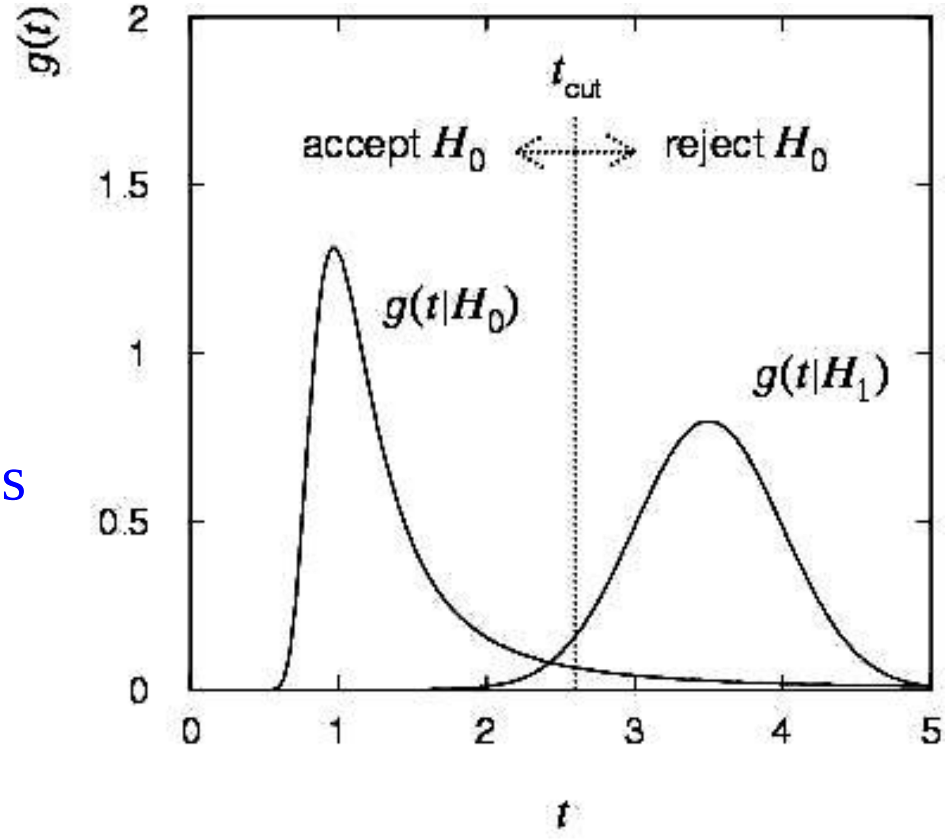


Significance level and power

Probability to reject H_0 if it is true (type-I error):

$$\alpha = \int_{t_{\mathrm{cut}}}^{\infty} g(t|H_0)\,dt$$

(significance level)



Probability to accept H_0 if H_1 is true (type-II error):

$$\beta = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|H_1)\,dt$$

($1 - \beta =$ power)

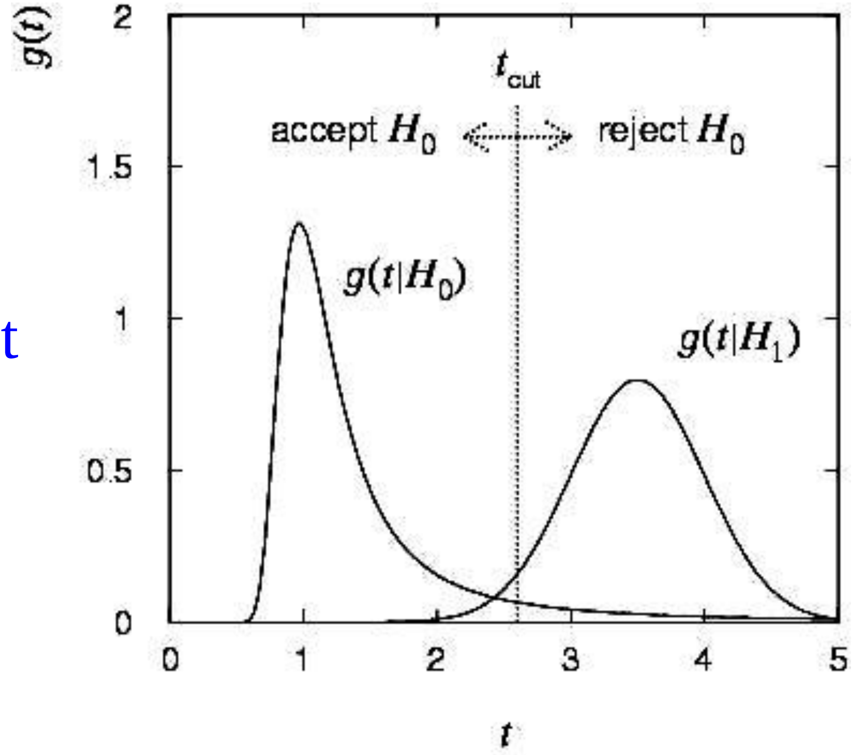
Signal/background efficiency

Probability to reject background hypothesis for background event (background efficiency):

$$\varepsilon_{\mathrm{b}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{b})\,dt = \alpha$$

Probability to accept a signal event as signal (signal efficiency):

$$\varepsilon_{\mathrm{s}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{s})\,dt = 1 - \beta$$



Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are π_s and π_b (prior probabilities).

Suppose we select events with $t < t_{\mathrm{cut}}$. What is the 'purity' of our selected sample?

Here purity means the probability to be signal given that the event was accepted. Using Bayes' theorem we find:

$$P(\mathrm{s}|t < t_{\mathrm{cut}}) = \frac{P(t < t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}}}{P(t < t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}} + P(t < t_{\mathrm{cut}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= \frac{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}}}{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}} + \varepsilon_{\mathrm{b}}\pi_{\mathrm{b}}}$$

So the purity depends on the prior probabilities as well as on the signal and background efficiencies.

