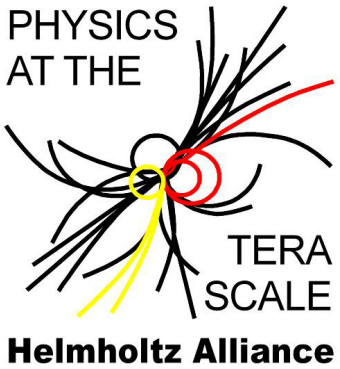
Statistical Methods for Discovery and Limits

Lecture 1: Introduction, Statistical Tests, Confidence Intervals

http://www.pp.rhul.ac.uk/~cowan/stat_desy.html
https://indico.desy.de/conferenceDisplay.py?confId=4489



School on Data Combination and Limit Setting
DESY, 4-7 October, 2011

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

Outline

Lecture 1: Introduction and basic formalism
Probability, statistical tests, confidence intervals.

Lecture 2: Tests based on likelihood ratios
Systematic uncertainties (nuisance parameters)

Lecture 3: Limits for Poisson mean
Bayesian and frequentist approaches

Lecture 4: More on discovery and limits
Spurious exclusion

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

Subjective (A = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

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

Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

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.)

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 $x \sim f(x|H)$.

x could represent e.g. observation of a single particle, a single event, or an entire "experiment".

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.,

$$P(x \in W \,|\, H_0) \leq \alpha$$

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

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

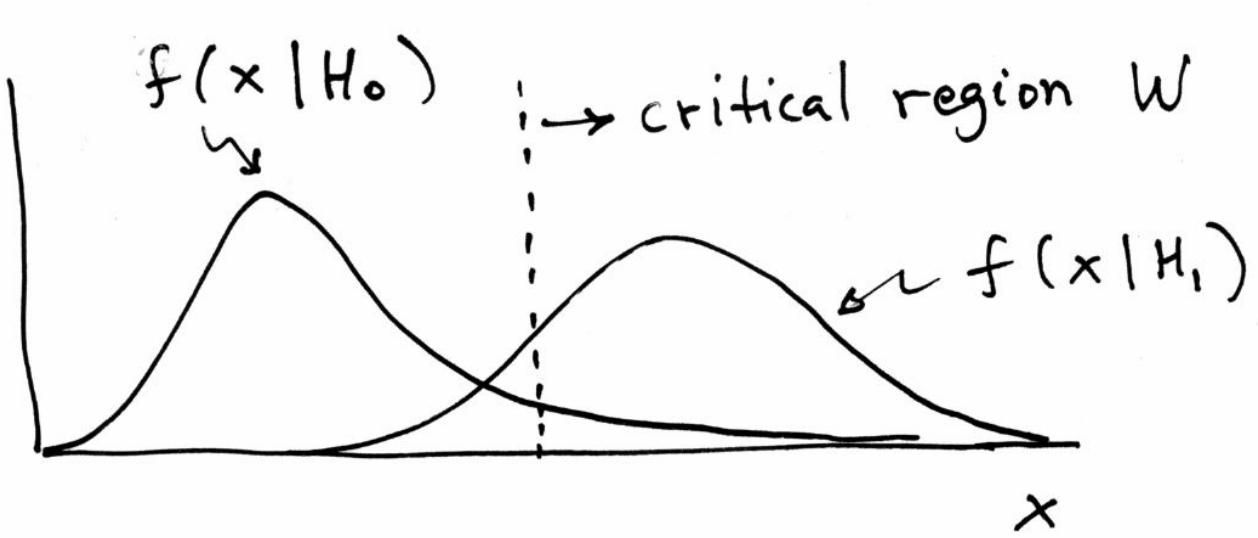
Critical region also called "rejection" region; complement is acceptance region.

Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



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 maximum probability for this is the size of the test:

$$P(x \in W \,|\, H_0) \leq \alpha$$

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

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

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

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

Physics context of a statistical test

Event Selection: the event types in question are both known to exist.

Example: separation of different particle types (electron vs muon) or known event types (ttbar vs QCD multijet).
Use the selected sample for further study.

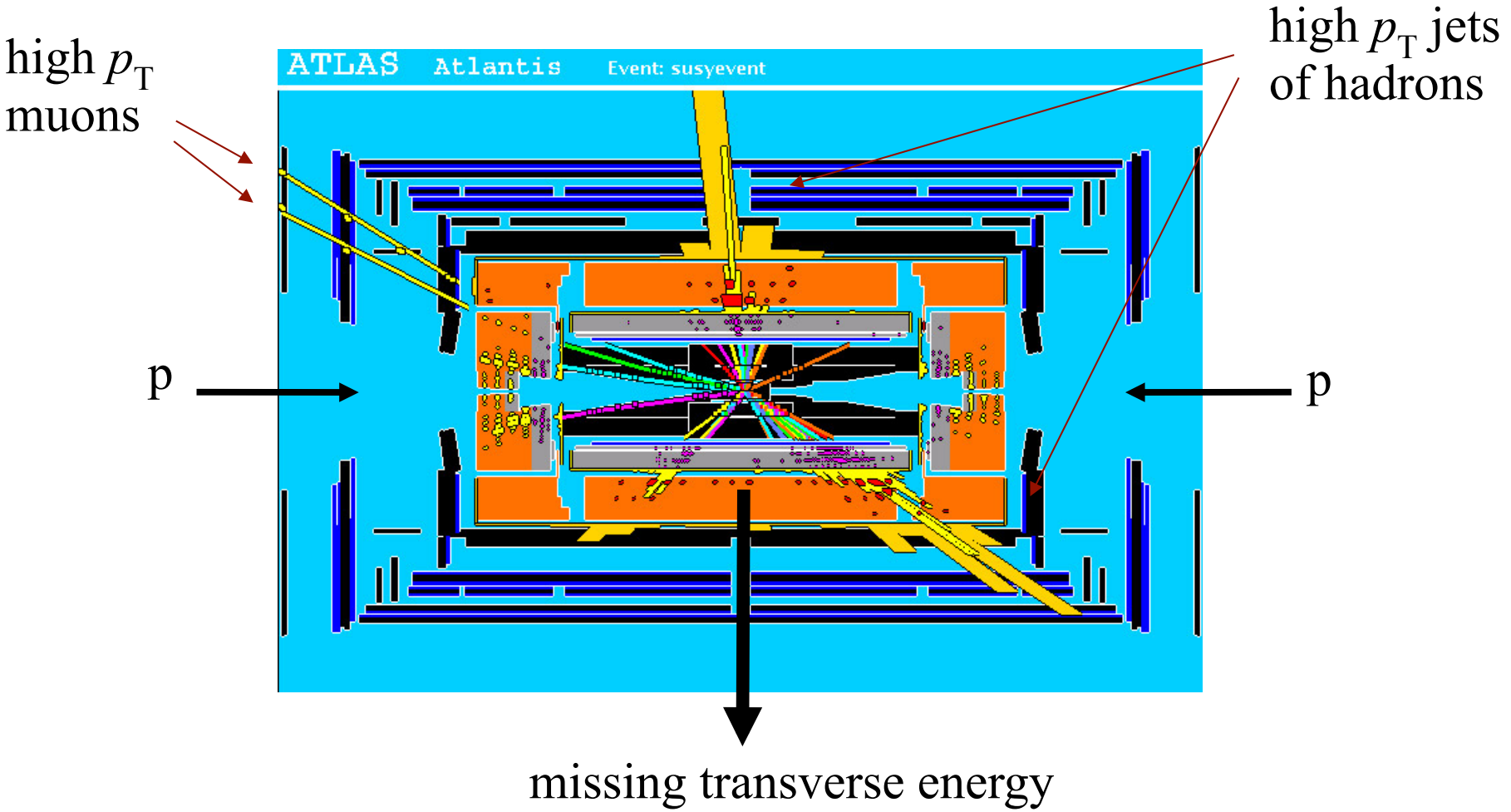
Search for New Physics: the null hypothesis H_0 means Standard Model events, and the alternative H_1 means "events of a type whose existence is not yet established" (to establish or exclude the signal model is the goal of the analysis).

