Statistics for Rare Event Searches
Lecture 1

Hands-on PhD School on
Astroparticle Physics
Gran Sasso, Italy
11 September 2025

https://agenda.infn.it/event/44808/

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

Outline

→ Lecture 1: Introduction
Probability
Hypothesis tests
Parameter estimation
Confidence limits

Lecture 2: Systematic uncertainties
Prototype analysis
Experimental sensitivity

Almost everything is a subset of the University of London course:

http://www.pp.rhul.ac.uk/~cowan/stat_course.html

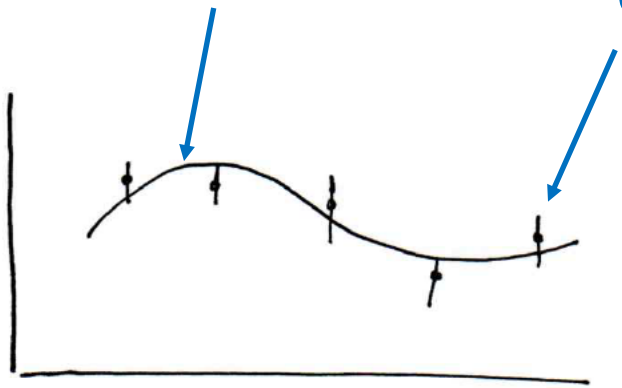
Theory ↔ Statistics ↔ Experiment

Theory (model, hypothesis):

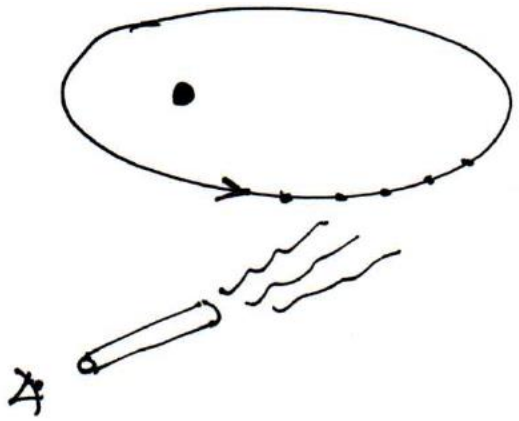
$$F = -G\frac{m_1 m_2}{r^2}, \quad \ldots$$

+ response of measurement apparatus

= model prediction



Experiment (observation):



Uncertainty enters on many levels

→ quantify with probability

A quick review of probability

Frequentist (A = outcome of repeatable observation)

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

Subjective (A = hypothesis)

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

Conditional probability:

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

E.g. rolling a die, outcome n = 1,2,…,6:

$$P(n \leq 3 | n \text{ even}) = \frac{P((n \leq 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

A and B are independent iff:

$$P(A \cap B) = P(A)P(B)$$

I.e. if A, B independent, then

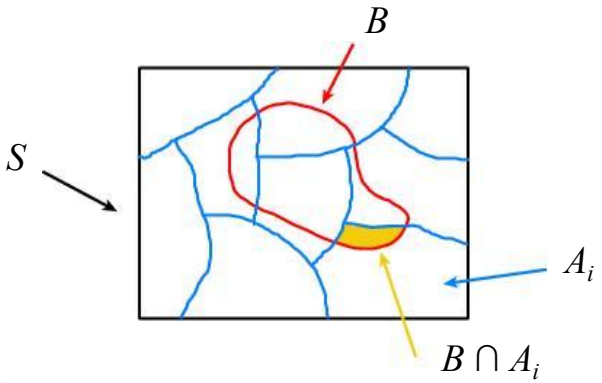
$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes' theorem

Use definition of conditional probability and $P(A \cap B) = P(B \cap A)$

$$\rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{(Bayes' theorem)}$$

If set of all outcomes $S = \cup_i A_i$ with A_i disjoint, then law of total probability for $P(B)$ says



$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

so that Bayes' theorem becomes $$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Bayes' theorem holds regardless of how probability is interpreted (frequency, degree of belief...).

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: $\boldsymbol{x}$).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P ($0.117 < \alpha_s < 0.119$),
P (Whitmer wins in 2028),

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.

Preferred theories (models, hypotheses, ...) are those that predict a high probability for data "like" the data observed.

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability 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

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$$P(\mathbf{x}|H) \quad = \quad \text{the likelihood of } H$$

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \qquad = \quad \text{the "likelihood function"}$$

Note:

1) For the likelihood we treat the data $\boldsymbol{x}$ as fixed.
2) The likelihood function $L(\boldsymbol{\theta})$ is not a pdf for $\boldsymbol{\theta}$.

Frequentist hypothesis tests

Suppose a measurement produces data $\boldsymbol{x}$; consider a hypothesis H_0 we want to test and alternative H_1

H_0, H_1 specify probability for $\boldsymbol{x}$: $P(\boldsymbol{x}|H_0), P(\boldsymbol{x}|H_1)$

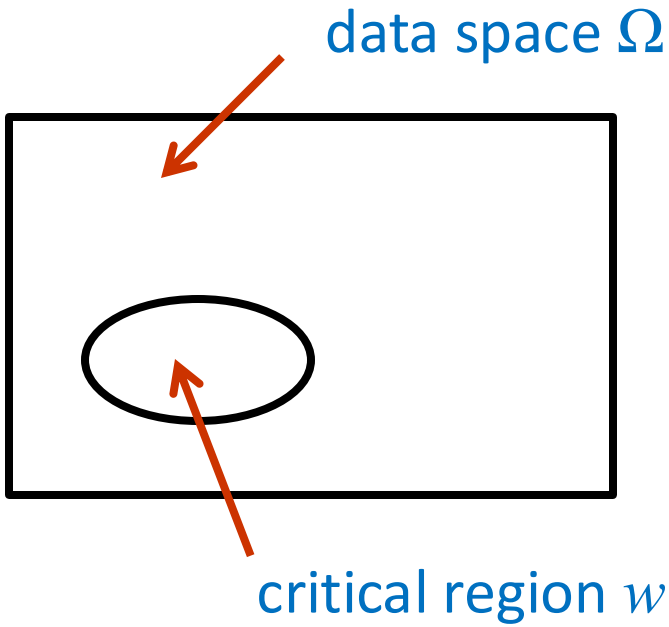
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(\boldsymbol{x} \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

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

If $\boldsymbol{x}$ is observed in the critical region, reject H_0.

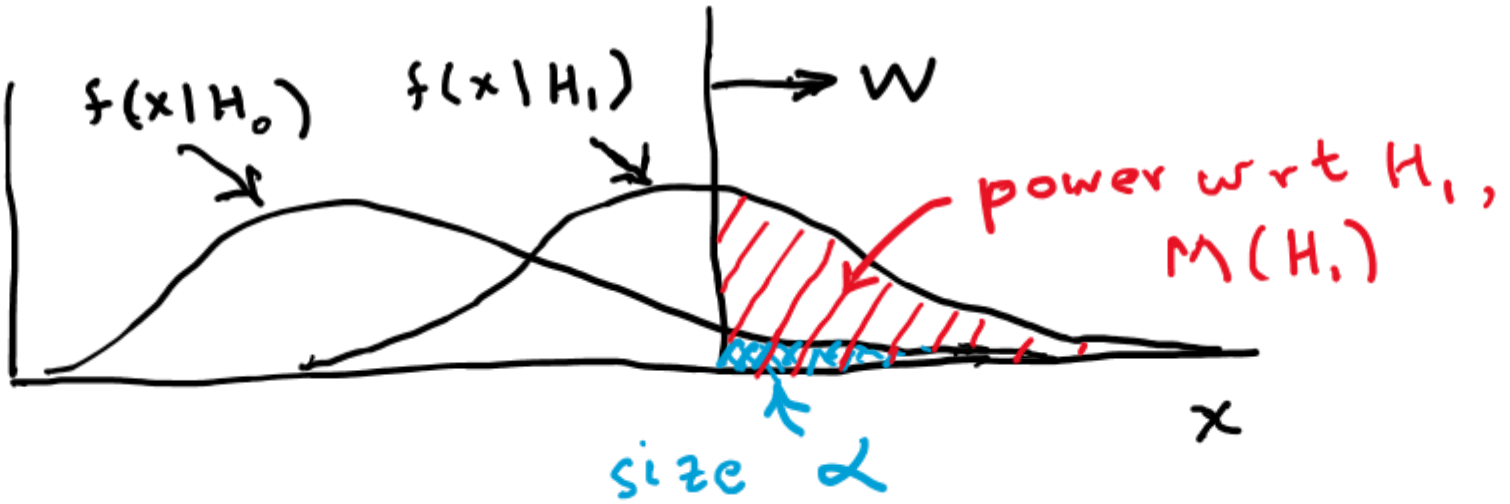


Definition of a test (2)

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

Use the alternative hypothesis H_1 to motivate where to place the critical region.

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



Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., H_0 = b.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the "true class label", which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as "candidate events of type s". Equivalent Particle Physics terminology:

background efficiency $\varepsilon_{\mathrm{b}} = \int_W f(\mathbf{x}|H_0)\, d\mathbf{x} = \alpha$

signal efficiency $\varepsilon_{\mathrm{s}} = \int_W f(\mathbf{x}|H_1)\, d\mathbf{x} = 1 - \beta = \mathrm{power}$

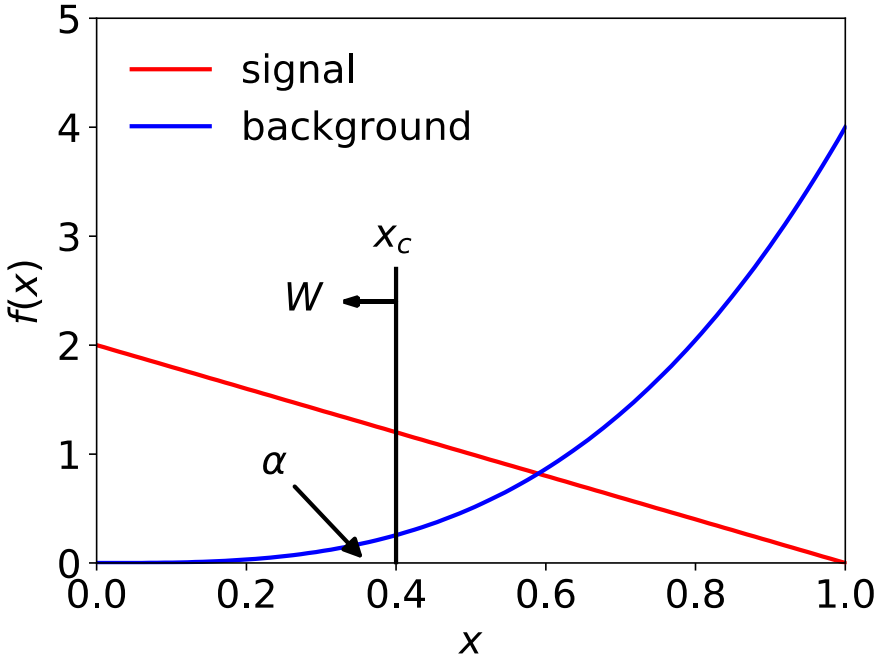
Example of a test for classification

Suppose we can measure for each event a quantity x, where

$$f(x|s) = 2(1-x)$$

$$f(x|b) = 4x^3$$

with $0 \le x \le 1$.



For each event in a mixture of signal (s) and background (b) test

H_0 : event is of type b

using a critical region W of the form: $W = \{x : x \le x_c\}$, where x_c is a constant that we choose to give a test with the desired size α.