Constructing a test statistic

How can we select events in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest ε_s for a given ε_b (highest power for a given significance level), choose acceptance region such that

$$\frac{f(\vec{x}|\mathsf{s})}{f(\vec{x}|\mathsf{b})} > c$$

where c is a constant which determines ε_s.

Equivalently, optimal scalar test statistic is

$$\boxed{t(\vec{x}) = \frac{f(\vec{x}|\mathsf{s})}{f(\vec{x}|\mathsf{b})}}$$

N.B. any monotonic function of this is leads to the same test.

Proof of Neyman-Pearson lemma

We want to determine the critical region *W* that maximizes the power

$$1 - \beta = \int_W P(x|H_1)\, dx$$

subject to the constraint

$$\alpha = \int_W P(x|H_0)\, dx$$

First, include in *W* all points where P($x|H_0$) = 0, as they contribute nothing to the size, but potentially increase the power.

Proof of Neyman-Pearson lemma (2)

For $P(x|H_0) \neq 0$ we can write the power as

$$1 - \beta = \int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx$$

The ratio of $1 - \beta$ to α is therefore

$$\frac{1 - \beta}{\alpha} = \frac{\int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx}{\int_W P(x|H_0)\, dx}$$

which is the average of the likelihood ratio $\mathrm{P}(x|H_1)$ / $\mathrm{P}(x|H_0)$ over the critical region W, assuming H_0.

$(1 - \beta) / \alpha$ is thus maximized if W contains the part of the sample space with the largest values of the likelihood ratio.

Purity vs. efficiency — optimal trade-off

Consider selecting *n* events:

expected numbers *s* from signal, *b* from background;

→ $n \sim$ Poisson $(s + b)$

Suppose *b* is known and goal is to estimate *s* with minimum relative statistical error.

Take as estimator: $\hat{s} = n - b\ .$

Variance of Poisson variable equals its mean, therefore

$$V[\hat{s}] = V[n - b] = V[n] = s + b \quad \rightarrow \quad \frac{\sigma_{\hat{s}}}{s} = \frac{\sqrt{s+b}}{s}$$

So we should maximize $\dfrac{s}{\sqrt{s+b}}$ (or $\varepsilon_s/\sqrt{b}$ if $s \ll b$),

equivalent to maximizing product of signal efficiency × purity.

Two distinct event selection problems

In some cases, the event types in question are both known to exist.

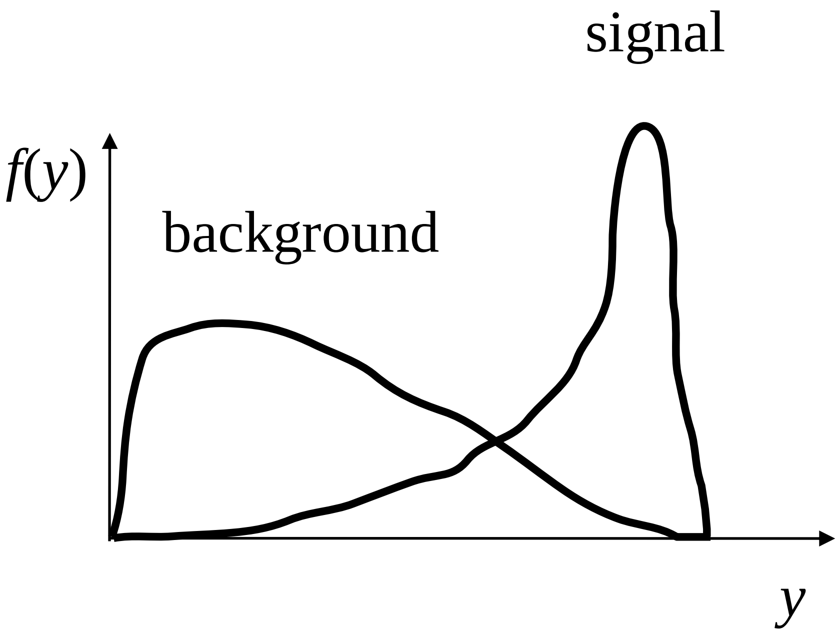
Example: separation of different particle types (electron vs muon)
Use the selected sample for further study.

In other cases, the null hypothesis H_0 means "Standard Model" events, and the alternative H_1 means "events of a type whose existence is not yet established" (to do so is the goal of the analysis).

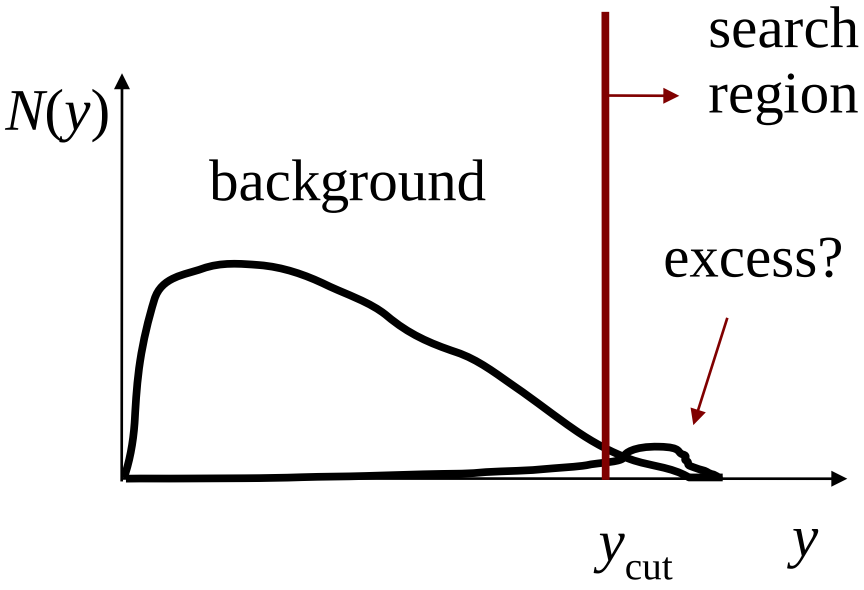
Many subtle issues here, mainly related to the heavy burden of proof required to establish presence of a new phenomenon.