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

The optimal statistical test for a search is closely related to that used for event selection.

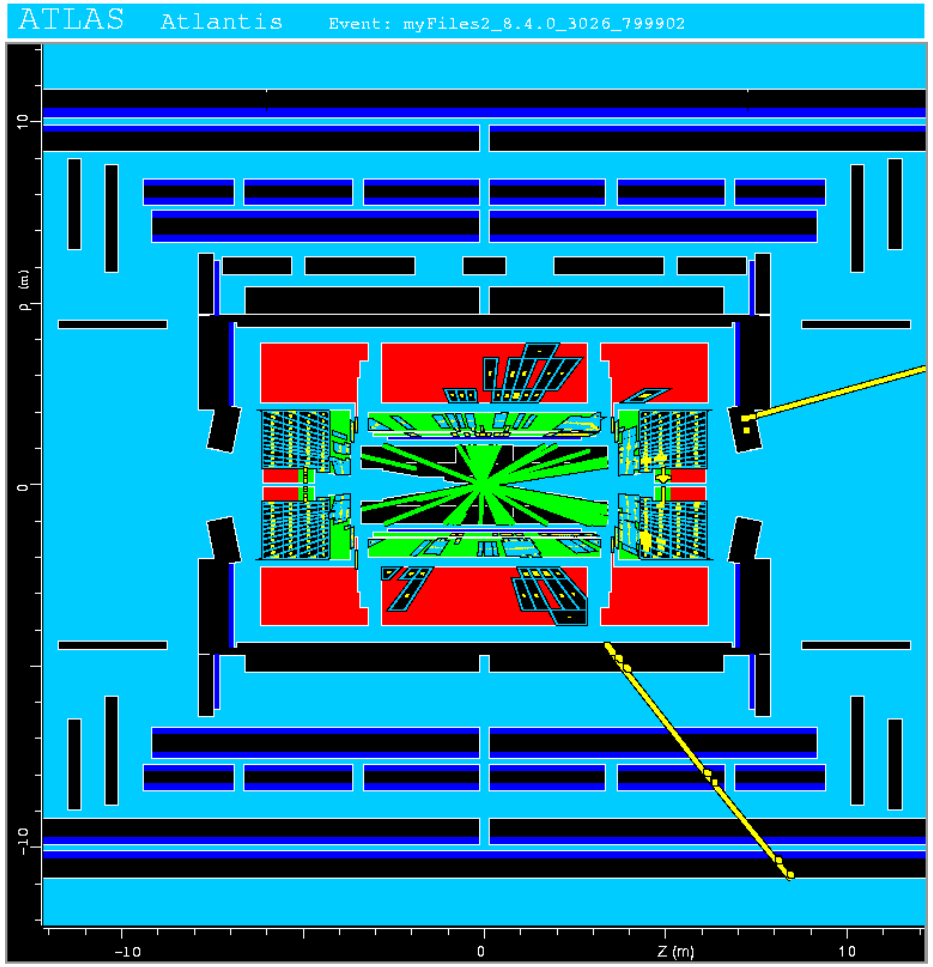
Suppose we want to discover this…

SUSY event (ATLAS simulation):



But we know we'll have lots of this…

ttbar event (ATLAS simulation)



SM event also has high p_T jets and muons, and missing transverse energy.

→ can easily mimic a SUSY event and thus constitutes a background.

Example of a multivariate statistical test

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

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

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.

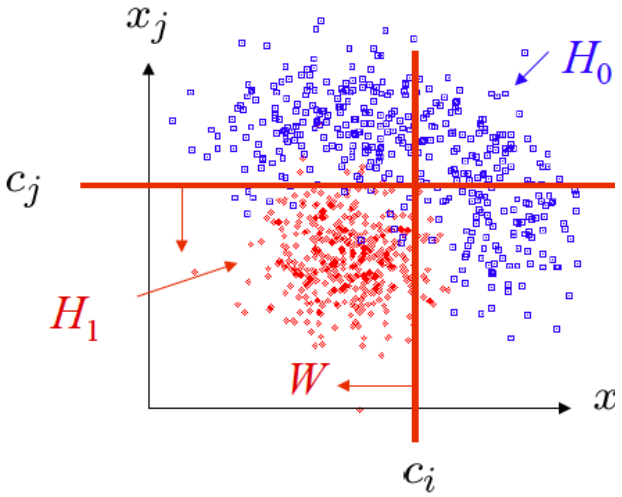
Often call H_0 the background hypothesis (e.g. SM events); H_1, H_2, ... are possible signal hypotheses.