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \leq x_c|\mathrm{b}) = \int_0^{x_c} f(x|b)\,dx = \left.\frac{4x^4}{4}\right|_0^{x_c} = x_c^4$$

and therefore $x_c = \alpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \leq x_c|\mathrm{s}) = \int_0^{x_c} f(x|\mathrm{s})\,dx = 2x_c - x_c^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

$$\pi_{\mathrm{s}} = 0.001$$
$$\pi_{\mathrm{b}} = 0.999$$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathrm{s}|x \leq x_{\mathrm{c}}) = \frac{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}}}{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}} + P(x \leq x_{\mathrm{c}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= 0.655$$

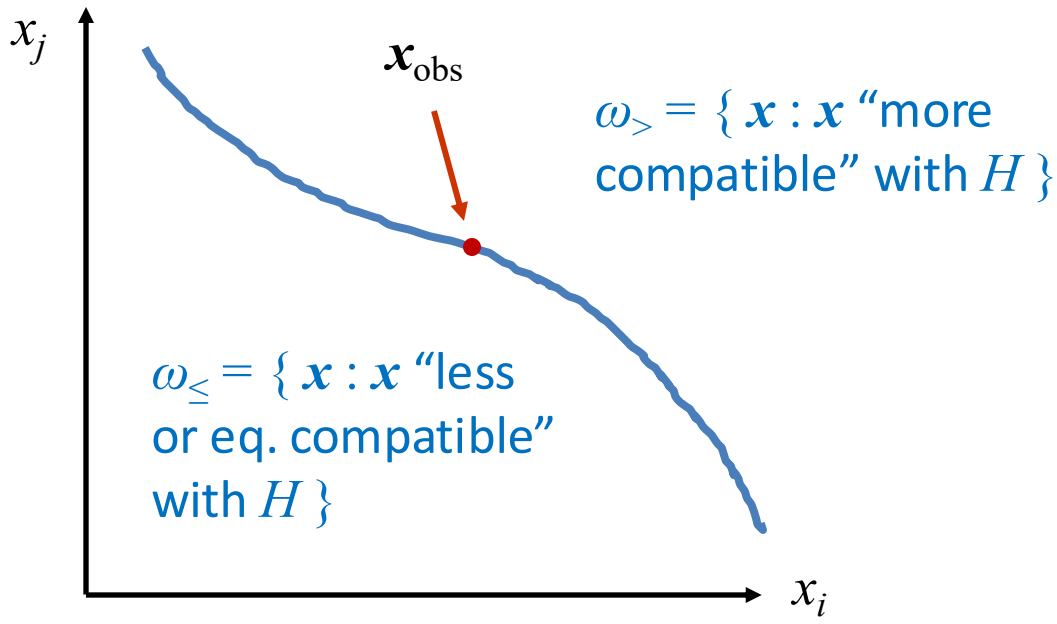
Testing significance / goodness-of-fit

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

We observe a single point in this space: $\boldsymbol{x}_{\text{obs}}$.

How can we quantify the level of compatibility between the data and the predictions of H?

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



p-values

Express level of compatibility between data and hypothesis (sometimes 'goodness-of-fit') by giving the p-value for H:

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\text{obs}})|H)$$

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

= probability, under assumption of H, to observe data as discrepant with H as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then H is "disfavoured by the data".

If the p-value is below a user-defined threshold α (e.g. 0.05) then H is rejected (equivalent to hypothesis test of size α as seen earlier).

p-value of H is not $P(H)$

The p-value of H is not the probability that H is true!

In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation).

If we do define $P(H)$, e.g., in Bayesian statistics as a degree of belief, then we need to 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 $P(H)$.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

Poisson counting experiment: discovery p-value

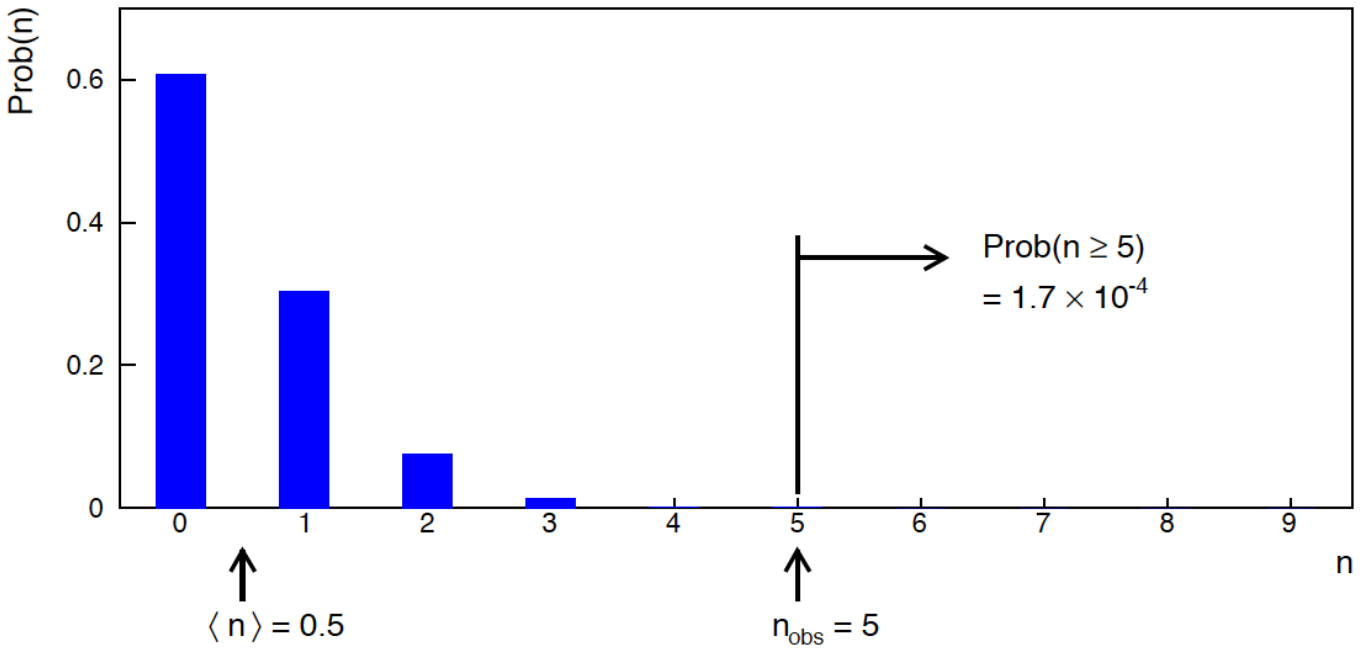
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$, suppose relevant alt. is $s > 0$.

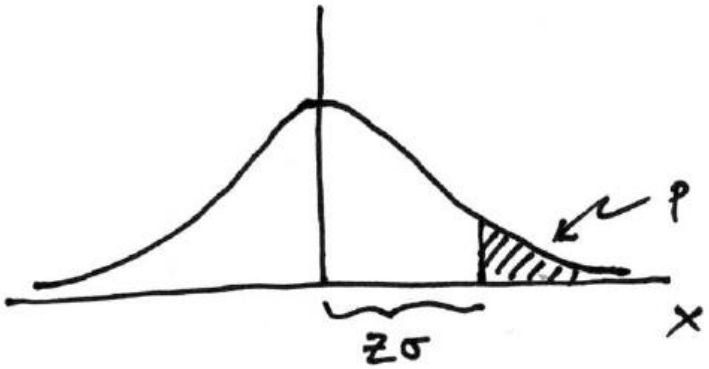
$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

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



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.



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

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

```
in ROOT:
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

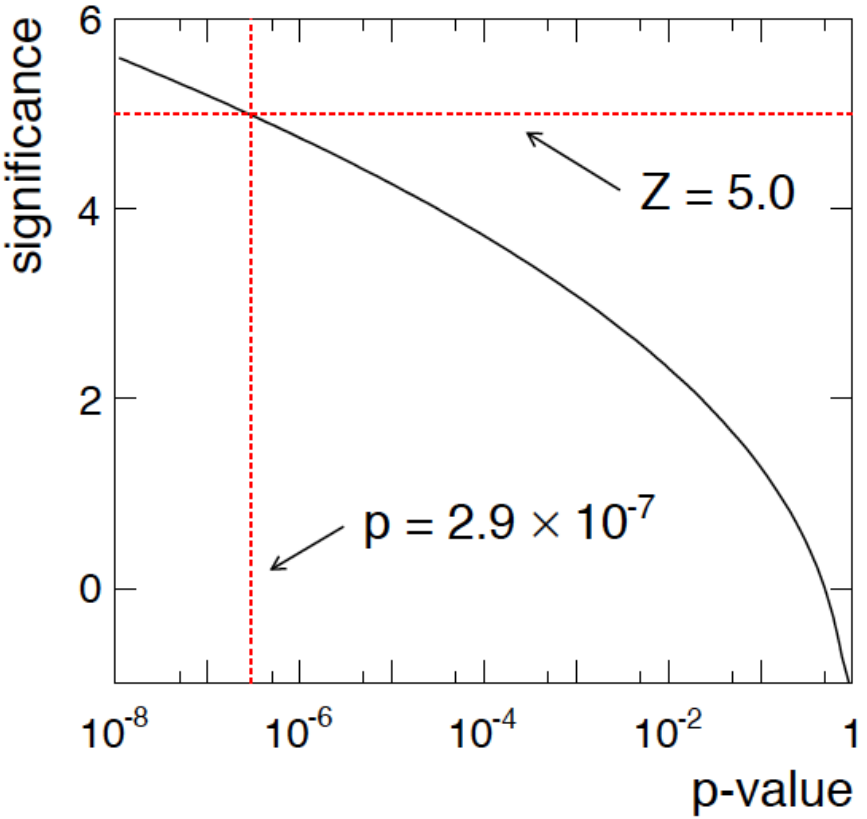
```
in python (scipy.stats):
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

Result Z is a "number of sigmas". Note this does not mean that the original data was Gaussian distributed.

Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1 - p) = 3.6$

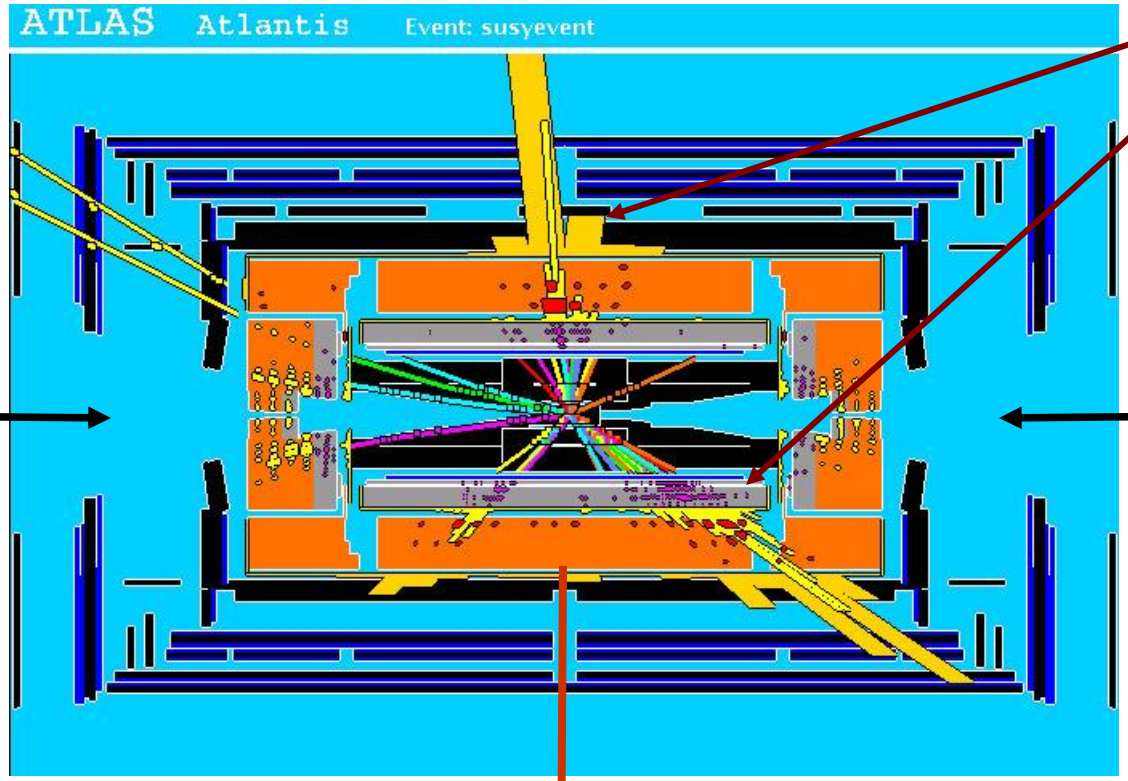
Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: p-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Particle Physics context for a hypothesis test