Typically require p-value of background-only hypothesis below ~ 10^{-7} (a 5 sigma effect) to claim discovery of "New Physics".

Using test for discovery



Normalized to unity



Normalized to expected number of events

Discovery = number of events found in search region incompatible with background-only hypothesis.

p-value of background-only hypothesis can depend crucially distribution $f(y|\mathrm{b})$ in the "search region".

Multivariate methods

Many new (and some old) methods for determining test:

- Fisher discriminant
- Neural networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

New software for HEP, e.g.,

TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

More on this in the lectures by Kegl, Feindt, Hirshbuehl, Coadou.

For the rest of these lectures, I will focus on other aspects of tests, e.g., discovery significance and exclusion limits.

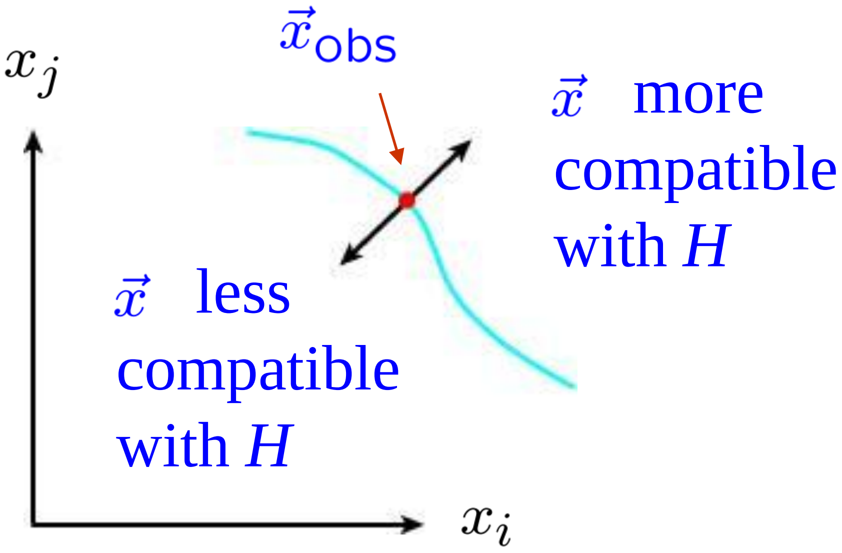
Testing significance / goodness-of-fit

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of *H* in light of the data?

Decide what part of the data space represents less compatibility with *H* than does the point $\vec{x}_{\mathrm{obs}}$. (Not unique!)



$\vec{x}$ more compatible with *H*

$\vec{x}$ less compatible with *H*

p-values

Express level of agreement between data and hypothesis by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

⚠ This is not the probability that *H* is true!

In frequentist statistics we don't talk about *P*(*H*) (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as *P*(*H*).

p-value example: testing whether a coin is 'fair'

Probability to observe *n* heads in *N* coin tosses is binomial:

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

Hypothesis *H*: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with *H* relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \; .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of *H*.

The significance of an observed signal

Suppose we observe *n* events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means *s*, *b*, then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

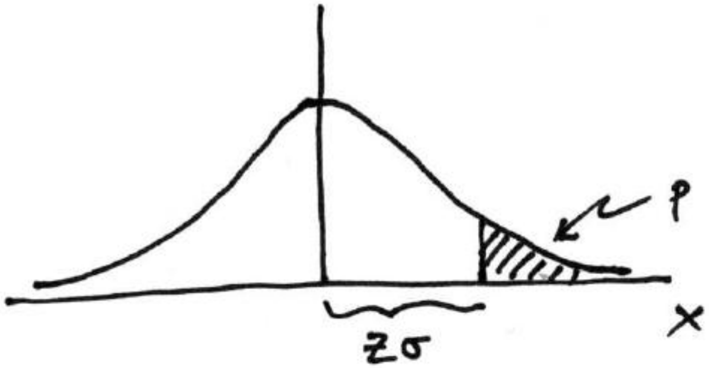
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

Give *p*-value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s=0)! \end{aligned}$$

Significance from *p*-value

Often define significance *Z* as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same *p*-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

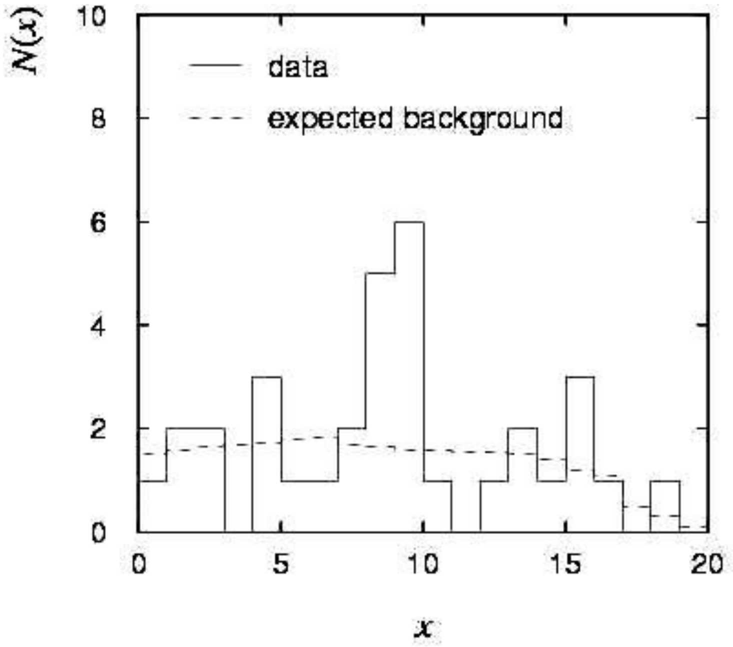
`1 - TMath::Freq`

$$Z = \Phi^{-1}(1 - p)$$

`TMath::NormQuantile`

The significance of a peak

Suppose we measure a value x for each event and find:



Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with b = 3.2. The p-value for the s = 0 hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0 D^0$ mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

One should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null hypothesis; p-value is only first step!