Defining a multivariate critical region

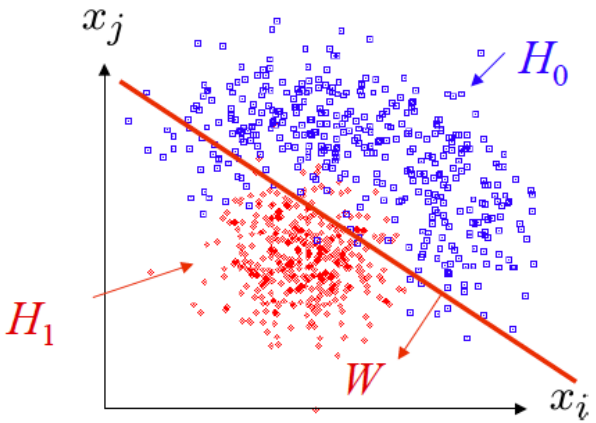
Each event is a point in $\boldsymbol{x}$-space; critical region is now defined by a 'decision boundary' in this space.

What kind of decision boundary is best?

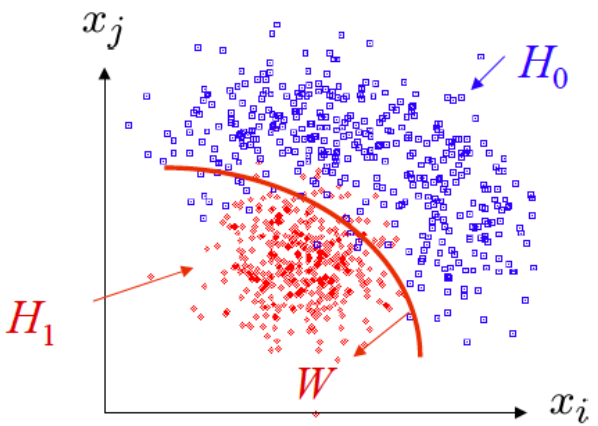
Cuts?



Linear?



Nonlinear?



Multivariate methods

Many new (and some old) methods for finding decision boundary:

- 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

For more see e.g. references at end of this lecture.

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

Test statistics

The decision boundary can be defined by an equation of the form

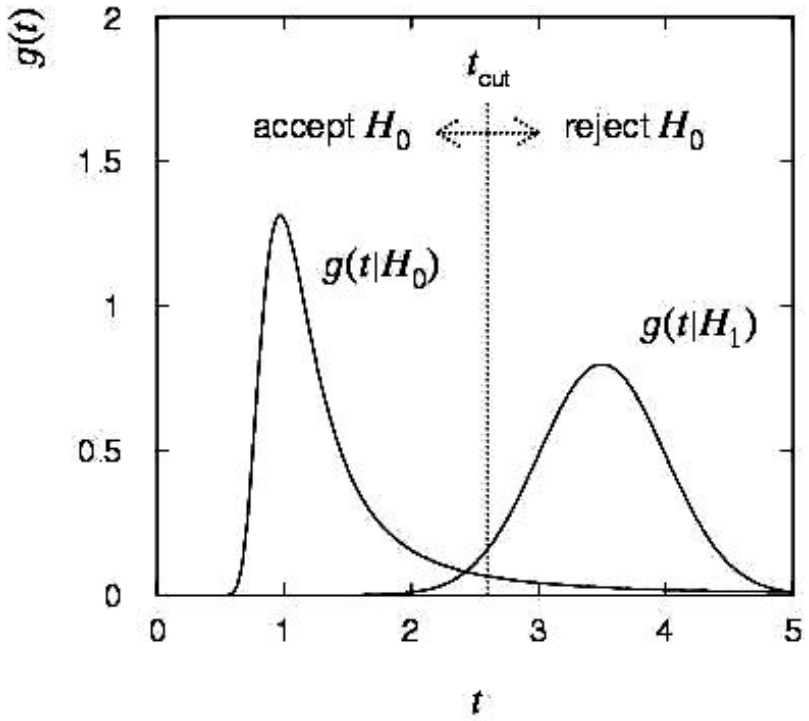
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

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

Decision boundary is now a single 'cut' on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Significance level and power

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

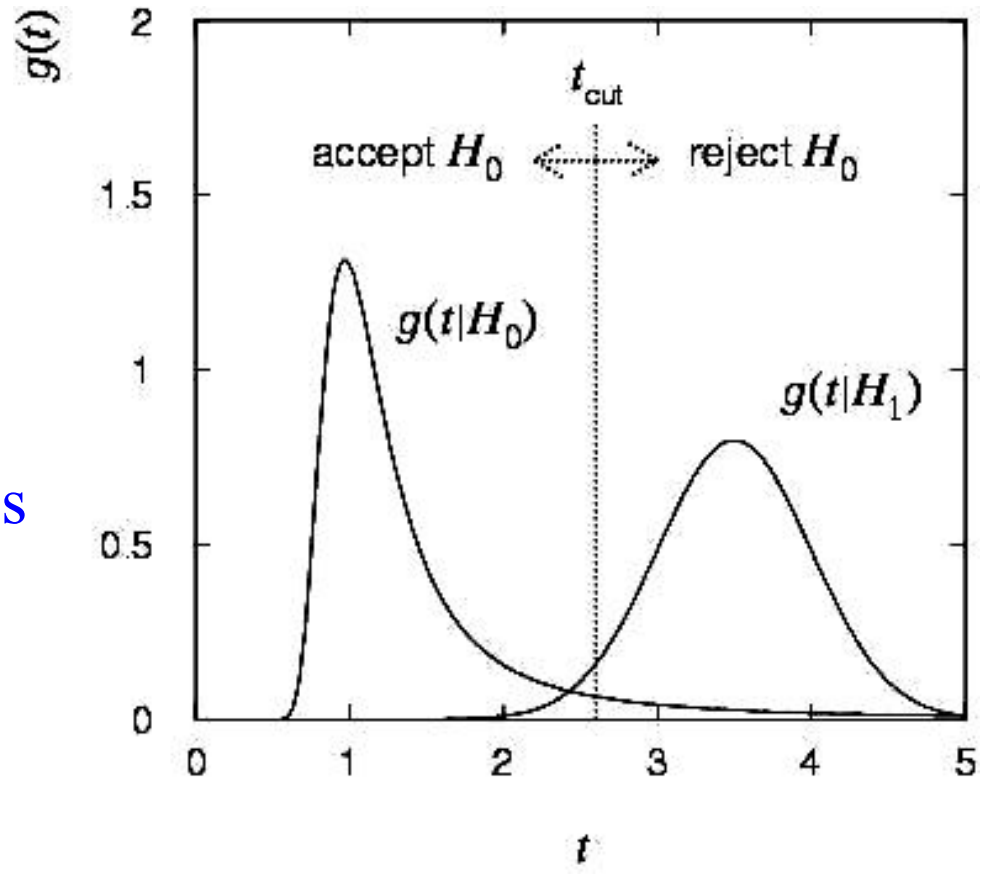
$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)\, dt$$

(significance level)

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

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

$(1 - \beta = \text{power})$



Constructing a test statistic

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test H_0, (background) versus H_1, (signal) (highest ε_s for a given ε_b) choose the critical (rejection) region such that

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

where c is a constant which determines the power.