A simulated SUSY event ("signal"):



high p_T muons

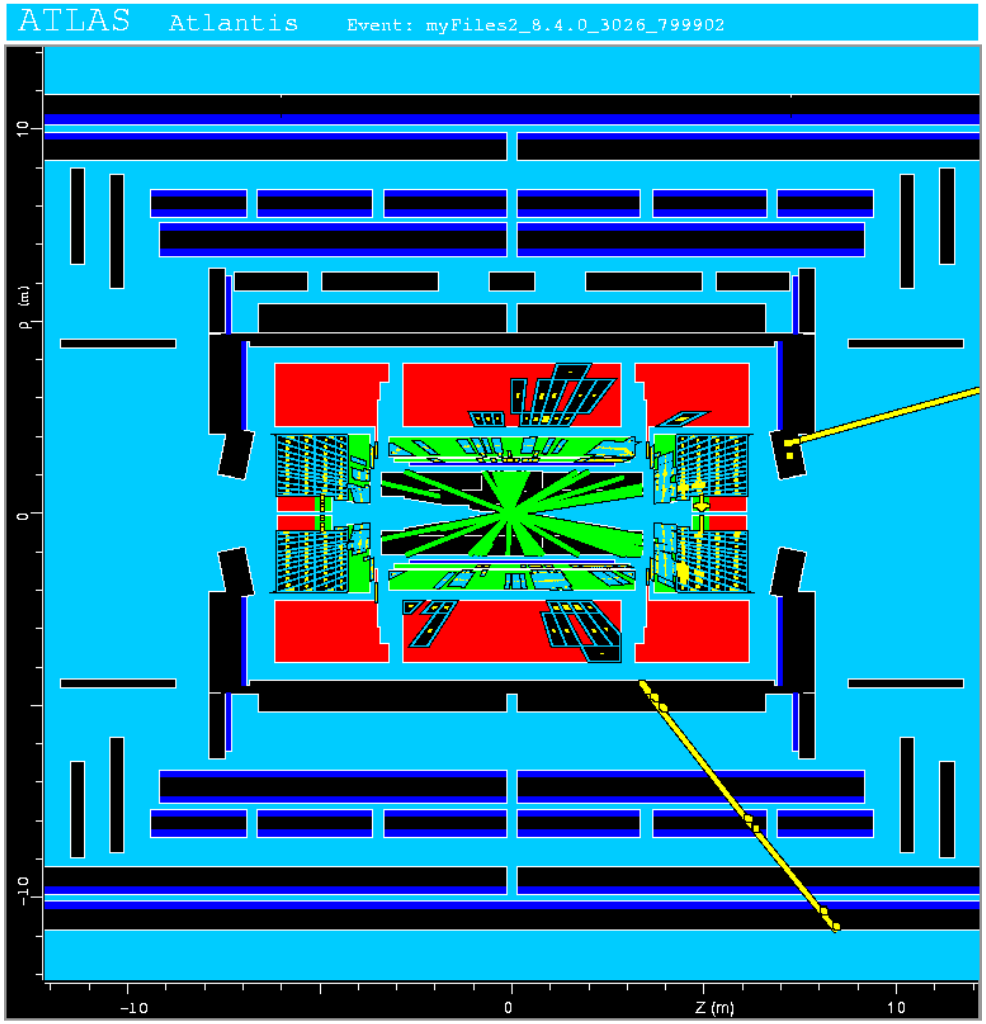
high p_T jets of hadrons

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 signal event.

Classification of measured events

Measured events can be considered to come in two classes:

signal (the kind of event we're looking for, $y = 1$)
background (the kind that mimics signal, $y = 0$)

For each event we measure a collection of features:

x_1 = energy of muon
x_2 = angle between jets
x_3 = total jet energy
x_4 = missing transverse energy
x_5 = invariant mass of muon pair
x_6 = ...

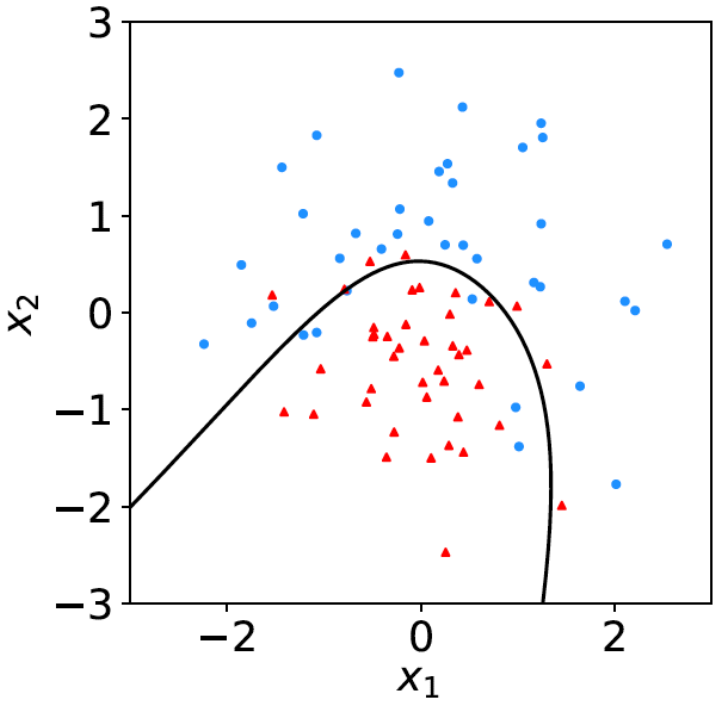
The real events don't come with true class labels, but computer-simulated events do. So we can have a set of simulated events that consist of a feature vector $\boldsymbol{x}$ and true class label y (0 for background, 1 for signal):

$$(\boldsymbol{x}, y)_1, (\boldsymbol{x}, y)_2, \ldots, (\boldsymbol{x}, y)_N$$

The simulated events are called "training data".

Distributions of the features

If we consider only two features $\boldsymbol{x} = (x_1, x_2)$, we can display the results in a scatter plot (red: $y = 0$, blue: $y = 1$).



For real events, the dots are black (true type is not known).

For each real event test the hypothesis that it is background.

(Related to this: test that a sample of events is *all* background.)

The test's critical region is defined by a "decision boundary" – without knowing the event type, we can classify them by seeing where their measured features lie relative to the boundary.