Prototype search analysis

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Search for signal in a region of phase space; result is histogram of some variable *x* giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \underset{\uparrow}{\mu} s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx \,, \qquad b_i = b_{\rm tot} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_{\mathrm{s}}$, $\boldsymbol{\theta}_{\mathrm{b}}$,$b_{\mathrm{tot}}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

maximizes L for specified μ

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximize L

The likelihood ratio gives optimum test between two point hypotheses (Neyman-Pearson lemma).

Should be near-optimal in present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. only regard upward fluctuation of data as evidence against the background-only hypothesis.

Large q_0 means increasing incompatibility between the data and hypothesis, therefore *p*-value for an observed $q_{0,\mathrm{obs}}$ is

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

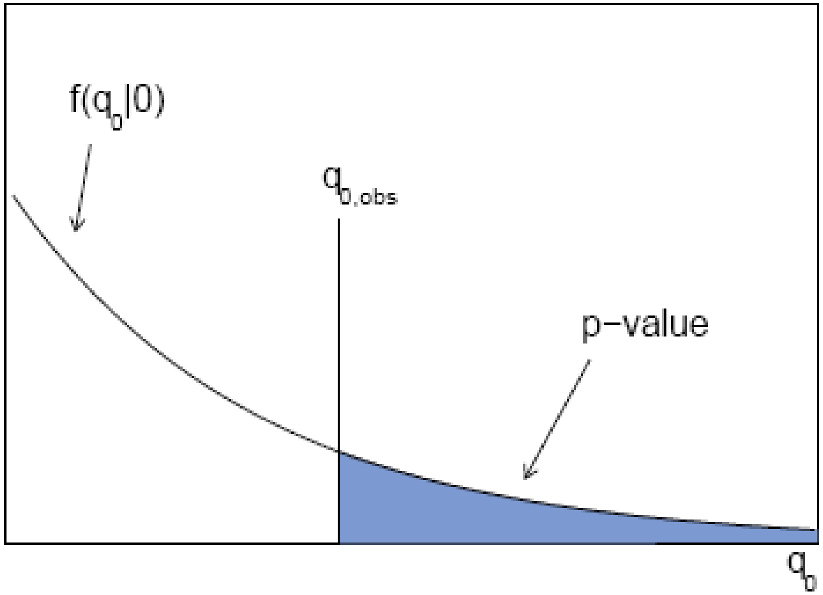
will get formula for this later

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore *p*-value for an observed $q_{0,\mathrm{obs}}$ is

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

will get formula for this later

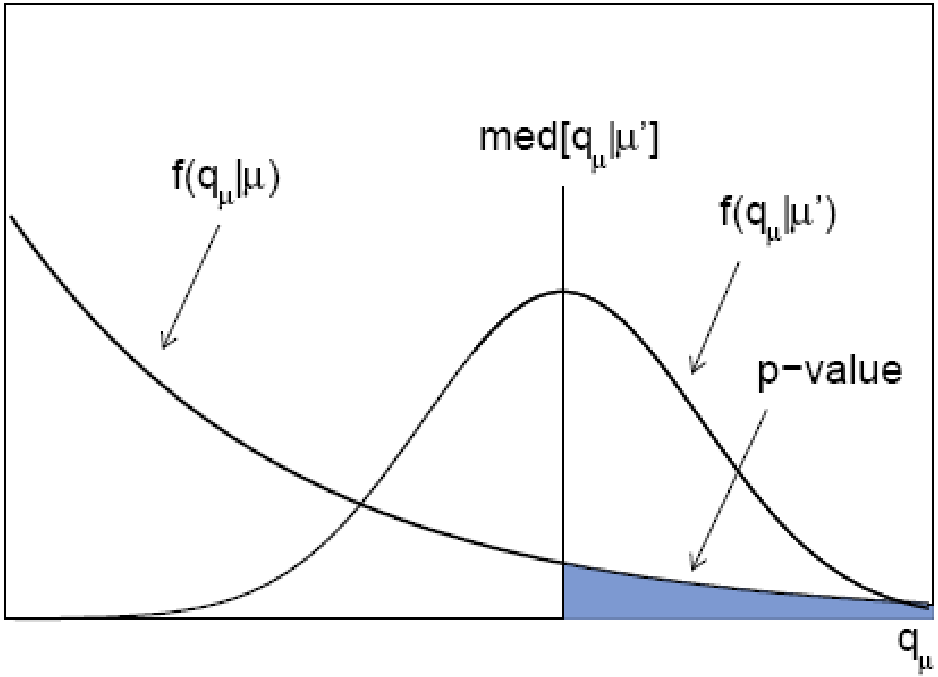


From *p*-value get equivalent significance,

$$Z = \Phi^{-1}(1-p)$$

Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter μ'.



So for p-value, need $f(q_0|0)$, for sensitivity, will need $f(q_0|\mu')$,

Wald approximation for profile likelihood ratio

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells, *Using the Profile Likelihood in Searches for New Physics*, in preparation.