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

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

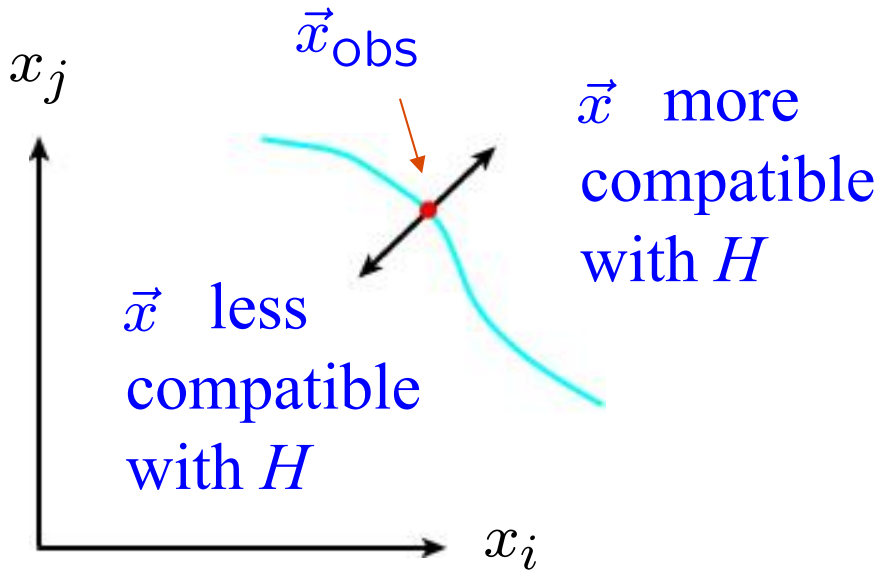
Testing significance / goodness-of-fit

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: $\vec{x}_{\text{obs}}$

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 $\vec{x}_{\text{obs}}$.
(Not unique!)



p-values

Express level of agreement between data and H with p-value:

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

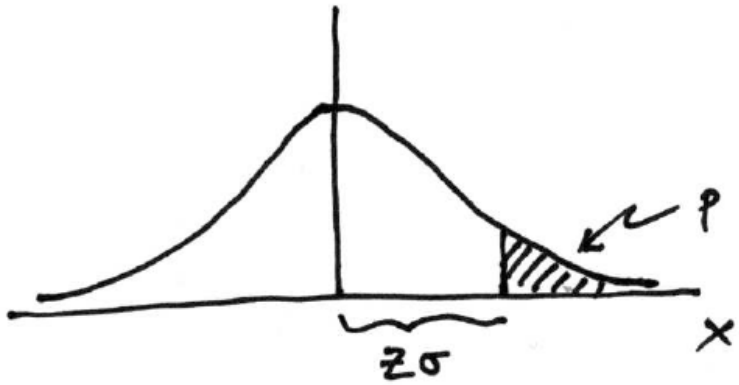
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

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 $P(H)$.

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 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 $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

Distribution of the p-value

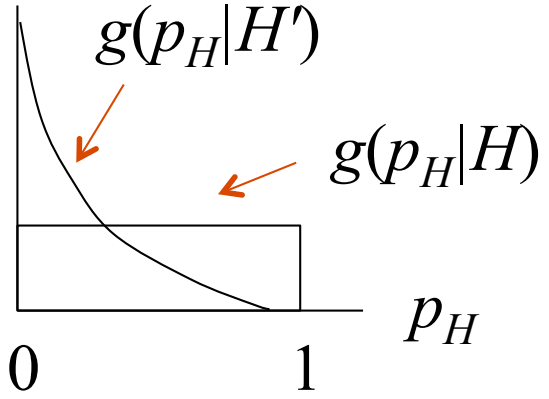
The p-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p-value of H is found from a test statistic $t(\boldsymbol{x})$ as

$$p_H = \int_t^\infty f(t'|H)dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \le p_H \le 1)$$

In general for continuous data, under assumption of H, $p_H \sim$ Uniform[0,1] and is concentrated toward zero for some (broad) class of alternatives.



Using a p-value to define test of H_0

So the probability to find the p-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We started by defining critical region in the original data space ($\boldsymbol{x}$), then reformulated this in terms of a scalar test statistic $t(\boldsymbol{x})$.

We can take this one step further and define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Interval estimation — introduction

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Desirable properties of such an interval may include:

communicate objectively the result of the experiment;
have a given probability of containing the true parameter;
provide information needed to draw conclusions about the parameter possibly incorporating stated prior beliefs.

Often use +/− the estimated standard deviation of the estimator.
In some cases, however, this is not adequate:

estimate near a physical boundary,
e.g., an observed event rate consistent with zero.

We will look briefly at Frequentist and Bayesian intervals.

Frequentist confidence intervals

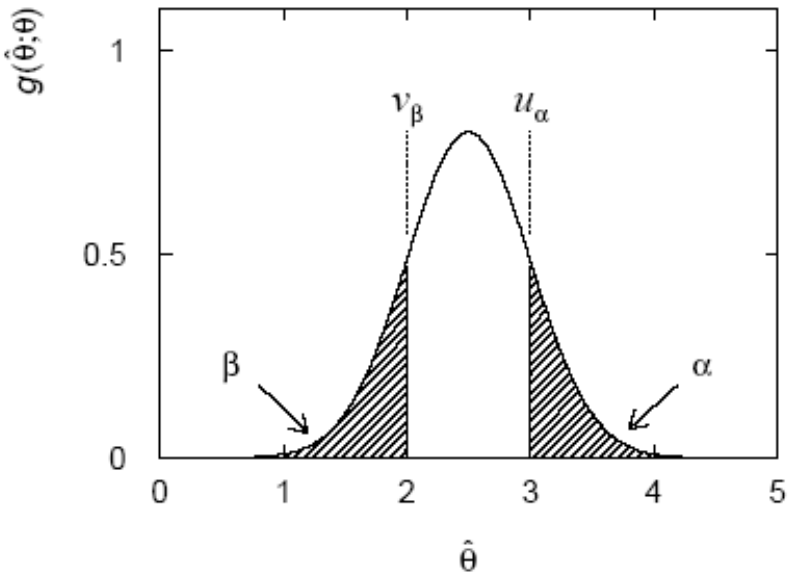
Consider an estimator $\hat{\theta}$ for a parameter θ and an estimate $\hat{\theta}_{\text{obs}}$.