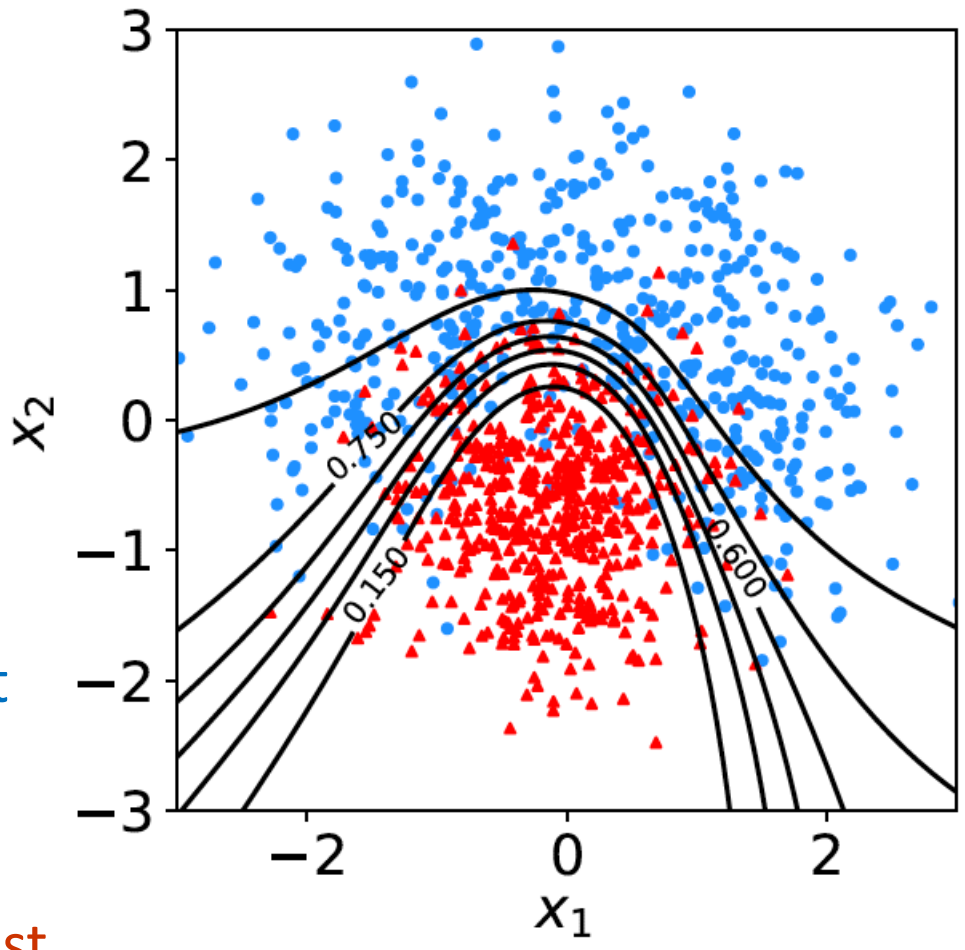
Decision function, test statistic

A surface in an n-dimensional space can be described by

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

scalar function

constant

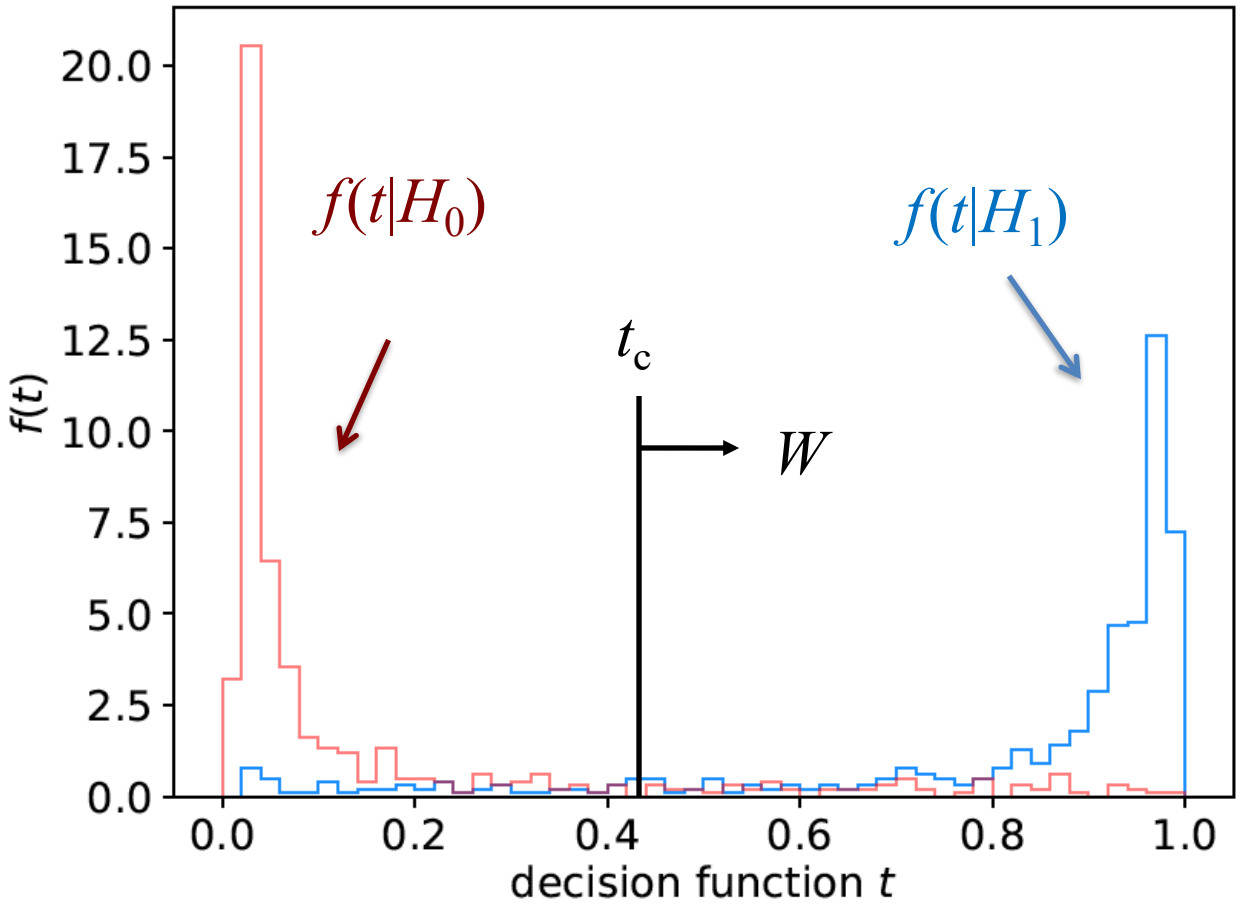


Different values of the constant t_c result in a family of surfaces.

Problem is reduced to finding the best decision function or test statistic $t(\boldsymbol{x})$.

Distribution of $t(\boldsymbol{x})$

By forming a test statistic $t(\boldsymbol{x})$, the boundary of the critical region in the n-dimensional $\boldsymbol{x}$-space is determined by a single single value t_c.

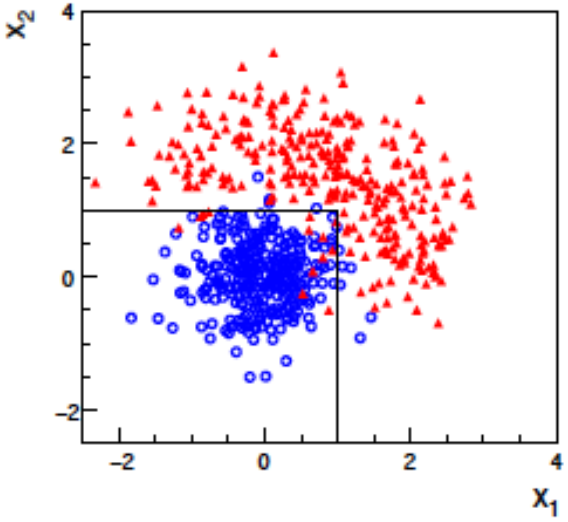


Types of decision boundaries

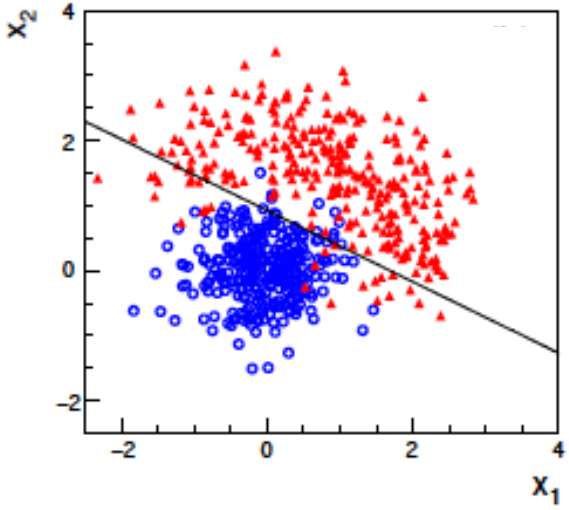
So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic $t(\boldsymbol{x})$?

First find best $t(\boldsymbol{x})$, later address issue of optimal size of test.

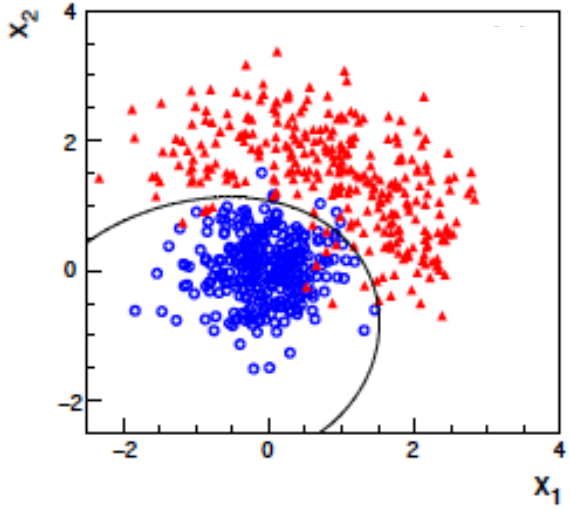
Remember $\boldsymbol{x}$-space can have many dimensions.



"cuts"



linear



non-linear

Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way', in particular if the data space is multidimensional?

Neyman-Pearson lemma states:

For a test of H_0 of size α, to get the highest power with respect to the alternative H_1 we need for all $\boldsymbol{x}$ in the critical region W

"likelihood ratio (LR)" →

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$$

inside W and $\leq c_\alpha$ outside, where c_α is a constant chosen to give a test of the desired size.

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.

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(\boldsymbol{x}|\mathrm{s}), f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{s}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

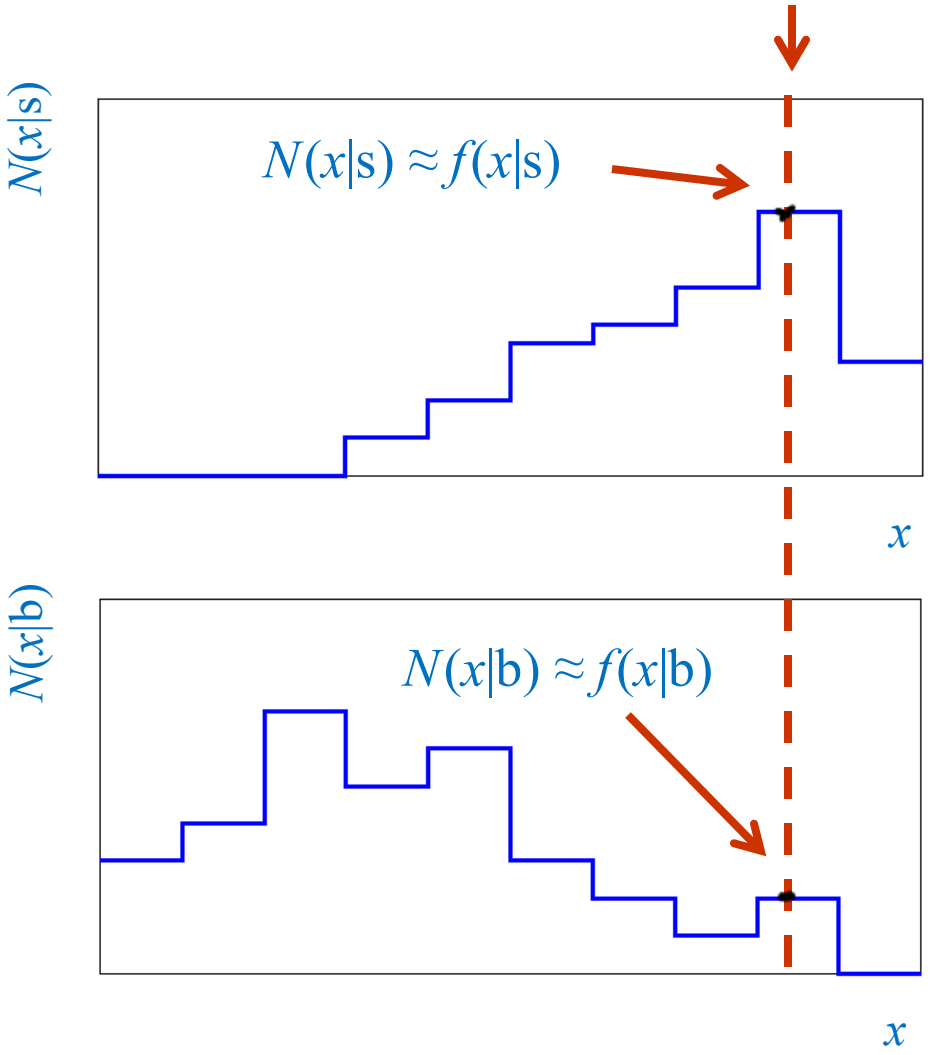
generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

This gives samples of "training data" with events of known type.

- Use these to construct a statistic that is as close as possible to the optimal likelihood ratio (→ Machine Learning).

Approximate LR from histograms

Want $t(x) = f(x|\mathrm{s})/f(x|\mathrm{b})$ for x here



One possibility is to generate MC data and construct histograms for both signal and background.

Use (normalized) histogram values to approximate LR:

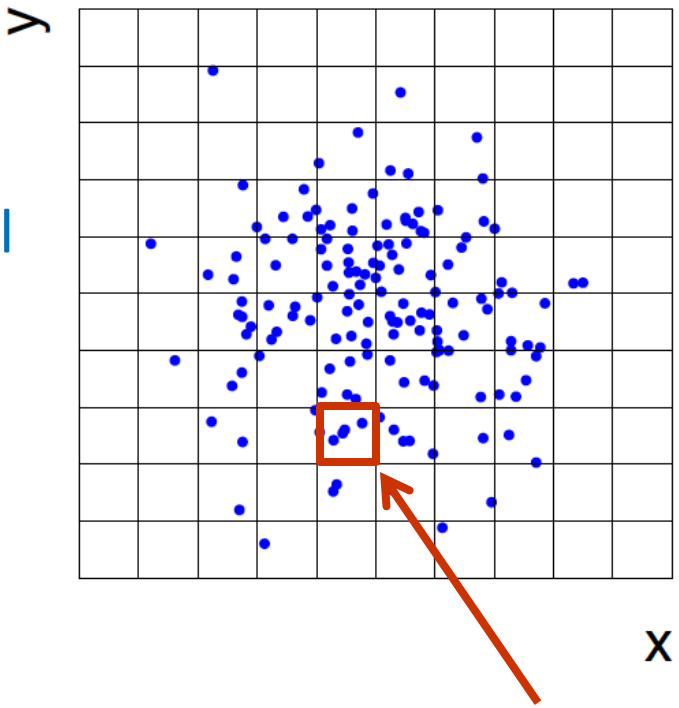
$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

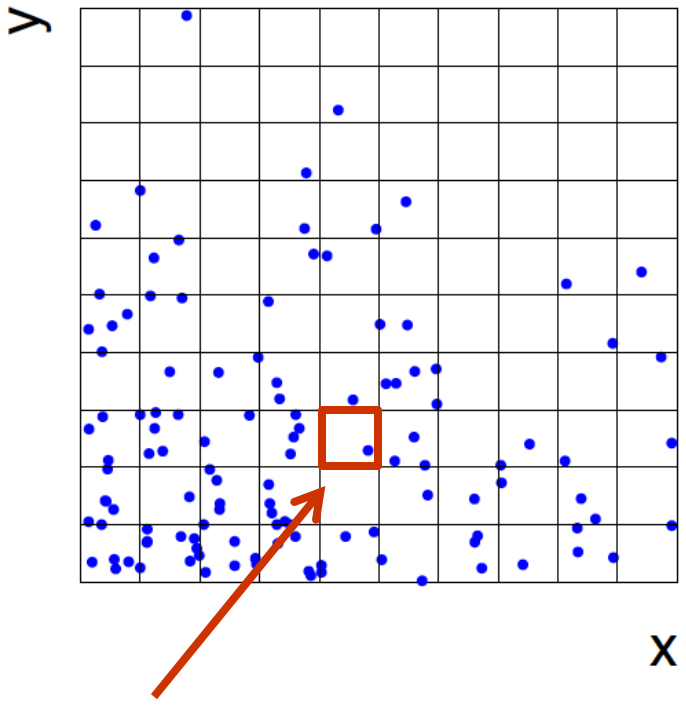
Can work well for single variable.