To find *p*-values, we need: $f(q_0|0), \quad f(q_\mu|\mu)$

For median significance under alternative, need: $f(q_\mu|\mu')$

Use approximation due to Wald (1943)

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N})$$

$$\hat{\mu} \sim \text{Gaussian}(\mu', \sigma)$$

sample size

i.e., $E[\hat{\mu}] = \mu'$

σ from covariance matrix V, use, e.g.,

$$V^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

Noncentral chi-square for $-2\ln\lambda(\mu)$

If we can neglect the $O(1/\sqrt{N})$ term, $-2\ln\lambda(\mu)$ follows a noncentral chi-square distribution for one degree of freedom with noncentrality parameter

$$\Lambda = \frac{(\mu - \mu')^2}{\sigma^2}$$

As a special case, if $\mu' = \mu$ then $\Lambda = 0$ and $-2\ln\lambda(\mu)$ follows a chi-square distribution for one degree of freedom (Wilks).

Distribution of q_0

Assuming the Wald approximation, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \Phi\left(\frac{\mu'}{\sigma}\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of *q*0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The *p*-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance *Z* is simply

$$\boxed{Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}}$$

The Asimov data set

To estimate median value of –2ln$\lambda(\mu)$, consider special data set where all statistical fluctuations suppressed and n_i, m_i are replaced by their expectation values (the “Asimov” data set):

$$\begin{aligned} n_i &= \mu' s_i + b_i \\ m_i &= u_i \end{aligned}$$

→ $\hat{\mu}_{\mathrm{A}} = \mu'$, $\boldsymbol{\theta}_{\mathrm{A}} = \hat{\boldsymbol{\theta}}$ from very large MC sample, so

$$\lambda_{\mathrm{A}}(\mu) = \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\mathrm{A}}(\hat{\mu}, \hat{\boldsymbol{\theta}})} \approx \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\mathrm{A}}(\mu', \boldsymbol{\theta}_{\mathrm{A}})}$$

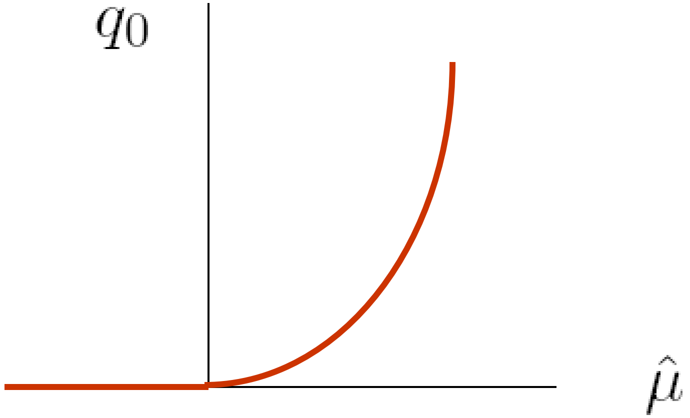
$$-2\ln\lambda_{\mathrm{A}}(\mu) = \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

Asimov value of –2ln$\lambda(\mu)$ gives non-centrality param. Λ, or equivalently, σ

Relation between test statistics and $\hat{\mu}$

Assuming Wald approximation, the relation between q_0 and $\hat{\mu}$ is

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$



Monotonic, therefore quantiles of $\hat{\mu}$ map one-to-one onto those of q_0, e.g.,

$$\mathrm{med}[q_0] = q_0(\mathrm{med}[\hat{\mu}]) = q_0(\mu') = \frac{\mu'^2}{\sigma^2} = -2\ln\lambda_A(0)$$

Higgs search with profile likelihood

Combination of Higgs boson search channels (ATLAS)

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Standard Model Higgs channels considered (more to be used later):

$H \rightarrow \gamma\gamma$

$H \rightarrow WW^{(*)} \rightarrow e\nu\mu\nu$

$H \rightarrow ZZ^{(*)} \rightarrow 4l \quad (l = e, \mu)$

$H \rightarrow \tau^+\tau^- \rightarrow ll, lh$

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

An example: ATLAS Higgs search

(ATLAS Collab., CERN-OPEN-2008-020)

Statistical Combination of Several Important Standard Model Higgs Boson Search Channels.