We also need for all possible θ its sampling distribution $g(\hat{\theta}; \theta)$.

Specify upper and lower tail probabilities, e.g., $\alpha = 0.05$, $\beta = 0.05$, then find functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ such that:

$$\begin{aligned} \alpha &= P(\hat{\theta} \geq u_\alpha(\theta)) \\ &= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta)\, d\hat{\theta} \end{aligned}$$

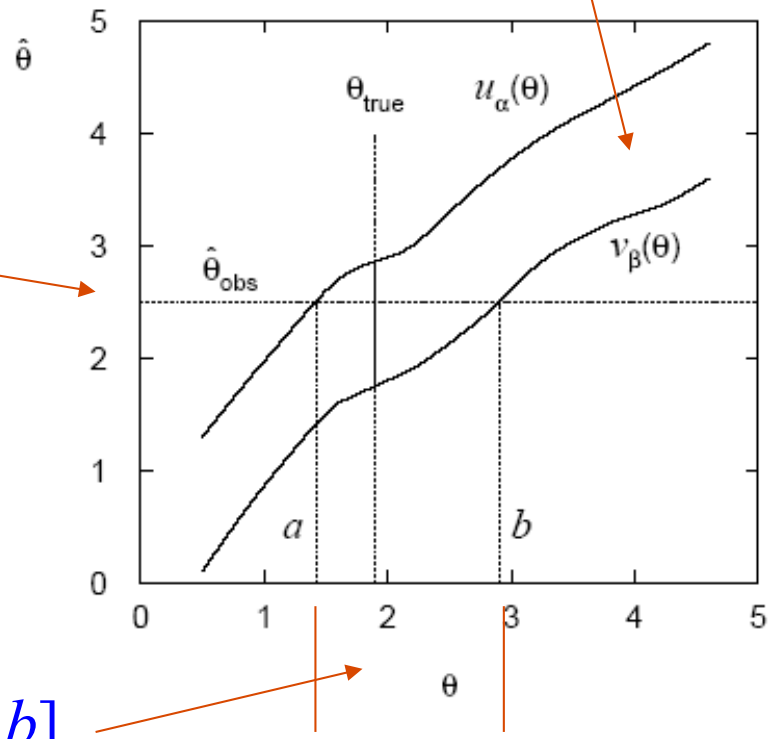
$$\begin{aligned} \beta &= P(\hat{\theta} \leq v_\beta(\theta)) \\ &= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta)\, d\hat{\theta} \end{aligned}$$



Confidence interval from the confidence belt

The region between $u_\alpha(\theta)$ and $v_\beta(\theta)$ is called the confidence belt.

Find points where observed estimate intersects the confidence belt.



This gives the confidence interval $[a, b]$

Confidence level = $1 - \alpha - \beta$ = probability for the interval to cover true value of the parameter (holds for any possible true θ).

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of θ with probability $\geq 1 - \gamma$.

Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ..

If $p_\theta < \gamma$, then we reject θ.

The confidence interval at CL = $1 - \gamma$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta \geq \gamma$.

In practice find by setting $p_\theta = \gamma$ and solve for θ.

Meaning of a confidence interval

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Test statistic for p-value of a parameter

One often obtains the p-value of a hypothesized value of a parameter θ using a test statistic $q_\theta(\boldsymbol{x})$, such that large values of q_θ correspond to increasing incompatibility between the data ($\boldsymbol{x}$) and hypothesis (θ).

The data result in a value $q_{\theta,\mathrm{obs}}$.

The p-value of the hypothesized θ is therefore

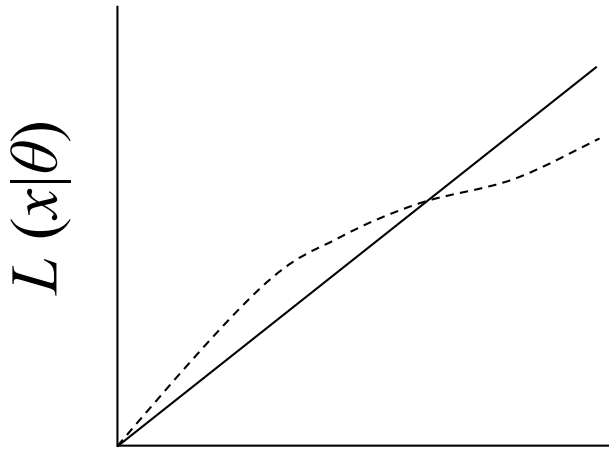
$$p_\theta = \int_{q_{\theta,\mathrm{obs}}}^{\infty} f(q_\theta|\theta)\, dq_\theta$$

So to find this we need to know the distribution of q_θ under assumption of θ.

For some problems we can write this down in closed form (at least approximately); other times need Monte Carlo.

Nuisance parameters

In general our model of the data is not perfect:



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \to L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Distribution of q_θ in case of nuisance parameters

The p-value of θ is now $\quad p_\theta = \int_{q_{\theta,\mathrm{obs}}}^{\infty} f(q_\theta|\theta,\nu)\, dq_\theta$

But what values of ν to use for $f(q_\theta|\theta, \nu)$?

Fundamentally we want to reject θ only if $p_\theta < \alpha$ for all ν.

→ "exact" confidence interval

We will see that for certain statistics (based on the profile likelihood ratio), the distribution $f(q_\theta|\theta, \nu)$ becomes independent of the nuisance parameters in the large-sample limit.

But in general for finite data samples this is not true; may be unable to reject some θ values if ν is assumed equal to some value that is strongly disfavoured by the data (resulting interval for θ "overcovers").

Profile construction ("hybrid resampling")

K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Compromise procedure is to reject θ if $p_\theta < \alpha$ where the p-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ:

$\hat{\hat{\nu}}(\theta)$

"double hat" notation means value of parameter that maximizes likelihood for the given θ.