Approximate LR from 2D-histograms

Suppose problem has 2 variables. Try using 2-D histograms:

signal





back-ground

Approximate pdfs using $N(x,y|\mathrm{s})$, $N(x,y|\mathrm{b})$ in corresponding cells.

But if we want M bins for each variable, then in n-dimensions we have M^n cells; can't generate enough training data to populate.

→ Histogram method usually not usable for $n > 1$ dimension.

Strategies for multivariate analysis

Neyman-Pearson lemma gives optimal answer, but cannot be used directly, because we usually don't have $f(\boldsymbol{x}|\mathrm{s}), f(\boldsymbol{x}|\mathrm{b})$.

Histogram method with M bins for n variables requires that we estimate M^n parameters (the values of the pdfs in each cell), so this is rarely practical.

A compromise solution is to assume a certain functional form for the test statistic $t(\boldsymbol{x})$ with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities $f(\boldsymbol{x}|\mathrm{s})$ and $f(\boldsymbol{x}|\mathrm{b})$ (with something better than histograms) and use the estimated pdfs to construct an approximate likelihood ratio.

Multivariate methods (Machine Learning)

Many new (and some old) methods:

Fisher discriminant

(Deep) Neural Networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting, Boosting

ML → AI, beyond the scope of these lectures but useful to keep in mind the connection to statistics. Recommend (free online):

C.M. Bishop and H. Bishop, Deep Learning, https://www.bishopbook.com/

James, Witten, Hastie, Tibshirani & Taylor, An Introduction to Statistical Learning, https://www.statlearning.com/

Parameter estimation

The parameters of a pdf are any constants that characterize it,

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

r.v. parameter

i.e., θ indexes a set of hypotheses.

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

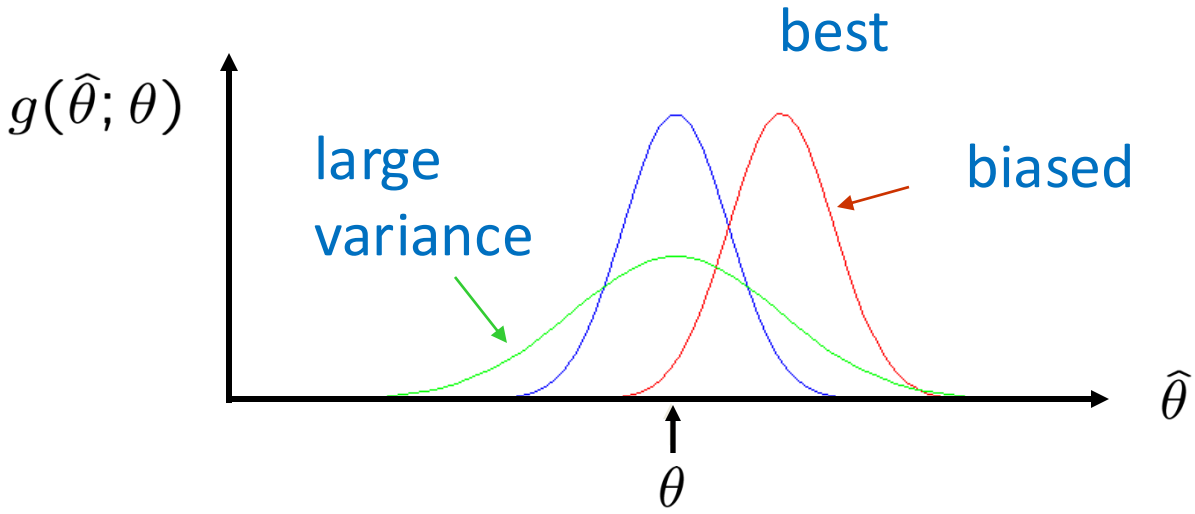
We want to find some function of the data to estimate the parameter(s):

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

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

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

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

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

Consider n independent observations of x: $x_1, \ldots, x_n$, 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 (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

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

MLE 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 $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

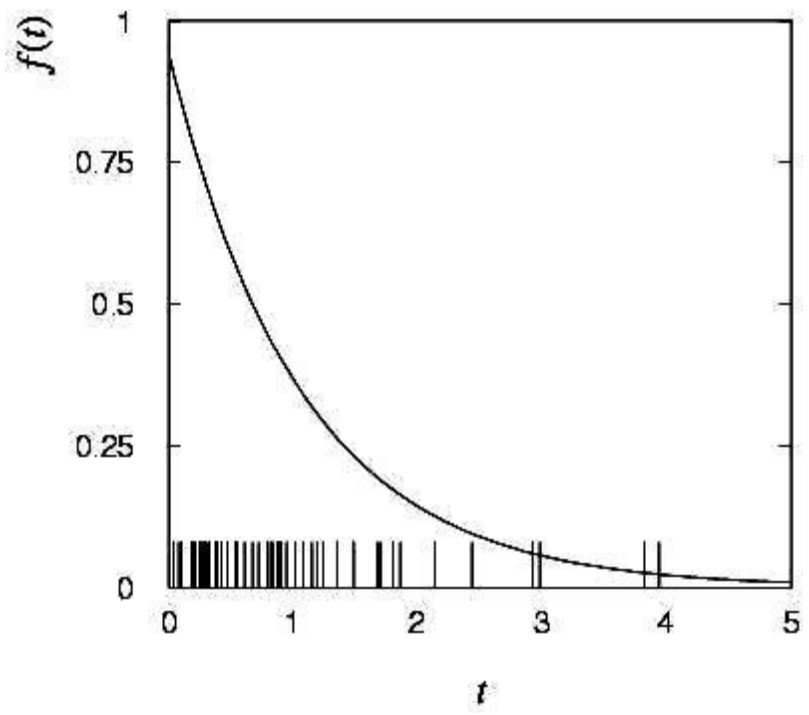
The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

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

MLE 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:

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

MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t-\tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

Variance of estimators: Monte Carlo method

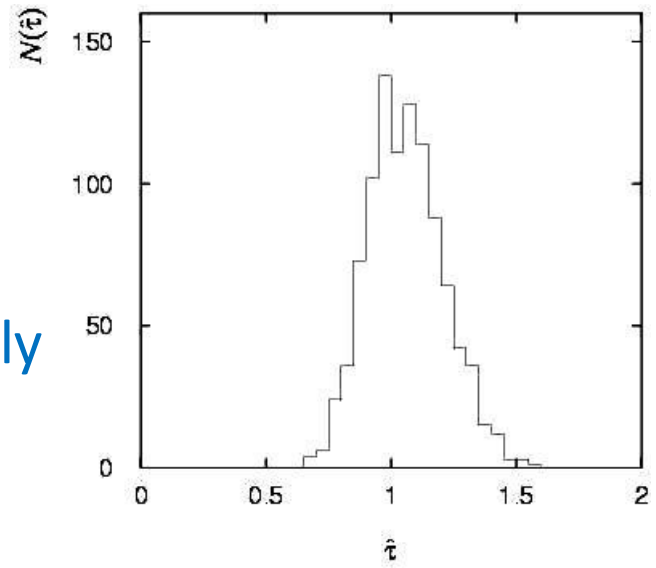
Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.



Variance of estimators from information inequality

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

Minimum Variance Bound (MVB)

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

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:

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

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial \tau}\right)^2 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

and we showed $b = 0$, hence $\partial b/\partial \tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $$E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}\,,$$

and therefore $$\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]\,.$$ (Here MLE is "efficient").

Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}}$$

i.e., $$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

$\rightarrow$ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2.

Example of variance by graphical method

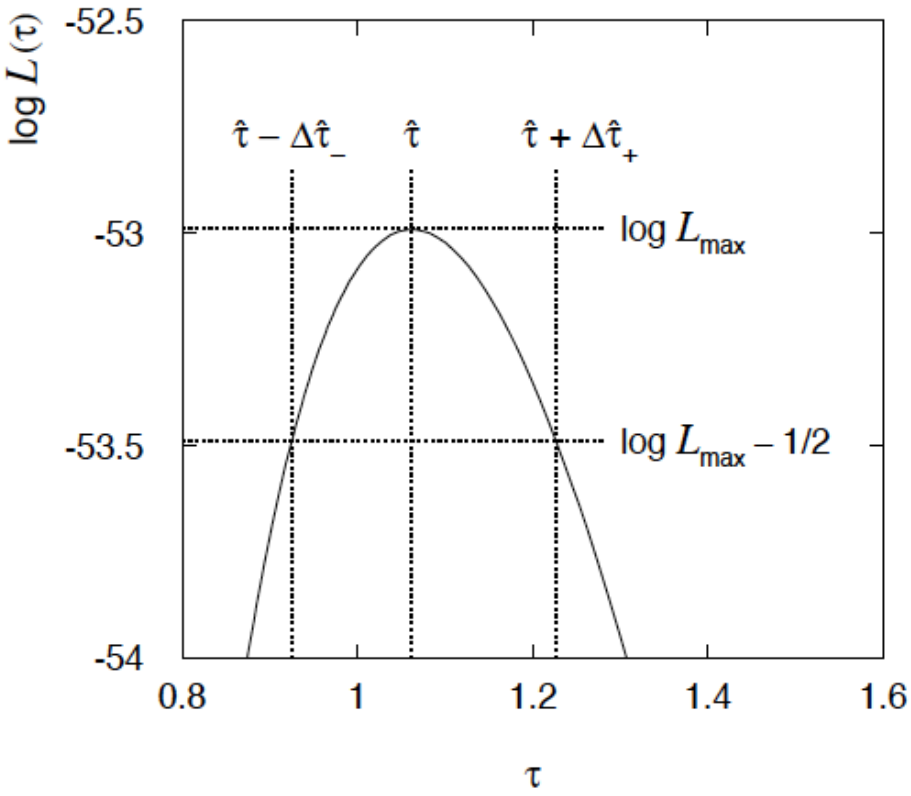
ML example with exponential:

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

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Not quite parabolic $\ln L$ since finite sample size (n = 50).

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,...,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l} \left(\delta_{ik} + \frac{\partial b_i}{\partial\theta_k}\right) I^{-1}_{kl} \left(\delta_{lj} + \frac{\partial b_l}{\partial\theta_j}\right)$$

is positive semi-definite:

$\boldsymbol{z}^{\mathrm{T}} M \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias $\longrightarrow$ 0

inequality $\longrightarrow$ equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

This can be estimated from data using $$\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$$

Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Confidence intervals by inverting a test

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

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 $P(\text{data in critical region}|\theta) \leq \alpha$ for a prespecified α, 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 are not rejected in a test of size α (confidence level CL is $1-\alpha$).