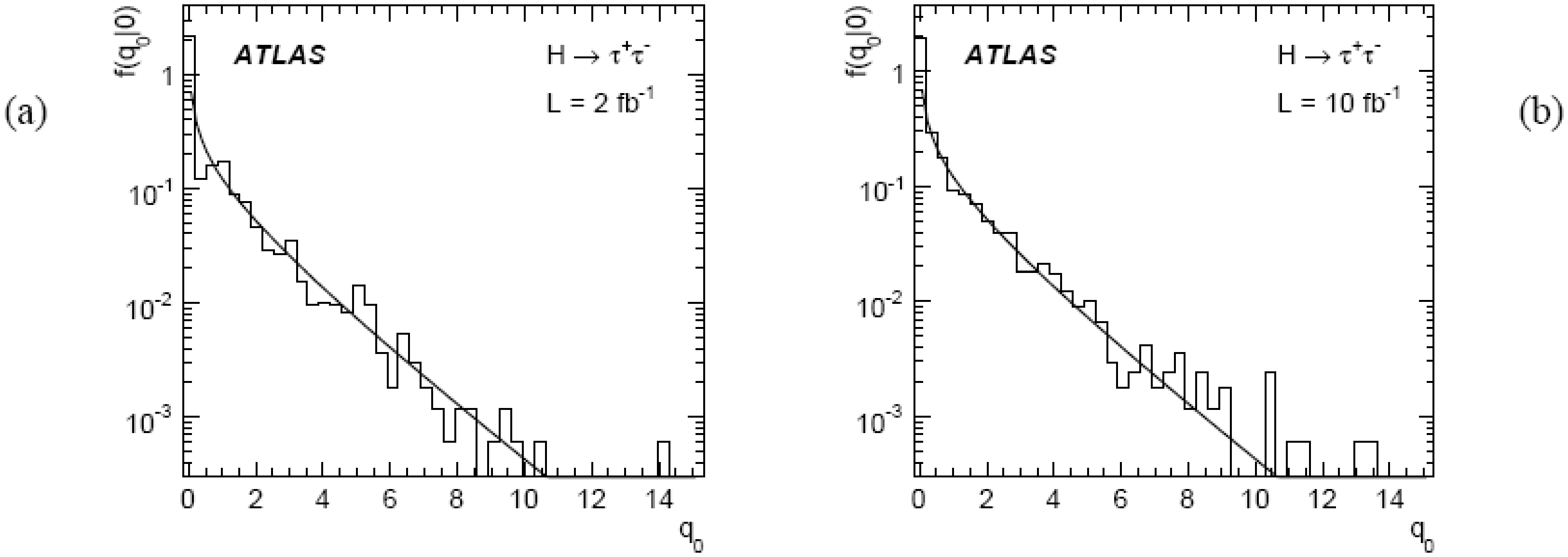
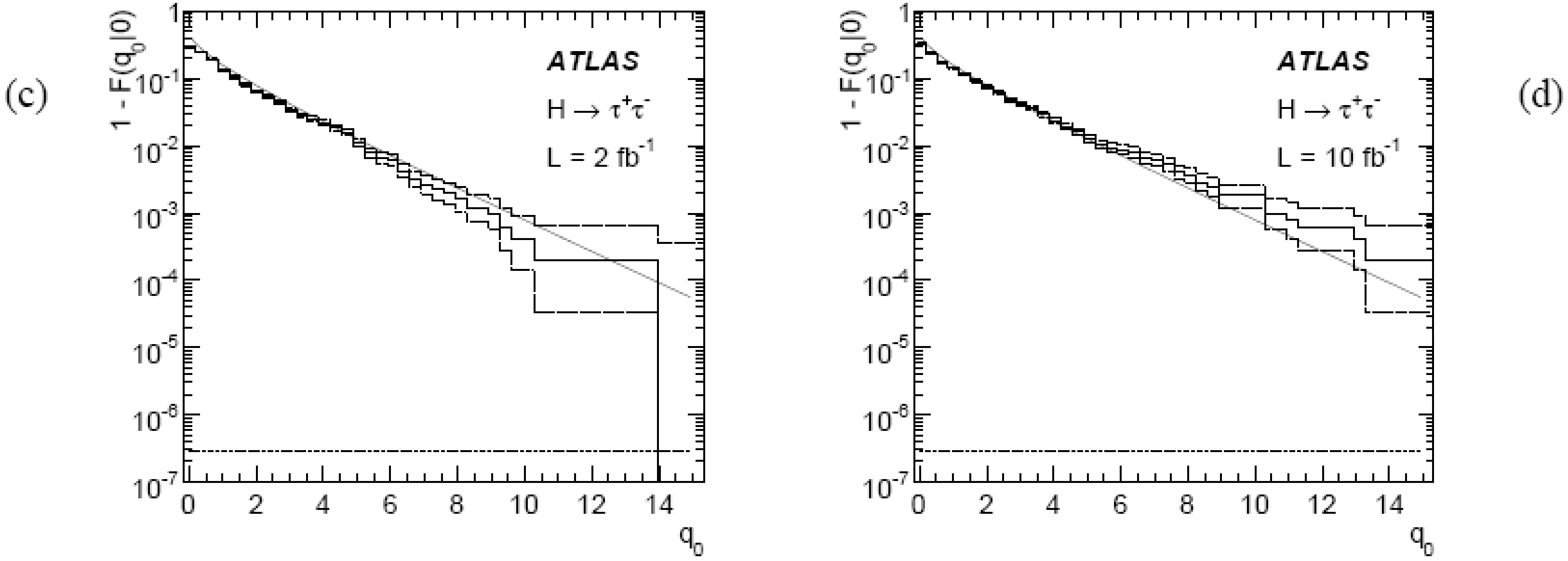


Figure 12: The distribution of the test statistic q_0 for $H \to \tau^+\tau^-$ under the null background-only hypothesis, for $m_H = 130\,\mathrm{GeV}$ with an integrated luminosity of 2 (a) and 10 (b) fb^{-1}. A $\frac{1}{2}\chi^2_1$ distribution is superimposed. Figures (c) and (d) show $1 - F(q_0)$ where $F(q_0)$ is the corresponding cumulative distribution. The small excess of events at high q_0 is statistically compatible with the expected curves, as can be seen by comparison with the dotted histograms that show the 68.3% central confidence intervals for $p = 1 - F(q_0|0)$. The lower dotted line at 2.87×10^{-7} shows the 5σ discovery threshold.