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{\nu}}(\theta))$.

Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial).

"Hybrid frequentist-Bayesian" method

Alternatively, suppose uncertainty in ν is characterized by a Bayesian prior $\pi(\nu)$.

Can use the marginal likelihood to model the data:

$$L_{\rm m}(x|\theta) = \int L(x|\theta,\nu)\pi(\nu)\,d\nu$$

This does not represent what the data distribution would be if we "really" repeated the experiment, since then ν would not change.

But the procedure has the desired effect. The marginal likelihood effectively builds the uncertainty due to ν into the model.

Use this now to compute (frequentist) p-values → result has hybrid "frequentist-Bayesian" character.

The "ur-prior" behind the hybrid method

But where did $\pi(\nu)$ come frome? Presumably at some earlier point there was a measurement of some data y with likelihood $L(y|\nu)$, which was used in Bayes' theorem,

$$\pi(\nu|y) \propto L(y|\nu)\pi_0(\nu)$$

and this "posterior" was subsequently used for $\pi(\nu)$ for the next part of the analysis.

But it depends on an "ur-prior" $\pi_0(\nu)$, which still has to be chosen somehow (perhaps "flat-ish").

But once this is combined to form the marginal likelihood, the origin of the knowledge of ν may be forgotten, and the model is regarded as only describing the data outcome x.

The (pure) frequentist equivalent

In a purely frequentist analysis, one would regard both x and y as part of the data, and write down the full likelihood:

$$L(x, y|\theta, \nu) = L(x|\theta, \nu)L(y|\nu)$$

"Repetition of the experiment" here means generating both x and y according to the distribution above.

In many cases, the end result from the hybrid and pure frequentist methods are found to be very similar (cf. Conway, Roever, PHYSTAT 2011).

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.

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.

Confidence intervals

Set of parameter values not rejected in a test of size α gives confidence interval at $1 - \alpha$ CL.

Systematic uncertainties ↔ nuisance parameters

Extra slides

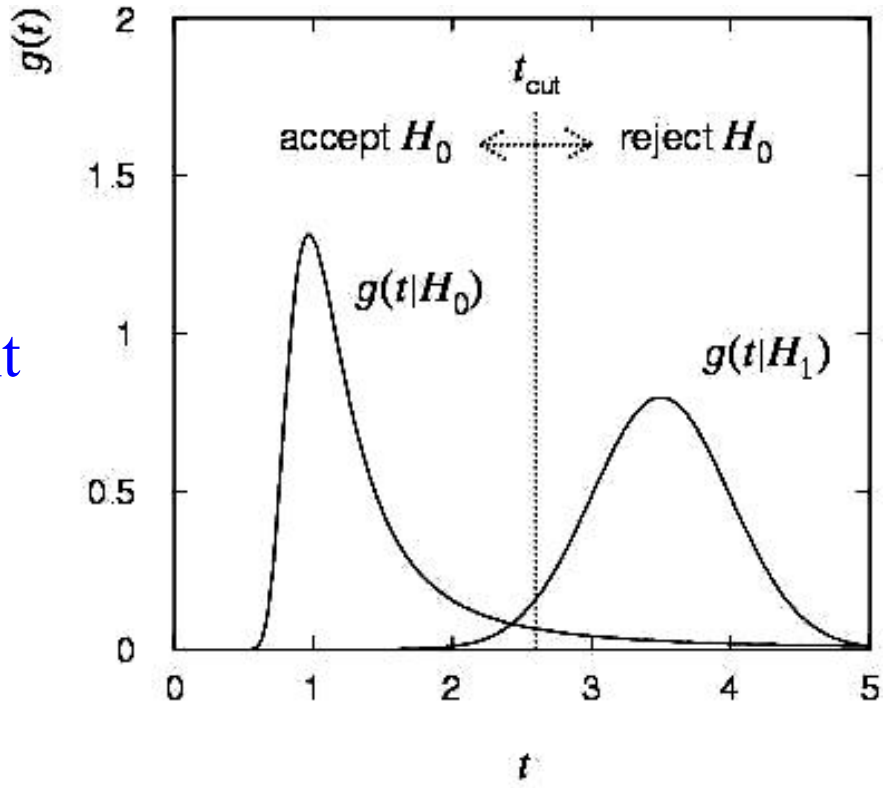
Signal/background efficiency

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

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

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

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



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 signal events with $t > t_{\rm cut}$. 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:

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

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

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

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 $P(x|H_1) / P(x|H_0)$ 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.

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.

Quick review of parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

random variable parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

$\widehat{\theta}(\vec{x})$ ← estimator written with a hat

Sometimes we say 'estimator' for the function of x_1, ..., x_n; 'estimate' for the value of the estimator with a particular data set.

The likelihood function

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$, and suppose the joint pdf for the data $\boldsymbol{x}$ is a function that depends on a set of parameters $\boldsymbol{\theta}$:

$$f(\vec{x};\vec{\theta})$$

Now evaluate this function with the data obtained and regard it as a function of the parameter(s). This is the likelihood function:

$$L(\vec{\theta}) = f(\vec{x};\vec{\theta})$$

($\boldsymbol{x}$ constant)

The likelihood function for i.i.d.*. data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: x_1, ..., x_n, where x follows $f(x;\theta)$. The joint pdf for the whole data sample is:

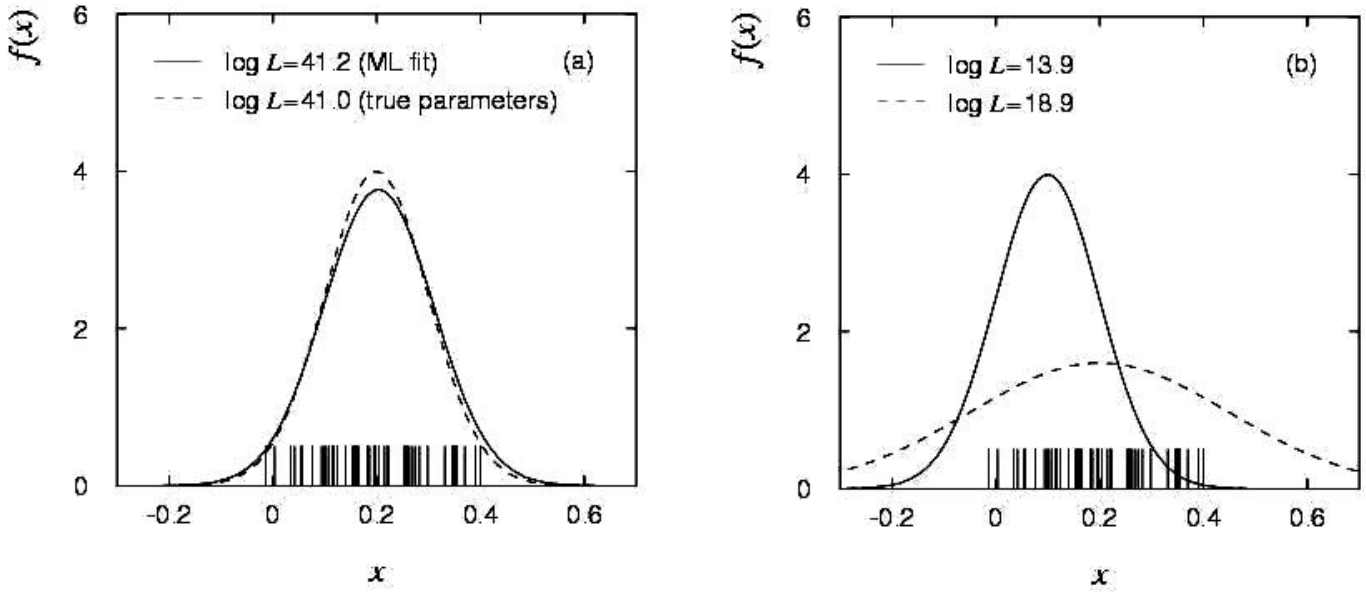
$$f(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i;\vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

ML estimators not guaranteed to have any 'optimal' properties, (but in practice they're very good).

ML example: parameter of exponential pdf

Consider exponential pdf, $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

ML example: parameter of exponential pdf (2)

Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

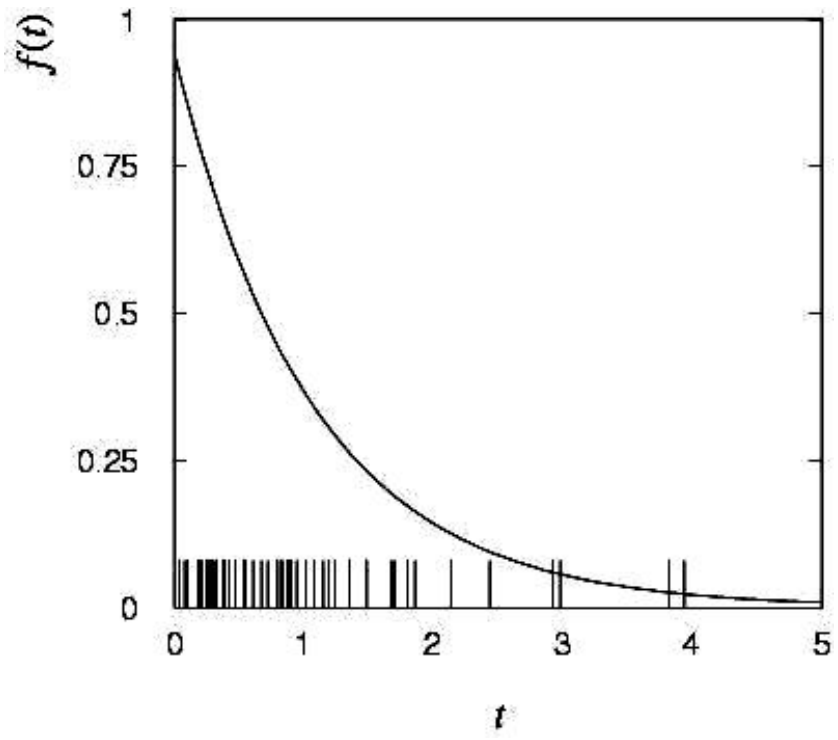
$$\rightarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values
using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$



Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \qquad (b = E[\hat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\hat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

Information inequality for n parameters

Suppose we have estimated n parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -n\int f(x;\vec{\theta})\frac{\partial^2 \ln f(x;\vec{\theta})}{\partial\theta_i \partial\theta_j}\,dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

Extended ML

Sometimes regard n not as fixed, but as a Poisson r.v., mean ν.

Result of experiment defined as: n, x_1, ..., x_n.

The (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

Suppose theory gives $\nu = \nu(\vec{\theta})$, then the log-likelihood is

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

where C represents terms not depending on $\vec{\theta}$.

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L\, dt$

where the total cross section $\sigma(\boldsymbol{\theta})$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x.

Extended ML uses more info → smaller errors for $\hat{\vec{\theta}}$

Important e.g. for anomalous couplings in $e^+e^- \to W^+W^-$

If $\boldsymbol{\nu}$ does not depend on $\boldsymbol{\theta}$ but remains a free parameter, extended ML gives:

$$\hat{\nu} = n$$

$$\hat{\theta} = \text{same as ML}$$

Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable x: $f_s(x)$ and $f_b(x)$.

We observe a mixture of the two event types, signal fraction = θ, expected total number = ν, observed total number = n.

Let $\mu_s = \theta\nu$, $\mu_b = (1-\theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x;\mu_s,\mu_b) = \frac{\mu_s}{\mu_s+\mu_b} f_s(x) + \frac{\mu_b}{\mu_s+\mu_b} f_b(x)$$

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

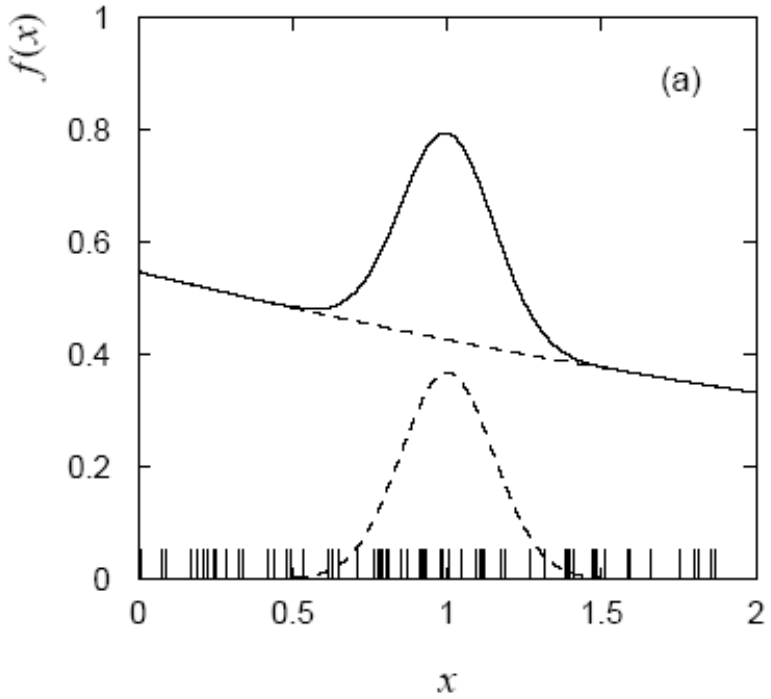
$$\rightarrow \quad \ln L(\mu_s,\mu_b) = -(\mu_s+\mu_b) + \sum_{i=1}^{n} \ln\left[(\mu_s+\mu_b) f(x_i;\mu_s,\mu_b)\right]$$