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 $p_\theta \leq \alpha$, then we reject θ.

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

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

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

For a multidimensional parameter space $\boldsymbol{\theta} = (\theta_1, \ldots \theta_M)$ *use* same idea – result is a confidence "region" with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or "cover" θ is

$$P(\text{conf. interval "covers" } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\boldsymbol{x}|H,\theta)$.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

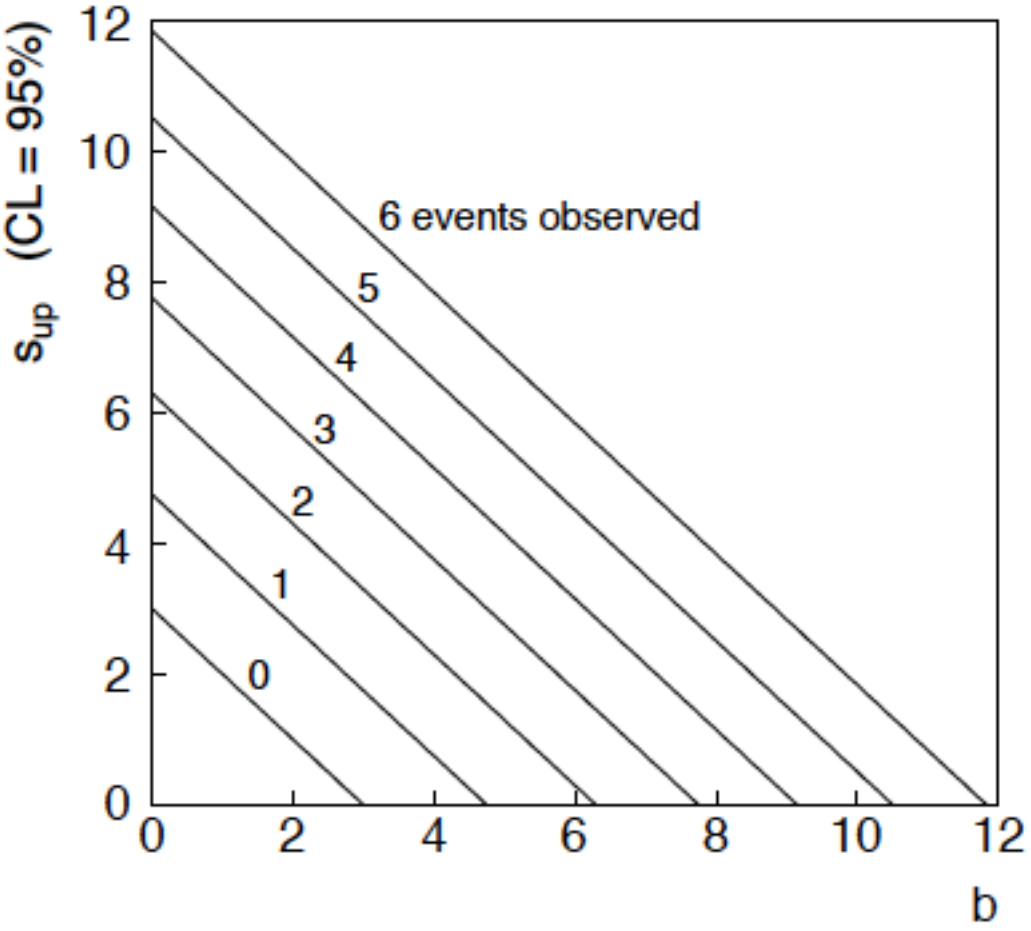
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

$n \sim$ Poisson(s+b): frequentist upper limit on s

For low fluctuation of n, formula can give negative result for $s_{\rm up}$; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL $= 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s.

Expected limit for $s = 0$

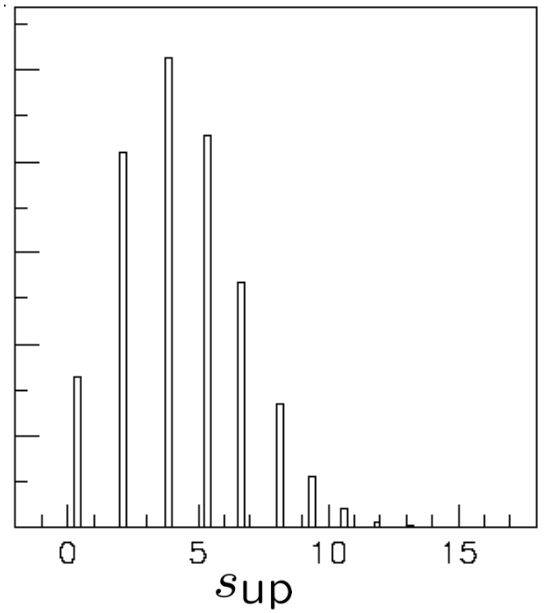
Physicist: I should have used CL = 0.95 — then $s_{\rm up} = 0.496$

Even better: for CL = 0.917923 we get $s_{\rm up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



Background-free vs background dominated limits

For $n \sim \text{Poisson}(s+b)$ $\rightarrow$ $s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2(n+1)) - b$

Upper limit on the rate of a signal process is $\Gamma_{\text{up}} = \frac{s_{\text{up}}}{Nt}$

targets, time

For $b \ll 1$, $\langle s_{\text{up}} \rangle_{s=0} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2)$ $\rightarrow$ $\langle \Gamma_{\text{up}} \rangle_{s=0} = \frac{3.00}{Nt} \propto \frac{1}{t}$

For $b \gg 1$, $n \rightarrow x \sim \text{Gauss}(\mu = s+b, \sigma = \sqrt{s+b})$

$(1-\alpha = 0.95)$

$\hat{s} = n - b$, $\sigma_{\hat{s}} = \sqrt{s+b}$, $s_{\text{up}} = \hat{s} + \sigma_{\hat{s}} \Phi^{-1}(1-\alpha)$

$\langle s_{\text{up}} \rangle_{s=0} = \sqrt{b}\Phi^{-1}(1-\alpha)$ $\rightarrow$ $\langle \Gamma_{\text{up}} \rangle_{s=0} = \frac{\sqrt{b}\Phi^{-1}(1-\alpha)}{Nt} = \frac{1.64\sqrt{b}}{Nt} \propto \frac{1}{\sqrt{t}}$

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad\qquad 0 \leq \lambda(\boldsymbol{\theta}) \leq 1$$

Lower $\lambda(\boldsymbol{\theta})$ means worse agreement between data and hypothesized $\boldsymbol{\theta}$. Equivalently, usually define

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher $t_{\boldsymbol{\theta}}$ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore $\qquad p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

chi-square dist. with # d.o.f. = # of components in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

Assuming this holds, the p-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_n}(t_{\boldsymbol{\theta}})$$ ← set equal to α

To find boundary of confidence region set $p_{\boldsymbol{\theta}} = \alpha$ and solve for $t_{\boldsymbol{\theta}}$:

$$t_{\boldsymbol{\theta}} = F^{-1}_{\chi^2_n}(1 - \alpha)$$

Recall also

$$t_{\boldsymbol{\theta}} = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

For example, for $1-\alpha = 68.3\%$ and $n = 1$ parameter,

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

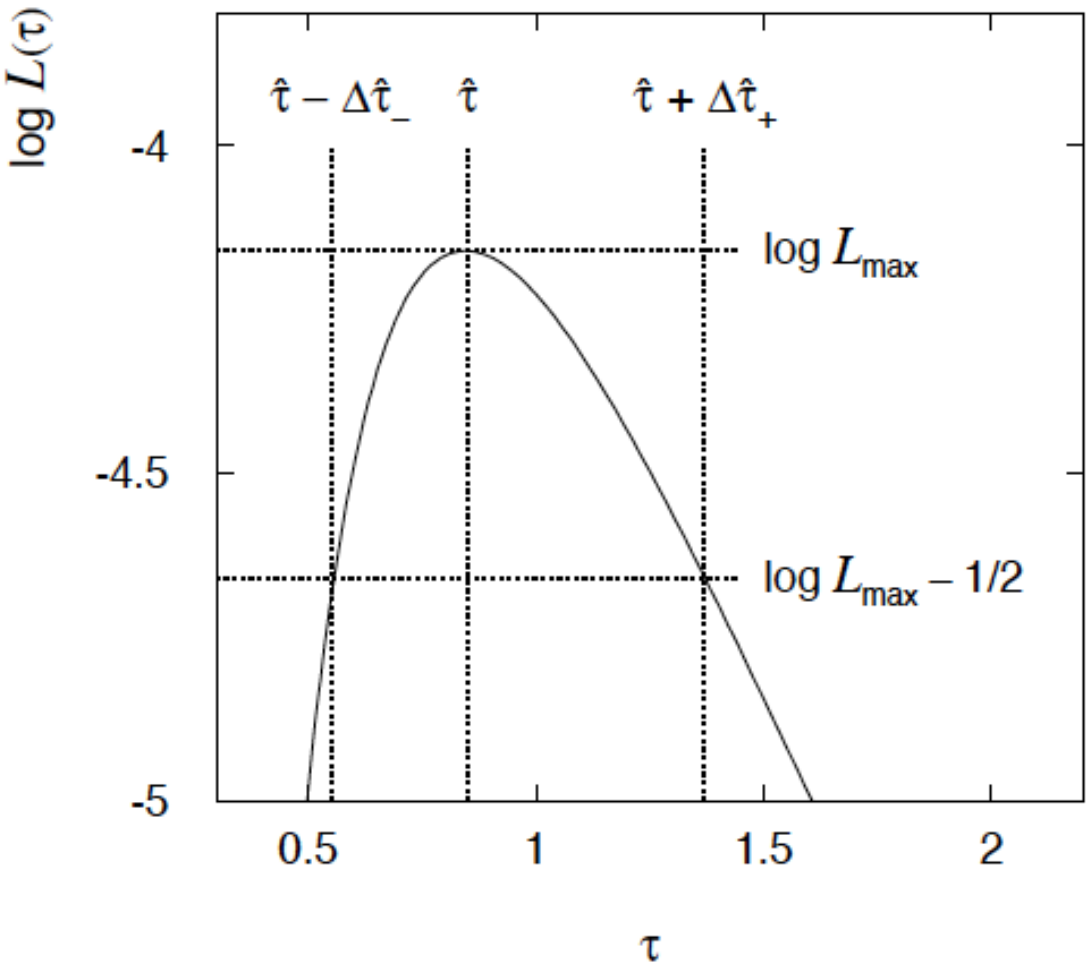
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from $\ln L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\alpha = 1$.



Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as "asymmetric error bar":

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

For increasing number of parameters, CL $= 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$

Q_α	$1-\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given CL $= 1 - \alpha$.

$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Extra slides

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

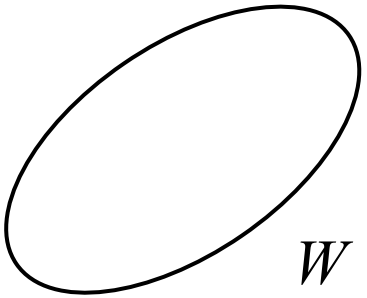
S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

S. Navas et al. (Particle Data Group), Phys. Rev. D 110, 030001 (2024); **pdg.lbl.gov** sections on probability, statistics, MC.