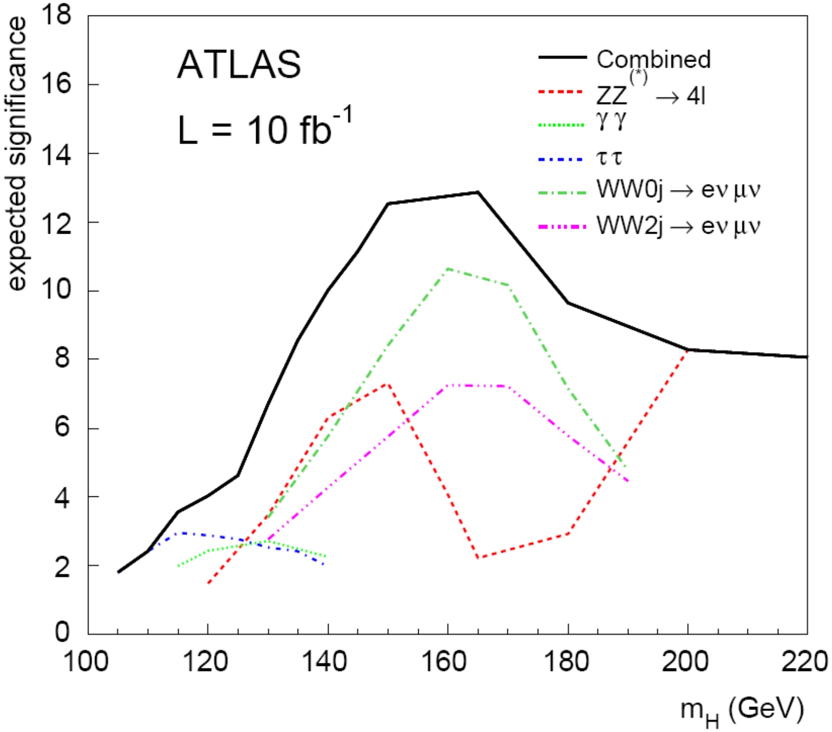
Cumulative distributions of q_0

To validate to 5σ level, need distribution out to $q_0 = 25$, i.e., around 10^8 simulated experiments.

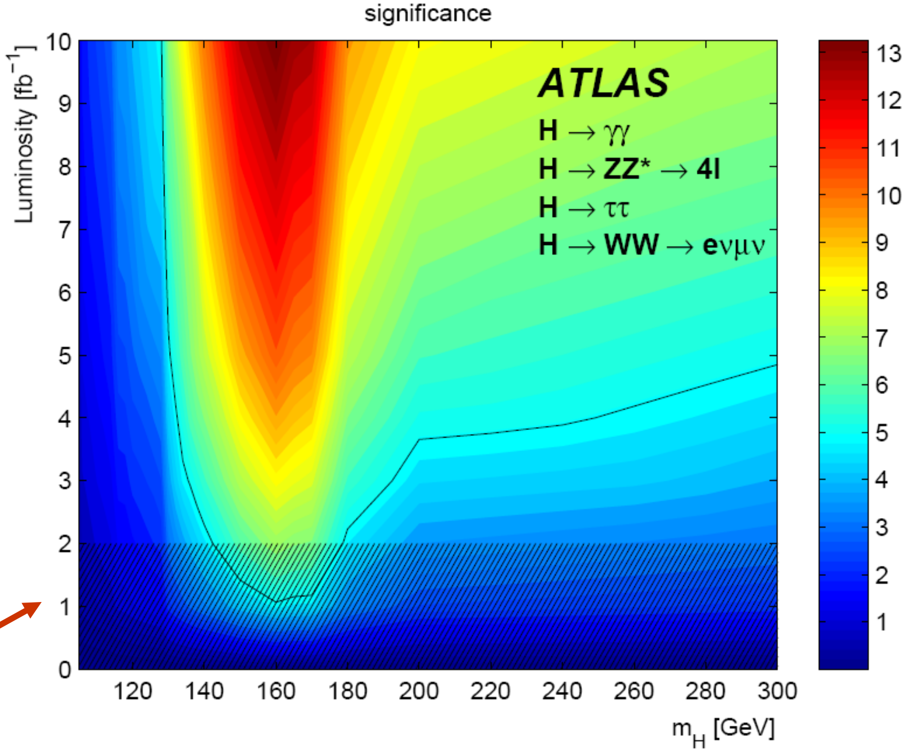
Will do this if we really see something like a discovery.



Combined discovery significance



Discovery signficance (in colour) vs. L, m_H:



Approximations used here not always accurate for $L < 2$ fb^{-1} but in most cases conservative.

Discovery significance for $n \sim \text{Poisson}(s + b)$

Consider again the case where we observe n events , model as following Poisson distribution with mean $s + b$ (assume b is known).

1) For an observed n, what is the significance Z_0 with which we would reject the $s = 0$ hypothesis?

2) What is the expected (or more precisely, median) Z_0 if the true value of the signal rate is s?

Gaussian approximation for Poisson significance

For large $s + b$, $n \rightarrow x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob$(x > x_{\mathrm{obs}} \mid s = 0)$,:

$$p_0 = 1 - \Phi\left(\frac{x_{\mathrm{obs}} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\mathrm{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\mathrm{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for Poisson significance

Likelihood function for parameter *s* is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

or equivalently the log-likelihood is

$$\ln L(s) = n \ln(s+b) - (s+b) - \ln n!$$

Find the maximum by setting $\frac{\partial \ln L}{\partial s} = 0$

gives the estimator for *s*: $\hat{s} = n - b$

Approximate Poisson significance (continued)

The likelihood ratio statistic for testing *s* = 0 is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \; 0 \text{ otherwise}$$

For sufficiently large *s* + *b*, (use Wilks' theorem),

$$Z_0 \approx \sqrt{q_0} = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b, \; 0 \text{ otherwise}$$

To find median[Z_0|*s*+*b*], let *n* → *s* + *b*,

$$\boxed{\text{median}[Z_0 | s + b] \approx \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}}$$

This reduces to *s*/√*b* for s << b.

Wrapping up lecture 1

General framework of a statistical test:

Divide data spaced into two regions; depending on where data are then observed, accept or reject hypothesis.

Properties:

significance level (rate of Type-I error)
power (one minus rate of Type-II error)

Significance tests (also for goodness-of-fit):

p-value = probability to see level of incompatibility between data and hypothesis equal to or greater than level found with the actual data.

Extra slides

Pearson's χ^2 statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$ (n_i independent) to predicted mean values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2}\,, \quad \text{where } \sigma_i^2 = V[n_i]\,.$$

(Pearson's χ^2 statistic)

χ^2 = sum of squares of the deviations of the *i*th measurement from the *i*th prediction, using σ_i as the 'yardstick' for the comparison.