Extended ML example (2)

Monte Carlo example with combination of exponential and Gaussian:

$$\mu_s = 6$$

$$\mu_b = 60$$



Maximize log-likelihood in terms of μ_s and μ_b:

$$\hat{\mu}_s = 8.7 \pm 5.5$$

$$\hat{\mu}_b = 54.3 \pm 8.8$$

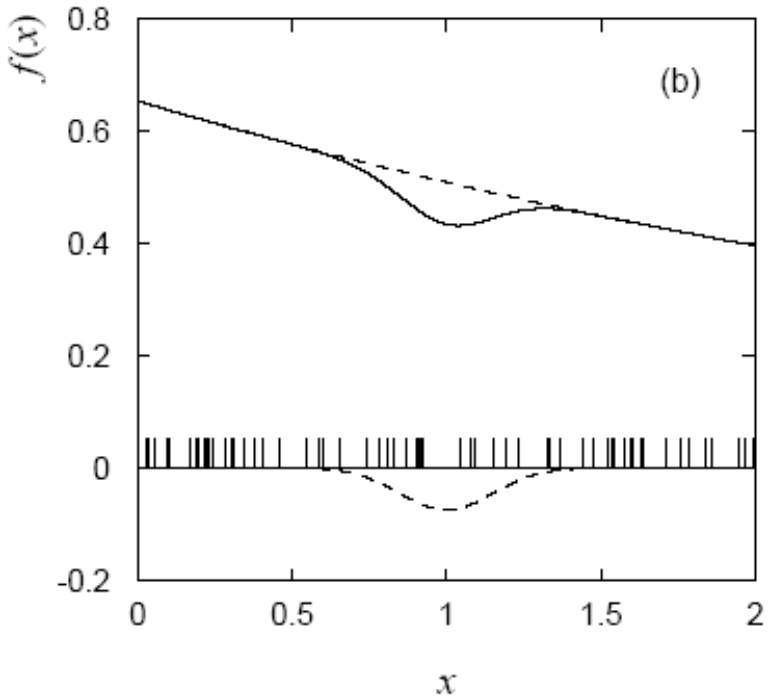
Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.

Extended ML example: an unphysical estimate

A downwards fluctuation of data in the peak region can lead to even fewer events than what would be obtained from background alone.

Estimate for μ_s here pushed negative (unphysical).

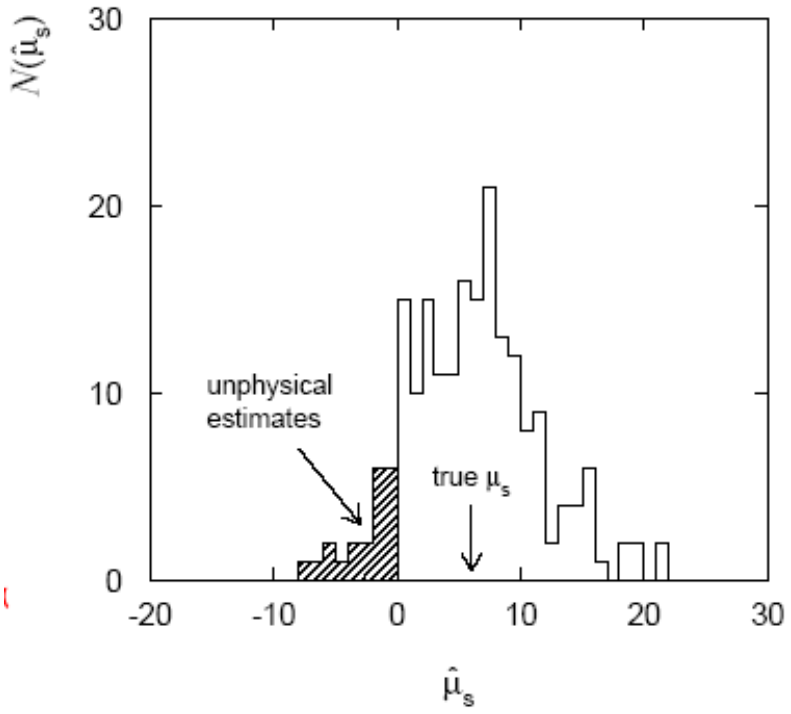
We can let this happen as long as the (total) pdf stays positive everywhere.



Unphysical estimators (2)

Here the unphysical estimator is unbiased and should nevertheless be reported, since average of a large number of unbiased estimates converges to the true value (cf. PDG).

Repeat entire MC experiment many times, allow unphysical estimates:



Resources on multivariate methods

Books:

C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001

R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., Wiley, 2001

A. Webb, *Statistical Pattern Recognition*, 2nd ed., Wiley, 2002

Materials from some recent meetings:

PHYSTAT conference series (2002, 2003, 2005, 2007,...) see
www.phystat.org

Caltech workshop on multivariate analysis, 11 February, 2008
indico.cern.ch/conferenceDisplay.py?confId=27385

SLAC Lectures on Machine Learning by Ilya Narsky (2006)
www-group.slac.stanford.edu/sluo/Lectures/Stat2006_Lectures.html

Software for multivariate analysis

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

From **tmva.sourceforge.net**, also distributed with ROOT
Variety of classifiers
Good manual

StatPatternRecognition, I. Narsky, physics/0507143

Further info from **www.hep.caltech.edu/~narsky/spr.html**
Also wide variety of methods, many complementary to **TMVA**
Currently appears project no longer to be supported