Proof of Neyman-Pearson Lemma

Consider a critical region W and suppose the LR satisfies the criterion of the Neyman-Pearson lemma:

$$P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \geq c_\alpha \quad \text{for all } \boldsymbol{x} \text{ in } W,$$
$$P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \leq c_\alpha \quad \text{for all } \boldsymbol{x} \text{ not in } W.$$



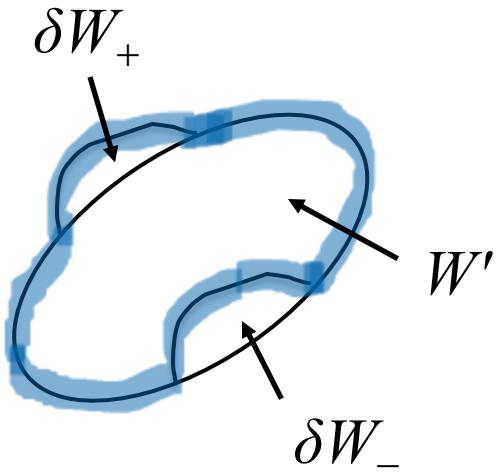
Try to change this into a different critical region W' retaining the same size α, i.e.,

$$P(\mathbf{x} \in W'|H_0) = P(\mathbf{x} \in W|H_0) = \alpha$$

To do so add a part δW_+, but to keep the size α, we need to remove a part δW_-, i.e.,

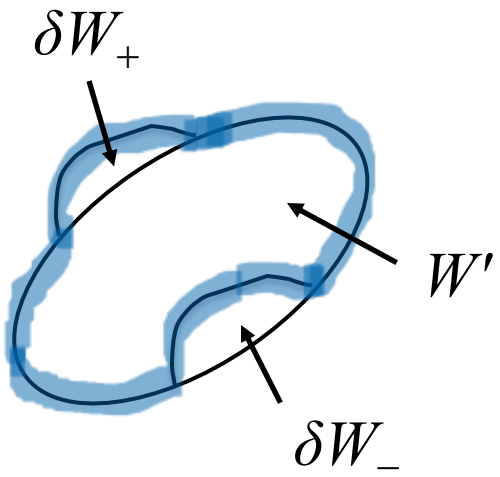
$$W \to W' = W + \delta W_+ - \delta W_-$$

$$P(\mathbf{x} \in \delta W_+|H_0) = P(\mathbf{x} \in \delta W_-|H_0)$$



Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all $\boldsymbol{x}$ in δW_- removed than for the $\boldsymbol{x}$ in δW_+ added, and therefore



$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_+ | H_0) c_\alpha$$

$$P(\mathbf{x} \in \delta W_- | H_1) \geq P(\mathbf{x} \in \delta W_- | H_0) c_\alpha$$

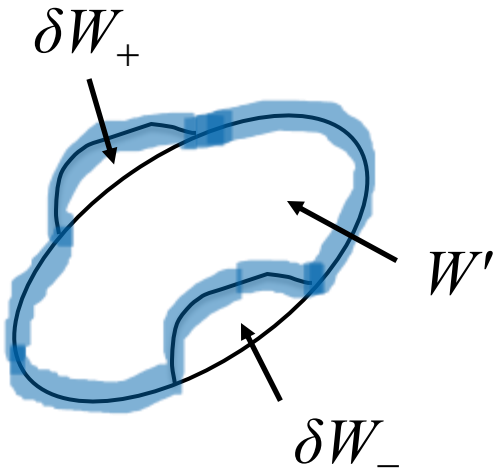
The right-hand sides are equal and therefore

$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_- | H_1)$$

Proof of Neyman-Pearson Lemma (3)

We have

$$W \cup W' = W \cup \delta W_+ = W' \cup \delta W_-$$



Note W and δW_+ are disjoint, and W' and δW_- are disjoint, so by Kolmogorov's 3rd axiom,

$$P(\mathbf{x} \in W') + P(\mathbf{x} \in \delta W_-) = P(\mathbf{x} \in W) + P(\mathbf{x} \in \delta W_+)$$

Therefore

$$P(\mathbf{x} \in W'|H_1) = P(\mathbf{x} \in W|H_1) + P(\mathbf{x} \in \delta W_+|H_1) - P(\mathbf{x} \in \delta W_-|H_1)$$

≤ 0

Proof of Neyman-Pearson Lemma (4)

And therefore

$$P(\mathbf{x} \in W' | H_1) \leq P(\mathbf{x} \in W | H_1)$$

i.e. the deformed critical region W' cannot have higher power than the original one that satisfied the LR criterion of the Neyman-Pearson lemma.

Large-sample (asymptotic) properties of MLEs

Suppose we have an i.i.d. data sample of size n: $x_1,...,x_n$

In the large-sample (or "asymptotic") limit ($n \to \infty$) and assuming regularity conditions one can show that the likelihood and MLE have several important properties.

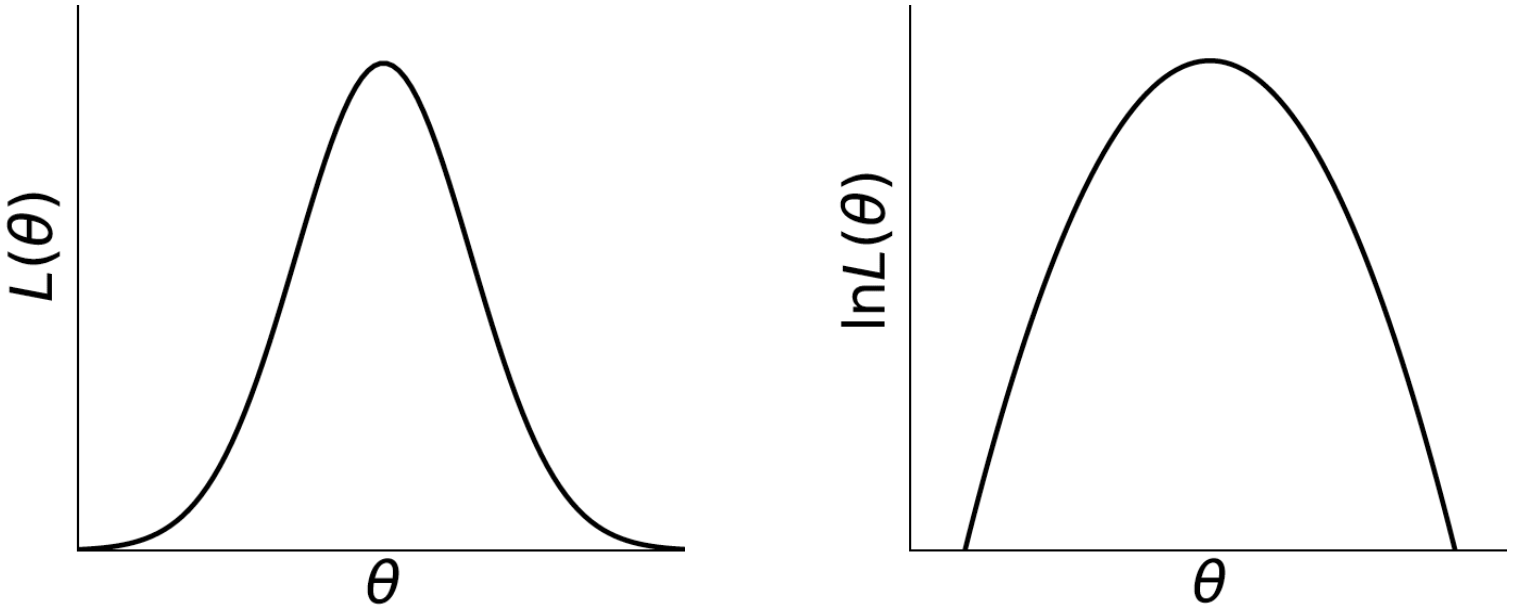
The regularity conditions include:

- the boundaries of the data space cannot depend on the parameter;
- the parameter cannot be on the edge of the parameter space;
- $\ln L(\theta)$ must be differentiable;
- the only solution to $\partial \ln L / \partial \theta = 0$ is $\hat{\theta}$.

In the slides immediately following, the properties are shown without proof for a single parameter; the corresponding properties hold also for the multiparameter case, $\boldsymbol{\theta} = (\theta_1,..., \theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).



The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak,
but $L(\theta) = P(\boldsymbol{x}|\theta)$ is the probability for $\boldsymbol{x}$, not a pdf for θ.

The MLE converges to the true parameter value

In the large-sample limit, the MLE converges in probability to the true parameter value.

That is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

The MLE is said to be *consistent*.

MLE is asymptotically unbiased

In general the MLE can be biased, but in the large-sample limit, this bias goes to zero:

$$\lim_{n\to\infty} E[\hat{\theta}] - \theta = 0$$

(Recall for the exponential parameter we found the bias was identically zero regardless of the sample size n.)

The MLE's variance approaches the MVB

In the large-sample limit, the variance of the MLE approaches the minimum-variance bound, i.e., the information inequality becomes an equality (and bias goes to zero):

$$\lim_{n\to\infty} V[\hat{\theta}] = -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

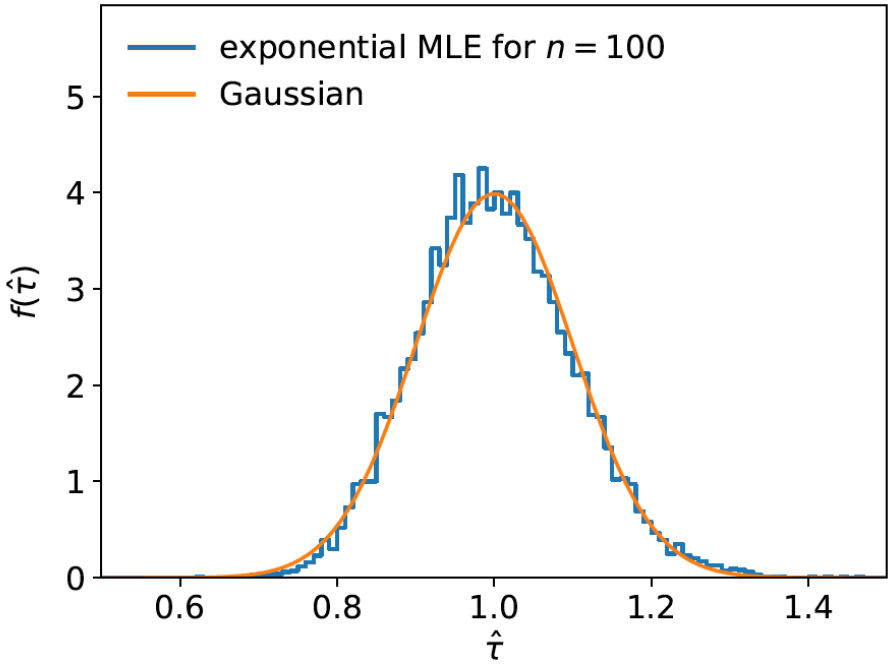
In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma^2_{\hat{\theta}}}$$

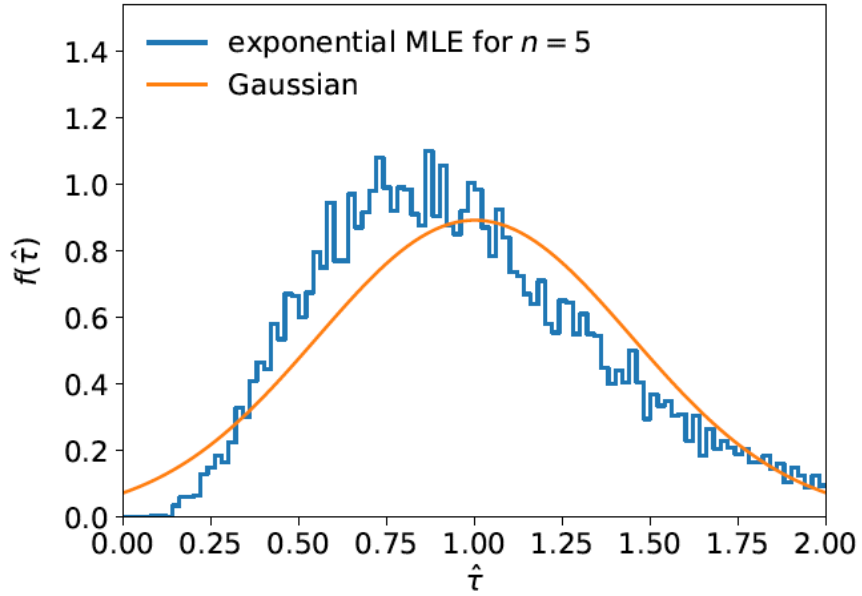
where $\sigma^2_{\hat{\theta}}$ is the minimum variance bound (note bias is zero).