For n_i ~ Poisson(ν_i) we have $V[n_i] = \nu_i$, so this becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}\,.$$

Pearson's χ^2 test

If n_i are Gaussian with mean ν_i and std. dev. σ_i, i.e., $n_i \sim \mathrm{N}(\nu_i\,,\,\sigma_i^2)$, then Pearson's χ^2 will follow the χ^2 pdf (here for $\chi^2 = z$):

$$f_{\chi^2}(z;N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

If the n_i are Poisson with $\nu_i >> 1$ (in practice OK for $\nu_i > 5$) then the Poisson dist. becomes Gaussian and therefore Pearson's χ^2 statistic here as well follows the χ^2 pdf.

The χ^2 value obtained from the data then gives the *p*-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z;N)\,dz\ .$$

The 'χ^2 per degree of freedom'

Recall that for the chi-square pdf for N degrees of freedom,

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

This makes sense: if the hypothesized ν_i are right, the rms deviation of n_i from ν_i is σ_i, so each term in the sum contributes ~ 1.

One often sees χ^2/N reported as a measure of goodness-of-fit. But... better to give $\chi^2$and N separately. Consider, e.g.,

$$\begin{aligned} \chi^2 &= 15, \; N = 10 \rightarrow p-\text{value} = 0.13 \,, \\ \chi^2 &= 150, \; N = 100 \rightarrow p-\text{value} = 9.0 \times 10^{-4} \,. \end{aligned}$$

i.e. for N large, even a χ^2 per dof only a bit greater than one can imply a small p-value, i.e., poor goodness-of-fit.

Pearson's χ^2 with multinomial data

If $n_{\text{tot}} = \sum_{i=1}^{N}$ is fixed, then we might model n_i ~ binomial

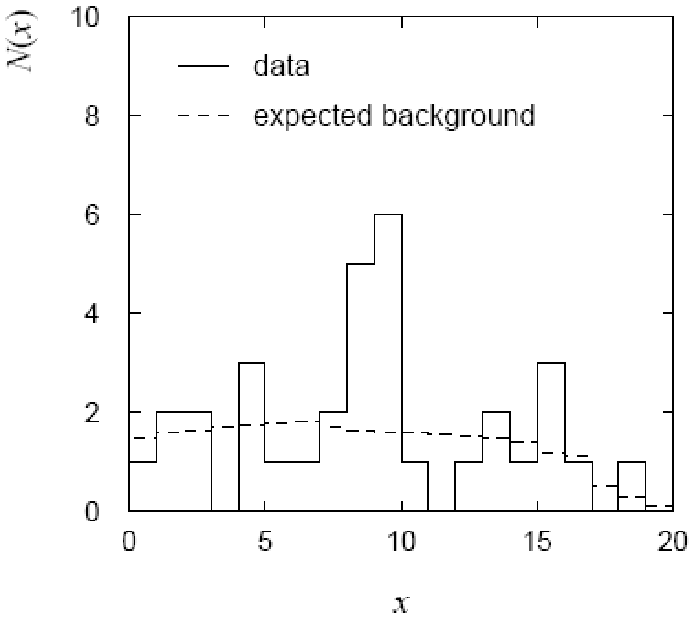
with $p_i = n_i / n_{\text{tot}}$. I.e. $\vec{n} = (n_1, \ldots, n_N)$ ~ multinomial.

In this case we can take Pearson's χ^2 statistic to be

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

If all $p_i\, n_{\text{tot}} >> 1$ then this will follow the chi-square pdf for $N - 1$ degrees of freedom.

Example of a χ^2 test



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for N = 20 dof.

Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect χ^2 to follow the chi-square pdf.

Using MC to find distribution of χ^2 statistic

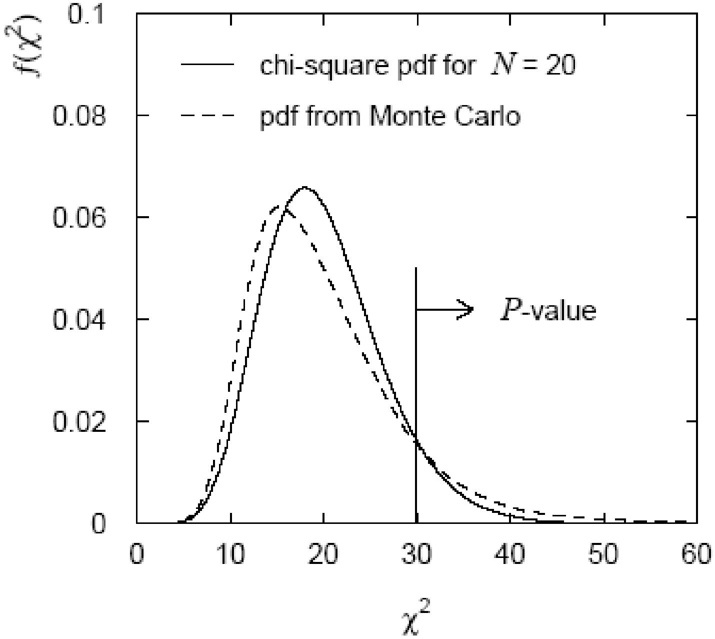
The Pearson χ^2 statistic still reflects the level of agreement between data and prediction, i.e., it is still a ‘valid’ test statistic.

To find its sampling distribution, simulate the data with a Monte Carlo program: $n_i \sim \text{Poisson}(\nu_i)\,,\ i = 1, N\,.$

Here data sample simulated 10^6 times. The fraction of times we find $\chi^2 > 29.8$ gives the *p*-value:

$$p = 0.11$$

If we had used the chi-square pdf we would find $p = 0.073$.


$f(\chi^2)$
chi-square pdf for $N = 20$
pdf from Monte Carlo
P-value
χ^2