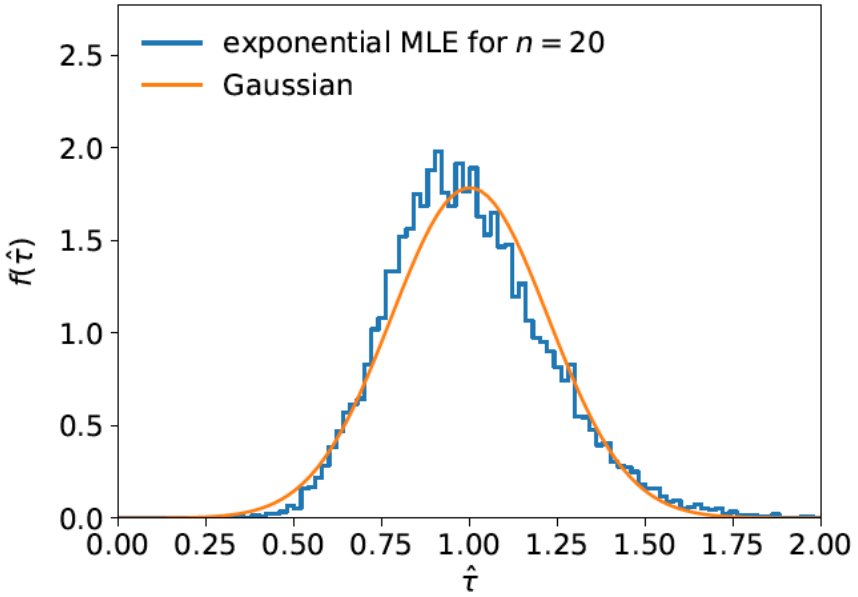
For example, exponential MLE with sample size $n = 100$.

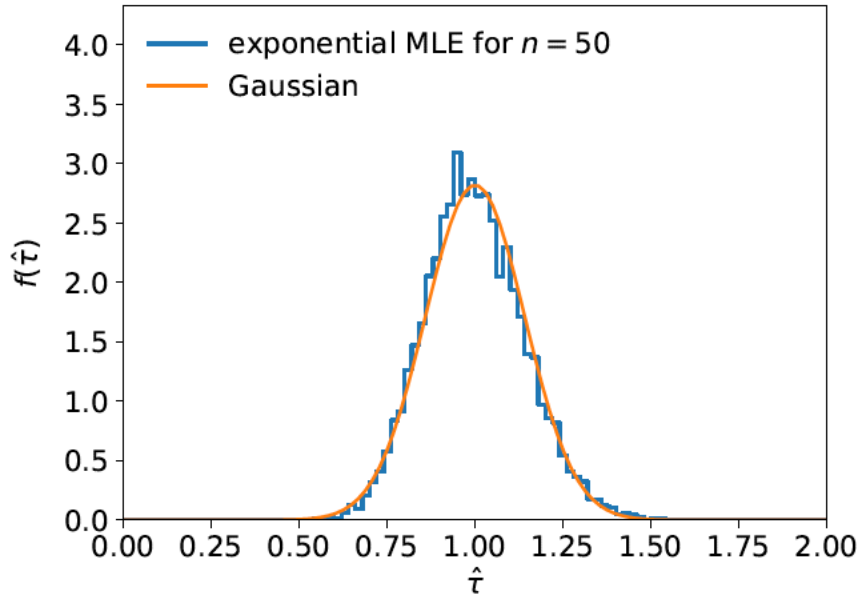
Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.

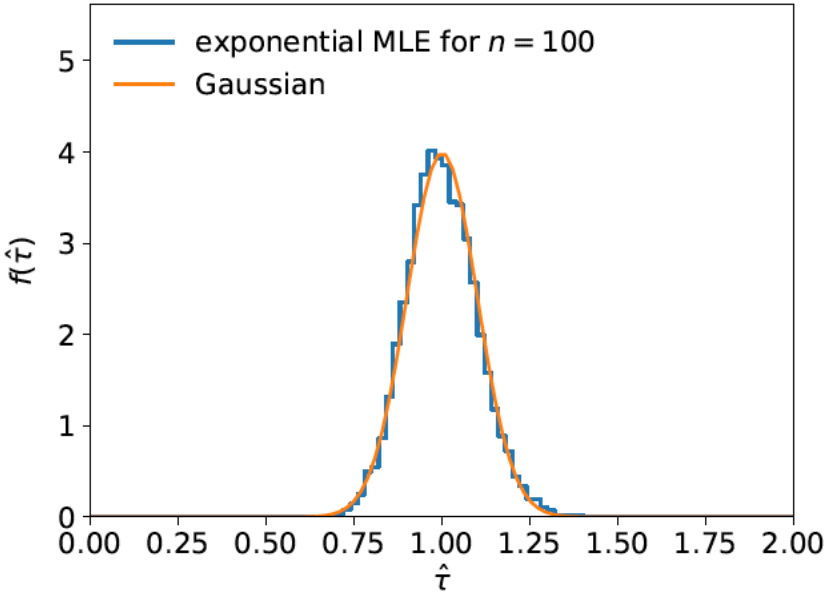


Distribution of MLE of exponential parameter









Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$\ln L_{\max}$

zero

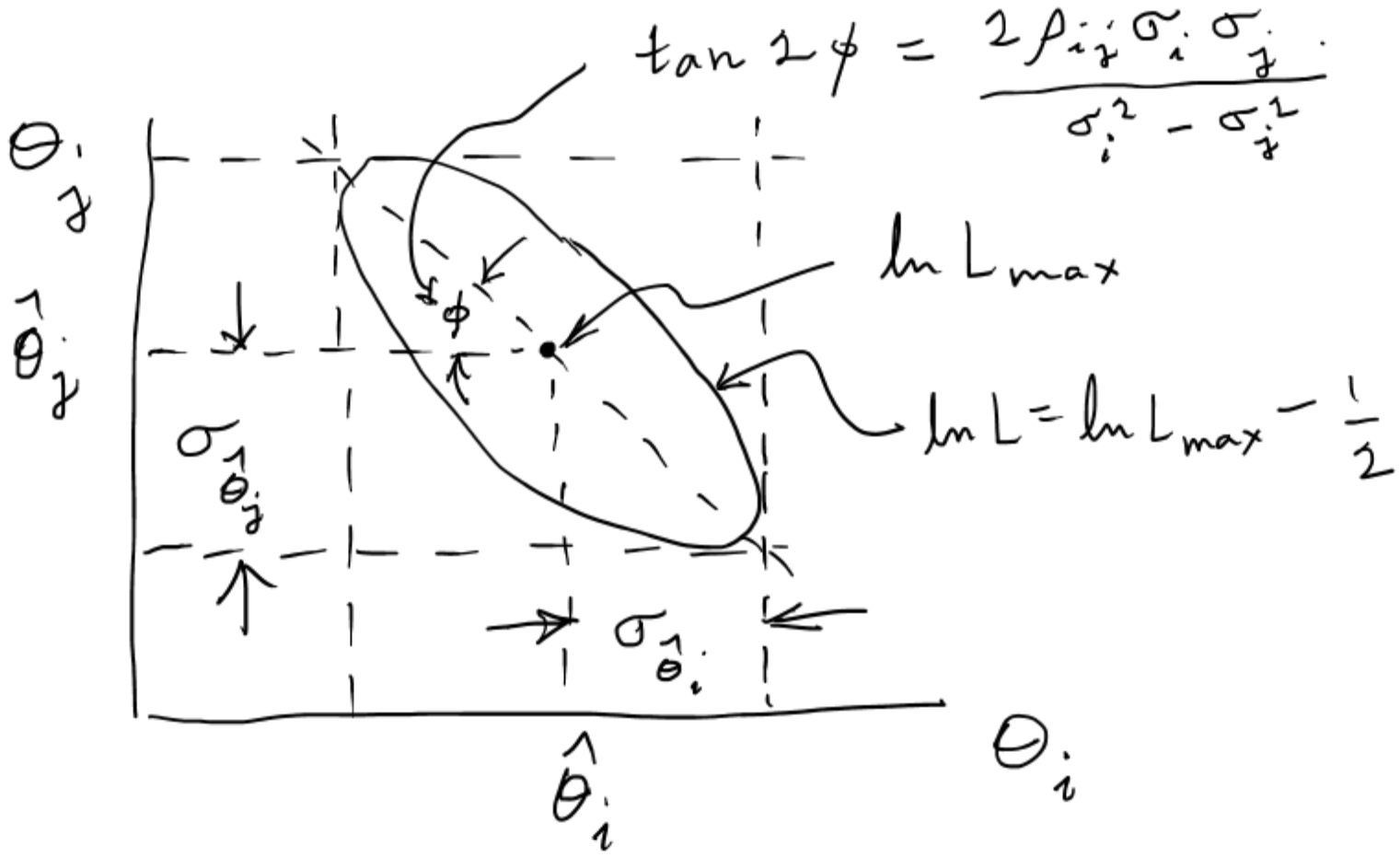
relate to covariance matrix of MLEs using information (in)equality.

Result: $$\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)



Distance from MLE to tangent planes gives standard deviations.

The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\,d\theta} \propto p(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta|x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson(s+b), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s\text{up}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized; can be OK provided $p(n|s)$ dies off quickly for large s.

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of s, then this would imply a non-flat prior for s.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference; or viewed as a recipe for producing an interval whose frequentist properties can be studied (e.g., coverage probability, which will depend on true s).

Bayesian upper limit with flat prior for s

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto p(n|s)\pi(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \times 1\,, \quad s \geq 0$$

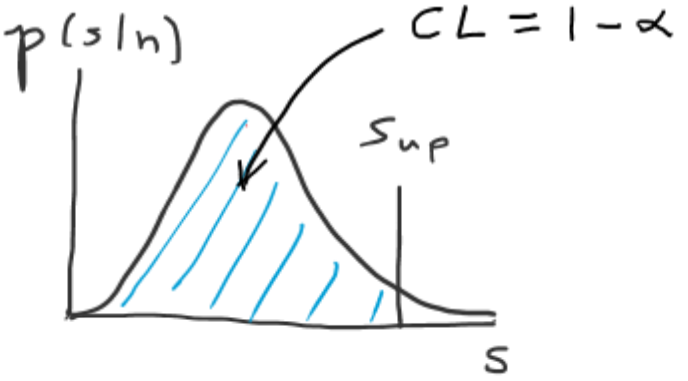
Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete gamma function

Upper limit $s_{\rm up}$ determined by

$$1 - \alpha = \int_0^{s_{\rm up}} p(s|n)\, ds$$



Bayesian interval with flat prior for s

Solve to find limit $s_{\rm up}$:

$$s_{\rm up} = \frac{1}{2} F_{\chi^2}^{-1}[p, 2(n+1)] - b$$

where

$$p = 1 - \alpha\left(1 - F_{\chi^2}[2b, 2(n+1)]\right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

Bayesian interval with flat prior for *s*

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

For $b = 0$, Bayesian and frequentist upper limits come out equal.

Never goes negative.

Doesn't depend on b if $n = 0$.